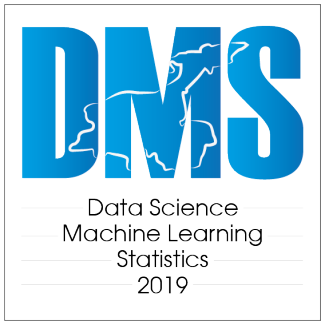


International Conference on Data Science, Machine Learning and Statistics - 2019 (DMS-2019)*

June 26-29, 2019

Proceedings Book

Conference Venue:
Faculty of Economics and Administrative Sciences,
Van Yuzuncu Yil University, 65080 Van, Turkey.

* This conference was supported by Scientific Research Projects Coordination Unit of Van Yuzuncu Yil University. Project number FTD-2019-7971.

Preface

Dear Colleagues,

Welcome to the international conference on “Data Science, Machine, Learning and Statistics-2019 (DMS-2019)” held by Van Yuzuncu Yil University from June 26-29, 2019. The DMS-2019 shall create a forum to discuss recent progress and emerging ideas in areas of interest. During the conference, participants will have opportunities to discuss new theoretical and practical issues in their fields and to establish new collaborations. Presentations can cover topics such as advance soft computing, heuristic algorithms, data infrastructures and analytics which are recent advances in Data Science, Machine Learning and Statistics.

The scientific program of DMS-2019 contains eight invited lectures of renowned personalities in the field. We wish you a productive, stimulating conference and a memorable stay in Olomouc.

Editors
H. Eray Celik
Cagdas Hakan Aladag

Scientific Committee

Abbas Mohamad Ali	Engin Avci	Mehmet Recep Minaz	Ramazan Tekin
Adem Kalinli	Erkan Aydar	Melih Kuncan	Rayimbek Sultanov
Adil Baykasoglu	Fatih Ozkaynak	Mete Celik	Resul Das
Ali Caglayan	Gratiela Dana Boca	Moayad Y. Potrus	Ridvan Saracoglu
Ali Karci	Guckan Yapar	Muhammed Baykara	Samer M. Barakat
Ali Rostami	Hamit Mirtagioglu	Murat Atan	Sedat Yerli
Alparslan A. Basaran	Hatice Hicret Ozkoc	Murat Demir	Serhat Omer Rencber
Alper Basturk	Ibrahim Kilic	Musa Atas	Sevil Senturk
Arash Kalami	Ibrahim Turkoglu	Mustafa Sevuktekin	Sinan Calik
Ashis SenGupta	Kambiz Majidzadeh	Naci Genc	Suat Ozdemir
Atilla Goktas	Kasirga Yildirak	Nasip Demirkus	Suat Toraman
Aydin Sipahioglu	Kazim Hanbay	Necmettin Sezgin	Sylwia Gwozdziewicz
Burak Uyar	Keming Yu	Novruz Allahverdi	Sakir sleyen
Bulent Batmaz	Kenan nce	Nuh Alpaslan	Sengul Cangur
Candan Gkceoglu	M.H. Fazel Zarandi	Nuri Almali	Tahir Hanalioglu
Carlos A. Coelho	M. Kenan Dnmez	Olcay Arslan	Timur Han Gur
Cetin Guler	M. Marinescu Mazurencu	Omed Salim Khalind	Ufuk Tanyeri
Davut Hanbay	M. Selim Elmali	Onur Koksoy	Veysel Yilmaz
Deniz Dal	Mahdi Zavvari	Orhan Ecemis	Yener Altun
Dervis Karaboga	Mehmet Kabak	Omer Faruk Ertugrul	Yildirim Demir
Ebru Akcapinar Sezer	Mehmet Karadeniz	Ozgur Yeniay	Yilmaz Kaya
Ebru Caglayan Akay	Mehmet Mendes	Ozgur Yilmazel	Zeki Yildiz

Organization Committee

Birdal Senoglu	Fatma Gul Akgul	Fikriye Ataman	Serpil Sevimli Deniz
Fevzi Erdogan	Hanifi Van	Kubra Bagci	Taner Uckan
Suat Sensoy	Murat Canayaz	Onur Camli	Ali Yilmaz
Hayrettin Okut	Recep Ozdag	Seda Basar Yilmaz	Israfil Celik
Cetin Guler	Sukru Acitas	Ebubekir Seyyarer	Mesut Kapar
Hayati Cavus	Talha Arslan	Erol Kina	Aksel Akyurek
Fatma Zehra Dogru	Kadir Emir	Faruk Ayata	Hayrullah Urcan
Fuat Tanhan	Asuman Yilmaz Duva	Firat Kapar	
Sinan Saracli	Emre Bicek	Necati Erdogan	
Alper Hamzadayi	Fatih Uludag	Serbest Ziyanak	

Table of Contents

Statistical Inference for the Moyal Distribution under Type-II Censoring

T. Arslan [1], S. Acitas [2], B. Senoglu [3]

[1] *Van Yuzuncu Yil University, Van, Turkey, mstalhaarslan@yyu.edu.tr*

[2] *Eskisehir Technical University, Eskisehir, Turkey, sacitas@eskisehir.edu.tr*

[3] *Ankara University, Ankara, Turkey, senoglu@science.ankara.edu.tr*

Abstract

This study aims to obtain maximum likelihood (ML) estimators of the location and scale parameters of the Moyal distribution under the Type-II censored sample. The ML estimators of the location and scale parameters can not be obtained in closed from since likelihood equation involve nonlinear functions of the unknown parameters. Therefore, iterative methods, such as Newton-Raphson, should be used in solving the likelihood equations simultaneously. Different from the ML methodology, here we use modified ML methodology which results in closed form estimators for the unknown parameters.

1. Introduction

The motivation of this study is to obtain explicit estimators of the location and scale parameters of the Moyal distribution under Type-II censoring. Proposed methodology in this study provides to avoid encountering computational difficulties in solving likelihood equations iteratively.

The Moyal distribution has the following probability density function (pdf) and cumulative distribution function (cdf):

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{2}\exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right]; x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+. \quad (1)$$

and

$$F_X(x;\mu,\sigma) = \Gamma\left[\frac{1}{2}\exp\left(-\left(\frac{x-\mu}{\sigma}\right), \frac{1}{2}\right)\right], \quad (2)$$

respectively. Here μ is the location parameter, σ is the scale parameter and

$$\Gamma\left[\frac{1}{2}\exp\left(-\left(\frac{x-\mu}{\sigma}\right), \frac{1}{2}\right)\right] = \frac{1}{\Gamma\left(\frac{1}{2}\right)}\int_{\frac{1}{2}\exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)}^{\infty} u^{0.5-1}\exp(-u)du$$

stands for the upper incomplete gamma function. For further information about the Moyal distribution see Arslan et al. [1] and references therein.

The rest of the paper is organized as follows. Estimators of the location and scale parameters of the Moyal distribution are obtained by using the ML and modified ML (MML) methodologies in Section 2. The paper is finalized with the some concluding remarks.

2. Parameter Estimation under Type-II Censored Samples

Let $x_{(r_1+1)} \leq x_{(r_1+2)} \leq \cdots \leq x_{(n-r_2-1)} \leq x_{(n-r_2)}$ be a Type-II censored sample. Here, $x_{(\cdot)}$ denotes the ordered observations in ascending ways. $r_1, r_2 \geq 0$ are the numbers of the censored observation(s) from below and above, respectively.

The log-likelihood ($\ln L$) function for the Moyal distribution under Type-II censored sample is written as follows:

$$\begin{aligned}\ln L(\mu,\sigma) = \quad & C - (n-r_1-r_2)\ln\sigma - \frac{1}{2}\sum_{i=r_1+1}^{n-r_2} z_{(i)} - \frac{1}{2}\sum_{i=r_1+1}^{n-r_2}\exp(-z_{(i)}) \\ & + r_1\ln\left[F\left(z_{(r_1+1)}\right)\right] + r_2\ln\left[1 - F\left(z_{(n-r_2)}\right)\right]\end{aligned} \quad (3)$$

where, $z_{(i)} = (x_{(i)} - \mu)/\sigma$, C is a constant and $F(\cdot)$ is the cdf of the Moyal distribution given in Equation 2.

The ML and MML estimators of the location and scale parameters of the Moyal distribution under Type-II censored samples are obtained in following subsections, respectively.

2.1. The ML estimation

After taking derivatives of the $\ln L$ with respect to the parameters μ and σ following likelihood equations are obtained:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{2\sigma}\sum_{i=r_1+1}^{n-r_2} g_1(z_{(i)}) - \frac{r_1}{\sigma} g_2(z_{(r_1+1)}) + \frac{r_2}{\sigma} g_3(z_{(n-r_2)}) = 0 \tag{4}$$

and

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} &= -\frac{n-r_1-r_2}{\sigma} + \frac{1}{2\sigma}\sum_{i=r_1+1}^{n-r_2} z_{(i)} g_1(z_{(i)}) - \frac{r_1}{\sigma} z_{(r_1+1)} g_2(z_{(r_1+1)}) \\ &\quad + \frac{r_2}{\sigma} z_{(n-r_2)} \frac{f(z_{(n-r_2)})}{g}_3 (z_{(n-r_2)}) = 0,\end{aligned} \tag{5}$$

respectively. Here, $g_1(z_{(i)}) = [1 - \exp(-z_{(i)})]$, $g_2(z_{(r_1+1)}) = \frac{f(z_{(r_1+1)})}{F(z_{(r_1+1)})}$, $g_3(z_{(n-r_2)}) = \frac{f(z_{(n-r_2)})}{1-F(z_{(n-r_2)})}$, and $f(\cdot)$ and $F(\cdot)$ stand for the pdf and cdf of the Moyal distribution.

It is clear that the ML estimators of the unkown parameters μ and σ cannot be obtained explicitly therefore iterative methods should be used to obtain the estimates of them. However, it should be noted that using numerical methods may have the following problems: (i) non-convergence of iterations (ii) convergence to multiple roots and (iii) convergence to wrong root, see e.g. Barnett [2], Puthenpura and Sinha [3] and Vaughan [5].

2.2. The MML estimation

In this subsection, Tiku's [4] MML methodology avoiding the computational difficulties encountered in iterative techniques is used to obtain the explicit estimators of the location and scale parameters of the Moyal distribution. Resulting estimators are called as the MML estimators and are asymptoticaly equivalent to the corresponding ML estimators.

The MML estimators of the location and scale parameters are obtained by linearizing the nonlinear functions of the unknown parameters around the expected values of the standardized ordered observations (i.e. $t_{(i)} = E(z_{(i)})$) using the first two terms of Taylor series expansion; see below:

$$\begin{aligned} g_1(z_{(i)}) &\cong \alpha_{1i} + \beta_{1i} z_{(i)}; \quad (i = r_1+1, \ldots, n-r_2), \\ g_2(z_{(r_1+1)}) &\cong \alpha_{2(r_1+1)} + \beta_{2(r_1+1)} z_{(r_1+1)} \\ g_3(z_{(n-r_2)}) &\cong \alpha_{3(n-r_2)} + \beta_{3(n-r_2)} z_{(n-r_2)}. \end{aligned} \tag{6}$$

After incorporating the linearized functions in (6) into the likelihood equations (4) and (5), we obtain the following modified likelihood equations:

$$\frac{\partial \ln L^\star}{\partial \mu} = \frac{1}{2\sigma}\sum_{i=r_1+1}^{n-r_2} (\alpha_{1i} + \beta_{1i} z_{(i)}) - \frac{r_1}{\sigma}(\alpha_{2(r_1+1)} + \beta_{2(r_1+1)} z_{(r_1+1)}) + \frac{r_2}{\sigma}(\alpha_{3(n-r_2)} + \beta_{3(n-r_2)} z_{(n-r_2)}) \tag{7}$$

and

$$\begin{aligned}\frac{\partial \ln L^\star}{\partial \sigma} &= -\frac{n-r_1-r_2}{\sigma} + \frac{1}{2\sigma}\sum_{i=r_1+1}^{n-r_2} z_{(i)}(\alpha_{1i} + \beta_{1i} z_{(i)}) - \frac{r_1}{\sigma} z_{(r_1+1)}(\alpha_{2(r_1+1)} + \beta_{2(r_1+1)} z_{(r_1+1)}) \\ &\quad + \frac{r_2}{\sigma} z_{(n-r_2)}(\alpha_{3(n-r_2)} + \beta_{3(n-r_2)} z_{(n-r_2)}).\end{aligned} \tag{8}$$

Here,

$$\begin{aligned} &\alpha_{1i} = 1 - \exp(-t_{(i)}) - t_{(i)}\exp(-t_{(i)}), \quad \beta_{1i} = \exp(t_{(i)}); \quad (i = r_1+1, \ldots, n-r_2), \\ &\alpha_{2(r_1+1)} = \frac{f(t_{(r_1+1)})}{F(t_{(r_1+1)})} - t_{(r_1+1)} \frac{f'(t_{(r_1+1)})F(t_{(r_1+1)}) - f^2(t_{(r_1+1)})}{F^2(t_{(r_1+1)})}, \\ &\beta_{2(r_1+1)} = \frac{f'(t_{(r_1+1)})F(t_{(r_1+1)}) - f^2(t_{(r_1+1)})}{F^2(t_{(r_1+1)})}, \\ &\alpha_{3(n-r_2)} = \frac{f(t_{(n-r_2)})}{1-F(t_{(n-r_2)}} - t_{(n-r_2)} \frac{f'(t_{(n-r_2)})(1-F(t_{(n-r_2)})) - f^2(t_{(n-r_2)})}{(1-F(t_{(n-r_2)}))^2}, \\ &\beta_{3(n-r_2)} = \frac{f'(t_{(n-r_2)})(1-F(t_{(n-r_2)})) - f^2(t_{(n-r_2)})}{(1-F(t_{(n-r_2)}))^2}. \end{aligned}$$

Solutions of the equations in (7) and (8) are the following MML estimators formulated by

$$\hat{\mu}_{MML} = \bar{x}_w + \frac{\Delta}{m}\hat{\sigma}_{MML} \quad and \quad \hat{\sigma}_{MML} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{A(A-1)}} \tag{9}$$

where

$$\bar{x}_w = [0.5(\sum_{i=r_1+1}^{n-r_2} \beta_{1i} x_{(i)}) - r_1\beta_{2(r_1+1)}x_{(r_1+1)} + r_2\beta_{3(n-r_2)}x_{(n-r_2)}]/m,$$

$$m = 0.5(\sum_{i=r_1+1}^{n-r_2} \beta_{1i}) - r_1\beta_{2(r_1+1)} + r_2\beta_{3(n-r_2)},$$

$$\Delta = 0.5(\sum_{i=r_1+1}^{n-r_2} \alpha_{1i}) - r_1\alpha_{2(r_1+1)} + r_2\alpha_{3(n-r_2)}, \quad A = n - r_1 - r_2,$$

$$B = 0.5\left[\sum_{i=r_1+1}^{n-r_2} \alpha_{1i}\left(x_{(i)} - \bar{x}_w\right)\right] - r_1\alpha_{2(r_1+1)}\left(x_{(r_1+1)} - \bar{x}_w\right) + r_2\alpha_{3(n-r_2)}\left(x_{(n-r_2)} - \bar{x}_w\right)$$

and

$$C = [0.5\sum_{i=r_1+1}^{n-r_2} \beta_{1i}(x_{(i)} - \bar{x}_w)^2] - r_1\beta_{2(r_1+1)}(x_{(r_1+1)} - \bar{x}_w)^2 + r_2\beta_{3(n-r_2)}(x_{(n-r_2)} - \bar{x}_w)^2.$$

Remark: The original denominator of $\hat{\sigma}_{MML}$ is $2A$, however it is replaced by $2\sqrt{A(A-1)}$ for bias correction.

It should be noted that $t_{(i)}$ values cannot be obtained exactly. We therefore use their approximate values using the following equality:

$$t_{(i)} = F^{-1}\left(\frac{i}{n+1}\right), \quad i = r_1 + 1, \ldots, n - r_2$$

where $F^{-1}(\cdot)$ is the inverse of the cdf of the Moyal distribution given in Equation 2.

3. Conclusion

In this study, estimators of the unknown parameters of the Moyal distribution are obtianed by using the ML methodology in which iterative techniques are used and the MML methodology which gives explicit solutions of the likelihood equations. It can be concluded that the MML estimators can be prefered to the ML estimators if our main concern is to avoid the computational complexities besides having high efficiencies.

References

[1] Arslan, T., Acitas, S. and Senoglu, B., 2019. Estimation of the location and scale parameters of Moyal Distribution. TWMS Journal of Applied and Engineering Mathematics, doi: 10.26837/jaem.590766.

[2] Barnett, V.D., 1966. Evaluation of the maximum likelihood estimator when the likelihood equation has multiple roots. Biometrika 53, 151–165.

[3] Puthenpura, S., Sinha, N.K., 1986. Modified maximum likkelihood method for the robust estimation of system parameters from very noise data. Automatica 22, 231–235.

[4] Tiku, M. L., 1967. Estimating the mean and standard deviation from a censored normal sample. Biometrika 54, 155–165.

[5] Vaughan, D.C., 1992. On the Tiku-Suresh method of estimation. Commun. Stat.-Theory Meth. 21(2), 451–469.

Prognostic Performance of Statistical and Machine Learning Methods on MIMIC-III Clinical Database

M.A. Ozalp[1], K. Yildirak[2], C.H. Aladag[3], I. Zor[4], N. Unal[5]

[1]*Hacettepe University, Actuarial Sciences, Ankara, Turkey, m.asim.ozalp@hacettepe.edu.tr*
[2]*Hacettepe University, Actuarial Sciences, Ankara, Turkey, kasirga@hacettepe.edu.tr*
[3] *Hacettepe University, Department of Statistics Ankara, Turkey, chaladag@gmail.com*
[4] *Hacettepe University, Department of Statistics Ankara, Turkey, ibrahimz@hacettepe.edu.tr*
[5]*Ankara University, Department of Anesthesiology and Intensive Care, Ankara, Turkey, necmettinunal@gmail.com*

Abstract

MIMIC-III is a data set composed of more than 60000 admissions made to Beth Israel Hospitals. For every de-identified critical care patients demographics, vital signs, laboratory tests, medications, and more are hold in this database. The most important cause of deaths in the hospital is considered as Sepsis. Sepsis is defined as 'life-threatening organ dysfunction caused by a dysregulated host response to infections. In medical literature, many scoring systems such as SOFA, LODS, SIRS, NEWS, etc. have been suggested for the early prediction/diagnosis of sepsis and evaluation of prognosis. Both machine learning and statistical learning methods have been applied to model survival/death status for intensive care unit patients in Mimic - III database. Used methods are Random Forest, Support Vector Machine, Logistic Regression, Naive Bayes, Adaboost and Artificial Neural Networks (ANN). It is a well-known fact that ANN approach is an effective prediction tool. And, it is very crucial to determine the best ANN model in order to get accurate predictions. In this study, different ANN models have been applied to MIMIC-III data set to determine the best ANN model. As a result of the implementation, all obtained prognostic results are presented and discussed.

1. Introduction

Sepsis is defined as 'life-threatening organ dysfunction caused by a dysregulated host response to infections, and leads to deaths (Mervyn Singer; Clifford S. Deutschman; ChristopherWarren Seymour; Manu Shankar-Hari. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) JAMA. 2016;315(8):801-810.)[1]. Mortality of sepsis is very high and ranges between 15-63% which depends to several factors (Carly J. Paoli, Mark A. Reynolds; Meenal Sinha; Matthew Gitlin; Elliott Crouser. Crit Care Med 2018; 46:1889–1897; Epidemiology of sepsis in intensive care units in Turkey: a multicenter, point-prevalence study Nur Baykara, Halis Akalın, Mustafa Kemal Arslantaş, Volkan Hancı, Çiğdem Çağlayan, Ferda Kahveci, Kubilay Demirağ, Canan Baydemir, Necmettin Ünal and Sepsis Study Group. Critical Care (2018) 22:93) [2,3]. One third or half of all hospital deaths are linked with sepsis [4](Chanu Rhee; Raymund Dantes; Lauren Epstein; David J. Murphy; ChristopherW. Seymour. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. JAMA. doi:10.1001/jama.2017.13836; Liu V, Escobar .J., Greene J.D. Hospital Deaths in Patients With Sepsis From 2 Independent CohortsJAMA 2014; 312: 90-92.

In addition to high mortality rate, hospital cost of each septic patient may be as high as 50.000 USD and it's generate huge total cost to healthcare systems (Carly J. Paoli, PharmD, MPH1; Mark A. Reynolds, PhD1; Meenal Sinha, MBA2; Matthew Gitlin, PharmD3; Elliott Crouser,.Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level. Crit Care Med 2018; 46:1889–1897) [2]. It has been estimated that yearly hospital cost of sepsis is higher than 30 billion dollars in USA and it increases approximately 9% each year ((Carly J. Paoli, Mark A. Reynolds; Meenal Sinha; Matthew Gitlin; Elliott Crouser. Crit Care Med 2018; 46:1889–1897) [2].

Early diagnosis and early treatment are main factors which decrease mortality rate in addition to cost tremendously. However, sepsis is a very complicated clinical syndrome and early diagnosis needs to evaluate several clinical and laboratory indices. Therefore, several scoring systems have been introduced to clinical practice for early prediction/diagnosis of sepsis and prognostic evaluation of patients in addition to updating the definitions of sepsis approximately 3 times in about 30 years period. Some of these scoring systems are SIRS, SOFA, LODS, NEWS, and etc.(https://www.england.nhs.uk/ourwork/clinical-policy/sepsis/nationalearlywarningscore/; ChristopherW. Seymour,MD, MSc; Vincent X. Liu, MD, MSc; Theodore J. Iwashyna, MD, PhD; Frank M. Brunkhorst. Assessment of Clinical Criteria for Sepsis for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-774.) [5]. We believe that artificial intelligence and machine learning methods, which can be performed rapidly by considering the weighted effectiveness of different combinations of current laboratory

and clinical parameters, will lead to more accurate and rapid prediction and diagnosis of sepsis. Because of reliability and facility in patient data availability, in this study we use the data from Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC). MIMIC-III is a data set composed of more than 60000 admissions made to Beth Israel Hospitals. For every de-identified critical care patients demographics, vital signs, laboratory tests, medications, and more are hold in this database.

In medical literature, rule-based sepsis severity scores (SOFA, LODS, SIRS, etc.) are suggested for the sepsis-related organ failure assessment score that are used to track a person's status during the stay in an intensive care unit. [5],[6] Due to the spread of electronic health records and developments in defined data through databases such as MIMIC, it is recently stepped forward the study of detection and prediction of sepsis using machine learning and statistical learning methods in literature. Some of these studies are given below. Sepsis severity models based on artificial neural networks [7], Sepsis severity models based on dynamic Bayesian network [8], Survival-analytic models for length of stay hospital of sepsis patient [9,10], and Custom feature engineering for sepsis severity scores [11,12]. The aim of the studies is to improve upon traditional rules-based scoring systems by counting in nonlinear relationships, patient's trends, correlations between variables.

This study uses a dataset comprised of 63510 intensive care unit (ICU) patient records gather from MIMIC-III which is a well-known dataset. In addition to the variables examined in calculating rule-based sepsis severity scores, all important variables for sepsis are also used in this study. In the implementation part of the study, MIMIC-III data set is analyzed both statistical and machine learning approaches such as Logistic Regression, Naive Bayes, Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), AdaBoost, Random Forest, Induction Algorithm (CN2) and Artificial Neural Networks (ANN). The death/survive situations of the patients are classified by employing these algorithms. Actually, we interest in the mortality of sepsis related patients in hospital. All of these algorithms are applied by using Orange program. Furthermore, the performance of both the methods based on statistical and machine learning algorithms are compared in terms of some measures such as Area Under the Curve (AUC), precision, recall and confusion matrix. And, all obtained results are presented and discussed.

2. MIMIC-III Data Set

MIMIC-III is a large, database comprising information relating to patients in critical care units at Beth Israel Hospitals. This study uses a dataset comprised of 63510 ICU patient records gather from MIMIC-III. Sepsis related patients aged between 1 and 85 years old are included in the data set. Besides the variables examined in calculating rule-based sepsis severity scores, all other important variables available in the related literature for sepsis are also included in this study. All variables used in the application can be divided into two sub-groups which are laboratory measurements and vital signs. While the lab measurements group includes 36 variables, the vital signs group has 14 variables. All variables included in these groups are presented in Table 1 and 2.

Table 1. The Laboratory Measurements

Variable name	**Variable name**	**Variable name**	**Variable name**	**Variable name**	**Variable name**
aniongap_min	aniongap_max	lactate_min	lactate_max	glucose_min	glucose_max
bands_min	bands_max	platelet_min	platelet_max	hemotocrit_min	hemotocrit_max
bicarbonate_min	bicarbonate_max	potassium_min	potassium_max	hemoglobin_min	hemoglobin_max
bilurbin_min	bilurbin_max	ptt_min	ptt_max	sodium_min	sodium_max
creatin_min	creatin_max	inr_min	inr_max	bun_min	bun_max
chloride_min	chloride_max	pt_min	pt_max	wbc_min	wbc_max

Table 2. The Vital Signs

Variable name	**Variable name**	**Variable name**	**Variable name**
heartrate_min	heartrate_max	urineouput_min	urineouput_max
sysbp_min	sysbp_max	mingcs_min	mingcs_max
resprate_min	resprate_max	ventilation_min	ventilation_max
tempc_min	tempc_max		

3. The implementation

By applying the methods mentioned above to MIMIC-III data set, deaths arising from sepsis are examined. The data is divided into two sets which are used training and test. %80 and %20 of the data compose of training and test sets, respectively. Determining the best ANN architecture is an important issue in order to get accurate results. Therefore, different ANN architectures are examined to obtain the best results. In this study, feed forward neural networks contains two hidden layers is used. By changing the number of neurons between 1 and 20 in both hidden layers, 400 architectures are totally examined and the best architectures which produce the best classification results are picked. The results obtained from all methods are compared. The results over the test set are evaluated by using some criteria and confusion matrices compose. Definition of confusion matrix is given below. In here, 0 and 1 represent the patient situations alive and dead, respectively.

Confusion Matrix		Actual	
		0	1
Prediction	0	TP	FP
	1	FN	TN

Table 3. Formulas for AC, F1, Precision and Recall

AC	(TP+TN)/(Total)
Recall	FP+FN /(TP+FP+FN+TN)
Precision	TP+TN/(TP+FP+FN+TN)
F1	2*(Precision*Recall)/(Precision+Recall)

Literally, recall is the ratio of the prediction values of positive in model to actual number of positive observation in data. Precision is the ratio by which percentage of the values that the model estimates as positive is accurate. Algebraically, the precision is the ratio of the positive predicted values to all the predicted positive values of the model. Based on the definition of confusion matrix, formulas for AC, F1, Precision and Recall are given in Table 3.

In this study, Precision_0/Precision_1 and recall_0/recall_1 values are compared in terms of both survival and death status (0/1). For example, Precision_1 represents the ratio of accurate prediction of dead people. In other words, Precision_1 is calculated by using the formula TN/(FN+TN). All obtained prognostic results for all methods are summarized in Table 4. In this table, for example ANN 11,13 represents a feed forward neural network in which 11 and 13 neurons are included in first and second hidden layers, respectively.

Table 4. The obtained results

Method	AUC	CA	F1	Precision	Recall	Precision_0	Recall_0	Precision_1	Recall_1
SVM	0.319	0.647	0.696	0.759	0.647	0.856	0.719	0.06	0.13
SGD	0.649	0.883	0.872	0.867	0.883	0.913	0.959	0.53	0.34
Random Forest	0.825	0.895	0.876	0.879	0.895	0.907	0.981	0.67	0.28
ANN 11,13	0.865	0.896	0.88	0.88	0.896	0.911	0.977	0.65	0.31
ANN 10,9	0.863	0.897	0.882	0.882	0.897	0.913	0.976	0.66	0.33
ANN 10,12	0.862	0.898	0.881	0.882	0.898	0.911	0.979	0.67	0.31
ANN 10,5	0.862	0.896	0.877	0.879	0.896	0.909	0.979	0.66	0.29
Naive Bayes	0.789	0.568	0.639	0.868	0.568	0.961	0.531	0.19	0.84
Logistic	0.838	0.895	0.873	0.878	0.895	0.905	0.983	0.68	0.26
CN2	0.757	0.856	0.851	0.848	0.856	0.911	0.927	0.39	0.34
AdaBoost	0.803	0.893	0.871	0.874	0.893	0.904	0.982	0.65	0.25

According to Table 4, ANN models, logistic regression, and random forest give the best prognostic performance in terms of almost every measure. On the other hand, SVM produce the worst results in terms of all measures. Also, confusion matrices for all methods are presented in Table 5. When all confusion matrices are examined, it can be said that ANN logistic regression, and random forest produces the most accurate prognostic results for MIMIC-III data set.

Table 5. Confusion matrices

			Prediction						Prediction		
			0	1	Σ				0	1	Σ
ANN 10,9	Actual	0	13731	337	14068	Logistic Regression	Actual	0	13835	233	14068
		1	1307	641	1948			1	1448	500	1948
		Σ	15038	978	16016			Σ	15283	733	16016

ANN 10,12	Actual	0	13766	302	14068	Random Forest	Actual	0	13801	267	14068
		1	1338	610	1948			1	1407	541	1948
		Σ	15104	912	16016			Σ	15208	808	16016
ANN 11,13	Actual	0	13739	329	14068	AdaBoost	Actual	0	13815	253	14068
		1	1337	611	1948			1	1467	481	1948
		Σ	15076	940	16016			Σ	15282	734	16016
ANN 10,5	Actual	0	13775	293	14068	SGD	Actual	0	13486	582	14068
		1	1378	570	1948			1	1285	663	1948
		Σ	15153	863	16016			Σ	14771	1245	16016
SVM	Actual	0	10048	4336	14384	CN2	Actual	0	13039	1029	14068
		1	1344	278	1622			1	1281	667	1948
		Σ	11392	4614	16016			Σ	14320	1696	16016

4. Conclusion

Although rule based sepsis score calculation approaches has been used in the literature, scores obtained from machine learning methods has received a considerable amount of attention recently because of electronic recording systems which can store large amount of patient records. In this study, statistical and machine learning methods such as logistic regression, SVM, SGD, naive Bayes, AdaBoost, random forest, CN2 and ANN are applied to MIMIC-III well-known data set. Prognostic performances of all methods are compared to each other in terms of accurate prediction ratios. As a result of the implementation, it is observed that ANN models, logistic regression, and random forest produce the most accurate results while SVM method gives the worst results.

References

[1] Singer, Mervyn, et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)." *Jama* 315.8 (2016): 801-810.

[2] Paoli, Carly J., et al. "Epidemiology and costs of sepsis in the United States—An analysis based on timing of diagnosis and severity Level." *Critical care medicine* 46.12 (2018): 1889.

[3] Baykara, Nur, et al. "Epidemiology of sepsis in intensive care units in Turkey: a multicenter, point-prevalence study." *Critical Care* 22.1 (2018): 93.

[4] Rhee, Chanu, et al. "Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014." *Jama* 318.13 (2017): 1241-1249.

[5] Seymour, Christopher W., et al. "Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)." *Jama* 315.8 (2016): 762-774.

[7] Lipton, Zachary C., et al. "Learning to diagnose with LSTM recurrent neural networks." *arXiv preprint arXiv:1511.03677*(2015).

[8] Nachimuthu, Senthil K., and Peter J. Haug. "Early detection of sepsis in the emergency department using Dynamic Bayesian Networks." *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association, 2012.

[9] Henry, Katharine E., et al. "A targeted real-time early warning score (TREWScore) for septic shock." *Science translational medicine* 7.299 (2015): 299ra122-299ra122.

[10] Nemati, Shamim, et al. "An interpretable machine learning model for accurate prediction of sepsis in the ICU." *Critical care medicine* 46.4 (2018): 547-553.

[11] Mao, Qingqing, et al. "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU." *BMJ open* 8.1 (2018): e017833.

[12] Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

[13] Calvert, Jacob, et al. "High-performance detection and early prediction of septic shock for alcohol-use disorder patients." *Annals of medicine and surgery* 8 (2016): 50-55.

[14] Desautels, Thomas, et al. "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics* 4.3 (2016): e28.

[15] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997).

[16] Aladag, C.H., Editor, (2017) Advances in Time Series Forecasting, Volume 2, Bentham Science Publishers Ltd., eISBN: 978-1-68108-528-9, ISBN: 978-1-68108-529-6.

Robust Estimators for the Parameters of the Power Lindley Distribution

B. Çakmak[1], F. Z. Doğru[2]

[1]Giresun University, Giresun, Turkey, brvnckmk@gmail.com
[2] Giresun University, Giresun, Turkey, fatma.dogru@giresun.edu.tr

Abstract

In general, the maximum likelihood (ML) estimation, the most known classical method, is used to estimate the parameters of a given distribution. The ML estimators give desired results when there is no outlier in the data, however; they are failed when the data contains outlier. Therefore, the robust estimators are proposed as an alternative estimation method to the ML estimation method. We consider to use power Lindley distribution, which was introduced by [1], is used to analyze survival data sets. In this study, we propose robust estimators for the parameters of the power Lindley distribution via the optimal B-robust (OBR) estimation method given by [2]. We conduct a small simulation study to illustrate the performance of the newly proposed estimators over the ML estimators in the outlier case.

1. Introduction

The Lindley distribution was first introduced by [3]. Its probability density function (pdf) is given by:

$$f(x|\beta) = pf_1(x|\beta) + (1-p)f_2(x|2,\beta), \tag{1}$$

where

$$p = \frac{\beta}{\beta+1}\ , \qquad f_1(x|\beta) = \beta e^{-\beta x} \qquad \text{and} \qquad f_2(x|2,\beta) = \beta^2 x e^{-\beta x}.$$

It can be seen from (1) that the Lindley distribution is obtained from a two-component mixture of an exponential distribution (with scale beta) and a gamma distribution (with scale 2 and scale beta), with mixing proportion $p = \beta/(\beta+1)$.

Some distributional properties and applications of Lindley distribution was also examined by [4]. Since this distribution is not too flexible for the application, the power Lindley distribution was proposed by [1] as a flexible extension of the Lindley distribution. The pdf of the power Lindley distribution is given by:

$$f(x|\alpha,\beta) = \frac{\alpha\beta^2}{\beta+1}(1+x^\alpha)x^{\alpha-1}e^{-\beta\alpha}\ , \quad x>0,\ \ \alpha,\beta>0. \tag{2}$$

The cumulative distribution of the power Lindley distribution are defined as:

$$F(x) = 1 - \left(1 + \frac{\beta}{\beta+1}x^\alpha\right)e^{-\beta x^\alpha}, \quad x>0, \qquad \beta>0. \tag{3}$$

2. Parameter Estimation

In this section, we will first summarize the ML estimation method for the parameters of the power Lindley distribution which have been proposed by [1]. Then, we will give the OBR estimation method to estimate the parameters of interest.

2.1 ML estimation method

Let $x_1, \dots, x_n$ be a random sample of size n from power Lindley distribution given in (2). The ML estimators of α and β can be obtained by maximizing the following log-likelihood function:

$$\log L(\alpha,\beta) = n(\log(\alpha) + 2\log(\beta) - \log(\beta+1)) + \sum_{i=1}^{n}\log(1+x_i^{\alpha}) + (\alpha-1)\sum_{i=1}^{n}\log(x_i) - \beta\sum_{i=1}^{n}x_i^{\alpha}.$$

Then, after taking derivative of the log-likelihood function with respect to the parameters and setting them to zero, the ML estimates for α and β will be the solutions of the following non-linear equations:

$$\frac{\partial}{\partial\alpha}lnL = \frac{n}{\alpha} + \sum_{i=1}^{n}\frac{x_i^{\alpha}\ln(x_i)}{1+x_i^{\alpha}} - \beta\sum_{i=1}^{n}x_i^{\alpha}\ln(x_i) = 0,$$

$$\frac{\partial}{\partial\beta}lnL = \frac{n(\beta+2)}{\beta(\beta+1)} - \sum_{i=1}^{n}x_i^{\alpha} = 0\,.$$

Note that a numerical algorithm should be used to solve the above two equations.

2.2 OBR estimation method

The OBR estimation method was introduced by [2]. Recently, this method used by [5] to estimate the shape parameters of Burr XII distribution and [6] proposed OBR estimators for the parameters of the generalized half-normal distribution. In this part, we will propose the OBR estimators for the parameters of power Lindley distribution which will be an alternative to the ML estimators.

Let $\boldsymbol{\theta} = (\alpha,\beta)$. The M-estimator for $\boldsymbol{\theta}$ can be defined as minimum of $\sum_{i=1}^{n}\rho(x_i,\boldsymbol{\theta})$. Alternatively, if the ρ function is differentiable, the M-estimator of $\boldsymbol{\theta}$ can be found by solving $\sum_{i=1}^{n}\psi(x_i,\boldsymbol{\theta}) = 0$. In this study, we will use the following Huber's ρ and ψ functions with a b robustness tunning constant to obtain the OBR estimators:

$$\rho_b = \begin{cases} \dfrac{x^2}{2}\,, & |x| \le b, \\ b|x| - \dfrac{1}{2}b^2\,, & |x| > b, \end{cases} \qquad , \qquad \psi_b(x) = \begin{cases} x\,, & |x| \le b, \\ sgn(x)b\,, & |x| > b, \end{cases}$$

The influence function (IF) of M estimator is given by $= \frac{\psi(x,\boldsymbol{\theta})}{-\int\frac{\partial}{\partial\boldsymbol{\theta}}\psi(x,\boldsymbol{\theta})dF_{\boldsymbol{\theta}}(x)}$, which is used to measure robustness of an estimator. Also, the IF for ML estimator is computed by $IF = J(\boldsymbol{\theta})^{-1}\,s(x,\boldsymbol{\theta})$, where

$$s(x,\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{1}{\alpha} + \dfrac{x^{\alpha}\log(x)}{1+x^{\alpha}} + \log(x) + \beta x^{\alpha}\log(x) \\ \dfrac{\beta+2}{\beta(\beta+1)} - x^{\alpha} \end{bmatrix}$$

is the score function and $J(\boldsymbol{\theta})^{-1}$ is the Fisher information matrix. It can be seen that the score functions are not bounded; the IF for ML will be very sensitive to the outliers. Thus, we should obtain the robust estimators for the parameters of power Lindley distribution.

In this study, we will use the standardized OBR estimation method which is given by:

$$\sum_{i=1}^{n}\psi\left(\boldsymbol{A}(\boldsymbol{\theta})\big(\boldsymbol{s}(x_i,\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\big)\right) = \sum_{i=1}^{n}W_b(x_i,\boldsymbol{\theta})\{\boldsymbol{s}(x_i,\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\} = \boldsymbol{0},$$

where $W_b(x,\boldsymbol{\theta}) = \min\left\{1; \frac{b}{\|\boldsymbol{A}(\boldsymbol{\theta})\{\boldsymbol{s}(x_i,\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\}\|}\right\}$ is the weight function and $\|.\|$ shows the Euclidian norm. Also the nonsingular $p \times p$ matrix $\boldsymbol{A}(\boldsymbol{\theta})$ and the $p \times 1$ vector $\boldsymbol{a}(\boldsymbol{\theta})$ are given as:

$E\{\boldsymbol{\psi}(x,\boldsymbol{\theta})\boldsymbol{\psi}(x,\boldsymbol{\theta})^T\} = \{\boldsymbol{A}(\boldsymbol{\theta})^T\boldsymbol{A}(\boldsymbol{\theta})\}^{-1},\ E\{\boldsymbol{\psi}(x,\boldsymbol{\theta})\} = \mathbf{0}.$

To obtain the OBR estimates the algorithm proposed by [7] can be applied.

3. Simulation Study

In this section, we will provide a simulation study to compare the ML estimators over the OBR estimators for the parameters of the power Lindley distribution when the data includes outliers. We generate the data from power Lindley distribution using the estimation procedure given in [1]. For the comparison, we use the bias and root mean square error (RMSE) using the 1000 replicates for the sample sizes $25, 50$ and 100. We set the true parameter values as $(\alpha,\beta) = (1.5,1),(1,2),(0.8,0.35)$. We take the stopping rule as 10^{-6} for the algorithm given by [7]. All the computations are conducted using MATLAB2017b. For the computation of the OBR estimation, we use the starting value procedure given in [7]. For the outlier case, we use the outlier model: $(n-r)PL(x;\alpha,\beta) + rUniform(\overline{\mathrm{X}}+5\sigma,\overline{\mathrm{X}}+5\sigma)$, where r is chosen by multiplying the sample sizes by 0.1, $\bar{X}$ is the mean of the data set, and σ is the standard deviation of the data set.

Table 1 shows the simulation results for the data set with outliers generated from outlier model. This table contains bias and RMSE values of the parameter estimates. We can observe from this table that the OBR estimators perform better for almost all cases than the ML estimators according to the RMSE values.

Table 1. Bias and RMSE values for different sample sizes with outliers.

			Parameter(α)		Parameter(β)	
			ML	OBR	ML	OBR
n	α	β	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)
25	1.5	1	-0.5494(0.5585)	-0.3327(0.3576)	-0.0612(0.01205)	0.0202(0.1416)
	1	2	-0.2683(0.2816)	-0.1177(0.1628)	-0.4695(0.5042)	-0.1786(0.3255)
	0.8	0.35	-0.1742(0.1880)	-0.0761(0.1160)	0.0604(0.09067)	0.0315(0.0895)
50	1.5	1	-0.6060(0.6096)	-0.4347(0.4424)	-0.0843(0.1085)	-0.0168(0.0895)
	1	2	-0.3107(0.3165)	-0.1856(0.2003)	-0.5547(0.5665)	-0.3192(0.3575)
	0.8	0.35	-0.2091(0.2147)	-0.1247(0.1369)	0.0680(0.0871)	0.0451(0.0750)
100	1.5	1	-0.6239(0.6256)	-0.4550(0.4585)	-0.0829(0.0946)	-0.0164(0.0620)
	1	2	-0.3254(0.3282)	-0.2026(0.2092)	-0.5690(0.5744)	-0.3390(0.3559)
	0.8	0.35	-0.2224(0.2251)	-0.1392(0.1447)	0.0761(0.0854)	0.0539(0.0684)

4. Conclusions

In this study, we have proposed robust estimators for the parameters of the power Lindley distribution using the OBR estimation method. We have also provided a simulation study. From this simulation study, we have observed that since our proposed estimators are not failed according to the outliers, they can be used as an alternative to the ML estimators.

References

[1] Ghitany, M.E., Al-Mutairi, D.K., Balakrishnan, N., and Al-Enezi, L.J. (2013). Power Lindley distribution and associated inference. *Computational Statistics & Data Analysis*, *64*, 20-33.

[2] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions.* New York: Wiley.

[3] Lindley, D.V. (1958).Fiducial distributions and Bayes'theorem. *Journal Royal Statistical Society* oft he 20(1):102-107

[4] Ghitany et al. (2008). Lindley distribution and its application. *Mathematics and Computers Simulation,* 78, 493-506

[5] Doğru, F.Z. and Arslan, O. (2016). Optimal B-robust estimators for the parameters of the Burr XII distribution. *Journal of Statistical Computation and Simulation*, 86(6), 1133-1149.

[6] Doğru, F.Z., Bulut, Y.M. and Arslan, O. (2017). Optimal B-robust estimators for the parameters of the generalized half-normal distribution. *REVSTAT-Statistical Journal*, 15(3), 455-471.

[7] Victoria-Feser, M.P. and Ronchetti, E.M. (1994). Robust methods for personal-income distribution models. *The Canadian Journal of Statistics*, 22(3), 247-258.

Estimation Methods for Three-Parameter Burr Type XII Distribution: Hydrological Application

F.G. Akgül[1], B. Şenoğlu[2]

[1]Artvin Çoruh University, Artvin, Turkey, ftm.gul.fuz@artvin.edu.tr
[2]Ankara University, Ankara, Turkey, senoglu@science.ankara.edu.tr

Abstract

In this paper, we consider the estimation of parameters of three-parameter Burr Type XII distribution by using maximum likelihood (ML), least squares (LS), weighted least squares (WLS), Cramer-von Mises (CM), Anderson Darling (AD) and modified Anderson Darling (MAD) methods. The performances of the estimators are compared via Monte-Carlo simulation study. The flexibility of this distribution is investigated by using streamflow data set.

1. Introduction

Burr Type XII distribution was first introduced by [1] as one of the Burr system of distributions. It has common usage in engineering, reliability, hydrology, survival analysis and actuarial science. Besides being symmetric or skew, according to different shape parameters settings, Burr Type XII distribution can also be short tailed or long tailed. These properties provide flexibility for modeling the data sets. The problem of the estimation of unknown parameters of Burr Type XII distribution has been considered by many authors, see [2-4].

The cumulative distribution function (cdf) and probability density function (pdf) of Burr Type XII distribution are

$$F(x) = 1 - \left(1 + \left(\frac{x}{\sigma}\right)^{c}\right)^{-k}, \quad x > 0, \quad c, k, \sigma > 0 \tag{1}$$

and

$$f(x) = \frac{kc}{\sigma}\left(\frac{x}{\sigma}\right)^{c-1}\left(1 + \left(\frac{x}{\sigma}\right)^{c}\right)^{-(k+1)}, \quad x > 0, \quad c, k, \sigma > 0 \tag{2}$$

where c and k are the shape parameters and σ is the scale parameter.

The main aim of this study is to estimate the unknown parameters of Burr Type XII distribution by using maximum likelihood (ML), least squares (LS), weighted least squares (WLS), Cramer-von Mises (CM), Anderson Darling (AD) and modified Anderson Darling (MAD) methods. It should be stated that among from these methods, the CM, AD and MAD methods are called as minimum distance methods. Therefore, we compare the performances of classical and minimum distance estimation methods for estimating the parameters of Burr Type XII distribution.

2. Estimation of the Parameters

In this section, we give brief description of the parameter estimators used in the study.

2.1 Maximum Likelihood Estimators

Let $x_1, x_2, \ldots, x_n$ be a random sample from the Burr Type XII distribution. Then, the log-likelihood $(\ln L)$ function is given by

$$\ln L\,(c, k, \sigma|\boldsymbol{x}) = n\ln k + n\ln c - n\ln\sigma + (c-1)\sum_{i=1}^{n}\ln\left(\frac{x_i}{\sigma}\right) - (k+1)\sum_{i=1}^{n}\ln\left(1 + \left(\frac{x_i}{\sigma}\right)^{c}\right). \tag{3}$$

In order to maximize $\ln L$ function with respect to variables of interest, we obtain the first derivatives of the (3) and equate them to zero as shown below

$$\frac{\partial \ln L}{\partial c} = \frac{n}{c} + \sum_{i=1}^{n} \ln\left(\frac{x_i}{\sigma}\right) - (k+1)\sum_{i=1}^{n} \frac{\left(\frac{x_i}{\sigma}\right)^c \ln\left(\frac{x_i}{\sigma}\right)}{\ln\left(1+\left(\frac{x_i}{\sigma}\right)^c\right)} = 0, \tag{4}$$

$$\frac{\partial \ln L}{\partial k} = \frac{n}{k} - \sum_{i=1}^{n} \ln\left(1+\left(\frac{x_i}{\sigma}\right)^c\right) = 0, \tag{5}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} - \frac{n(c-1)}{\sigma} + \frac{c(k+1)}{\sigma}\sum_{i=1}^{n} \frac{\left(\frac{x_i}{\sigma}\right)^c}{1+\left(\frac{x_i}{\sigma}\right)^c} = 0. \tag{6}$$

It is obvious that the ML estimators of c, k and σ cannot be obtained in explicit form. Therefore, we resort to iterative methods.

2.2 Least Squares Estimators

Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ be the order statistics of a random sample of size n from Burr Type XII distribution. The LS estimators of c, k and σ are found by minimizing following equation

$$\sum_{i=1}^{n} \left(F(x_{(i)}) - \frac{i}{n+1}\right)^2 \tag{7}$$

with respect to the parameters of interest. Here, $\frac{i}{n+1}\ (i = 1, \ldots, n)$ are the expected values of $F(x_{(i)})$. By incorporating the cdf of Burr Type XII distribution given in (1) into (7), the LS estimators of c, k and σ are obtained by minimizing following equation

$$\sum_{i=1}^{n} \left(1 - \left(1+\left(\frac{x_{(i)}}{\sigma}\right)^c\right)^{-k} - \frac{i}{n+1}\right)^2. \tag{8}$$

2.3 Weighted Least Squares Estimators

The WLS estimators of c, k and σ are found by minimizing following equation with respect to the parameters of interest

$$\sum_{i=1}^{n} w_i \left(F(x_{(i)}) - \frac{i}{n+1}\right)^2. \tag{9}$$

Here, $w_i = (n+1)^2(n+2)/i(n-i+1)\ \ (i = 1, \ldots, n)$. By incorporating the cdf of Burr Type XII distribution given in (1) into (9), the WLS estimators of c, k and σ are obtained by minimizing following equation

$$\sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left(1 - \left(1+\left(\frac{x_{(i)}}{\sigma}\right)^c\right)^{-k} - \frac{i}{n+1}\right)^2. \tag{10}$$

2.4 Cramer-von Mises Estimators

The CM estimators of c, k and σ are obtained by minimizing following equation with respect to the parameters of interest

$$\frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_{(i)}) - \frac{2i-1}{2n}\right)^2. \tag{11}$$

By incorporating the cdf of Burr Type XII distribution given in (1) into (11), the CM estimators of c, k and σ are obtained by minimizing following equation

$$\frac{1}{12n}+\sum_{i=1}^{n}\left(1-\left(1+\left(\frac{x_{(i)}}{\sigma}\right)^{c}\right)^{-k}-\frac{2i-1}{2n}\right)^{2}. \tag{12}$$

2.5 Anderson Darling Estimators

The AD estimators of c, k and σ are obtained by minimizing following equation with respect to the parameters of interest

$$-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\log\left[F\left(x_{(i)}\right)\left(1-F\left(x_{(j)}\right)\right)\right], \tag{13}$$

where $j=n-i+1$. By incorporating the cdf of Burr Type XII distribution given in (1) into (13), the AD estimators of c, k and σ are obtained by minimizing following equation

$$-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\log\left[\left(1-\left(1+\left(\frac{x_{(i)}}{\sigma}\right)^{c}\right)^{-k}\right)\left(\left(1+\left(\frac{x_{(j)}}{\sigma}\right)^{c}\right)^{-k}\right)\right]. \tag{14}$$

2.6 Modified Anderson Darling Estimators

The MAD estimators of c, k and σ are obtained by minimizing following equation with respect to the parameters of interest

$$\frac{n}{2}-2\sum_{i=1}^{n}F\left(x_{(i)}\right)-\sum_{i=1}^{n}\left(\frac{2n-2i-1}{n}\right)\log\left(1-F\left(x_{(i)}\right)\right). \tag{15}$$

By incorporating the cdf of Burr Type XII distribution given in (1) into (15), the MAD estimators of c, k and σ are obtained by minimizing following equation

$$\frac{n}{2}-2\sum_{i=1}^{n}\left(1-\left(1+\left(\frac{x_{(i)}}{\sigma}\right)^{c}\right)^{-k}\right)-\sum_{i=1}^{n}\left(\frac{2n-2i-1}{n}\right)\log\left(\left(1+\left(\frac{x_{(i)}}{\sigma}\right)^{c}\right)^{-k}\right). \tag{16}$$

3. Simulation Study

In this section, we present the results of Monte-Carlo simulation study to compare the performances of the different estimation methods. Comparisons are done based on defficieny criterion defined as $Def\left(\hat{c},\hat{k},\hat{\sigma}\right)=MSE(\hat{c})+MSE\left(\hat{k}\right)+MSE(\hat{\sigma})$. Here, MSE presents the mean squares error of the estimators.

Table 1. Defficiencies of the estimators

(c,k,σ)	n	ML	LS	WLS	CM	AD	MAD
(2,1,1)	50	0.0108	0.1209	0.1301	0.1201	0.0104	0.0366
	100	0.0067	0.0764	0.0839	0.0745	0.0101	0.0368
(2,3,3)	50	0.0117	0.2888	0.3618	0.2916	0.0111	0.4151
	100	0.0113	0.2703	0.3066	0.2687	0.0106	0.2169

It is obvious from Table 1 that the ML estimators demonstrate the best performances with the lowest defficieny values. They are followed by AD estimators. It should be stated that the WLS and MAD estimators do not perform well for all sample sizes and parameter settings.

4. Real Data Application

Now, we model Meriç river (Turkey) streamflow data which is recorded from 1982 to 2012 using Burr Type XII distribution. Before starting analysis, we first ensure that Burr Type XII distribution provides good fit for the data set. Then, we obtain the ML, LS, WLS, CM, AD and MAD estimates of the parameters and determine best

fitting model which is constructed based on these estimates via model selection criteria. They are Akaike information criterion (AIC), Bayesian information criterion (BIC) and corrected AIC (AICc).

Table 2. Parameter estimates and model selection criteria values for streamflow data set.

Method	$\hat{c}$	$\hat{k}$	$\hat{\sigma}$	AIC	BIC	AICc
ML	3.541	0.592	490.945	465.823	470.125	466.712
LS	2.978	0.606	484.433	466.897	471.199	467.785
WLS	3.155	0.596	486.865	466.345	470.647	467.234
CM	3.077	0.626	497.188	466.368	470.670	467.257
AD	3.223	0.617	497.671	466.058	470.360	466.946
MAD	2.460	1.082	668.514	466.785	471.087	467.674

According to Table 2 that the ML estimates have the smallest AIC, BIC and AICc values. In other words, Burr Type XII model based on ML estimates is the most appropriate model among the others. They are followed by AD, WLS and CM estimates.

References

[1] Burr, I.W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13, 215-232.

[2] Shao, Q. (2004). Notes on maximum likelihood estimation for the three-parameter Burr XII distribution, *Computational Statistics & Data Analysis*, 45, 675-687.

[3] Akgül, F.G., Şenoğlu, B. (2018). The comparisons of the parameter estimation methods for Burr XII distribution: Hydrological Applications. ACUBAP, 2016.F13.02.02.

[4] Akgül, F.G., Acıtaş, Ş. Şenoğlu, B. (2019). Estimation of the location and the scale parameters of Burr Type XII distribution. *Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat.*, 68(1), 1030-1044.

On the Asymptotic Stability of Riemann–Liouville Fractional Neutral Type Neural Networks with Time Delay

Y. Altun[1]

[1]Yuzuncu Yil University, Van, Turkey, yeneraltun@yyu.edu.tr

Abstract

This paper, investigates the asymptotic stability of Riemann–Liouville fractional neutral type neural networks with time delay. It is assumed that activation functions are globally Lipschitz continuous. The Lyapunov-Krasovskii functional is utilized to achieve the desired results. The linear matrix inequalities (LMIs) approach which can be easily solved are developed to derive sufficient conditions ensuring the Riemann–Liouville fractional neutral type neural networks is asymptotic stable. Numerical examples are provided to illustrate that the proposed method is flexible and efficient in terms of computation and to demonstrate the feasibility of the theoretical results by using MATLAB-Simulink.

1. Introduction

Recently fractional calculus has been of considerable interest in many scientific areas. In particular, the interest in stability analysis of various fractional differential systems and stability analysis of neural networks with time delays have been widely studied much attention in many scientific areas such as engineering techniques fields, economics, physics, signal processing, pattern recognition, power systems, parallel computing, associative memories, mechanics of structures and materials, and other scientific areas (see [1-7] and the references therein).

As known, neutral type fractional systems have a more general class than those of the delayed type. Hence stability of these systems proves to be a more complex question because the system includes the derivative of the retarded state. When one checks the relative literature, especially, in the past a few years increased attention has been devoted to the problem of delay dependent or independent stability via different approaches for fractional neutral systems. In this direction, the stability issue of solutions to systems considered is one of the important problems both theoretically and practically.

The aim of the present work is to investigate the asymptotic stability of Riemann–Liouville fractional neutral type neural networks with time delay. In study, the technique of the proof involves some basic inequalities, LMIs and Lyapunov- Krasovskii functional method.

2. Problem description

In the current paper, motivated by above discussions, we consider a class of Riemann–Liouville (RL) fractional neutral type neural networks with time delay defined by

$$ {}_{t_0}^{RL}D_t^q x(t) = -Ax(t) + Bg(x(t)) + Cg(x(t-\tau)) + E\,{}_{t_0}^{RL}D_t^q x(t-\tau),\ t \geq 0, \tag{2.1} $$

with the initial condition

$$ {}_{t_0}D_t^{q-1} x(t) = \vartheta(t), \quad t \in [-\tau, 0], \tag{2.2} $$

where $x = [x_1, x_2, \ldots, x_n]^T \in \Re^n$ is the state vector of neural network, $\tau > 0$ is a constant delay, ${}_{t_0}^{RL}D_t^q x(.)$ denotes a q order Riemann–Liouville fractional derivative of $x(.)$ with $q \in (0,1)$, $A = diag\{a_1, a_2, \ldots, a_n\}$ is a positive definite diagonal $a_i > 0$. The matrices $B = (b_{ij})_{n\times n}, C = (c_{ij})_{n\times n}$ and $E = (e_{ij})_{n\times n}$ are the connection matrices representing the weight coefficients of the neurons. $g(x(t)) = [g_1(x(t)), g_2(x(t)), \ldots, g_n(x(t))]^T$ denotes the activation functions of the neurons with $g_i(0) = 0$ and satisfies the following Lipschitz condition,

$$ |g_i(u_1) - g_i(u_2)| \leq M_i |u_1 - u_2|, \quad \forall u_1, u_2 \in \Re, \quad i = 1, 2, \ldots, n, \tag{2.3} $$

where $M_i \in \Re^{n\times n}$ are known positive constant matrices.

Definition 2.1 ([5]) The Riemann–Liouville fractional integral and derivative for a function g are defined as, respectively

$${}_{t_0}D_t^{-q}\{g(t)\} = \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}g(s)ds, \quad t \geq t_0,$$

$${}_{t_0}^{RL}D_t^{p}\{g(t)\} = \frac{1}{\Gamma(n-p)}\frac{d^n}{dt^n}\int_{t_0}^{t}\frac{g(s)}{(t-s)^{p-n+1}}ds, \quad (0 \leq n-1 \leq p < n),$$

where $n \in Z^+$ and Γ denotes the Gamma function.

Lemma 2.1 ([3]) If $p > q > 0,$ then the formulas:
${}_{t_0}^{RL}D_t^{p}\left\{{}_{t_0}D_t^{-q}g(t)\right\} = {}_{t_0}^{RL}D_t^{p-q}g(t)$
holds for "sufficiently good" functions $g(t)$. In particular, this relation holds if $g(t)$ is integrable.

Lemma 2.2 ([5]) Let $x(t) \in \Re^n$ be a vector of differentiable function for $\forall t \geq t_0.$ Then the following relationship holds

$$\frac{1}{2}{}_{t_0}^{RL}D_t^{q}\left\{x^T(t)Kx(t)\right\} \leq x^T(t)K\,{}_{t_0}^{RL}D_t^{q}\{x(t)\}, \quad 0 < q < 1,$$

where $K \in \Re^{n\times n}$ is a constant, square, symmetric and positive semi-definite matrix.

3. Main results

Theorem 3. 1. The zero solution of system (2.1) is asymptotic stable, if $\|E\| < 1$ and there exist symmetric positive definite matrices P, Q, R, S and two diagonal matrices $M > 0$ and $N > 0$ such that the following LMI holds

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} & \Omega_{14} \\ \Omega_{12}^T & \Omega_{22} & \Omega_{23} & \Omega_{24} \\ \Omega_{13}^T & \Omega_{23}^T & \Omega_{33} & \Omega_{34} \\ \Omega_{14}^T & \Omega_{24}^T & \Omega_{34}^T & \Omega_{44} \end{bmatrix} < 0, \tag{3.1}$$

where $\Omega_{11} = -PA - A^TP + \Sigma M\Sigma + 2\Sigma N + A^T(Q+\tau^2 S)A, \Omega_{12} = PB + N - A^T(Q+\tau^2 S)B,$
$\Omega_{13} = PC - A^T(Q+\tau^2 S)C, \Omega_{14} = PE - A^T(Q+\tau^2 S)E, \Omega_{22} = -M + R + B^T(Q+\tau^2 S)B,$
$\Omega_{23} = B^T(Q+\tau^2 S)C, \Omega_{24} = B^T(Q+\tau^2 S)E, \Omega_{33} = -R + C^T(Q+\tau^2 S)C, \Omega_{34} = C^T(Q+\tau^2 S)E,$
$\Omega_{44} = -Q + E^T(Q+\tau^2 S)E, \Sigma = diag(\sigma_1, \sigma_1, \ldots, \sigma_n).$

Proof. Let us consider a Lyapunov–Krasovskii functional as follows:

$$V(t) = {}_{t_0}D_t^{q-1}\left(x^T(t)Px(t)\right) + \int_{t-\tau}^{t}({}_{t_0}^{RL}D_t^{q}x(s))^T Q({}_{t_0}^{RL}D_t^{q}x(s))ds$$
$$+ \int_{t-\tau}^{t} g^T(x(s))Rg(x(s))ds + \tau\int_{-\tau}^{0}\int_{t+\beta}^{t}({}_{t_0}^{RL}D_s^{q}x(s))^T S({}_{t_0}^{RL}D_s^{q}x(s))ds d\beta. \tag{3.2}$$

It follows from (2.2) and Lemmas 2.1 and 2.2 that we get the time derivative $V(t)$ along the trajectories of system fractional (2.1), as follows

$$\dot{V}(t) = {}_{t_0}^{RL}D_t^{q}(x^T(t)Px(t)) \leq 2x^T(t)P\,{}_{t_0}^{RL}D_t^{q}(x(t)) + ({}_{t_0}^{RL}D_t^{q}x(t))^T Q({}_{t_0}^{RL}D_t^{q}x(t))$$
$$-({}_{t_0}^{RL}D_t^{q}x(t-\tau))^T Q({}_{t_0}^{RL}D_t^{q}x(t-\tau)) + g^T(x(t))Rg(x(t))$$
$$- g^T(x(t-\tau))Rg(x(t-\tau)) + \tau^2({}_{t_0}^{RL}D_t^{q}x(t))^T S({}_{t_0}^{RL}D_t^{q}x(t))$$

$$
\begin{aligned}
&-\tau\int_{t-\tau}^{t}({}^{RL}_{t_0}D_s^q x(s))^T S({}^{RL}_{t_0}D_s^q x(s))ds\\
&\le x^T(t)(-PA-A^TP)x(t)+2x^T(t)PBg(x(t))+2x^T(t)PCg(x(t-\tau))\\
&\quad+2x^T(t)PE\,{}^{RL}_{t_0}D_t^q x(t-\tau)+({}^{RL}_{t_0}D_t^q x(t))^T Q({}^{RL}_{t_0}D_t^q x(t))\\
&\quad-({}^{RL}_{t_0}D_t^q x^T(t-\tau))Q({}^{RL}_{t_0}D_t^q x(t-\tau))+g^T(x(t))Rg(x(t))\\
&-g^T(x(t-\tau))Rg(x(t-\tau))+g^T(x(t))Mg(x(t))-2x^T(t)Ng(x(t))\\
&\quad-g^T(x(t))Mg(x(t))+2x^T(t)Ng(x(t))+\tau^2({}^{RL}_{t_0}D_t^q x(t))^T S({}^{RL}_{t_0}D_t^q x(t)).
\end{aligned}
\tag{3.3}
$$

From (3.3), noting that $M>0$ and $N>0$ are diagonal and using (2.3), we obtain

$$
g^T(x(t))Mg(x(t))\le x^T(t)\Sigma M\Sigma x(t),
\tag{3.4}
$$

and

$$
-x^T(t)Ng(x(t))\le x^T(t)\Sigma N x(t).
\tag{3.5}
$$

Taking into account (3.3), we get

$$
\begin{aligned}
&({}^{RL}_{t_0}D_t^q x(t))^T Q({}^{RL}_{t_0}D_t^q x(t))+\tau^2({}^{RL}_{t_0}D_t^q x(t))^T S({}^{RL}_{t_0}D_t^q x(t))\\
&\quad=[-Ax(t)+Bg(x(t))+Cg(x(t-\tau))+E\,{}^{RL}_{t_0}D_t^q x(t-\tau)]^T(Q+\tau^2S)\\
&\quad\times[-Ax(t)+Bg(x(t))+Cg(x(t-\tau))+E\,{}^{RL}_{t_0}D_t^q x(t-\tau)]\\
&\quad=x^T(t)A^T(Q+\tau^2S)Ax(t)-x^T(t)A^T(Q+\tau^2S)Bg(x(t))\\
&\quad-x^T(t)A^T(Q+\tau^2S)Cg(x(t-\tau))-x^T(t)A^T(Q+\tau^2S)E\,{}^{RL}_{t_0}D_t^q x(t-\tau)\\
&\quad-g^T(x(t))B^T(Q+\tau^2S)Ax(t)+g^T(x(t))B^T(Q+\tau^2S)Bg(x(t))\\
&\quad+g^T(x(t))B^T(Q+\tau^2S)Cg(x(t-\tau))+g^T(x(t))B^T(Q+\tau^2S)E\,{}^{RL}_{t_0}D_t^q x(t-\tau)\\
&\quad-g^T(x(t-\tau))C^T(Q+\tau^2S)Ax(t)+g^T(x(t-\tau))C^T(Q+\tau^2S)Bg(x(t))\\
&\quad+g^T(x(t-\tau))C^T(Q+\tau^2S)Cg(x(t-\tau))+g^T(x(t-\tau))C^T(Q+\tau^2S)E\,{}^{RL}_{t_0}D_t^q x(t-\tau)\\
&\quad-({}^{RL}_{t_0}D_t^q x(t-\tau))^T E^T(Q+\tau^2S)Ax(t)+({}^{RL}_{t_0}D_t^q x(t-\tau))^T E^T(Q+\tau^2S)Bg(x(t))\\
&\quad+({}^{RL}_{t_0}D_t^q x(t-\tau))^T E^T(Q+\tau^2S)Cg(x(t-\tau))+({}^{RL}_{t_0}D_t^q x(t-\tau))^T E^T(Q+\tau^2S)E\,{}^{RL}_{t_0}D_t^q x(t-\tau).
\end{aligned}
\tag{3.6}
$$

From (3.3)-(3.6), we obtain $+g^T(x(t-\tau))C^T(Q+\tau^2S)E\,{}^{RL}_{t_0}D_t^q x(t-\tau)$

$$
\dot V(t)\le\xi^T(t)\Omega\xi(t),
$$

where $\xi^T(t)=[x^T(t)\quad g^T(x(t))\quad g^T(x(t-\tau))\quad({}^{RL}_{t_0}D_t^q x(t-\tau))^T]$ and Ω is defined (3.1). If $\Omega<0,\dot V(t)$ is negative definite for $\xi(t)\neq 0.$ This implies that the system fractional (2.1) is asymptotic stable. Thus, the proof is complete.

4. Numerical examples

Example 4.1 As a special case of system (2.1), we consider the following Riemann–Liouville fractional neutral type neural networks with time delay

$$
{}^{RL}_{t_0}D_t^q x(t)=-Ax(t)+Bg(x(t))+Cg(x(t-\tau))+E\,{}^{RL}_{t_0}D_t^q x(t-\tau),\ t\ge 0,
\tag{4.1}
$$

where $0<q\le 1,\ \tau=0.15,\ \ x(t)=[x_1(t),\ x_2(t)]^T,$

$$
A=\begin{bmatrix}4.1 & 0\\ 0 & 5\end{bmatrix},\ B=\begin{bmatrix}0.8 & -1.4\\ -1.3 & 1.3\end{bmatrix},\ C=\begin{bmatrix}0.8 & 0.6\\ 0.4 & 0.7\end{bmatrix},E=\begin{bmatrix}0.12 & 0\\ 0 & 0.08\end{bmatrix}.
$$

Let us choose $\sigma_1=0.42,\ \sigma_2=0.18,$

$$
M=\begin{bmatrix}18.6 & 0\\ 0 & 18.6\end{bmatrix},N=\begin{bmatrix}0.24 & 0\\ 0 & 0.24\end{bmatrix},P=\begin{bmatrix}7.6 & 0\\ 0 & 7.8\end{bmatrix},\ Q=\begin{bmatrix}0.42 & 0.14\\ 0.14 & 0.42\end{bmatrix},
$$

$$R=\begin{bmatrix}6.4 & 0\\ 0 & 6.4\end{bmatrix},\ \text{and}\ S=\begin{bmatrix}6.9 & 0.1\\ 0.1 & 6.9\end{bmatrix}.$$

Under the above assumptions, all eigenvalues of the LMI described by (3.1) are $\lambda_{\max}(\Omega)\le -0.2573$ by using MATLAB-Simulink. As a result, condition (3.1) holds, which implies that the zero solution of system (4.1) is asymptotic stable according to Theorem 3.1.

Example 4.2 As a special case of system (2.1), we consider the following Riemann–Liouville fractional neutral type neural networks with time delay

$$ {}_{t_0}^{RL}D_t^q x(t) = -Ax(t) + Bg(x(t)) + Cg(x(t-\tau)) + E\,{}_{t_0}^{RL}D_t^q x(t-\tau),\ t\ge 0, \qquad (4.2)$$

where $0<q\le 1,\ \tau=0.2,\ x(t)=[x_1(t),\ x_2(t)]^T$,

$$A=\begin{bmatrix}5.6 & 0\\ 0 & 6.2\end{bmatrix},\ B=\begin{bmatrix}1.32 & -1.2\\ -0.8 & 1.28\end{bmatrix},\ C=\begin{bmatrix}0.6 & 0.4\\ 0.9 & 0.7\end{bmatrix},\ E=\begin{bmatrix}0.24 & 0\\ 0 & 0.16\end{bmatrix}.$$

Let us choose $\tilde{\sigma}_1=0.32,\ \tilde{\sigma}_2=0.28,$

$$\tilde{M}=\begin{bmatrix}16.2 & 0\\ 0 & 16.2\end{bmatrix},\tilde{N}=\begin{bmatrix}2.2 & 0\\ 0 & 2.5\end{bmatrix},\tilde{P}=\begin{bmatrix}5.6 & 0\\ 0 & 7.8\end{bmatrix},\ \tilde{Q}=\begin{bmatrix}0.3 & 0.26\\ 0.26 & 0.6\end{bmatrix},$$

$$\tilde{R}=\begin{bmatrix}5.8 & 0.1\\ 0.1 & 5.8\end{bmatrix}\ \text{and}\ \tilde{S}=\begin{bmatrix}8.2 & 0.2\\ 0.2 & 6.4\end{bmatrix}.$$

Under the above assumptions, all eigenvalues of the LMI described by (3.1) are $\lambda_{\max}(\Omega)\le$ -0.0925 by using MATLAB-Simulink. As a result, condition (3.1) holds, which implies that the zero solution of system (4.2) is asymptotic stable according to Theorem 3.1.

References

[1] Altun, Y. (2019). New Results on the Exponential Stability of Class Neural Networks with Time-Varying Lags, BEU Journal of Science, 8(2), 443-450.

[2] Arik, S. (2004). An analysis of exponential stability of delayed neural networks with time varying delays. *Neural Networks*, 17, 1027–1031.

[3] Kilbas, A. A., Srivastava, H. M., Trujillo, J. J. (2006). Theory and Application of Fractional Differential Equations, Elsevier, New York, USA.

[4] Li, T., Luo, Q., Sun, C., and Zhang, B. (2009). Exponential stability of recurrent neural networks with time-varying discrete andd istributed delays. *Nonlinear Anal. Real World Applications*, 10(4), 2581–2589.

[5] Liu, S., Wu, X. Zhang, Y. J., Yang, R. (2017). Asymptotical stability of Riemann–Liouville fractional neutral systems, Appl. Math. Lett. 69, 168–173.

[6] Podlubny, I. (1999). Fractional Differential Equations, Academic Press., New York, USA.

[7] Tian, J. K., Zhong, S. and Wang, Y. (2012). Improved exponential stability criteria for neural networks with time-varying delays, *Neurocomputing*, vol. 97, pp. 164–173.

Game Based Computational Thinking Skills in Preschool Period

Ezgi Pekel[1], Çetin Güler[2]

[1]Van Yüzüncü Yıl University, Van, Turkey, ezgipekel.ep@gmail.com
[2]Van Yüzüncü Yıl University, Van, Turkey, cetin.guler@gmail.com

Abstract

One of the recent topics in the field of education is the computational thinking skills of individuals. The necessity of using electronic technology in individuals' Computational Thinking (CT) skills development process can be seen as an acceptable perception. In addition to electronic environments, concrete objects can be considered as effective in terms of acquiring CT skills in students. The use of different practices and environments can be suggested to gain CT skills in preschool period. Play can be thought of as a useful activity to improve CT skill in preschool students. In this study, it is aimed to make a game-based application and to examine the application in order to gain CT skills by using concrete tools at preschool level. In this study, a single group pre-test post-test semi-experimental design was employed. Twenty-two students attending a public kindergarten of Van İpekyolu District Directorate of National Education were included in the study. In the study, In-Game Experience Observation and Post-Game Evaluation scales and Evaluation Form were used to collect data. In addition, an "Information Form" was prepared and used in order to obtain the demographic characteristics of the students and information about the completion of the game. It is hoped that the findings obtained in this study will contribute to the studies and applications to be made to gain computational thinking skills at preschool level.

1. Introduction

One of the recent topics in the field of education is the computational thinking skills of individuals. For the first time, the concept of computational thinking (CT) was used in the form of Wing [1] Computational Thinking". There are equivalents such as computational thinking [2], computational thinking [3], computational thinking [4] and computational thinking [5]. Computational Thinking (CT) is a solution-oriented, multi-angle and formal thinking. CT skills are defined by various concepts such as problem solving, technology, thinking types, creativity, and cooperative learning and communication skills [6].

Computational thinking skills are seen as a basic skill that must be gained to individuals in the 21st century [5, 7, 8, 9, 10, 11, 12]. In order to acquire and develop CT skills, it is advocated that it can be continued from preschool to university years [7, 8, 10, 12, 13]. The ideas that CT skills will become important skills for the individual bring many arguments and suggestions about how to gain them.

The use of different practices and environments can be suggested to gain CT skills in preschool period. It has been argued that CT skills can be acquired in preschoolers by using concrete storytelling [14], user interface [15, 16, 17], robotic coding [18, 19], visual programming tool ScratchJr [20, 21] and mathematics learning [22]. Teaching technological products can be a difficult process for teachers. At the same time, the learning and introduction of these products may require a time-consuming and labor-intensive process for students of all levels. This may mean more time and effort as preschool students cannot think abstractly. In this period, since the students understand the concrete experiences more easily, the use of real objects is emphasized in the curriculum and the strategy of learning by doing is experienced in the forefront in order to make teaching more effective. Kanaki and Kalogiannakis [16] argue that the game enables children to work in groups, learn from their mistakes, and acquire skills such as classification, analysis, synthesis, evaluation and problem solving. Examples of games for this development are puzzles, pairings and group games [16]. Therefore, play can be considered as a useful activity to improve CT skill in preschool students.

2. Method

In this study, a single group pre-test post-test semi-experimental design was employed.

3. Study Group

This study was conducted with 22 students attending a public kindergarten in Van İpekyolu District Directorate of National Education. 54% of the study group was male and 46% was female. The ages of the students ranged

between 5-6 years. In the selection of this study group, it is considered that these students know the directions and numbers.

4. Data Collection Tools

In the study, In-Game Experience Observation and Post-Game Evaluation scales and Evaluation Form were used to collect data. In addition, an Information Form was prepared and used in order to obtain the personal characteristics of the students and information about the completion of the game.

5. Results

It was revealed that students had an average score of approximately three on the in-game (X = 2.48) and post-game (X = 2.74) scales. When the data about the time spent by the students during the game was examined, it was seen that there were seven students in 3-6 minutes, nine students in 7-9 minutes and six students in 10-16 minutes. When the data on how many trials the students found the right combination were examined, it was seen that they generally completed the right combination in two trials.

5. Conclusion and Discussion

Measurements revealed that students were enjoying the game, according to the In-Game Experiences and the Post-Game Evaluations. When the In-Game Experience and the Post-Game Evaluation scores of the students were examined, it was observed that they scored close to three points. The scores of the students can be interpreted as not having too much difficulty in the game and enjoying the game.

In-Game Experience and the Post-Game Evaluation forms were filled with the answers of the students and the observation of the teacher. A significant high relationship was observed between the mean scores of the two forms. This situation supports the idea that students enjoy.

References

[1] Wing, J.M., Computational thinking. Communications of the ACM, 2006. 49(3): p. 33-35.

[2] Korkmaz, Ö., et al., Bireylerin Bilgisayarca Düşünme Becerilerinin Farklı Değişkenler Açısından İncelenmesi. Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 2015. 34(2): p. 68-87.

[3] Yağcı, M., A valid and reliable tool for examining computational thinking skills. Education and Information Technologies, 2019. 24: p. 929-951.

[4] Oluk, A., Ö. Korkmaz, and H.A. Oluk, Effect of Scratch on 5th Graders' Algorithm Development and Computational Thinking Skills. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2018.

[5] Şahiner, A. and S.B. Kert, Komputasyonel düşünme kavramı ile ilgili 2006-2015 yılları arasındaki çalışmaların incelenmesi. EJOSAT: European Journal of Science and Technology, Avrupa Bilim ve Teknoloji Dergisi, 2016. 5(9).

[6] Haseski, H.I., U. İlic, and U. Tugtekin, Defining a New 21st Century Skill-Computational Thinking: Concepts and Trends. International Education Studies, 2018. 11(4).

[7] Wing, J.M., Computational thinking and thinking about computing. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2008. 366(1881): p. 3717-3725.

[8] Wing, J., Research notebook: Computational thinking—What and why. The Link Magazine, 2011: p. 20-23.

[9] Korkmaz, Ö., C. Karaçaltı, and R. Çakır, Öğrencilerin Programlama Başarılarının Bilgisayarca-Eleştirel Düşünme ile Problem Çözme Becerileri Çerçevesinde İncelenmesi. Amasya Üniversitesi Eğitim Fakültesi Dergisi, 2018. 7(2): p. 343-370.

[10] Haseski, H.İ., U. İlic, and U. Tuğtekin, Defining a New 21st Century Skill-Computational Thinking: Concepts and Trends. International Education Studies, 2018. 11(4): p. 29-42.

[11] Doleck, T., et al., Algorithmic thinking, cooperativity, creativity, critical thinking, and problem solving: exploring the relationship between computational thinking skills and academic performance. Journal of Computers in Education, 2017. 4(4): p. 355-369.

[12] Kert, S.B., S. Yeni, and A. Şahiner, Komputasyonel Düşünme İle İlişkilendirilen Alt Becerilerin İncelenmesi.

[13] Atman Uslu, N., F. Mumcu, and F. Eğin, Görsel Programlama Etkinliklerinin Ortaokul Öğrencilerinin Bilgi-İşlemsel Düşünme Becerilerine Etkisi. Ege Eğitim Teknolojileri Dergisi, 2018. 2(1): p. 19-31.

[14] Soleimani, A., D. Herro, and K.E. Green, CyberPLAYce – A Tangible, Interactive Learning Tool Fostering Children's Computational Thinking through Storytelling 2019.

[15] Gomes, T.C.S. and T.P. Falcão, Exploring an approach based on digital games for teaching programming concepts to young children. International Journal of Child-Computer Interaction, 2018. 16: p. 77-84.
[16] Kanaki, K. and M. Kalogiannakis, Introducing fundamental object-oriented programming concepts in preschool education within the context of physical science courses. Educ Inf Technol 2018.
[17] Wang, D., T. Wang, and Z. Liu, A tangible programming tool for children to cultivate computational thinking. The Scientific World Journal, 2014. 2014.
[18] Demertzi, E., et al., Online Learning Facilities to Support Coding and Robotics Courses for Youth. International Journal of Engineering Pedagogy (iJEP), 2018. 8(3): p. 69-80.
[19] Lieto, M.C.D., et al., Educational Robotics intervention on Executive Functions in preschool children: A pilot study. 2017. 71: p. 16-23.
[20] Papadakis, S., M. Kalogiannakis, and N. Zaranis, Developing fundamental programming concepts and computational thinking with ScratchJr in preschool education: A case study. International Journal of Mobile Learning and Organisation, 2016. 10(3): p. 187-202.
[21] Sullivan, A., M. Bers, and A. Pugnali, The Impact of User Interface on Young Children's Computational Thinking. Journal of Information Technology Education: Innovations in Practice, 2017. 16: p. 171-193.
[22] Palmér, H., Programming in preschool—with a focus on learning mathematics. International Research in Early Childhood Education 2017. 8(1): p. 75-87.

Modeling Evaluation Criteria for Resilient IT Project Supplier Selection

M. Nasrollahi[1], M. Sadraei[2]

[1]*Imam Khomeini International University, Qazvin, Iran, m.nasrollahi@soc.ikiu.ac.ir*
[2]*Islamic Azad University, Science and Research Branch, Tehran, Iran, mahmoodsadraie@gmail.com*

Abstract

The purpose of this paper is modeling the evaluation criteria for resilient IT project supplier selection and analysis of the interactions among them. To do this, we have identified 19 criteria from literature and consequent deliberations with experts from the industry. Interpretive Structural Modeling (ISM) has been utilized to distinguish driving criteria. The criterion modeling using ISM demonstrates that research and development and network scale are the most influential/driving criteria. Then again, the supplier's brand value and quality have been generally identified as the most dependent criteria for resilient IT project supplier selection. In this research, the MCDM model of evaluation criteria has been formulated based on the inputs from a few domain experts which may not reflect the opinion of the whole practitioners' community. The findings can enable the decision makers to appropriately choose the desired and drop undesired criterion in for resilient IT project supplier selection to improve the performance of the organization.

1. Introduction

Implementation of IT by organizations has been associated with many advantages such as increased integration in the production of data and information, improved operation speed, improved access and exchange of processed information, etc. Additionally, in today's competitive market, companies have realized the importance of selecting proper suppliers that can supply their requirements with the desired quality and within a scheduled timeframe. One of the critical challenges faced by the managers for maintaining the competitive advantage is the selection of strategic partners for supplying hardware and software products in a timely and cost-effective manner. Buyer-Supplier relationships based solely on price are no longer acceptable for suppliers and organizations that wish to practice the latest innovations in supply chain management. Recent emphasis has also been on other important strategic factors such as quality, delivery, and flexibility. On this basis, the number of outsourced projects in the design and implementation of IT-based systems are on the rise and a considerable portion of IT projects are being outsourced to the suppliers. Owing to this in mind, identifying the evaluation criteria and diagnosing their inter-relationships for supplier selection with the goal of outsourcing IT services seems a necessity. Due to many reasons, organizations seek a market that can supply the IS/IT resources of their business requirements; a solution which, in practice, refers to outsourcing. Some researchers have defined outsourcing as the delegation, through a contractual arrangement, of all or any part of the technical resources, the human resources and the management responsibilities associated with providing IT services, to an external supplier (Peppard & Ward, 2016). Disruptions could occur in inter-organizational relationships due to a variety of reasons. Such disruptions are usually caused by external or internal factors. In order to reduce the adverse effects of such risks, suppliers should be chosen which are prepared to face such complications and are able to come up with effective and efficient solutions. A resilient supplier, thus, is one that has the ability to recover quickly from disruptions and ensure customers are minimally affected. Studies suggest that the more recovery capability increases within the organizations, the quicker they return to the normality and the less severe the out-of-schedule event i.e. disruption will likely be. Resilience focuses on adaptability and developing the capability to be prepared against unforeseeable events, to be responsive against disruptions, and to recover quickly from such difficulties (Ponomarov & Holcomb, 2009). System resilience depends on inherent characteristics of the system and, in particular, the system's absorptive capacity, adaptive capacity, and recovery capacity (Proag, 2014).

In supplier selection problems, identifying a set of industry-appropriate criteria to help make this decision is of the essence. As many properties of the organization including its resilience could be affected by the suppliers, especially their resilience, failure on the part of suppliers to meet the required specifications could potentially leave the organization in turbulent and disruptive conditions that expose it to financial or intellectual loss. Managers need to be precise in selecting their supplier to minimize the potentially adverse effects that the supplier may have on the organization. In order to ensure that, recognizing the most fundamental set of selection criteria is necessary. The present study endeavors to model the criteria for the resilient IT project supplier selection. In particular, the study attempts to achieve three objectives: 1) Establishing hierarchy and inter-relationship among criteria using ISM, 2) Using MICMAC analysis for analyzing the driving power and dependence of criteria, and 3) Discussing the theoretical and managerial implication of this research and suggesting future research directions.

2. Modeling the evaluation criteria for resilient IT project supplier selection using ISM

Interpretive Structural Modelling (ISM) is a well-established method for identifying relationships among specific items and dealing with complex issues, which was proposed by J. Warfield in 1974. It is an interactive process that helps to impose order and direction on the complexity of the relationships among the variables of a system into a comprehensive model. In this technique, the main input is the judgments of the decision makers about pairwise dependence relationship between the criteria, represented in a structural self-interaction matrix. For problems, such as understanding the criteria that impact resilient IT project supplier selection, several of them may be impacting each other at different levels. However, the direct and indirect relationships between the criteria impacting the resilient IT project supplier selection are far more accurate than the individual criterion considered in seclusion.

The ISM methodology has been extensively utilized in various areas of management by researchers. Chen & Chen (2014), used ISM to identify the relationship between the resistance factors encountered during the organization innovation process in implementing cloud-based medical systems in hospitals. In another research, a structural model was developed to reflect the interactions among various critical factors impact implementing sustainable construction practice in HOPSCA projects (Yu, Shi, Zuo, & Chen, 2018). Govindan, Kannan, and Haq (2010) present a procedure to hierarchize the criteria used for supplier development based on the relationship between them, according to ISM. Another study began with identifying the criteria to evaluate agile suppliers. Then these factors were ranked and categorized using the ISM (Beikkhakhian, Javanmardi, Karbasian, & Khayambashi, 2015). The purpose of a study was the application of ISM integrated with MCDM techniques for enabling the sustainability supplier selection (Girubha, Vinodh, & Kek, 2016). As the literature shows, probably, modeling of evaluation criteria for resilient IT project supplier selection is neglected and so, in this study, we use ISM for identifying the interrelationships between these criteria. Various steps leading to the development of the ISM model for resilient IT project supplier selection criteria are discussed below.

1- Identification of evaluation criteria for resilient IT project supplier selection. Nineteen criteria for the evaluation of resilient IT project supplier have been identified through literature review and deliberations with domain experts. These are: quality (1), cost-efficiency (2), reputability (3), work experience (4), risk awareness (5), minimum vulnerability against disruptions (6), dispersion of key resources, production and market capacity (7), agility (8), commitment to contract (9), research and development (10), technical capability to adapt to the latest innovations (11), compliance with standards (domestic and international) (12), development of new and alternative technologies (13), prioritizing environmental concerns (14), backup energy resources (15), management stability and specialized staff (16), personnel, information and cyber security (17), network scale (18), and brand value (19).

2- Structural self-interaction matrix (SSIM). Once the criteria had been identified, it was necessary to determine the contextual relationships of 'influences' between the criteria to develop the SSIM. This means one criterion influences another. In total, ten experts were chosen to provide their views and the SSIM (Tab. 1) has developed based on their opinions.

Table 1. SSIM for evaluation criteria for resilient it project supplier selection

C.No.	Criterion	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
1	Quality	V	O	O	O	O	O	A	A	O	A	A	O	O	V	A	O	A	X
2	Cost-efficiency	O	O	O	A	A	O	A	O	V	A	A	X	A	X	A	A	O	
3	Reputability	X	O	A	A	O	A	O	A	O	A	X	A	O	O	O	A		
4	Work experience	V	O	O	O	O	O	O	V	O	A	O	O	O	O	V			
5	Risk awareness	O	O	V	A	V	V	A	O	A	A	O	X	O	V				
6	Minimum vulnerability against disruptions	V	A	X	A	A	O	A	A	A	A	A	V	V					
7	Dispersion of key resources, production and market capacity	O	A	A	V	X	O	A	O	O	O	O	V						
8	Agility	V	O	A	A	A	A	A	X	A	X	V							
9	Commitment to contract	V	A	O	A	O	O	O	O	A	A								
10	Research and development	O	V	V	O	V	V	V	O	V									
11	Technical capability to adapt to the latest innovations	V	O	V	A	V	X	X	O										
12	Compliance with standards (domestic and international)	V	O	A	A	O	O	A											
13	Development of new and alternative technologies	O	X	V	O	V	O												
14	Prioritizing environmental concerns	V	O	O	O	V													
15	Backup energy resources	O	O	O	O														
16	Management stability and specialized staff	O	O	V															
17	Personnel, information and cyber security	V	O																
18	Network scale	O																	

3- Reachability matrix from the SSIM. The reachability matrix is derived from the SSIM. It contains the relationships between the factors in binary form. The SSIM is transformed into the reachability matrix using four rules. Applying the ISM rules to the SSIM, an initial reachability matrix is obtained. From this matrix, a final reachability matrix (Tab. 2) is constructed taking into account the transitivity rule, which states that if a variable 'A' is related to 'B' and 'B' is related to 'C', then 'A' is necessarily related to 'C'.

Table 2. Final reachability matrix (including transitivity)

C.No.	Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Driving power
1	Quality	1	1	1*	0	0	1	1*	1*	0	0	1*	0	0	0	0	0	0	0	1	**7**
2	Cost-efficiency	1	1	0	0	1*	1	1*	1*	1*	0	1	0	1*	1*	1*	0	1*	0	1*	**13**
3	Reputability	1	1*	1	0	0	1*	0	0	1	0	0	0	0	0	0	0	0	0	1	**6**
4	Work experience	1*	1	1	1	1	1*	0	1*	1*	0	1*	1	1*	1*	1*	0	1*	0	1	**15**
5	Risk awareness	1	1	1*	0	1	1	1*	1	1*	1*	1*	1*	0	1	1	0	1	0	1*	**15**
6	Minimum vulnerability against disruptions	1*	1	1*	0	1*	1	1	1	1*	1*	1*	1*	0	0	1*	1*	1	0	1	**15**
7	Dispersion of key resources, production and …	1*	1	1*	0	1*	1*	1	1	1*	1*	1*	1*	0	0	1	1	0	0	1*	**14**
8	Agility	1*	1	1	1*	1	1*	0	1	1	1	1*	1	1*	1*	1*	0	1*	1*	1	**17**
9	Commitment to contract	1	1	1	0	0	1	1*	1*	1	0	0	0	0	0	0	0	1*	0	1	**9**
10	Research and development	1	1	1	1	1	1	1*	1	1	1	1	1*	1	1	1	0	1	1	1*	**18**
11	Technical capability to adapt to the latest …	1*	1*	1*	0	1	1	1*	1	1	1*	1	1*	1	1	1	0	1	1*	1	**17**
12	Compliance with standards	1	1*	1	0	1*	1	1*	1	1*	1*	0	1	0	0	0	0	0	0	1	**11**
13	Development of new and alternative technologies	1	1	1*	0	1	1	1	1	1*	1*	1	1	1	1*	1	1*	1	1	1*	**18**
14	Prioritizing environmental concerns	1*	1*	1	0	1*	1*	1*	1	1*	1*	1	1*	1*	1	1	0	1*	0	1	**16**
15	Backup energy resources	1*	1	1*	0	1*	1	1	1	1*	0	1*	1*	0	0	1	0	1*	0	1*	**13**
16	Management stability and specialized staff	1*	1	1	0	1	1	1*	1	1	1*	1	1	1*	1*	1*	1	1	0	1*	**17**
17	Personnel, information and cyber security	1*	1*	1	0	1*	1	1	1	1*	1*	0	1	0	0	1*	1*	1	1*	1	**15**
18	Network scale	1*	1*	1*	0	0	1	1	1*	1	1*	1*	0	1	0	1*	1*	1*	1	1*	**15**
19	Brand value	1*	0	1	0	0	0	0	0	1*	0	0	0	0	0	0	0	0	0	1	**4**
Dependence		**19**	**18**	**18**	**3**	**14**	**18**	**15**	**17**	**18**	**12**	**14**	**13**	**9**	**9**	**14**	**6**	**14**	**5**	**19**	

* Values which are changed due to transitivity

4- Partitioning the reachability matrix in different levels. The reachability matrix is partitioned into different levels through successive iterations. For this, the reachability set and the antecedent set of each enabler is found. The reachability set of an enabler consists of itself and all the other enablers which are influenced by it whereas the antecedent set of an enabler consists of itself and all the enablers which influence it. The intersection of these sets is derived for all the enablers. The criteria for which the reachability set and intersection set are same, are assigned the topmost level in the hierarchy. In iteration 1 quality (1) and brand value (19) have the same reachability and intersection sets. Hence they will occupy the topmost level in the ISM hierarchy. Once the hierarchy of an individual or a group of criteria is set, they are not considered for analysis in subsequent iterations. Using this procedure, the hierarchy of each criterion has been set. The iterations (1-8) are presented in Tab. 3.

Table 3. Iteration 1-8

C.No.	Reachability Set	Antecedent Set	Intersection set	Level
1	1,2,3,6,7,8,11,19	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19	1,2,3,6,7,8,11,19	I
2	2,5,6,7,8,9,11,13,14,15,17	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	2,5,6,7,8,9,11,13,14,15,17	II
3	3	3,4,5,7,8,10,11,12,13,14,15,16,17,18	3	III
4	4	4	4	VII
5	5,7,8,10,11,12,14,15,17	4,5,7,8,10,11,12,13,14,15,16,17	5,7,8,10,11,12,14,15,17	IV
6	2,3,5,6,7,8,9,10,11,12,15,16,17	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	2,3,5,6,7,8,9,10,11,12,15,16,17	II
7	5,7,8,10,11,12,15,16	5,7,10,11,12,13,14,15,16,17,18	5,7,10,11,12,15,16	IV
8	4,5,8,10,11,12,13,14,15,17,18	4,5,7,8,10,11,12,13,14,15,16,17,18	4,5,8,10,11,12,13,14,15,17,18	IV
9	2,3,6,7,8,9,17	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	2,3,6,7,8,9,17	II
10	10,18	10,18	10,18	VIII
11	10,11,13,14,18	4,10,11,13,14,16,18	10,11,13,14,18	VI
12	5,7,8,10,12	4,5,7,8,10,11,12,13,14,15,16,17	5,7,8,10,12	IV
13	10,11,13,14,16,18	4,10,11,13,14,16,18	10,11,13,14,16,18	VI
14	10,11,13,14	4,10,11,13,14,16	10,11,13,14	VI
15	11,15,17	4,10,11,13,14,15,16,17,18	11,15,17	V
16	16	16,18	16	VII
17	10,15,16,17,18	4,10,11,13,14,15,16,17,18	10,15,16,17,18	V
18	18	18	18	VIII
19	1,3,9,19	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19	1,3,9,19	I

5- Developing ISM model. The ISM model presented in Fig. 1 is constructed by utilizing the final reachability matrix (Tab. 2) and the hierarchical level of the criteria shown in Tab. 3. The model demonstrates that for the resilient IT project supplier selection, research and development (10) and network scale (18) are undoubtedly the most driving enabler.

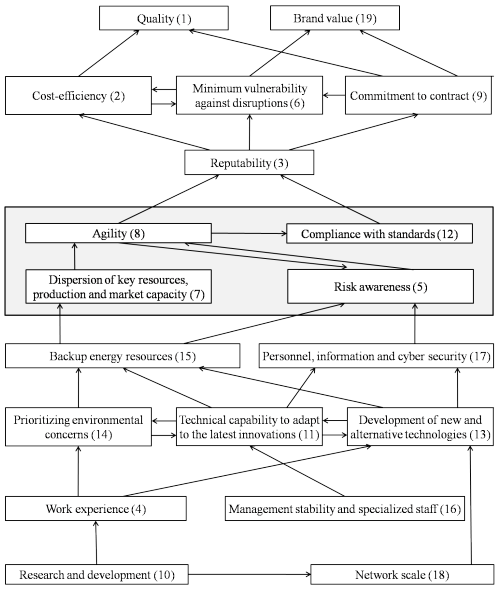


Figure1. ISM model of evaluation criteria for resilient it project supplier selection

3. MICMAC Analysis

The MICMAC analysis is used to classify the criteria in accordance with their driving power and dependence. The criteria are classified into four clusters namely autonomous, dependent, linkages, and independent as shown in Fig. 2. The grouping of variables in these clusters is as follow. **I) Autonomous variables:** These variables have weak dependence as well as weak driving power. They are relatively isolated from the system. There are no variables in this cluster. **II) Dependent variables:** The variables in this cluster have weak driving power but have a strong dependence on other factors to drive them. Enablers 1, 3, 9, 18, and 19 form this cluster. **III) Linkage variables:** The variables in this cluster have high dependence as well as high driving power. Ten criteria are in this cluster including 2, 5, 6, 7, 8, 10, 11, 12, 15, and 17. **IV) Independent variables:** The variables in this cluster have strong drive power and weak dependence. Criteria 4, 13, 14, and 16 fall in this cluster.

4. Conclusion

The ultimate goal of the resilient IT project supplier evaluation and selection problem is to select suitable suppliers who are largely adaptable to the company's resilience. As suppliers are one of the main sources of vulnerability for a company in outsourced projects, assessing supplier resilience is considered one of the most crucial requirements for building organizational resilience. Therefore, the predefined objective of this study was to model and analyze the supplier resilience evaluation criteria for IT projects using ISM. To achieve this, a set of criteria were identified by literature review and deliberations with domain experts.

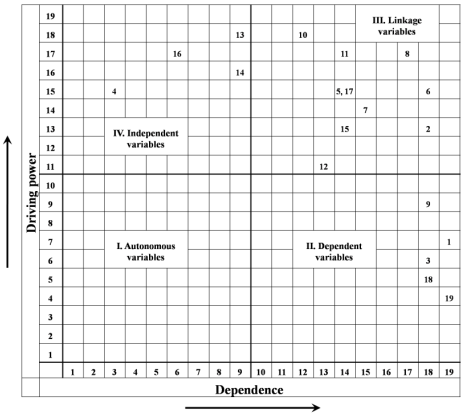


Figure 2. MICMAC diagram of evaluation criteria for resilient it project supplier selection

The prime focus of this paper is to establish a hierarchy of criteria along with classifying them according to their driving power and dependence so as to equip decision makers with a clear framework for focusing their efforts on few key criteria rather than all the criteria. The study is unique in its approach as it enlists the most prominent evaluation criteria for resilient IT project supplier and categorizes them based on their interrelationships into four categories viz., autonomous, driving, linkage and dependent variables. The ISM model (Fig. 1) and MICMAC diagram (Fig. 2) developed in the paper classifies various evaluation criteria for resilient IT project supplier selection according to their hierarchical levels and driving power/ dependence. From the MICMAC analysis, it can be seen that the autonomous cluster has no variables, which means that no variable can be assumed to be isolated from the entire structure and the management has to pay attention to all the identified criteria. The next cluster contains dependent variables which have high dependence and weak driving power. These variables though are critical to resilient IT project supplier selection but their high dependence means that they need other criteria to drive them. The next cluster represents linkage variables. Such variables are highly driving as well as highly dependent. There are ten variables for this cluster in the developed model. The last cluster contains independent variables. These criteria have high driving power but very weak dependence. These criteria play a prominent role in ensuring that the resilient IT project supplier selection is successfully adopted. Based on the ISM digraph model and MICMAC analysis, it can be seen that the factor research and development (10) being highest in driving power and lowest independence plays a vital role in driving the entire resilient IT project supplier selection problem. Owing to the fierce market competition, selection of the resilient supplier has been steadily gaining prominence in various organizations. Hence, the findings of the study can serve as a decision-making tool for top management. In this study, the ISM model has been developed to analyze the interactions among the evaluation criteria. The MICMAC matrices provide some valuable insights about the relative importance and interdependencies among the evaluation criteria. The theoretical frameworks developed in this study can greatly assist the decision makers in visualizing the evaluation criteria of resilient IT project supplier through a system approach, abridging a complex system of criteria into a structured format. Decision makers further can prioritize common driving criteria as per the need of implementation on the basis of their driving powers and dependence. This study will also assist managers in identifying highly sensitive variables having high driving power as well as dependence that require continued supervision for the effective and thorough the selection of resilient IT project supplier.

The present study employs ISM to model the criteria for selection of resilient IT project supplier. The contextual relationships used in the modeling have been established based on the opinion of a few domain experts which can have a certain degree of bias. The developed models have not been statistically validated. To alleviate these limitations, structured equation modeling (SEM) analysis can be performed based on the adequate number of responses collected using a structured questionnaire. More qualitative techniques like analytical hierarchy process (AHP) and best-worst method (BWM) can be used to determine the importance of criteria and ranking them in order to focus on the critical few criteria.

References

[1] Beikkhakhian, Y., Javanmardi, M., Karbasian, M., & Khayambashi, B. (2015). The application of ISM model in evaluating agile suppliers selection criteria and ranking suppliers using fuzzy TOPSIS-AHP methods. Expert Systems with Applications, 42(15–16), 6224–6236.

[2] Chen, P. T., & Chen, J. H. (2014). Implementing cloud-based medical systems in hospitals and strategic implications. Technology Analysis and Strategic Management, 27(2), 198–218.

[3] Girubha, J., Vinodh, S., & Kek, V. (2016). Application of interpretative structural modelling integrated multi criteria decision making methods for sustainable supplier selection. Journal of Modelling in Management, 11(2), 358–388.

[4] Govindan, K., Kannan, D., & Haq, A. N. (2010). Analyzing supplier development criteria for an automobile industry. Industrial Management and Data Systems, 110(1), 43–62.

[5] Peppard, J., & Ward, J. (2016). The strategic management of information systems: Building a digital strategy. John Wiley & Sons.

[6] Ponomarov, S. Y., & Holcomb, M. C. (2009). Understanding the concept of supply chain resilience. The International Journal of Logistics Management (Vol. 20).

[7] Proag, V. (2014). Assessing and Measuring Resilience. Procedia Economics and Finance, 18(September), 222–229.

[8] Yu, T., Shi, Q., Zuo, J., & Chen, R. (2018). Critical factors for implementing sustainable construction practice in HOPSCA projects: A case study in China. Sustainable Cities and Society, 37(November 2017), 93–103.

Electricity Consumption Forecast by Artificial Neural Networks- The Case of Turkey

R. Yumuşak[1], E.C. Özcan[2], T. Danışan[3], T. Eren[4]

[1]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, rabiayumusak95@gmail.com
[2] Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, enmcam@gmail.com
[3]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, tugbadanisan@gmail.com
[4]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, teren@kku.edu.tr

Abstract

Energy is needed in many countries from industry to heating, lighting to daily use in a country. The demand for energy in so many different areas also brings uncertainty. This uncertainty is the lack of awareness of which consumers, when and how much electricity they will demand. This results in consequences that cannot be expressed in terms of currency, rather than cost. Governments are obliged to provide the best service to society and to take measures against possible adverse situations. This responsibility brings along the necessity to manage energy policies very carefully. Sustainable energy policies include strategies that ensure the use of all available resources efficiently and with minimum cost without sacrificing quality specifications, environmental requirements and uninterrupted supply in the generation, transmission and distribution of energy. The first step in the development and implementation of these strategies is to determine the probable energy demand. Although many methodologies have been applied for the estimation of electricity consumption, Artificial Neural Networks (ANN) are prominent among these methods in the literature. In this context, in this study, Turkey's electricity consumption estimation problem is solved with ANN method. Although there is an estimation on the basis of different time periods in the literature, there is no study that takes the process monthly and takes an annual estimation. Thus, this study is thought to contribute to the literature.

1. Introduction

Energy is one of the basic indicators of a country's level of development. This makes the energy issue very important for countries. Since the systematicity in energy production will reduce the risk in the national economy and increase the efficiency, each planning made in this sector is of great importance [1]. The operating principle of many technologies used today is critical for the world of production and technology, as it requires electrical energy. In addition, electrical energy cannot be stored in large sizes. However, demand varies over time. This change has led to a difficult process to manage. Demand or consumption estimation studies have emerged in order to determine the demand on the difficulty of the process. Many different estimation methods are used in the literature for estimating energy demand. For example; gray prediction [2], regression model [3], particle herd optimization [4], genetic algorithm [5], fuzzy logic [6], artificial neural networks [7] and so on. energy demand or consumption is estimated using forecasting techniques. This study estimated electricity consumption in Turkey is discussed for the general problems. Artificial neural network has been used in non-linear problems such as energy demand estimation. In this study, two-year estimation was made unlike the literature. This estimate includes monthly periods. It is expected that the study, which takes a different time period from other studies, will contribute to the literature.

In the second part of this study, artificial neural network which is used is explained. In the third section, problem definition and application steps are explained. In the last section, the results of the study are evaluated.

2. Artificial Neural Networks

Artificial Neural Networks are computer systems that perform the learning function, which is the most basic feature of the human brain. They perform the learning process with the help of examples. These networks are composed of connected process elements (artificial nerve cells). Each link has a weight value. The information that the artificial neural network possesses is hidden in these weight values and has spread to the network [8].

Artificial neural networks offer a different calculation method than the known calculation methods. It is possible to see the successful applications of this calculation method that is adaptive to their environment, adaptable, can work with incomplete information, can make decisions under uncertainties, and is tolerant to errors. Although there is no specific standard in the determination of the structure of the network to be formed, the selection of network parameters, problems can be shown only with numerical information, it is not known how to finish the training and it cannot explain the behavior of the network, the interest in these networks is increasing with each passing day.

Particularly, artificial neural networks are among the most powerful techniques in classification, pattern recognition, signal filtering, data compression and optimization studies. Data mining, optical character handling, optimum routing, fingerprint recognition, material analysis, job scheduling and quality control, medical analysis can be seen in many areas of successful examples can be seen in daily life [8].

3. Application

Electrical energy has become one of the most preferred and consumed energy types due to its easy transportation, usage and clean energy. Electricity consumption varies depending on various social and economic variables such as population, economic growth and gross domestic product, as well as climatic variables such as temperature, precipitation and humidity. The electricity used for heating and cooling needs has a great effect on electricity consumption. Weather conditions cause an increase and decrease in electricity consumption, while the most effective meteorological variable is temperature. It is a difficult problem to estimate consumption by considering so many independent parameters. For this reason, in this study, artificial neural network model was created with minimum parameter of consumption estimation problemIn the literature estimate that Turkey's electricity consumption for many years and Kutay Hamzah in 2013 [9] model it was working. In this study, year and month are considered as input parameters. 80% of 528 data obtained from TEİAŞ database between 1975 and 2018 were used in education and 106% of it consisted of 20%. Monthly consumption forecast for 2019 and 2020 is made Feed forward artificial neural network model was used. There are two intermediate layers in the established network architecture. The number of neurons in the layers is 50, 50, respectively. Trainlm function was used as training function and tansig and logsig functions were used as transfer function respectively. MATLAB program was used for problem solving. In the learning phase, successful test results were obtained. An example of the results of the test phase is given in Table 1. MAPE value was found to be 4.9%.

Table 1: Test results

Year	Months	Actual consumption	Estimated consumption
32	5	16284	16488,02962
32	6	16527,1	16540,03785
32	7	18308,5	18565,73274
32	8	18391,8	19384,73339
32	9	16045,1	16032,15587
32	10	14917	14463,41385
32	11	15446	15397,26077
32	12	15816,4	15516,11325
33	1	16851,4	16225,89074
33	2	15010	15120,61018
33	3	15983,7	16254,21943
33	4	14849,1	14905,46229
33	5	15297,7	15618,79061

Since the error rate calculated in the test phase is acceptable, the estimation phase has started. Electricity consumption forecast for 2019 and 2020 has been made. The estimation results for 2020 are given in Table 2.

Table 2: Forecast results for 2020

Forecast Inputs		Forecast Outputs
Year	Months	Consumption
2020	1	21001,53949
2020	2	22748,08902
2020	3	23727,96869
2020	4	23592,17819
2020	5	23454,45595
2020	6	23833,12229
2020	7	24344,65757
2020	8	24548,19613
2020	9	24737,6264
2020	10	24657,34658
2020	11	24266,23988
2020	12	24292,26123

4. Result

The study is a prediction study and, unlike traditional prediction models, artificial neural networks have proven their superiority in nonlinear problems since they can learn the deviations in real life problems with the data. The reason for this is that they learn the system from problem examples and find a fine functional relationship between the parameters even if the problem is difficult. Therefore, artificial neural networks are a suitable method for problems where solutions are difficult and sufficient data or observation can be obtained [10]. Therefore, artificial neural network method was used for estimation.

In this study, 80% of the total 504 data were used in the test phase and 104 data, 20% of which were used in education, and the MAPE value was found to be 4.9%. Forecasting results in Turkey was estimated electricity consumption by 2019 and to 2020 for the general. As a result of the estimation, consumption has increased gradually.The study is expected to contribute to the literature in terms of time periods. In the future, the effect of different parameters on consumption can be analyzed with data mining.

References

[1] Özcan, E.C., Yumuşak, R. and Eren, T. (2019). Risk Based Maintenance in the Hydroelectric Power Plants. Energies, 12 (8): 1502-1523.
[2] Akay, D., and Atak, M. (2007). Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. *Energy*, 32(9), 1670-1675.
[3] Bessec, M., & Fouquau, J. (2008). The non-linear link between electricity consumption and temperature in Europe: a threshold panel approach. *Energy Economics,* 30(5), 2705-2721.
[4] Ünler, A. (2008). Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025. *Energy Policy*, 36(6), 1937-1944.
[5] Ceylan, H., and Ozturk, H. K. (2004). Estimating energy demand of Turkey based on economic indicators using genetic algorithm approach. *Energy Conversion and Management,* 45(15), 2525-2537.
[6] Kucukali, S., and Baris, K. (2010). Turkey's short-term gross annual electricity demand forecast by fuzzy logic approach. *Energy Policy*, 38(5), 2438-2445.
[7] Hamzaçebi, C. (2007). Forecasting of Turkey's net electricity energy consumption on sectoral bases. *Energy policy*, 35(3), 2009-2016.
[8] Öztemel, E. (2003). Yapay Sinir Ağları. *PapatyaYayincilik*, Istanbul.
[9] Hamzaçebi, C., & Kutay, F. (2004). Yapay sinir ağlari ile türkiye elektrik enerjisi tüketiminin 2010 yilina kadar Tahmini. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 19(3).
[10] White, H. , "Learning in artificial neural networks: A statistical perspective", Neural computation, 1(4), 425-464, (1989).

Generation Estimation with Artificial Neural Network in the Natural Gas Combined Cycle Power Plants

R. Yumuşak[1], E.C. Özcan[2], T. Danışan[3], T. Eren[4]

[1]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, rabiayumusak95@gmail.com
[2] Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, enmcam@gmail.com
[3]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, tugbadanisan@gmail.com
[4]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, teren@kku.edu.tr

Abstract

Energy is one of the key indicators reflecting the level of development of a country. In addition, energy, which is the decisive parameter for socio-economic progress, is one of the main factors determining the place of countries in the global world. For this reason, it is a necessity to follow the developments in the energy sector and produce alternative solutions to the problems in today's competitive environment. One of the main problems in the energy sector is the sudden changes in supply and demand. Although this problem exists in other sectors, as the final product in the energy sector cannot be stored on an industrial scale, electrical energy generation becomes more critical. For this reason, it is a must to manage electricity generation within a system. In addition to this, natural gas is one of the most important sources used in electrical energy obtained from power plants with different generation processes. Natural gas combined cycle power plants with 28% share in Turkey's total installed capacity are one of the keystones in terms of energy supply security. Taking all these into consideration, the impact of the production planning of electricity produced in these plants on the costs becomes an indisputable fact. The first stage of this planning should be the demand forecast. In this context, in this study, generation forecast is made in a natural gas combined cycle power plant. This estimation problem is solved by the Artificial Neural Network model which has taken place in the literature for nonlinear problems. This study is expected to contribute to the literature in terms of reflecting the characteristic features of a power plant to the model and solving the problem with the Artificial Neural Network.

1. Introduction

Energy is one of the main indicators reflecting the level of development of a country. And therefore, the issue of energy is very important for countries (Özcan et al., [1]). Therefore, in an important field such as energy, meeting the demand on time is a critical factor that needs attention. Electrical energy is an energy that cannot be stored and in order to meet the demand, various production plans are made at the power plants. Many studies have been conducted in the literature for various forecasting problems in the energy sector. Artificial neural network (ANN) are frequently used in these problems. For detailed information, Suganthi and Samuel [2] estimate electricity demand, Weron [3]'s electricity price, and Kuster et al. [4]'s electrical load estimation, Wang and Srinivasan [5]'s literature reviews on energy use will be guiding. Within the scope of this study, one-year production quantity estimation in a natural gas combined cycle power plant is realized by using ANN method. As a result of this estimation, it is aimed to make personnel, material and maintenance planning in the enterprise. In the second part of this study, ANN method knowledge, in the third part the application of the problem and in the fourth part the results are given.

2. ANN Method

ANN is a processor that has a natural tendency to store and make available information based on experience and this method is an artificial intelligence method capable of establishing a functional relationship between the input and output parameters. ANN is widely used in the literature and non-linear problems can be reflected in a more real way. ANN is the preferred method in this study because it is widely used in the literature and non-linear problems can be reflected to the reality (White, [6]; Ripley, [7]; Cheng and Titterington, [8]).

3. Application

In the study, production quantity estimation for 2019 was made by ANN using the production data taken from 2000 year. MATLAB program was used in this process and the working steps are as shown in Figure 1.

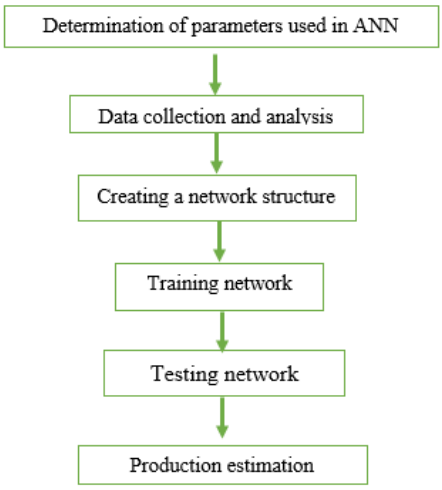


Figure 1. Application steps

The parameters of the network used in ANN are given below.

- Months
- Temperature
- Group A energy losses
- Group B energy losses
- Working hours

There are two intermediate layers in the established network architecture. Trainlm function was used as training function and purelin and logsig functions were used as transfer function respectively. Figure 2 shows the structure reflecting the established network structure.

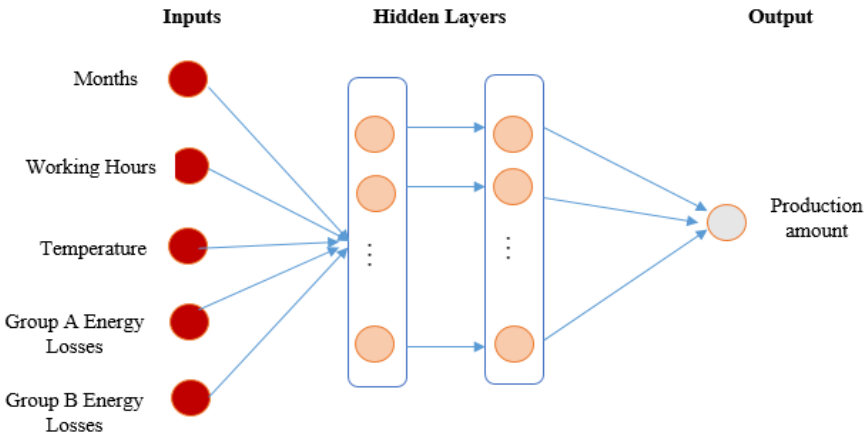


Figure 2. ANN Architecture

After the network was established, training and testing stages were carried out. The graph obtained during the training phase is as in Figure 3.

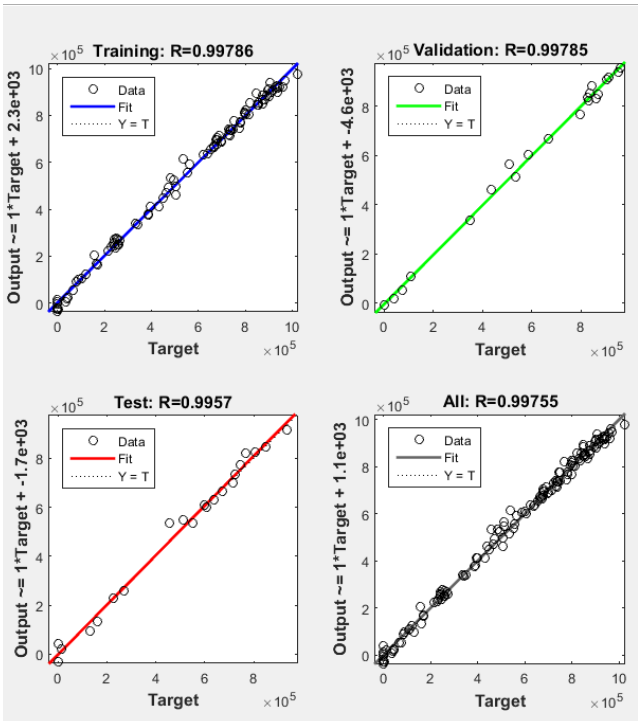


Figure 3. Regression graphs

As a result, it is reflected in the graphs that the learning and testing process in the network is successful and these rates are around 99%. Two different attempts were made in the learning process of the network. The first one was made on learning data and the error was found to be 0.0567. The second one was made on the test data and the error was found to be 0.0732.

Since the above results are seen to be consistent, the estimation is made for 2019 after this stage. The result of the estimation for 2019 is given in Table 1.

Table 1. The result of the estimation for 2019

Year	Month	Estimation	Year	Month	Estimation
2019	1	404612,5	2019	7	570714,9
2019	2	475238,3	2019	8	708351,3
2019	3	401237,1	2019	9	610215,2
2019	4	334187,4	2019	10	533945,1
2019	5	349138,7	2019	11	539231,7
2019	6	514502,8	2019	12	506038,9

4. Result

As a result of this study, monthly production amounts for year 2019 were estimated in a natural gas combined cycle plant. This study contributed to the literature in terms of the type of plant and parameters affecting production. In line with the results obtained, material, personnel and maintenance plans of the enterprise can be realized and will be able to plan the timing in the best way to meet the demand.

References

[1] Özcan, E. C., Danışan, T., & Eren, T. (2019). Hidroelektrik Santralların En Kritik Elektriksel Ekipman Gruplarının Bakım Stratejilerinin Optimizasyonu için Matematiksel Bir Model Önerisi. Pamukkale University Journal of Engineering Sciences, (in press).

[2] Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting—A review. Renewable and sustainable energy reviews, 16(2), 1223-1240.
[3] Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International journal of forecasting, 30(4), 1030-1081.
[4] Kuster, C., Rezgui, Y., & Mourshed, M. (2017). Electrical load forecasting models: A critical systematic review. Sustainable cities and society, 35, 257-270.
[5] Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renewable and Sustainable Energy Reviews, 75, 796-808.
[6] White, H. (1989). Learning in artificial neural networks: A statistical perspective. Neural computation, 1(4), 425-464.
[7] Ripley, B. D. (1993). Statistical aspects of neural networks. Networks and chaos—statistical and probabilistic aspects, 50, 40-123.
[8] Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. Statistical science, 2-30.

Combined Multi Criteria Decision Making Model for Localization Problem of the Main Systems in the Hydroelectric Power Plants

Betül Demirelli[1], Nermin Avşar Özcan[2], Evrencan Özcan[3], Tamer Eren[4]

[1]*Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, betulddemirelli@gmail.com*
[2]*Strategy Development Department, Electricity Generation Corporation, Ankara, Turkey, nerminavsar.ozcan@euas.gov.tr*
[3]*Department of Industrial Engineering, Kırıkkale University, Kırıkkle, Turkey, enmcan@gmail.com*
[4]*Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, teren@kku.edu.tr*

Abstract

Developing countries tend to use renewable energy resources and localize its technology and follow a strategy from import to export. Despite decreasing its dependence on energy gradually, Turkey makes generation from foreign resources with a rate of 51,03%. Dependence on foreign resources is a critical element which treats energy supply security. In order to gain global competitive power and protect supply security, countries need to immediately realize the active use of domestic resources in energy generation and so localize its technology which is the main directive of this important study. Energy is an indispensable element for human being with many factors such as growing economy, industrialization, international trade, technology. The share of renewable energy resources in Turkey increases gradually due to preserving sustainability and providing more positive results compared to fossil resources. Hydroelectric power plants are among the renewable energy resources and have the largest share in Turkey with a generation rate of 21%. Therefore, localization of hydroelectric power plant technology is considered as an important gain for Turkey and the lack of localization problem for hydroelectric power plants in literature has been realized. In this study, when the main systems are examined for hydroelectric power plants, although many equipment exist, 14 main systems from penstocks to turbine, circuit breakers to main power transformers and 11 criteria such as profit share, equipment criticality, and industry capacity in Turkey which influences these systems have been examined. In this study, the localization problem for the main systems in the hydroelectric power plants is handled and after weighting the criteria with AHP and comparing the relation of criteria with alternatives, the order of priority for main systems (equipment) is obtained with TOPSIS. Consequently, while the criteria with the highest rank is the development level of technology in the weighting process, consecutively main power transformers, then current measurement transformers and penstocks are acquired in the order of priority of the main system localization.

1. Introduction

Energy has been among the most critical issues of the countries for more than centuries such as policy, economy, development level and even unfortunately war. The main objective of energy strategies and policies is to ensure energy supply security and to provide sustainable and uninterrupted energy with minimum cost which does not treat environment conditions. Energy security is a concept which covers the existence, accessibility, affordability and sustainability of energy resources and energy supply security is provided through these four factors [1]. There is a strong relation between energy and development level and in the definition of the developed country, the development level of the mentioned countries are evaluated according to their energy resources and the amount they use. The common characteristic of developed countries in being powerful is that they don't have an energy resource problem and have the technology to use energy. Localization in energy resources is the basic factor for the countries in global competition, in increasing social welfare levels, socio-economic development, converging towards import without any dependence on foreign resources and advancing. As developing countries don't have the technology to use energy resources within their borders, dependence on foreign resources is inevitable. Foreign dependence on technology treats supply security and disrupts global developments of these countries. The opportunity to gain global power and meet the energy need which increases in parallel with population increase may be possible by generating energy with sustainable energy resources within the scope of "domestic resources, domestic generation and domestic technology". Developing Turkey meets 48,97% of electricity generation from domestic resources and aims to further increase this rate [2]. Renewable energy is a sustainable resource gained from endless natural resources and leads less environment pollution compared to fossil resources. Solar, wind, water, biomass and hydroelectric are among renewable energy. Among renewable energy resources, hydroelectric is used to generate power in approximately 150 countries around the world [3]. The importance of renewable energy resources is increasing day by day and has a growth rate of 9,8% annually and the hydroelectric energy has the largest rate of 1,8% within this share [4]. Despite the continuous increase in energy demand and energy security, it is inevitable for Turkey to adopt "domestic technology and domestic generation" vision and continue the works like the success achived by Turkey in domestic generation. Due to its geopolitical importance, Turkey uses various resources in energy generation. Turkey provides electricity generation using energy mostly from fossil (natural gas and imported coal) resources and then hydroelectric energy

which is a renewable resource with a generation of 20,08% kWh [5]. Fossil resources are exhausted rapidly and development level and technology ownership of renewable energy which will replace these resources in the future directly affect energy supply security. Turkish economy ranks 17th worldwide and proves its position as developing country [6]. In order to improve development level and protect its existence in global competition, Turkey needs foreign dependence using its resources, and establish the technology infrastructure to use own resources and realize localization in technology and start export. The development of the country is foreseen by leading the growing economy and technology.

In literature study, 11 main system equipment for coal fired thermal power plants are evaluated with Analytical Hierarchy Process (AHP) with criteria such as technology, design, labour force, market, generation capability by Levent for the analysis of energy technologies localization and the order of priority is obtained for domestic manufacturing [7]. AHP method used in this study is weaker in sorting compared to the other decision making methods. In another study in literature, the criteria weighting is performed and priority analysis is obtained with Analytical Network Process (ANP) method handled by Ozcan and colleagues for hydroelectric power plants [8]. In terms of the importance of the localization in technology and based on the lack of the criticality of hydroelectric power plants in Turkey in renewable energy in literature, considering 11 critical criteria such as the feasibility in Turkish industry, 14 main system equipment of a hydroelectric power plant are weighted with AHP method which is among the most appropriate methods suggested for the solution of this kind of problem in literature and then for providing development in ranking, the priority analysis is obtained using TOPSIS method (Technique for Order of Preference by similarity to Ideal Solution) by comparing the proximity to ideal solution and benefit maximization and the distance to ideal solution and loss minimization.

2. Hydroelectric Power Plants

Among renewable energy resources, hydroelectric energy covers the systems which generate electricity in hydroelectric power plants (HPP). Hydroelectric energy resources are widely available on the world. HPP are cheaper, investment and operation costs are lower, the capability of meeting the need for peak and base load is higher and, they don't harm environment. All these factors make hydroelectric power plants indispensable compared to the other renewable energy resources. According to the International Energy Agency (IEA) forecasts, average share of hydroelectric power is envisaged to be 1,8 % in renewable energy in energy generation of 6,9% until 2040 [4]. Turkey has a hydroelectric potential of 140 billion kWh annually and ranks 7th in hydraulic power usage. Turkey is rich in terms of this important resource mostly used in renewable energy generation and hydroelectric power accounts for 21% of total generation [9]. General working principles of hydroelectric power plants, water in the reservoir or river bed is transmitted through power conduits which are penstocks or through transmission channels to spiral case and then from spiral case to turbine. The water transmitted to turbine with pressure gives mechanic energy to turbine and thus the rotor combined with turbine starts to rotate and rotating rotor provides continuous energy generation sending exciting current to generator. In this study, 14 system equipment (main power transformers, separators, power conduits, generator, speed regulator, butterfly valves, protective relays, engines, surge arresters, SCADA and PLC, turbine, oil containers, instrument transformers) are determined. This is because in case this main equipment are not available due to compulsory maintenance or planned maintenance, large financial losses may occur, and instabilities may arise as they cannot meet the supply. In this context, the study aimed to solve the problem of localization priority order for the said critical equipment.

3. Methods

Although decision making concept is defined differently by the researchers according to the circumstances, it is generally defined as: Decision making is choosing the most appropriate alternative taking into account the determined criteria among the existing alternatives in order to realize the identified goal. Multi criteria decision making (MCDM) is the set of methods required to choose, sort and classify one or more alternatives from the set of choices with the same characteristics according to criteria with generally different weights, to identify numerous and generally conflicting qualitative and quantitative criteria to help a decision process [10]. The implementation process of all MCDM techniques is composed of three basic steps [11]: 1) Identifying related criteria and alternatives, 2) Determining the weights of criteria which demonstrate their significance level and evaluating the alternatives according to these criteria, 3) Handling numerical values according to the identified method and identifying the order of each alternative.

In this study, weighted significance level of 11 criteria is determined using Analytical Hyrerachy Process (AHP) after identifying the alternatives and criteria. Following this, the order of 14 alternatives is determined with TOPSIS

method which has easier implementation compared to the other sorting algorithms and produces effective results. Step algorithms of AHP and TOPSIS methods are shown in Figure 1.

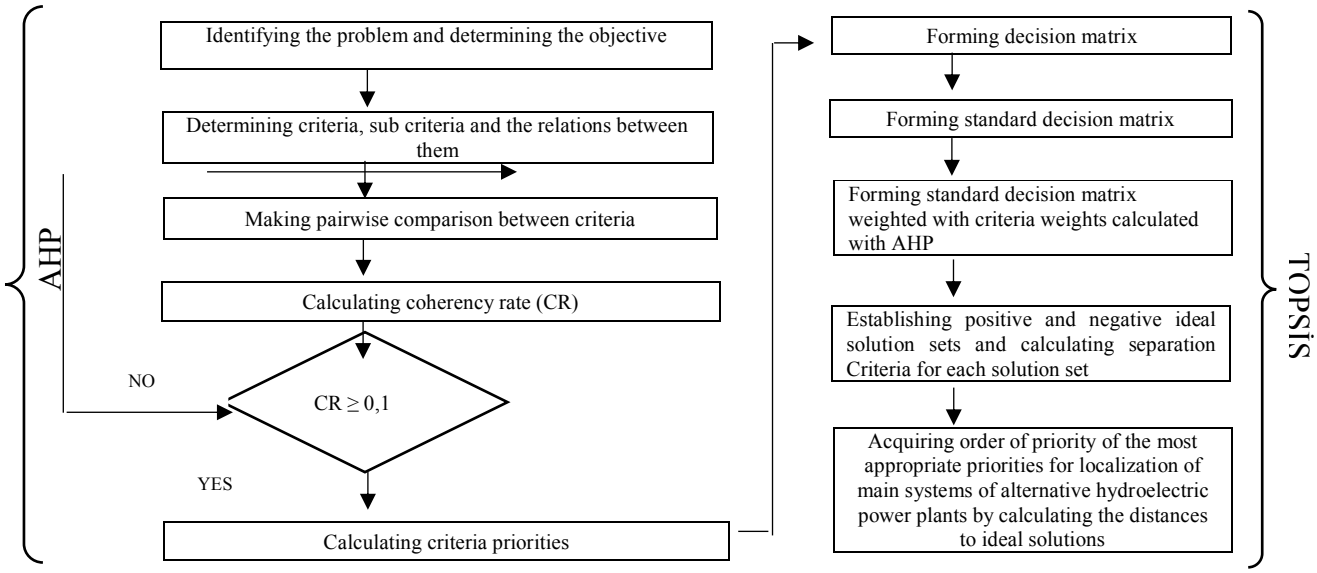


Figure 1. Implementation stages

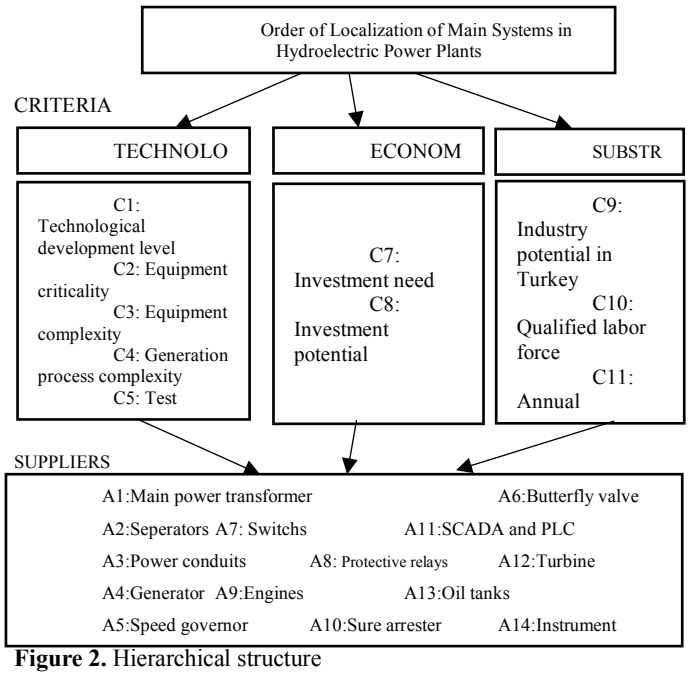


Figure 2. Hierarchical structure

Table 1. *Criteria weights with AHP*

Main criteria	Sub criteria	Weights
TECHNOLOGY	Technological development Level	0,246
	Equipment criticality	0,066
	Equipment complexity	0,132
	Complexity of generation process	0,047
	Test procedures	0,066
	Special condition need for generation	0,18
SUBSTRUCTURE	Industry capacity in Turkey	0,101
	Qualified labour force	0,034
ECONOMY	Annual producible figure	0,024
	Investment need	0,061
	Profit potential	0,042

Table 2.

Alternatives	Priority Value	Priority Order	Alternatives	Priority Value	Priority Order
Turbine	0,347	14	**Instrument transformer**	0,712	2
Engines	0,602	5	**Surge arrester**	0,584	6
Power conduits	0,668	3	**Protective relays**	0,495	11
Butterfly valves	0,609	4	**Speed regulator**	0,501	10
Main Power Transformers	0,839	1	**Oil tanks**	0,57	9
Switch	0,575	8	**SCADA and PLC**	0,474	13
Separators	0,578	7	**Generator**	0,491	12

4. Application

In this study, critical equipment which cease electricity generation in case of a breakdown and/or maintenance are defined as main power transformers, separators, power conduits, generator, speed regulator, butterfly valve, switches, protective relays, engines, surge arresters, SCADA and PLC, turbine, oil tanks and instrument transformers. In the studies in the literature, the localization problem of main system equipment are collected under three main criteria as substructure, economy, technology and divided into sub criteria within the scope of hydroelectric power plant operation and maintenance principles, generation processes regarding equipment, their costs and other requirements and economic realities of Turkey. 11 sub criteria are shown under the main criteria in the hierarchical structure shown in Figure 2.

The hierarchical structure for criteria and alternatives is shown in Figure 2 based on the problem "priority analysis for localization problem of the main system in hydroelectric power plants". The criteria which are under technology, economy and substructure in terms of level are effective criteria such as technological development level, equipment criticality, profit potential, qualified labour force. It is foreseen that determining the weights among these substructures and the development of the country in this direction may contribute to localization. The weighted

priority values are acquired taking into account AHP steps given in Figure 1. The priority weights of AHP criteria are shown in Table 1 and technological development level is shown to have the highest weight among the criteria. The weighted normalized matrix is formed using normalized decision matrix attained as a result of decision matrix normalization with criteria weights calculated with AHP. Then ideal positive and ideal negative solution sets are prepared and separation criteria are calculated from these sets and the proximities to ideal positive and negative solutions, meanly priority order for localization alternative equipment are attained. Priority values and priority order of the alternatives are given in Table 2. As seen in Table 2, main system localization takes the lead with 0,839 priority value in terms of priority order and main power transformers follow with 0,712 priority order.

5. Results and Discussion

As a result of the evaluations carried out within the scope of this study, the main power transformer has the highest score and the suitability of the subsequent measurement transformers and penstock localization has been determined with higher weight. The main power transformer is at the top of the priority list with the criteria such as the development level of the technology and the less complexity of the equipment. Generator, SCADA and PLC, turbine main systems are in the back rows of the priority order as special condition requirement for generation is high, technological development level is not enough to produce these equipment yet and in this context, Turkish Government had better invest required to produce these main systems and support the necessary criteria. From this point of view, new studies can be conducted for solar, wind, biomass, geothermal that are renewable energy resources and for technologies to be localized and thus foreign dependence problem can be overcome and more powerful Turkey can be achieved which is moving towards being a developing country.

References

[1] Erdal L., Karakaya E. (2012). Economic, Political and Geographical Factors Affecting Energy Supply Security, Uludağ Journal of Economy and Societyi, 1111(1), 107-136.
[2] The Atlas of Electricity in the Energy Atlas, 2018, Erişim adresi: https://www.enerjiatlasi.com/elektrik-uretimi/, Erişim tarihi: 19 Ocak 2019.
[3] WECTNC (World Energy Council Turkish National Committee), Hydraulic Renewable Energy Working Group Hydraulic Energy Sub-Working Group Report, 2007, Erişim adresi: https://www.dunyaenerji.org.tr/?s=hidroelektrik, Erişim tarihi: 19 Ocak 2019.
[4] World and Turkey's Energy and Natural Resources Outlook Report, (2017), Strategy Development, Erişim adresi: http://www.enerji.gov.tr/Resources/Sites/1/Pages/Sayi_15/mobile/index.html, Erişim tarihi: 19 Ocak 2019.
[5] Daily Report Electricity Generation by Energy Resource Atlas of Turkey, Erişim adresi: http://www.enerjiatlasi.com/elektrik-uretimi/, Erişim tarihi: 19 Ocak 2019.
[6] World's Largest Economies, Erişim adresi: https://egezegen.com/ekonomi/dunyanin-en-buyuk-ekonomileri-siralamasi/ , Erişim tarihi: 19 Ocak 2019.
[7] Yağmur L. (2016). Multi-criteria evaluation and priority analysis for localization equipment in a thermal power plant using the AHP (analytic hierarchy process), Energy, 94, 476-482.
[8] Özcan E.C., Demirelli B., Özder E.H., Eren T. (2018). Hidroelektrik Santrallarda Ana Sistemlerin Yerlileştirilmesi Problemi İçin Analitik Ağ Süreci İle Öncelik Analizi, 2nd International Symposium on Scientific and Professional Studies (BILMES) Full Text Proceedings.
[9] Özcan E.C., Varlı E., Eren T., "Hedef programlama yaklaşımı ile hidrolik santrallarda personel çizelgeleme", International Journal of Informatics Technologies, 10(4):363-370, 2017.
[10] Zopounidis, C., "Multicriteria decision aid in financial management", European Journal of Operational Research, 119(2), 404-415, 1999.
[11] Triantaphyllou E. (2000) Multi-Criteria Decision Making Methods. In: Multi-criteria Decision Making Methods: A Comparative Study. Applied Optimization, vol 44. Springer, Boston, MA.

Maintenance Strategy Optimization with Analytical Hierarchy Process and Integer Programming Combination

R. Yumuşak[1], E.C. Özcan[2], T. Eren[3]

[1]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, rabiayumusak95@gmail.com
[2]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, enmcam@gmail.com
[3]Department of Industrial Engineering, Kırıkkale University, Kırıkkale, Turkey, teren@kku.edu.tr

Abstract

In the study, maintenance strategy selection problem is handled based on the negative effects of electricity generation on the environment, and the direct effect of maintenance planning on uninterrupted supply of energy in power plants and the reduction of generation costs in one of the large-scale hydroelectric power plants which directly affect the security of energy supply to approximately 20% share of total generation in Turkey. The problem was made on the turbine, which is the most critical equipment of a hydroelectric power plant. Although the plant is made up of identical units, the wear rate of the units is different from each other. These factors were analyzed and AHP (Analytic Hierarchy Process) multi-criteria decision-making method was used to determine the rate of wear. These wear rates and the system's specific constraints are reflected in the Integer Programming model and the problem of maintenance strategy selection for the turbines has been solved. The proposed solution methodology contributes to the literature in terms of calculating the wear rate by the AHP method and reflecting this ratio to the mathematical model.

1. Introduction

Energy, the country plays a fascinating role in the economic and social progress. Energy is one of the most important factors that increase the competitiveness of countries in a globalizing world [1]. Population growth, industrialization and urbanization phenomena increase the demand for energy with each passing day. In order to meet the increasing demand in the desired quality and quantity at the desired time, production, personnel, material and maintenance legs must be managed efficiently. Maintenance, one of the four basic pillars, it has a direct impact on the uninterrupted, reliable, efficient and economic production required for sustainability and is therefore critical. In addition, maintenance is costly in terms of time, labor requirement and material and is difficult to manage due to the unique constraints of these three components. For this reason, different maintenance management policies have emerged. Revision maintenance, periodic maintenance, predictive maintenance and corrective maintenance are the main maintenance strategies [2]. There are many studies dealing with the selection of maintenance strategy, which is the first stage of maintenance planning. The most comprehensive and up-to-date updates were made by Shafiee [3] and Ding and Kamaruddin [4]. In the study deals with the hydroelectric power plant which is one of the renewable energy sources. By the end of January 2019 the total installed capacity reached 89,132.0 MW in Turkey, 31.74% share (28,291.4 MW) hydroelectric power facilities are located in the first [5]. Therefore, it has a direct impact on energy supply security. There are only three studies dealing with optimization of maintenance strategy in hydroelectric power plant. Firstly, Özcan et al. [6] In 2017, it was the study of 7 equipments. AHP-TOPSIS-goal programming was used in the study. In the second study, Yumuşak et al. [7] In 2018, they proposed a model for generator equipment in the electrical equipment group. The third study was conducted by Özcan et al. [8] discussed 7 equipment in the electrical equipment group. In the study, for the first time, 3 turbine maintenance strategy optimization was made by AHP-Integer programming combination considering wear rate. It is thought that it will contribute to the literature with the depredation rate limit and application area.

In the second part of the study, AHP method used for calculating the wear rate is explained. In the third part, Integer programming method is explained. The fourth section includes the application. Finally, in the fifth chapter, the results are evaluated.

2. Analytic Hierarchy Process

AHP, which is a multi-criteria decision making method, is used in many areas. AHP method developed by Saaty [9] was used to calculate the wear rate of turbines in three units in this study. Generally, the application steps are as follows:

Step 1: Determination of objective, criteria, sub-criteria, alternatives and hierarchical structure

Step 2: Perform binary comparisons between criteria and alternatives for each criterion

Step 3: Normalization and calculation of importance weights

Step 4: Calculation and control of consistency ratio (CR)

Step 5: Analysis of AHP scores

Consistency ratios should be less than 0.1. Otherwise, it is necessary to continue the same steps until consistency is obtained by reconstructing the compared matrices.

3. Application

The problem deals with a hydroelectric power plant with three units. Each of these units has one turbine. The turbine is considered to be the most critical mechanical equipment for production. These units are identical on the basis of material and construction. But because they are commissioned in different time periods and the difference is the production plans, the wear rates are different from each other. This has necessitated the implementation of different maintenance strategies. Then, the wear rate was calculated by AHP method considering three criteria. These criteria are the time of operation of the unit, the working hours and the amount of production. When these criteria are weighted with AHP method, it is seen that the most important criterion is when it was commissioned with 0.63 points. Then the working hour is 0.26. Finally, the production amount was calculated with 0.11. Three turbines were evaluated according to this criterion weight. Results are calculated over 100 points. The wear rates of the first, second and third units were 68.64, 62.84 and 100 respectively. After calculating the wear rate, the integer programming model for maintenance strategy optimization, which is the second stage, was adopted. The notations are as follows:

CP_{ij}: i. for equipment j. labor cost of maintenance strategy
CM_{ij}: i. for equipment j. material cost of maintenance strategy
D_{ij}: i. equipment j. Production downtime when maintenance strategy is applied
T_{ij}: i. equipment j. Maintenance time when maintenance strategy is applied
Y_i: i. equipment wear rate
Tp: Budget for labor costs
Tm: Budget allocated for material costs
Tt: Total time allocated for maintenance

Decision Variables:

$$X_{ij} = \begin{cases} 1, & \text{i. equipment j. status of maintenance strategy assignment} \\ 0, & \text{other situation} \end{cases}$$

i=1,…,3 j=1,…,4

Mathematical Model:

$$\text{Min } Z = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} * X_{ij} \quad (1)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} CP_{ij} * X_{ij} \leq T_p \quad (2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} CM_{ij} * X_{ij} \leq T_m \quad (3)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} T_{ij} * X_{ij} \leq T_t \quad (4)$$

$$\sum_{j=1}^{n} X_{ij} \geq 1 \qquad i=1,...,m \quad (5)$$

if (Y_i >=90)

X_{i2} =1 i=1,…,m (6)

X_{i3}=1 i=1,…,m (7)

else if (Y_i<90)

X_{i1}=1 i=1,…,m (8)

X_{i3}=1 i=1,…,m (9)

As a result of the mathematical model, optimal results given in Table 1 were obtained.

Table 1: Mathematical model results

Equipment	Revision Maintenance	Periodic maintenance	Predictive Maintenance	corrective maintenance
Turbine 1	✓		✓	
Turbine 2	✓		✓	
Turbine 3		✓	✓	

4. Result

In this study, firstly the wear rate of three turbines was calculated by AHP method. Then these wear rates are reflected as a special constraint in the integer programming model and the model is solved. When the results are examined, it is seen that Turbine 3, whose wear rate is 100, is assigned to periodic maintenance instead of revision maintenance. This results in the need for more frequent maintenance of equipment with high wear rate. In the continuation of this study, the problem size can be increased by including the turbines in the model with other mechanical equipment. When the problem size increases, applications can be made with different solution methods.

5. Acknowledgements

This study was supported by Kırıkkale University Scientific Research Projects Coordination Unit within the scope of the 2019/087 scientific research project.

References

[1] Özcan, E.C., Bulut, M. (2011). Güneş enerjisi teknolojileri ve bu teknolojilerin Türkiye'deki geleceği. *VI. Yeni ve Yenilenebilir Enerji Kaynakları Sempozyumu,* Kayseri, 247-262.
[2] Yumuşak, R., Özcan, E.C., Danışan, T. and Eren, T. (2018, May) AHP-TOPSIS kombinasyonu ile hidroelektrik santrallarda bakım planlaması. *Uluslararası GAP Yenilenebilir Enerji ve Enerji Verimliliği Kongresi*, 76-79, Şanlıurfa, 10-12 Mayıs 2018.
[3] Shafiee, M. (2015). Maintenance strategy selection problem: an MCDM overview. *Journal of Quality in Maintenance Engineering*, 21(4), 378-402.
[4] Ding, S. H., and Kamaruddin, S., "Maintenance policy optimization—literature review and directions." The International Journal of Advanced Manufacturing Technology, 76(5-8), 1263-1283,2015.
[5] Özcan, E.C., Yumuşak, R. and Eren, T. (2019). Risk Based Maintenance in the Hydroelectric Power Plants. *Energies*, 12 (8): 1502-1523.
[6] Özcan, E., C., Ünlüsoy, S. and Eren, T. (2017). A combined goal programming – AHP approach supported with TOPSIS for maintenance strategy selection in hydroelectric power plants. *Renewable and Sustainable Energy Reviews*, 78, 1410-1423.
[7] Yumuşak, R., Özcan, E.C., Danışan, T. and Eren, T. (2018, May). AHP-TOPSIS-tam sayılı programlama entegrasyonu ile hidroelektrik santrallarda bakım strateji optimizasyonu. *Uluslararası GAP Yenilenebilir Enerji ve Enerji Verimliliği Kongresi*, 80-84, Şanlıurfa.
[8] 35. Özcan, E.C., Danışan, T. and Eren, T. (2019). Hidroelektrik Santrallarin En Kritik Elektriksel Ekipman Gruplarinin Bakim Stratejilerin Optimizasyonu icin bir Matematiksel Model Onerisi, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, basımda.
[9] Saaty T. (1980) The Analytic Hierarchy Process: Planning, Priority Setting, *Resource Allocation*. New York: McGraw-Hill.

Using forecastHybrid Package to Ensemble Forecast Functions in the R

Z. Pala[1]

[1] *Muş Alparslan University, Computer Eng Dept. Muş, Turkey, z.pala@alparslan.edu.tr*

Abstract

The prediction process for time series is used in many applications. The prediction process can be performed with different algorithms using the forecast library of the R language. However, the forecastHybrid library is intended to enable multiple functions to play a role in the forecasting process with a certain weight. In this study, future estimations were made with forecastHybrid package support by using monthly series in R datasets and the results obtained were compared with individual estimation results.

1. Introduction

Time series prediction is a closed box in the field of data science. Therefore, it is one of the most widely used data science techniques in the business world. Time series are widely used in engineering, environmental science, physical sciences [1], finance, supply chain management, production and inventory planning. Time series also constitute a theoretical basis in the theory of statistics and dynamic systems [2].

The purpose of this study is to estimate the future with the help of the ensemble model using the forecastHybrid package of the R programming language and compare the results with the individual results.

2. ForecastHybrid Package

The ForecastHybrid library provides a common estimate by combining the predictive capabilities of many functions individually used in the R programming language and environment and in the forecast library [2, 3]. Auto.arima, ets, thetaf, nnetar, stlm, tbats and snaive predictive functions for the predetermined error metric gives the opportunity to estimate the weight or with a certain weight [4].

The ForecastHybrid includes Auto.arima, Nnetr, Stlm, Thetam, Ets, Tbats and Snaive time series models. We can briefly explain these models:

Auto.arima: Estimates AR, MA and difference models using the best parameters. It can be used for both stationary and non-stationary time series.

Nnetr: It is a forward-feed artificial neural network with only a hidden layer. It can also work on time series with seasonal features. With this model, data transformation is performed using the Box-Cox method.

Stlm: Model a series of time, seasonal, trend and irregular components after modeling and makes an estimate behind. Performs seasonal transformation using the last year of the seasonal component. The time series is expected to have a seasonal component for the use of this model.

Thetam: This model trains an exponential smoothing condition space model.

Ets: Exponential time series is known as smoothing.

Tbats: State-space model is exponentially softened by using the Box-Cox conversion model, ARMA errors, trend and seasonal components.

Snaive: Returns an estimate and prediction time intervals from an ARIMA (0,0,0) (0,1,0) m model. Here m represents the seasonal period.

A variety of statistical measures have been calculated, including the Average Absolute Percent Error (MAPE), the root mean square error (RMSE), and the mean absolute error (MAE) to examine the performance of the models using our time series numerical values [5]. MAPE represents the percentage of the mean absolute error.

$$\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}\left|\frac{e_t}{y_t}\right| x100. \quad (1)$$

Here n is the number of observations. RMSE, which is limited to excessive errors in estimation, is used to evaluate models:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}(e_{t+h})^2} \quad (2)$$

The MAE is the mean absolute deviation of the predicted values from the original values. It is the average of all absolute errors and is also called mean absolute deviation (MAD). MAE can be defined as:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}|\boldsymbol{e_{T+\Delta}}|, \quad (3)$$

Here, the meat used as the prediction error of the model can be defined as follows:

$$e_t = \boldsymbol{y_t} - \boldsymbol{\hat{y}_t} \quad (4)$$

Here $\boldsymbol{y_t}$ shows the actual value of the model, $\boldsymbol{\hat{y}_t}$ is the estimated value.

3. Material and Method

The dataset, which forms the basis of this study, was taken directly from the datasets library of the R programming language (version 3.5.1) used for analysis.

For the time series to be trained and to evaluate the results realistically, part of the series is reserved for the training and the remaining part is for testing. For this purpose, the first 132 records (92%) were reserved for training, while the last 12 records (8%) were reserved for the test procedure.

4. Results and Discussion

If the equal value is used for the weights parameter used in the HybridModel model, the weights of the predictive models are equal. However, if this parameter is **insample.errors** or **cv.errors**, it is necessary to use the **errorMethod** parameter. In this case, one or more of the RMSE, MAE and MAPE metrics can be used as the evaluation criteria.

Air passengers dataset was used for the common prediction of multiple models. The 12 and 120-month results obtained as a result of the estimation are shown in Figure 1 and Figure 2, respectively.

As shown in Table I, different usage approaches are shown for the weights of the algorithms used in the estimation. In a weight-based approach, the weight of auto.Arima, Ets, Thetam, Nnetar, Stlm and Tbats models were 0.200, 0.211, 0.051, 0.138, 0.185 and 0.214, respectively. The performance ratio of the weights to be equal was lower than that of the model. The values of the RMSE, MAE and MAPE measurement metrics obtained in the approach where the Weights parameter for the weights were set to, insample.errors ”were 20.17, 15.73, and 0.51, respectively.

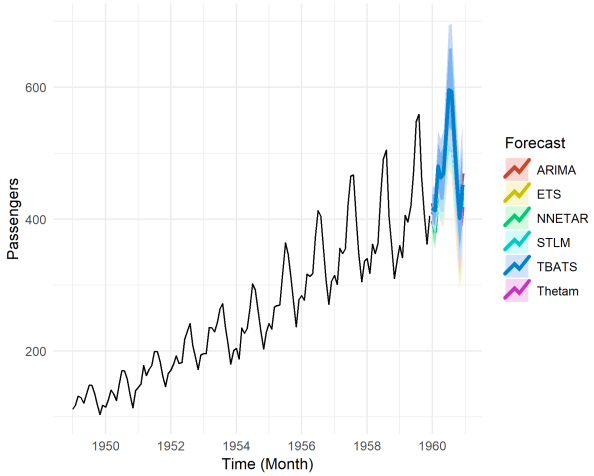


Figure 1: 12-month estimates obtained by using forecasthybrid ensemble models.

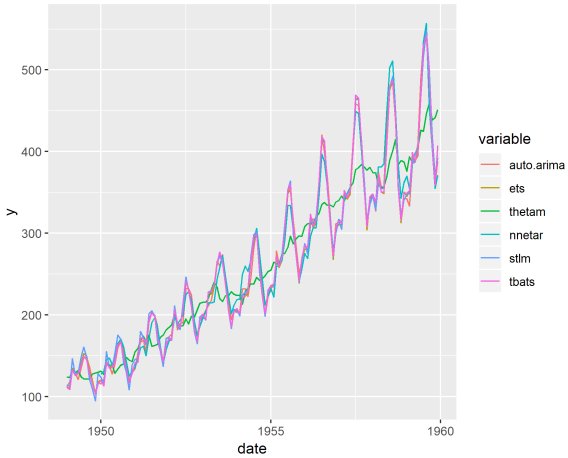


Figure 2: 120-month estimates obtained by using forecasthybrid ensemble models.

Table 1: Comparison metrics of forecasting methods

Weights/Metric	**RMSE**	**MAE**	**MAPE**	**Weights**
equal	20.65	16.26	3.32	0.166
insample.errors	20.17	15.73	3.27	auto.arima (0.200), ets(0.211), thetam(0.052), nnetar(0.138), stlm(0.185), tbats(0.214)

As shown in Figure 2, the lowest performance in the prediction process was obtained with the Thetam model. Without the use of Thetam model, the RMSE value obtained at the end of the estimation process was lower than the previous situation.

5. Conclusions

In this study, future estimations were made with forecastHybrid package support by using monthly time series in R datasets and the results were compared with individual estimation results. In the common estimation process, the performance achieved was higher when the models with low success were not used.

References

[1] Pala, Z., Atici. R. (2019). Forecasting Sunspot Time Series Using Deep Learning Methods. Solar Phys 294 (50), 1-14.
[2] Tattar, P. N. (2018). Hands-on Ensemble Learning with R, Packt Publishing Ltd., Birmingham, UK.
[3] Shaub, David, Ellis, P. (2018). forecastHybrid: Convenient Functions for Ensemble Time Series Forecasts. R package version 2.0.10. https:// CRAN.R- project.org/package=forecastHybrid.
[4] Bates, J. M., Granger, C. W. J. (1969). The Combination of Forecasts, Journal of the Operational Research Society, 20(4), 451-468. doi=10.1057/jors.1969.103.
[5] Namin , S. S., Namin ,A. S. (2018). Forecasting economic and financial time series: Arima vs. LSTM, Texas Tech University, Lubbock, TX, USA.

Multi-robot Path Planning Using Fractional Order Darwinian Particle Swarm Optimization Algorithm

F. Okumuş[1], A. F. Kocamaz[2]

[1]*Malatya Turgut Ozal University, Malatya, Turkey, fatih.okumus@ozal.edu.tr*
[2] *Inonu University, Malatya, Turkey, fatih.kocamaz@inonu.edu.tr*

Abstract

The use of autonomous robots in industrial logistics increases day by day. This increase is accelerating, especially with the industry 4.0 revolution. The autonomous use of robots in the industry also requires them to work in harmony. One of the critical issues facing this requirement is the Path Planning problem. Path planning is the determination of the optimal path traveled with avoiding obstacles by the robot from one point to another. In multi-robot applications, Path Planning requires a collaborative environment. It is also an important issue that robots do not collide with each other in addition to known obstacles. In this study, the Fractional Order Darwinian PSO algorithm is applied for optimal Path Planning of industrial robots with avoiding obstacles and each other. The algorithm was tested in a simulation environment and compared to classical PSO algorithm. Simulation results showed that more successful results were obtained with the Fractional Order Darwinian PSO algorithm.

1. Introduction

Robots, which are incorporated into our lives with technology, now work in many areas for human. With Industry 4.0 and Society 5.0, international steps are being taken to increase and control the effectiveness of robots that will replace people in working life. One of these steps, Cyber Physical Production Systems (CPPS), enables the acquisition of common knowledge and acquisition and ensures the continuity of industries in the developing world [1]. In CPPS, since all of the work is carried out almost exclusively by robots, it is important that the robots work in harmony. Internal logistics operations are autonomous in these systems and robots in logistics are expected to work in harmony. In order for robots to work harmoniously in autonomous logistics, challenges such as positioning, robot path planning and multi-task identification must be addressed. The motivation of this publication is to ensure that industrial robots can reach from one point to another in an optimum path in an environment of obstacles.

Path planning creates a geometric path between a starting point and an end point, avoiding obstacles in the environment. Path planning in robotics is important for one or more robots to work in harmony and with high efficiency. It is a difficult task for the robot to decide which path to go like a human being. The map of the area where the robot works is important in path planning. Known, semi-known and completely unknown maps is different research areas for path planning [2]. In this paper, on a known map, optimum path planning research has been studied.

In the literature, several algorithms are used for path planning on a known map. These algorithms have generally been graph-based search algorithms such as A*, Dijkstra and D* [3]. Votion and Cao proposed a collaborative path planning algorithm for diversity-based multi-vehicle systems [4]. Xiong et al. used a Voronoi-based ant colony algorithm for path planning of multiple autonomous marine vehicles [5]. Zang and Li suggested that they could do fast path planning by their method. They created a hybrid method using Dijkstra, A * and rolling window principle in their study. [6]. Qian and Yinfa proposed an improved Dijkstra algorithm that first finds the length of the path piece by piece and then compares them to determine the best global path [7]. Meta-heuristic algorithms are not preferred for path planning alone because of the problem of inability to achieve global best in some cases. Meta-heuristic algorithms are mostly used to develop iterative search algorithms such as A * and Dijkstra, but there are also studies using mainly meta-heuristic methods. [8]–[14].

In this study, a new solution for path planning using Fractional Order Darwinian Particular Swarm Optimization (FODPSO) was proposed. With a fractional approach, FODPSO can avoid the problem of the traditional PSO being stuck in the local optimal solution due to the early convergence of swarms [15]. The proposed method is compared with the traditional PSO algorithm and the results are given in detail.

In the first part of the publication, Fractional Order Darwinian PSO algorithm is explained and the effects of fractional approach are discussed. In the second part, Path Planning algorithm is explained by giving a pseudo code. In the third chapter, the experimental results of the algorithm are shown and compared with the traditional PSO algorithm.

2. Fractional Darwinian PSO

FODPSO is the developed version of the PSO and the Darwinian PSO, and seeks the solutions to the problem of early convergence of swarms [15]. FODPSO removes swarms from the solution pool that approach optimal ahead of time or non-optimal solution. It also creates a new swarm group of particles with the best solution, providing an environment where information is inherited. Couceiro and Ghamisi have repeatedly compared FODPSO with traditional PSO and DPSO and have clearly demonstrated superiority in every respect [16]. In conventional PSO, each n particle moves in a multidimensional field with respect to positional x_n^s and velocity v_n^s in t[th] iteration to model the particle. These values vary depending on $x_{1_n}^s$ known as local best and $x_{2_n}^s$ known as global best. This dependency is expressed as:

$$v_n^s[t+1] = wv_n^s[t] + \sum_{i=0}^{2} p_i r_i\left(x_{i_n}^s[t] - x_n^s[t]\right) \quad (1)$$

$$x_n[t+1] = x_n[t] + v_n[t+1] \quad (2)$$

Fraction calculation (FC) can produce successful results when integer-regular calculation fails. Couceiro and Ghamisi stated that fractional derivatives are an excellent tool in defining memory and hereditary properties of processes. [16]. Assuming the inertia effect value (w) as 1 for the speed calculation in Equation (1), the following equation will be obtained:

$$v_n^s[t+1] = v_n^s[t] + \sum_{i=0}^{2} p_i r_i\left(x_{i_n}^s[t] - x_n^s[t]\right) \quad (3)$$

This equation can be written as follows:

$$v_n^s[t+1] - v_n^s[t] = \sum_{i=0}^{2} p_i r_i\left(x_{i_n}^s[t] - x_n^s[t]\right) \quad (4)$$

Here $v_n[t+1] - v_n[t]$ appears to be a derivative expression. This expression is a discrete time version of the fractional calculation. When we consider this expression as α = 1, the following equation will be obtained:

$$D^{\alpha}[v_n^s[t+1]] = \sum_{i=0}^{2} p_i r_i\left(x_{i_n}^s[t] - x_n^s[t]\right) \quad (5)$$

Based on fraction calculation, the derivative of the velocity value can be expressed as a real number between 0 and 1. This results in a softer variability and a more permanent memory effect. Therefore, when the fractional approach is taken, Equation (5) can be rewritten as follows:

$$v_n^s[t+1] = -\sum_{k=0}^{\infty} \frac{(-1)^k \Gamma(\alpha+1) v_n^s[t+1-kT]}{\Gamma(k+1)\,\Gamma(\alpha-k+1)} + \sum_{i=0}^{2} p_i r_i\left(x_{i_n}^s[t] - x_n^s[t]\right) \quad (6)$$

3. Path Planning

In this study, since the path planning is carried out on a known map, the coordinates of all crossing points and obstacles on the map are taken as input to the system. The aim of the algorithm is to minimize the path from the start point to the goal point. Of course, this path should also be created by avoiding obstacles. It should be considered as a constraint that the path does not cross over obstacles. Thus, the pseudocode in Algorithm 1 is realized.

```
Set xmin, ymin, xmax, ymax                    // Initialize Map Size
Set xobs, yobs, robs                          // Set obstacle coordinates and radius
Initialize Array[S]                           // Array of Swarm Groups
Set default values => vn^s[1], xn^s[1], X1^s[1], X2^s[1]

Loop (Main Program)
   Foreach S in List[S]
      Initialize swarms in S (Go to Initialize Swarm Function)
      Compute group (Go to Compute Swarm Group Function )
      Spawn new groups from the group
      Kill 'failed' groups
   End
Until iteration

Compute Swarm Group Function
   For each particle n in S
      Compute the fitness of n
      Find out the best distribution of fitness
      Update X1^s[t] and X2^s[t]
      Update vn^s[t+1]                        // fractional order computation
```

```
            Update x_n^s[t+1]
            If group S gets better
                Extend the life of the group and generate a new group
             Else
                Punish by shortening the life of the group
        End
    Return

    Initialize Swarm Function (xmin, ymin, xmax, ymax, xobs, yobs, robs)
        For 1 to N                          // N = Count of Population
            For 1 to N_COUNT                // N_ COUNT = Swarm count
                Swarms[i] = Random coordinate matris of path from start to goal
                            (coordinates without obstacle points)
             End
        End
    Return Swarms
```

Algorithm 1. Pseudocode of FODPSO

4. Experimental Results and Comparison

To see the results of the study, random obstacles were placed in a *20x20* area in a simulation environment and a test map was created. The starting point was taken as $x = 0$ and $y = 0$, and the target point was determined as $x = 19$ and $y = 19$. The position and radius of the obstacles on the map are randomly selected. After the map was created, an optimal path calculation was performed with PSO and FODPSO algorithms. The test showed that in some cases both the traditional PSO algorithm and the FODPSO algorithm could not escape the local solution and produced results far from the global solution (optimal path). Figure 1 shows an example of test results that produce local and global results.

Figure 1. Optimal path calculation from start point to goal point with FODPSO. a) Local solution (path length = 32.023). b) Global solution, optimum path (path length = 26.932).

In Figure 1, the starting point (x = 0, y = 0) is shown in yellow. The goal point (x = 19, y = 19) is shown in green. Blue circles represent obstacles on the map. The algorithm has been run 50 times and has guaranteed the achievement of the goal, avoiding obstacles each time. However, the optimum path was not achieved in all tests. This is due to the inability to avoid local solutions due to early convergence of swarms in PSO. As can be seen in Figure 1.a., the optimum path has not been found despite avoiding obstacles. In Figure 1.b, both obstacles were avoided and the optimum path was found. The advantage of the algorithm over traditional PSO is that it is more stable in achieving a global solution. Figure 2 shows the performance of the global solution of PSO and FODPSO algorithms running 50 times each in the same environment.

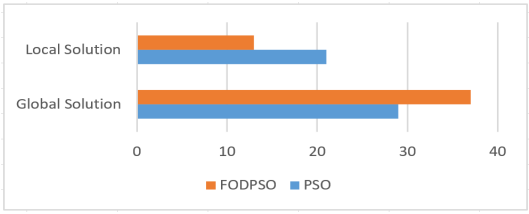


Figure 2. Number of local solutions and global solutions for PSO and FODPSO run 50 times.

As shown in Figure 2, in 50 times run, FODPSO has achieved 37 and PSO has achieved 29 times global solution. In addition, tests have shown that FODPSO showed more stable behavior while producing the right solution. Stability results are shown in Figure 3. After 50 runs, the PSO algorithm showed a fluctuating behavior, while the FODPSO algorithm produced more stable results.

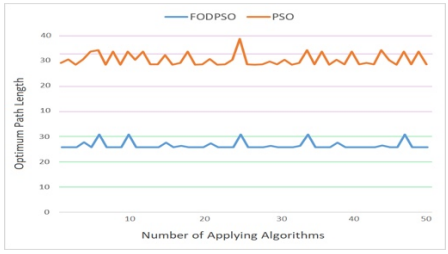


Figure 3. Stability graph for PSO and FODPSO run 50 times

5. Conclusions

In this study, Fractional Order Darwinian PSO algorithm has been used to solve the path planning problem in industrial autonomous robots. The algorithm has been applied to find the optimal path between a starting point and a target point in a simulation environment with obstacles. In order to measure the performance of the algorithm, both FODPSO algorithm and traditional PSO algorithm were applied 50 times in the same environment and the results were compared. According to the comparison, the FODPSO algorithm produced more stable results than the traditional PSO algorithm. However, the test results showed that; Meta-heuristic algorithms that are unlikely to reach a global solution may not always produce optimal results, such as graph-based algorithms such as A * and Dijkstra.

References

[1] H. Panetto, B. Iung, D. Ivanov, G. Weichhart, and X. Wang, "Challenges for the cyber-physical manufacturing enterprises of the future," *Annu. Rev. Control*, no. xxxx, 2019.
[2] G. Klančar *et al.*, "Path Planning," *Wheel. Mob. Robot.*, pp. 161–206, Jan. 2017.
[3] F. Okumuş and A. F. Kocamaz, "Comparing Path Planning Algorithms for Multiple Mobile Robots," *2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018*, no. 2013, pp. 1–4, 2019.
[4] J. Votion and Y. Cao, "Diversity-Based Cooperative Multivehicle Path Planning for Risk Management in Costmap Environments," *IEEE Trans. Ind. Electron.*, vol. 66, no. 8, pp. 6117–6127, Aug. 2019.
[5] C. Xiong, D. Chen, D. Lu, Z. Zeng, and L. Lian, "Path planning of multiple autonomous marine vehicles for adaptive sampling using Voronoi-based ant colony optimization," *Rob. Auton. Syst.*, vol. 115, pp. 90–103, May 2019.
[6] H. Zhang and M. Li, "Rapid path planning algorithm for mobile robot in dynamic environment," *Adv. Mech. Eng.*, vol. 9, no. 12, p. 168781401774740, Dec. 2017.
[7] L. Qian and Z. Yinfa, "A Shortest Path Algorithm Under Specified Nodes Constraint," in *2018 International Conference on Sensing,Diagnostics, Prognostics, and Control (SDPC)*, 2018, pp. 686–689.
[8] E. Masehian and D. Sedighizadeh, "A multi-objective PSO-based algorithm for robot path planning," in *2010 IEEE International Conference on Industrial Technology*, 2010, pp. 465–470.
[9] P. I. Adamu, H. I. Okagbue, and P. E. Oguntunde, "Fast and Optimal Path Planning Algorithm (FAOPPA) for a Mobile Robot," *Wirel. Pers. Commun.*, vol. 106, no. 2, pp. 577–592, May 2019.
[10] S. B. Germi, M. A. Khosravi, and R. FesharakiFard, "Adaptive GA-based Potential Field Algorithm for Collision-free Path Planning of Mobile Robots in Dynamic Environments," in *2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM)*, 2018, pp. 28–33.
[11] U. Orozco-Rosas, O. Montiel, and R. Sepúlveda, "Mobile robot path planning using membrane evolutionary artificial potential field," *Appl. Soft Comput.*, vol. 77, pp. 236–251, Apr. 2019.
[12] J. Zhang, Y. Zhou, and Y. Zhang, "Multi-objective Robot Path Planning based on Bare Bones Particle Swarm Optimization with Crossover Operation," in *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2018, pp. 330–335.
[13] T. Xiao-Ling, "Trajectory Stability Control of the lidar tracking robot Based on Improved Particle Swarm Filter Algorithm," *J. Phys. Conf. Ser.*, vol. 1168, p. 022107, Feb. 2019.
[14] A. S. Al-Araji, A. K. Ahmed, and M. K. Hamzah, "Development of a Path Planning Algorithms and Controller Design for Mobile Robot," in *2018 Third Scientific Conference of Electrical Engineering (SCEE)*, 2018, pp. 72–77.
[15] O. Akdağ, F. Okumuş, A. F. Kocamaz, and C. Yeroğlu, "Fractional Order Darwinian PSO with Constraint Threshold for Load Flow Optimization of Energy Transmission System," *Gazi Univ. J. Sci.*, vol. 31, no. 3, pp. 831–844, Sep. 2018.
[16] M. Couceiro and P. Ghamisi, *Fractional order Darwinian particle swarm optimization : applications and evaluation of an evolutionary algorithm*. .

A Parallel Comparison of Several String Matching Algorithms Employing Different Strategies

M. Nazli[1], O. Cankur[2], A. Ozsoy[3]

[1]Hacettepe University, Ankara, Turkey, mengu@hacettepe.edu.tr
[2]Hacettepe University, Ankara, Turkey, onurcankur@hacettepe.edu.tr
[3]Hacettepe University, Ankara, Turkey, adnan.ozsoy@hacettepe.edu.tr

Abstract

String matching is one of the classic problems in Computer Science and constitutes as a basic building block in many applications. In this paper, we present a parallel approach for improving the performance of exact string matching algorithms via Graphic Processing Unit (GPU). We present the GPU parallel implementations of Brute force (BF), Boyer Moore (BM), Backward Nondeterministic DAWG Matching (BNDM) and Backward Oracle Matching (BOM) algorithms and compare their performances with their CPU counterparts. These algorithms are selected to examine different string matching strategies like comparison, automaton and bit-parallel. We also discuss possible optimization strategies and examine their impact on the performance of these parallel GPU algorithms. The experimental results based on the tests using different parameters are presented in the paper.

1. Introduction

The use of a GPU instead of a CPU to perform general purpose computations is known as General Purpose computing on Graphics Processing Units (GPGPU). CUDA is a parallel programming platform created by NVIDIA Corporation, one of the leading GPU manufacturers worldwide. It allows software developers to use CUDA capable GPUs for general purpose processing[1]. Many studies have demonstrated that the GPU based solutions can be superior compared to CPU based ones. Huang et. al. proposed a modified Wu-Manber algorithm that run efficiently on GPUs for network intrusion detection.[2] Xu et. al. presented a bit-parallel method variation for multiple approximate string matching on GPU.[3]

This paper focuses on the parallelization of the Bruteforce[4], Boyer-Moore[5], Backward Nondeterministic DAWG Matching[6] and Backward Oracle Matching[7] algorithms through their implementations on GPU using CUDA toolkit. These algorithms are chosen because of their relevance in the literature and to gain insight on how various string matching strategies behave under similar circumstances and how efficient they are with a parallel implementation.

The rest of this paper is organized as the following. Section II provides background information on string matching and the existing approaches on the topic. The details about the methodology followed for the experiments and specifications of the experimentation rig will be shared in Section III. The results obtained by the GPU versions of algorithms and the comparison with their CPU counterparts are examined in Section IV. Section V draws a conclusion based on the experimental results and provides potential research directions for future work.

2. Methodology

The experiments are performed on a high end workstation with 64-bit ten-core Xeon processors (Intel Xeon Silver 4114) running at 2.2Ghz,128 GB of system memory and Nvidia 1080ti GPU (with 28 multiprocessors, 2584 CUDA cores and 11 GB of GDDR5X memory). This GPU is capable of running up to 57,344 active threads. The operating system used on the workstation is Ubuntu 18.04.2 LTS. To decrease random variation caused by external factors, the time results are averaged over 100 runs. String matching problem is inherently amenable to divide-and-conquer. Parallel versions of selected string matching algorithms are implemented using the divide-and-conquer strategy. Memory operations are another big factor on the performance of GPGPU applications. To mitigate the performance loss, pinned memory feature of CUDA library is used. Using this feature allows the allocation of page-locked memory on the host side (CPU) which reduces memory transactions and speeds up the operation up to 2-3 times. Tests are performed with both this feature enabled and disabled.

Tests on CPU versions of the algorithms are performed using the code obtained from SMART project, an open source toolkit for string matching research. These algorithms are implemented in ANSI C language and compiled using gcc tool (version 7.3.0) on our linux system without any optimization arguments. The GPU versions of

algorithms are implemented in C/C++ language and compiled using nvcc tool (version 10.0.130) for the compute capability of v3.5 and above.

A recent version of global wikipedia dump is used as the dataset to perform structured text tests. Smaller test files with the size of 1MB, 10MB, 100MB and 1GB are generated to study the effect of text size on performance. The query string is constructed from randomly chosen subsequences of text file with a length of 3, 6, 10, 20 and 32 characters. The patterns are limited to 32 characters so the corresponding search information can fit into a single 32-bit wide word variable if necessary.

3. Experimental Results

Figure I presents the speedup achieved by the utilization of pinned memory. On average, a speedup of 3x is achieved for memory transactions which shaves off a significant amount of time from our operations. Using this

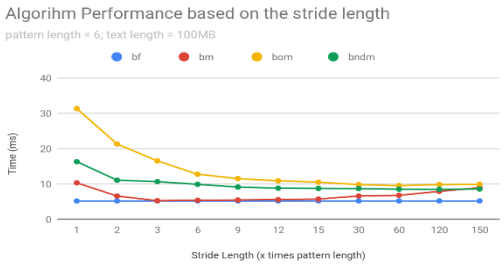


Figure 2-stride length optimization

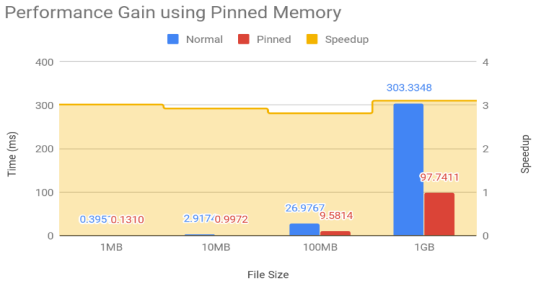


Figure 1-using pinned memory

option is almost always desirable because it is independent of application and the only downside is that it reserves the used memory portion on CPU side. This reduced memory may result in degraded performance for CPU operations, but the inclusion of a large enough memory or the scheduling of CPU operations can solve this problem. Our other tests are performed using pinned memory for apparent performance gain. Our test results for the granularity is presented in Figure II. We call sub-task length parameter stride length based on the idea of each new thread starting its search at one step away from previous thread. Sub-task test sizes are selected based on the length of our pattern string. These algorithms are performing full optimal shifts on most of their search cycles which means they are doing pattern sized leaps at a time. By selecting stride lengths at the multiples of pattern lengths, we aim to take advantage of these full jumps as much as possible.

As expected, brute force algorithm is not affected by the changes on stride length because it has no mechanism to do longer shifts. Also, the other algorithms all have some kind of state memories which has to be populated at the beginning of operation before the process gets better. Keeping stride length short increases the instances of these "cold-start" operations and denies the algorithms from their full potential. Other algorithms are affected by the changes on stride length similarly with slight differences. All other algorithms signal improvement as the stride length increases with the exception of Boyer-Moore getting worse after a certain point. Boyer-Moore algorithm shows a sweet spot between 3x and 12x stride lengths before degrading in performance. Automata based algorithms show improvement without any degradation on longer strides to a point where the performance gain becomes insignificant. Backward Nondeterministic DAWG Matching gets to this point around 12x while Backward Oracle Matching keeps improving up to around 30x. It should be noted that there is an inverse relation between the stride length and parallelism, the lower stride lengths should be preferred in case of similar performance to achieve high parallelism.

			text=100M				
	Tag	Name	p=3	p=6	p=10	p=20	p=32
GPU	bf	Bruteforce	16.0577	15.8879	14.6908	14.8692	15.8929
	bm	Boyer Moore	16.2718	14.4336	13.158	12.56	11.83
	bndm	Backward nondet. DAWG Matching	20.8194	16.5395	14.17	13.0333	12.0685
	bom	Backward Oracle Matching	23.7204	22.0784	18.8943	12.2762	12.3532

Figure 4-GPU performance table

			text=100M				
	Tag	Name	p=3	p=6	p=10	p=20	p=32
CPU	bf	Bruteforce	650.925	656.4636	578.9435	561.5968	657.2983
	bm	Boyer Moore	605.6879	422.2653	326.4597	289.4253	258.0711
	bndm	Backward nondet. DAWG Matching	596.3634	444.9036	324.793	300.2087	256.781
	bom	Backward Oracle Matching	821.9662	679.3488	551.6334	410.9978	318.0936

Figure 3-CPU performance table

Realistically, largest performance improvement becomes apparent near 2x-6x band. From the Figures 3,4, and 5, you

can see that after parallelizing the algorithms. 20x-30x performance gain is achieved. Results for 100 MB file can be seen from figures, and similar results are gained after testing with 1 MB, 10 MB, and 1 GB files.

			text=100M				
	Tag	Name	p=3	p=6	p=10	p=20	p=32
Speed Up	bf	Bruteforce	40.54	41.32	39.41	37.77	41.36
	bm	Boyer Moore	37.22	29.26	24.81	23.04	21.81
	bndm	Backward nondet. DAWG Matching	28.64	26.9	22.92	23.03	21.28
	bom	Backward Oracle Matching	34.65	30.77	29.2	33.48	25.75

Figure 5-speed up table

4. Conclusion

Performance of the parallelization of Bruteforce, Boyer-Moore, Backward Nondeterministic DAWG Matching, and Backward Oracle Matching string matching algorithms are testes with 4 different sized(1MB, 10MB, 100MB, 1GB) files. Results shows that significantly high speed improvement can be achieved by parallelizing. In addition to that, as explained in detail, pinned memory is used and stride length optimization is researched to improve the performance. It can be seen that gaining improvement is also possible with these approaches. Further development can be achieved using constant or shared memory to store the pattern and working on memory access patterns.

Acknowledgement

This research was supported by TUBITAK (The Scientific and Technological Research Council of Turkey) grant number 117E142.

References

[1] V. Simek and R. R. Asn, “Gpu acceleration of 2d-dwt image compression in Matlab with cuda,” Computer Modeling and Simulation, UKSIM European Symposium on, vol. 0, pp. 274–277, 2008.

[2] N.-F. Huang, H.-W. Hung, S.-H. Lai, Y.-M. Chu, andW.-Y. Tsai. A gpu-based multiple-pattern matching algorithm for network intrusion detection systems. In Advanced Information Networking and Applications Workshop, pages 62–67, March 2008.

[3] K. Xu, W. Cui, Y. Hu, and L. Guo. Bit-Parallel Multiple Approximate String Matching based on GPU. Procedia Computer Science, 17(0):523 – 529, 2013.

[4] T. H. Cormen and C. E. Leiserson and R. L. Rivest and C. Stein. Introduction to Algorithms. MIT Press, 2001.

[5] R. S. Boyer and J. S. Moore. A fast string searching algorithm. Commun. ACM, vol.20, n.10, pp.762--772, 1977.

[6] G. Navarro and M. Raffinot. A Bit-parallel Approach to Suffix Automata: Fast Extended String Matching. n.TR/DC--98--1, 1998.

[7] C. Allauzen and M. Crochemore and M. Raffinot. Factor oracle: a new structure for pattern matching. SOFSEM'99, Theory and Practice of Informatics, Lecture Notes in Computer Science, n.1725, pp.291--306, Springer-Verlag, Berlin, Milovy, Czech Republic, 1999.

Delayed Constraint Generation for the Type-II Assembly Line Balancing Problem

A. Hamzadayı[1], T. Bektaş[2]

[1] *Department of Industrial Engineering, Van Yüzüncü Yil University, Van, Turkey, alperhamzadayi@yyu.edu.tr*
[2] *University of Liverpool Management School, Liverpool, United Kingdom, T.Bektas@liverpool.ac.uk*

Abstract

SALBP-2 is an NP-hard problem and consists of assigning tasks to a given number of work stations aiming at maximizing the production rate by considering the possible precedence constraints between the tasks. For solving SALBP-2, we develop an iterative approach that use the general structure of Benders decomposition algorithm with a simple yet effective enhancement mechanism over the basic generic formulation of the problem. The algorithms are tested on a 96 set of benchmark instances and numerically compared with the best mixed-integer linear programming (MILP) formulation of the problem in literature, solved using a commercial optimizer and results reveal the superiority of Benders decomposition like algorithm against MILP.

Introduction-I

Assembly lines are the special case of flow-line production system consist of succeeding workstations, connected by a material handling system, performing a set of tasks on the product passing through them. Each workstation must complete the tasks within a fixed time period called the cycle time, and each task has an execution time and precedence relationships with other tasks. The simple assembly line balancing problem (SALBP) deals with the allocation of tasks among workstations while respecting the partial ordering of the tasks in order to optimize a given objective function. The problem is known as SALBP-1 when the goal is to minimize the number of workstations in a predefined cycle time, SALBP-2 when the goal is to minimize the cycle time by considering the given a predefined number of workstations.

There are more successful applications of Benders Decomposition and its variants to solve some combinatorial optimization problems like presented by [1]. However, there are very few studies to solve the varied assembly line balancing problems. Only studies we are aware of are [2-4], but they tackle different problems to what we study here.

In this paper, we mainly contribute to the literature by developing a Benders decomposition algorithm for solving SALBP-2. The algorithm is tested on a set of benchmark instances and numerically compared with the best mixed-integer linear programming formulation of the problem and the well-known iterative approaches, and solved using a commercial optimizer.

The remainder of this paper is structured as follows. Section 2 presents a basic integer programming models for SALBP-2. The proposed algorithm is described in Section 3. Computational results are presented in Section 4, followed by conclusions in Section 5.

Integer Programming Models for SALBP-2-II

A basic version of the mathematical model proposed by [5]. For a task $i \in N$, let F_i denote the set of its immediate followers, and P_i the set of its immediate predecessors. Let S be the set of workstations. The number of workstations m is part of the problem instance. In this case we set $S = \{1, \ldots, m\}$.

$$\text{Minimize } c \tag{1}$$

Subject to:

$$\sum_{s \in S} x_{si} = 1, \qquad \forall i \in N \tag{2}$$

$$x_{tj} \leq \sum_{(s \in S | s \leq t)} x_{si}, \qquad \forall i \in N, j \in F_i, t \in S \tag{3}$$

$$\sum_{i \in N} t_i x_{si} \leq c, \quad \forall s \in S \tag{4}$$

$$x_{si} \in \{0,1\}, \quad \forall i \in N, s \in S \tag{5}$$

$$c \in \mathbb{R} \tag{6}$$

In this basic model, x_{si} is a binary variable which is set to 1 if and only if task $i \in N$ is assigned to workstation $1 \leq s \leq m$. The objective function (1) minimizes the cycle time. The constraint sets (2) ensure that each task $i \in N$ is assigned to exactly one workstation $1 \leq s \leq m$. Constraint sets (3) ensure that a task has to be assigned to an earlier workstation than its followers. More specifically, if task $i \in N$ is assigned to a workstation $1 \leq s \leq m$, all tasks $j \in F_i$ must be assigned to workstations $1 \leq t \leq m$ with $s \leq t$. The cycle time constraint sets (4) guarantee that the cycle time is not exceeded by workstation time of any workstation. The constraint sets (5) forces x_{si} to take a binary value.

Application of Benders Decomposition-III

Let $M(c; x)$ denote the formulation (1)-(6) where $x = \{x_{si} | i \in N,\ s \in S\}$ and $c = \{c | \forall s \in S\}$ are the vectors of the variables. Suppose that variables x are now fixed as $x = \hat{x} \in X = \{x | x \text{ satisfies } (2), (3) \text{ and } (5)\}$. The resulting formulation, shown by $M(c; \hat{x})$, consists only of the variables c, and the constraints of which are assigned the dual variables $\alpha = \{\alpha_s \geq 0 | s \in S\}$ corresponding to constraints (4). The dual $D(\alpha; \hat{x})$ of $M(c; \hat{x})$ is given by the following:

$$\text{Maximize} \sum_{s \in S} \alpha_s \sum_{i \in N} \hat{x}_{si} t_i \tag{7}$$

$$\sum_{s \in S} \alpha_s = 1 \tag{8}$$

$$\alpha \in \mathbb{R} \tag{9}$$

The following Benders optimality cuts are obtained. where z is a lower bound on the optimal solution value of $M(c; x)$.

$$z \geq \sum_{s \in S} \sum_{i \in N} \alpha_s x_{si}$$

Using this result, we are now ready to present the following reformulation of $M(c; x)$, referred to as the master problem constructed using the set P_D of extreme points of $D(c; x)$ and shown as MP(P_D) below:

$$\text{Minimize } z \tag{10}$$

Subject to:

$$\sum_{s \in S} x_{si} = 1, \qquad \forall i \in N \tag{11}$$

$$x_{tj} \leq \sum_{(s \in S | s \leq t)} x_{si}, \qquad \forall i \in N, j \in F_i, t \in S \tag{12}$$

$$z \geq \sum_{s \in S} \sum_{i \in N} \alpha_s x_{si}, \qquad \alpha \in P_D \tag{13}$$

$$x_{si} \in \{0, 1\}, \quad \forall i \in N, s \in S \tag{14}$$

It is clear that the number of constraints on the cycle time in the original model will be as much as the number of workstations used in the problem.

Computational Results-IV

This section presents a computational study to assess the performance of the Benders decomposition based algorithm. The algorithm is coded in Visual C++, using CPLEX 12.7.1 as the solver. We have used an Intel Core i5-2450M computer with a 2.5 GHz CPU and 4 GB of memory. The tests are conducted on 15 instances of the SALBP-2 available at https://assembly-line-balancing.de/.

Table 1. Comparisons of the methods.

Problem Name	NF2 CPU	ABD(NF2) CPU	SMM CPU	IEC CPU
Buxey(7)	1.73	13.09	2.09	**1.07**
Buxey(8)	1.61	39.16	0.98	**0.49**
Buxey(9)	2.73	77.83	2.55	**1.29**
Buxey(10)	1.50	40.31	1.75	**1.16**
Buxey(11)	8.02	153.66	8.45	**2.87**
Buxey(12)	3.83	117.05	2.47	**1.60**
Buxey(13)	16.83	290.48	20.84	**3.54**
Buxey(14)	2.25	104.41	1.25	**0.16**
Gunther(6)	0.38	4.00	0.67	**0.22**
Gunther(7)	1.28	10.41	2.22	**1.23**
Gunther(8)	2.16	22.13	3.02	**2.03**
Gunther(9)	**1.30**	26.14	2.02	2.00
Gunther(10)	6.30	596.81	7.69	**2.67**
Gunther(11)	7.25	600.00	8.14	**2.43**
Gunther(12)	6.89	600.00	5.16	**3.73**

We compare benders implementation of basic model (shown by IEC) with basic mathematical model given in Section 2 (shown by SMM), the best-known formulations reported in the current literature [6] (shown by NF2) and the automated Benders decomposition of NF2 that is available within the software (shown by ABD). All of the models were run in a deterministic mode with two threads for a maximum time of 600 seconds.

In Table 1, the first columns shows the instance names (parentheses given within the instance names show the number of workstation used as a problem parameter) and the other columns show the computational time reached by the each method separately. As can be seen from Table 1, IEC is the fastest in all instances except for instance Gunther (9) for which NF2 shows a better performance. These results suggest that a benders implementation for basic model without any enhancements is highly effective and seem to provide encouraging results for the SALBP-2.

Conclusions-V

In this paper, we present an iterative approach that uses the general structure of Benders decomposition algorithm on the basic generic formulation of the problem. The algorithm is tested on a set of benchmark instances and numerically compared with the best mixed-integer linear programming formulation of the problem, solved using a commercial optimizer. Results reveal the superiority of Benders decomposition algorithm against the best mixed-integer linear programming formulation of the problem.

References

[1] Bektas, T. (2012). Formulations and benders decomposition algorithms for multidepot salesmen problems with load balancing. *European Journal of Operational Research,* 216(1), 83–93.

[2] Hazır, Ö., and Dolgui, A. (2013). Assembly line balancing under uncertainty: Robust optimization models and exact solution method. *Computers & Industrial Engineering*, 65(2), 261–267.

[3] Hazır, Ö., and Dolgui, A. (2015). A decomposition based solution algorithm for u-type assembly line balancing with interval data. *Computers & Operations Research*, 59, 126–131.

[4] Akpinar, S., Elmi, A., and Bektaş, T. (2017). Combinatorial benders cuts for assembly line balancing problems with setups. *European Journal of Operational Research*, 259(2), 527–537.

[5] Baybars, I. (1986). A survey of exact algorithms for the simple assembly line balancing problem. *Management science*, 32(8), 909–932.

[6] Ritt, M., and Costa, A.M. (2018). Improved integer programming models for simple assembly line balancing and related problems. *International Transactions in Operational Research*, 25(4), 1345–1359.

Application-Layer Dos Attack Detection Using Machine Learning

M. Alauthman[1], G. Albesani[2]

[1]Faculty of information technology, Zarqa University, Zarqa, Jordan,, malauthman@zu.edu.jo
[2] Faculty of information technology, Zarqa University, Zarqa, Jordan, ghbesani@zu.edu.jo

Abstract

The greatest risk to internet networks is denial of service (DoS) attacks. Where availability for network security is the primary goal, DoS attacks aim at breaking down servers to stop or diminish the availability of services by preventing or limiting the use of these services by legitimate users. Different approaches to application-layer Dos attack detection have been suggested, but this type of attack remains a major challenge as the current approaches to DoS attack detection such as lack of early detection. This research aims to increase the rate of detection by investigating the best features of the traffic network that can achieve high detection rates. In this research, Random forest, Support Vector Machine, Logistic Regression and Gaussian NB are tested to be as a malicious activity detector. Our method achieves around 99.0% detection rate using Random forest.

1. Introduction

The massive number of machines connected to the Internet has rapidly increased from 500 million in 2003 to 12.5 billion in 2010 and is expected to grow to 50 billion by 2020 [1]. Although convenient, the use of Internet-based services poses many security challenges. The main security threat is malicious software, also known as malware. Denial of Service (DoS) is a type of attack that effort to stop legitimate traffic from reaching network resources. The word DoS points to an individual system starting the attack toward its targeted victim. DDoS points to when various systems are launching out DoS attacks concurrently.

Application layer Denial of Service (DoS) attacks differs from classical DoS attacks because they are seen as quiet and secretive, complicated and practically not able to be detected at the network layer. Several characteristics make application layer DoS attacks stick out from the other traditional DoS attacks 1)Limited resources consuming, 2) Targeted damage, 3) Stealthiness [2].

This paper is structured as follows: in Section 2, a brief introduction about Dos attack. In Section. 3. an overview of DoS Detection approach using machine learning is explained . Dataset and experimental results are discussed in Section 4. Finally, conclusion and future research direction are presented in section 5.

Literate review

Recent years have testified many DoS and DDoS detection methods which can be classified as anomaly-based, signature-based and data mining-based [3]. Other researchers such as [4] have classified DDoS detection systems based on attack environment such as cloud, SDN network and the Application layer. Due to the high availability of malicious data and rapid progress in cyber-attacks, machine learning, data mining, and any related field have been adopted to meet the challenges of cybersecurity. Machine learning can be put to practical use in scan detection, signature detection, network traffic profiling, anomaly detection, and privacy-preserving data mining.

Zhang et al.[5] use DARPA MIT KDD Cup 1999 dataset, and they applied Random Forests (RFs) and evaluated their method performance using ROC. The excellent detection rate was reached by maintaining a lower FP rate compared with other uncontrolled anomalies systems.

Kacha et al. [6] introduced a unique manner matching method for enhancing the efficiency of Snort IDS. To decrease the percentage of false alarms, it joined anomaly and misuse detection methods. This approach gives a quick packet checking with less consuming of the IDS resource comparison to the regular one. Finally, the anomaly detection is used widely in network security and there are many introduced methods such [7-11] in order to reduce the effect of network attacks.

2. Overview of DoS Detection Approach using Machine Learning

There are three essential parts for the proposed DoS detection system: network traffic capturing, traffic feature extraction and DoS detection as stated in Figure 1. In features extraction, we extracted connection features between IPs. Then the DoS detection system, distinguishes activity into two classes: malicious activity or normal legitimate activity. Therefore, RF, Support Vector Machine (SVM), Logistic Regression (LR) and Gaussian NB [12] are tested to be as a malicious activity detector.

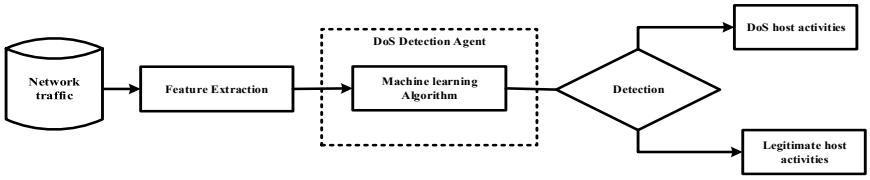


Figure 1. Overview of DoS detection phases.

The feature extractor extracts the essential features to be useful in action analysis engines including the order of system calls, start time, a period of network flow, source IP, source port, destination IP, destination port, protocol, number of bytes, and number of packets. In our research, we selected and evaluated the following features with aiming at finding the features that have a high detection rate for DoS application layer attacks. Table 1 list the selected features.

Table 1. Selected features

Features	Descriptions	Features	Descriptions
F1	# of packets in time-window.	F14	# of sent push packets in time-window.
F2	# of sent packets in time-window.	F15	# of received push packets in time-window.
F3	# of received packets in time-window.	F16	Ratio of sent packets in time-window.
F4	# of sent SYN packets (Initiates a connection) in time-window.	F17	Ratio of received packets in time-window.
F5	# of received SYN packets in time-window.	F18	Ratio of SYN packets in time-window.
F6	# of sent reset packets in time-window.	F19	Ratio of reset packets in time-window.
F7	# of receive reset packets in time-window.	F20	Ratio of sent payload packets in time-window.
F8	# of sent ACK packets (Acknowledges received data) in time-window.	F21	Ratio of received payload packets in time-window.
F9	#of received ACK packets in time-window.	F22	# of source ports in time-window.
F10	# of sent reset ACK packets in time-window.	F23	# of destination ports in time-window.
F11	#of received reset ACK packets in time-window.	F24	Connection duration.
F12	# of sent push ACK packets in time-window.	F25	Ratio of time between sent packets.
F13	# of received push ACK packets in time-window.		

3. Experiment and results

CIC DoS dataset (ISCX) collected for two weeks by the University of New Brunswick, Canada (Canadian Institute for Cybersecurity datasets). This dataset was created by [13] who have set up a victim web server in a testbed environment. In this accuracy (ACC), area under the ROC (AUC), and Mean Square Error (RMSE) are used for evolution the result of classifier:Accuracy (ACC) shows the rate of correct predictions of all cases.

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - t_i)^2}{N}} \quad (2)$$

Where N indicates the number of input samples, yi represents the real output of the model, and ti is the samples targets. To ensure the quality of the learned neural network agent, K-fold cross-validation method is used to estimate the error rate of classifiers. In addition, the standard deviation between cross-validation folds' results was estimated to measure the stability of the results in the offline phase. As shown in table 2. We conduct our experiments using classical machine learning algorithms that are used for DoS detection such as LR, Linear Discriminant Analysis (LDA), KNN Classifier, Decision Tree (DT) Classifier, Gaussian NB and RF. Table 2. shows the comparison results of a various machine-learning algorithm based on the different machine learning algorithm.

Table 2. Comparison of machine learning algorithm time-window based on 10-fold cross-validation.

Algorithm	ACC	AUC	MSE
Logistic Regression	0.909±0.001	0.886±0.019	0.039±0.000
Linear Discriminant Analysis	0.949±0.001	0.808±0.014	0.033±0.001
K- Nearest neighbour	0.959±0.001	0.894±0.016	0.034±0.002
Decision Tree	0.934±0.004	0.816±0.015	0.040±0.003
Gaussian NB	0.883±0.003	0.790±0.024	0.046±0.002
Random Forest	0.985±0.001	0.972±0.024	0.030±0.001

In Table 2. the RF achieve the best ACC, AUC and MSE at 99.0%, 97.2%, and 0.030. Meanwhile, the lowest standard deviation in RF for ACC. At the same time, Gaussian NB reached the worst standard deviation in Gaussian NB for ACC, AUC, and MSE at 0.88.3%, 9.9%, and 0.046 respectively.

4. Conclusion

In this research, we presented the feature set and the main part of the suggested DoS detection mechanism. Moreover, the introduced features set were extracted from real network traffic. In addation, several machine-learning algorithms are tested. The results of the experiments showed that the proposed features with Random forest achive better performance in detecting Dos attacks. In the future, we utilize our selected features for deep learning model in order to achieve a high detection rate.

References

[1] D. Evans, "The internet of things: How the next evolution of the internet is changing everything," *CISCO white paper,* vol. 1, pp. 1-11, 2011.

[2] E. Adi, Z. Baig, and P. Hingston, "Stealthy Denial of Service (DoS) attack modelling and detection for HTTP/2 services," *Journal of Network and Computer Applications,* vol. 91, pp. 1-13, 2017.

[3] H. R. Zeidanloo, A. B. Manaf, P. Vahdani, F. Tabatabaei, and M. Zamani, "Botnet detection based on traffic monitoring," in *Networking and Information Technology (ICNIT), 2010 International Conference on*, 2010, pp. 97-101.

[4] P. Kamboj, M. C. Trivedi, V. K. Yadav, and V. K. Singh, "Detection techniques of DDoS attacks: A survey," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, 2017, pp. 675-679.

[5] J. Zhang and M. Zulkernine, "A hybrid network intrusion detection technique using random forests," in *null*, 2006, pp. 262-269.

[6] C. C. Kacha, K. A. Shevade, and K. S. Raghuwanshi, "Improved Snort intrusion detection system using modified pattern matching technique," *Int. J. Emerg. Technol. Adv. Eng,* vol. 3, pp. 81-88, 2013.

[7] M. Alkasassbeh, "A Novel Hybrid Method for Network Anomaly Detection Based on Traffic Prediction and Change Point Detection," *Journal of Computer Science,* vol. 14, pp. 153-162, 2018.

[8] G. Al-Naymat, M. Al-Kasassbeh, and E. Al-Harwari, "Using machine learning methods for detecting network anomalies within SNMP-MIB dataset," *International Journal of Wireless and Mobile Computing,* vol. 15, pp. 67-76, 2018.

[9] M. Alauthaman, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "A P2P Botnet detection scheme based on decision tree and adaptive multilayer neural networks," *Neural Computing and Applications,* vol. 29, pp. 991-1004, 2018/06/01 2018.

[10] M. Alkasassbeh and M. Almseidin, "Machine Learning Methods for Network Intrusion Detection," *arXiv preprint arXiv:1809.02610,* 2018.

[11] M. Alkasassbeh, M. M. Hashim, M. S. M. Rahim, O. Kyungyoung, J. Hong, K.-O. Lee*, et al.*, "An Empirical Evaluation For The Intrusion Detection Features Based On Machine Learning And Feature Selection Methods," *Journal of Theoretical and Applied Information Technology,* vol. 95, 2017.

[12] C. Robert, "Machine learning, a probabilistic perspective," ed: Taylor & Francis, 2014.

[13] H. H. Jazi, H. Gonzalez, N. Stakhanova, and A. A. Ghorbani, "Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling," *Computer Networks,* vol. 121, pp. 25-36, 2017.

Comparative Analysis of LSTM and Kalman Filter in Time Series Estimation

M. Canayaz[1], S. Şehribanoğlu[2]

[1]University of Van Yuzuncu Yil, Van, Turkey, mcanayaz@yyu.edu.tr
[2] University of Van Yuzuncu Yil, Van, Turkey, sanem@yyu.edu.tr

Kalman filtering is a classic state estimation technique in the literature. The Kalman Filter, which can be applied to stationary and non-stationary states, calculates the state using the previous estimate, and new input data. This eliminates the need to store all the data observed in the past. Kalman filter is an algorithm that provides an optimized estimate of the state of the system. Instead of estimating a single measurement to obtain the state of the system, it does so by looking at several successive measurements to minimize the (recursive) error. Kalman Filter has been used in many areas such as econometric data, weather forecasts. Long-term memory (LSTM) is a recurrent neural network (RNN) architecture that recalls values at random intervals. The saved values do not change when the learned progress is made. An LSTM is well suited for classifying, processing and predicting time series considering the time delays of unknown size and time between significant events. In this study, the effects of Kalman filter and LSTM on estimation in time series were compared. Data sets obtained from the UCI database and open source data sets on the internet were used for the study. In the performance analysis, it was observed that the Kalman filter method obtained minimum RMSE(Root Mean Square Error) values in data sets.

1. Introduction

Statistical estimation, which is expressed as predicting the future by using historical data with mathematical models, is widely used today when deep learning applications have become popular. Especially in the industrial sector, demand forecasting, stock forecasting, and price forecasting are important as it is a tool that affects the future activities of firms. In addition to being used in the field of industry, there is a need for systems with advanced ability to predict in making important decisions in fields such as medicine and engineering. For these processes, long-term memory (LSTM), a recurrent neural network, a type of memory-using network structure is encountered in solving many problems. Cao et al (2019) used the LSTM and CEEMDAN models in a hybrid way in the analysis of financial time series. Li et al (2019) proposed a new approach for estimating time series with LSTM based on evolutionary computation [2]. Yıldırım et al. (2019) used LSTM to classify arrhythmia in the medical field [3]. Although Kalman Filter [4,5] is an old method, it is frequently encountered in estimation processes. It is widely used especially in the fields of computer vision.

This study is organized as follows. First, LSTM and Kalman Filter are described respectively. Then, experimental results on data sets were given using these methods and performance comparisons were shown.

2. LSTM

LSTM[6-8] is a special type of recurrent neural network that includes memory blocks, memory cells, and gate units. These networks try to model time or sequence-dependent behaviors.These networks have LSTM cell blocks instead of neural network layers. These cell blocks consist of three units such as entrance gate, forget gate and exit gate.These units control the status of the cells. These gates determine what is stored in the cell, when it is allowed to read, write or delete.These gates have a network structure and activation function. Just like in neurons, it passes or stops the incoming information according to its weight. These weights are calculated during learning of the recurrent neural network. With this structure, it is determined whether the cell will receive or delete the data.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{3}$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right) \tag{4}$$

$$h_t = o_t * \tanh\left(C_t\right)$$

3. Kalman filter

Kalman filtering[9-14] is a classical state prediction technique. Kalman filtering is very powerful in many ways: it supports the prediction of past, present and future situations, and can do so even if the exact structure of the modeled system is unknown. The state vector is represented by xk and k indicates discrete time. yk represents unobserved data.The equations of this filter are in equation 5, process and measurement equations.

1. Process equation $:x_{k+1} = A_k x_k + B_k u_k + G_k w_k$ (5)

2. Measurement equation $:y_k = H_k x_k + v_k$

$x_k \in R^q$ represents the system state vector, $y_k \in R^m$ represents the system observation vector, and $U_k \in R^p$ represents the system control vector. A_k, shows the transition matrix (qxq size) and H_k shows the observation matrix (mxq size). B_k (qxm) and G_k (qxq) qxm) dimensional matrices. $w_k \in R^q$and $v_k \in R^m$ represent zero-mean, white noisy error terms. The covariance matrices of w_k and v_k are as in (6) and (7).

$$E[w_n w_k^T] = \begin{cases} Q_k \text{ for } n = k \\ 0 \quad \text{for } n \neq k \end{cases} \quad (6) \qquad E[v_n v_k^T] = \begin{cases} R_k \text{ for } n = k \\ 0 \quad \text{for } n \neq k \end{cases} \quad (7)$$

The updates for running the filter are executed as follows eq. (8) and (9).

Prediction (time update), (8)

$$\hat{x}_{k|k-1} = A_{k-1}\hat{x}_{k-1|k-1} + B_{k-1}u_{k-1}$$

$$P_{k|k-1} = A_{k-1}P_{k-1}A'_{k-1} + G_{k-1}Q_{k-1}G'_{k-1}$$

Correction (Measurement update) (9)

$$\hat{x}_{k|k-1} = \hat{x}_{k|k-1} + K_k[y_k - H\hat{x}_{k|k-1}]$$
$$K_n = P_{n|n-1}H'_n\left(H_n P_{n|n-1}H'_n + R_n\right)^{-1}$$
$$P_k = (I - K_k H_k)P_{k|k-1}$$

Here, K_k is known as kalman gain. $\hat{x}_{k|k-1}$ and $\hat{x}_{k|k}$ are called prior and posterior estimates of the state vector.

4. Experimental results

To compare these two methods, UCI[15] database and open data sets on the internet were used. The layer structure for LSTM network and dataset are located in Table 1. The application made on these datasets was developed using Python programming language. In addition, the application was tested on a computer with ubuntu operating system with i7 processor integrated with graphics card with Geforce 1070 GPU.

Table 1.Datasets and LSTM Network Structure

Datasets	**Samples size**	**LSTM Network Structure**
AirQuality	10000	LSTM(50),Dense(1),Train Size:0.70, Test Size:0.30, Epoch size:50
Metro Traffic	10000	
IBM	1000	

Table 2.Results

Datasets	**Methods**			
	LSTM		**Kalman Filter**	
	RMSE	**R^2**	**RMSE**	**R^2**
AirQuality	0.76	0.74	0.19	0.98
Metro Traffic	760.83	0.86	161.48	0.99
IBM	4.31	0.86	0.52	1

When the results are analyzed in Table 2, it is seen that Kalman Filter is more successful than LSTM in estimating the time series on all three data sets. Especially in the IBM dataset, both the RMSE(Root Mean Square Error) value is lower and R^2 value is 1, which means that Kalman Filter is a perfect model for this data set. Figure 1 shows the results of the IBM dataset.

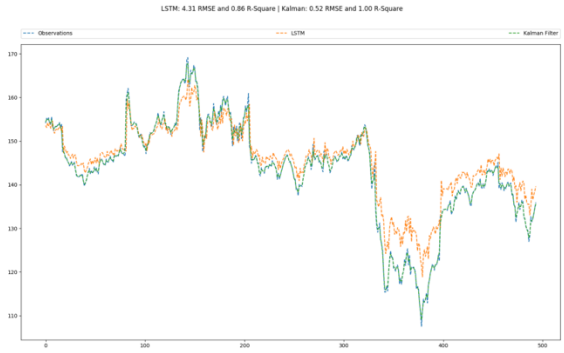


Figure1.IBM dataset results

5. Conclusions

In this study, it is tried to show performance of LSTM and Kalman Filter on various data sets in time series estimation. LSTM requires high CPU and memory. To overcome this, Gerforce GTX1070 GPU graphics card was used in the study. When we look at the results, it is seen that Kalman Filter gives better results than LSTM on the data sets. In the future studies, it is planned to make estimation with the data obtained from the sensors.

Acknowledgements

This study is supported Scientific Research Projects Department Project No. FBA-2018-6915 by Van Yuzuncu Yıl University.

References

[1]Cao J, Li Z, Li J, (2019). Financial time series forecasting model based on CEEMDAN and LSTM. Physica A 519 127–139.
[2]Li Y, et al.(2019). EA-LSTM: Evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems.Article inPress.
[3]Yildirim O, et al(2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. Computer Methods and Programs in Biomedicine 176, 121–133.
[4]Yan P, Swarnendu B, Donald S. F., Keshav, P.(2017). An Elementary introduction to kalman filtering. arXiv:1710.04055.
[5]Simon. D. (2001). Kalman Filtering. Embedded Systems Programming, 14(6), 72-79.
[6]Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network, arXiv:1808.03314.
[7]Sepp H, Schmidhuber, J. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
[8]Internet,https://adventuresinmachinelearning.com/keras-lstm-tutorial/.
[9]Welch, G. and Bishop,G.,(2006.) An Introduction to the Kalman Filter. https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.
[10]Chen,Z. (2003) . Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond.
[11]Yan Pei,Y., Biswas,S., Fussell,D.S., Pingali, K. (2017). An Elementary Introduction to Kalman Filtering.
[12]Youngjoo K, Hyochoong B, (2018). Introduction to Kalman Filter and Its Applications. 10.5772/intechopen.80600.
[13]Çayırlıoğlu İ,(2012),Kalman Filtresi ve Programlama. Fen ve Teknoloji Bilgi Paylaşımı, 2012-1.
[14] Satterthwaite C.P. (1999). Aircraft Computers ,Wiley Online Library.
[15]UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php

A New Meta-heuristic Approach for 3D Placement of Multiple Unmanned Aerial Vehicle Base Stations In Wireless Networks

R. Özdağ[1], H. Yanikomeroglu[2]

[1]*Van Yüzüncü Yıl University, Van, Turkey, rozdag@yyu.edu.tr*
[2]*Carleton University, Ottawa, Canada, halim@sce.carleton.ca*

Abstract

The use of Unmanned Aerial Vehicle as a mobile Base Stations (UAV-BSs) in wireless communication is expected to be an important component for the next generation wireless networks. Nowadays 3D placement of base stations mounted on Unmanned Aerial Vehicles (Drones), also called Low-altitude Aerial Platforms (LAPs), enable wireless communication to be performed effectively in the related environment. In this study, the location optimization of multiple UAV-BSs was performed in order to cover all users (receivers) that are assumed to be located at specified distance intervals in the urban environment defined according to the Air-to-Ground (ATG) model. For this purpose, a new meta-heuristic approach was improved called as Maximum Drone Deployment Algorithm based on the ElectroMagnetism-Like (MDDA-EML) algorithm to make efficient dynamic deployments of multiple 3D UAV-BSs. In addition, Pure-EML was designed based on the basic EML algorithm to compare the performance of the developed MDDA-EML. According to simulation results; it has been determined that the effective placement of multiple 3D Drone-BSs at near optimum height is provided by MDDA-EML which is covering the maximum number of users on the ground, thus reaching optimum coverage rates compared to Pure-EML.

1. Introduction

The use of UAVs that can be mobilized as flying base stations in environments where wireless communication is required, such as access to the Internet or emergency situations, is considered an important approach in increasing the coverage rate of the wireless network. Drone-BSs which can used as aerial base station have several advantages. First, due to their LAP properties, they can link with the users in the ground by providing high Line-of-Sight (LoS). Secondly, since Drones are easily mobilized, they can be deployed flexibly and thus communicate quickly with receivers [1]. Drone-BSs, also called UAV-BSs, has several important problems to solve. These problems; Drone 3D deployment, energy consumption and road planning can be considered. The deployment problem is considerable importance among these problems as it significantly affects the energy consumption of the drones. Due to technical limitations, the number of Drone-BSs that can be distributed after a disaster, for example, may be limited. Therefore, Drone-BSs altitude should be optimized to achieve the optimum possible coverage rate, since Drone-BSs distributed at very high and very low altitudes will reduce coverage in ground.

In this study, it is aimed to make efficient dynamic deployment of Drone-BSs by using meta-heuristic EML algorithm in order to reach the optimum altitude of randomly distributed multiple Drone-BSs in urban environment where is based on distribution environment. Simulations of MDDA-EML and Pure-EML approaches developed based on EML algorithm were performed by the MATLAB and their performances were compared. According to these results; It has been found that when 3D deployment of multiple Drone-BSs distributed by MDDA-EML are performed, compared to other algorithms optimum coverage rates are achieved and more stable simulation results are obtained.

2. System Model

LAPs are semi-static aerial platforms, such as Quadcopters, helicopters and balloons, which can be characterized at an altitude within the troposphere, which is the lowest layer where the atmosphere touches the ground. System modeling of Drone-BSs classified as LAPs is realized based on the ATG [2] channel model. ATG communication is performed according to two main propagation groups: LoS and NLoS (None Line-of-Sight). As shown in Fig. 1 [3]; Radio signals emitted by the LAP bring additional pathloss (*PL*) on the ATG link in the free space until they reach the urban environment in which they are exposed to shadowing caused by buildings or structures. When the radio signal reaches in urban environment, they are subject to excessive *PL* due to structures.

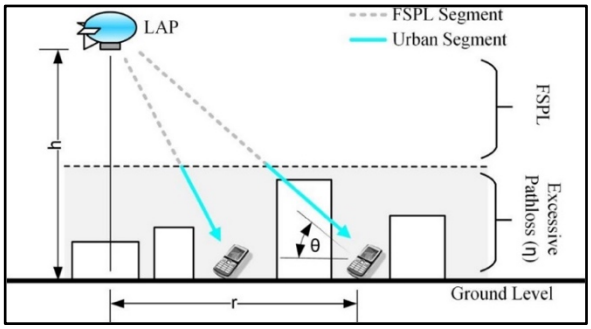


Figure 1. Propagate of LAP radio signal in urban environment.

For each propagation group in the ATG model, the average *PL* value which is exposed to the signal emitted by the LAP can be modeled according to *Eq*. (1) [3].

$$PL_{\delta} = FSPL + \eta_{\delta} \tag{1}$$

where *FSPL* is the *PL* value in the free space, η is the average of the excessive *PL* value in the urban environment, and δ is the propagation group. Therefore, according to δ; the probabilistic mean *PL* value between the *i* th receiver having the θ elevation angle (Fig. 1) on the ground and the LAP can be modeled [4] as following:

$$L(h, r_i) = \frac{\eta_{LOS} - \eta_{NLOS}}{1 + a \exp\left(-b\left(\theta_i - a\right)\right)} + 20 \log\left(\frac{4\pi f d_i}{c}\right) + \eta_{NLOS} \tag{2}$$

where η_{LOS} and η_{NLOS} are the average additional losses defined depending on the environment [3] for the propagation groups. *a* and *b* [3] are constants defined according to different environments where urban, rural, etc. *f* (in Hz) is the carrier frequency value. d_i refers distance between LAP and the *i* th receiver, which is calculated by $\sqrt{h^2 + r_i^2}$. *c* (in m / s) represents the speed of light. θ_i indicates the angle of elevation between LAP and the *i* th receiver, which is calculated by $tan^{-1}\left(\frac{h}{r_i}\right)$.

3. Proposed Method

If the *PL* value of the signal having *i* th receiver is less than threshold *PL* (L_{th}), it is determined as following that the receiver is covered by Drone-BS.

$$CS_i = \begin{cases} 1, & if\ L(h, r_i) \leq L_{th} \\ 0, & otherwise \end{cases} \tag{3}$$

where CS_i indicates Coverage Status (*CS*) of *i* th receiver. L_{th} is indicates threshold *PL* calculated according to parameters h_{opt} which is optimum Drone-BS altitude and R_{max} which is maximum coverage radius of Drone-BS on ground.

Fitness function (*fx*) of the Drone-BS is calculated summing the number of receivers who is covered by the Drone-BSs deployed for each *j* as follows:

$$fx^{j}_{DroneBS} = \sum CS_i \tag{4}$$

It is provided reach to maximum *fx* value of the Drone-BS providing that proposed algorithm updates *fx* values in each Drone-BS deployment.

4. Simulation Results

In this study, Monte-Carlo simulations were performed in order to coverage of each receiver located with ranges of 20 meters within 1 km^2 (1000 × 1000) area in the urban environment. it were performed by MDDA-EML and Pure-EML algorithms dynamic deployments of Drone-BSs, which was initially randomly distributed in the range of 2-13.

When R_{max} is defined as 200 meters for determining the optimum altitude of Drone-BS, h_{opt} is calculated to be 130 meters (Fig. 2). As soon as the Drone-BS starts to rise by increasing its altitude, the *PL* value of the signal which reaching the receivers gradually decreases. When the Drone-BS reaches the optimum altitude, the *PL* value is reached the threshold value and as the Drone-BS altitude is increased, *PL* value increases continuously.

The average coverage rates of the urban area according to the average number of iterations which each of the Drone-BSs deployed by MDDA-EML and Pure-EML are located in the optimum position are shown in Fig. 3.

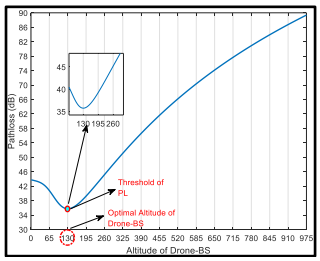
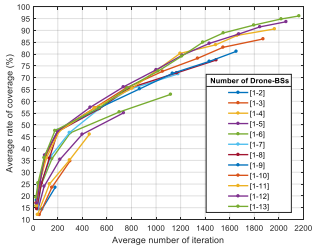
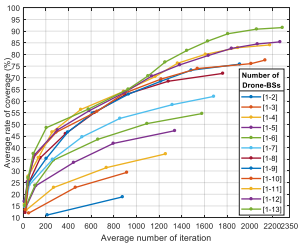


Figure 2. *PL* and L_{th} values calculated according to altitude of the Drone-BSs.



(a) MDDA-EML



(b) Pure-EML

Figure 3. Average coverage rates reached by MDDA-EML and Pure-EML.

5. Conclusion

In this study, MDDA-EML and Pure-EML algorithms developed in order to cover maximum number of receivers on the ground by making dynamic deployment of multiple Drone-BSs in urban environment are compared. According to Monte-Carlo simulations; MDDA-EML algorithm has been identified to be more optimal than Pure-EML both in terms of the average number of iterations and in terms of the average coverage rate of area by multiple Drone-BSs.

References

[1] Mozaffari, M., Saad, W., Bennis, M., and Debbah, M. (2015, December). Drone small cells in the clouds: Design, deployment and performance analysis. *In Proceedings IEEE Global Commun. Conf.* (San Diego, CA, USA, pp. 1–6).

[2] Kalantari, E., Yanikomeroglu, H., and Yongacoglu, A. (2016). On the number and 3d placement of drone base stations in wireless cellular networks. *In IEEE 84th Vehicular Technology Conference* (pp. 1–6).

[3] Hourani, A. A, Kandeepan, S., and Lardner, S. (2014, December). Optimal LAP altitude for maximum coverage. *IEEE Wireless Commun. Lett.* (Vol. 3, No. 6, pp. 569–572).

[4] Alzenad, M., El-keyi, A., Lagum, F., and Yanikomeroglu, H. (2017). 3D placement of an unmanned aerial vehicle base station (UAV-BS) for energy efficient maximal coverage. *IEEE Wireless Commun. Lett.* (pp. 1–4).

Number and Diversity of Requested Computerized Tomography from Emergency Service

A. M. Gündüz[1]

[1]University of Yüzüncü Yıl, Van, Turkey, alimahir72@hotmail.com

Abstract

Computed tomography is one of the most widely used diagnostic imaging modalities in patients presenting to the emergency department. In this study, we want to create a statistical data by determining the number and variety of computed tomography examinations performed to the patients who applied to the Emergency Department of the YYU Medical Faculty Hospital within the last 1 year. With the data we will obtain, we aim to prevent excessive CT scans other than clinical requirements.

Between May 1, 2018 and May 1, 2019 all computerized tomographies taken in the Radiology Department of the YYU Medical Faculty Hospital were classified according to the desired body region and their numbers were calculated. Accordingly, the number of all computed tomography scans in 1 year period is 44.232, and the number of those originating from the emergency department is 19.296 (%44,63). In the same period, the total number of patients admitted to the hospital and emergency department was 755.535 and 175.086, respectively. Although the number of patients admitted to the emergency department is 23.17% of the total number of patients, the number of emergency CT examinations constitutes 44.63% of all CT examinations. CT examination was performed in 5.85% of all patients and 11.02% of those who applied to the emergency department. The increase in the number of CT examinations from the emergency department can be attributed to the high number of trauma-related patients, workload, agitation of patients and their relatives, doctor's experience, desire to avoid risk and forensic concerns.

Excessive use of diagnostic imaging in emergency medicine is highly debated. Use of advanced diagnostic imaging has increased in the United States exponentially since the advent of computed tomography and magnetic resonance imaging. CT use has increased from 3 million scans in 1980 to greater than 60 million in 2005, and is still increasing (1). The use of CT increased in emergency department admissions, from 3.2% in 1996 to 13.9% in 2007 (2).

1. Introduction

Computed tomography is one of the most widely used diagnostic imaging modalities in patients presenting to the emergency department. In the computed tomography (CT) procedure, X-rays are used and it is possible to examine almost all organs in the body with a desired thickness. The advantage of computed tomography over other diagnostic imaging methods is the shortness of the examination time, the possibility of 3D imaging and the high image quality. The high radiation dose is the disadvantage (1). In this study, we want to create a statistical data by determining the number and variety of computed tomography examinations performed to the patients who applied to the Emergency Department of the YYU Medical Faculty Hospital within the last 1 year. With the data we will obtain, we aim to prevent excessive CT scans other than clinical requirements.

2. Material Method

Between May 1, 2018 and May 1, 2019 all computerized tomographies taken in the Radiology Department of the YYU Medical Faculty Hospital were classified according to the desired body region and their numbers were calculated. For example; brain CT, maxillofacial CT, thorax CT, all abdominal CT, extremity CT. Among these, the number of CT requests originating from the emergency department was determined and their rates were calculated both in itself and in all CT's.

3. Result

According to the data obtained from YYU Medical Faculty Hospital Data Processing Center; The number and distribution of all computed tomographies taken between May 1, 2018 and May 1, 2019 at YYU Medical Faculty Hospital are as follows **(Table 1)**.

The number and distribution of computed tomography originating from the emergency department at the same dates are as follows **(Table 2)**.

Accordingly, the number of all computed tomography scans in 1 year period is 44.232, and the number of those originating from the emergency department is 19.296 (%44,63). In the same period, the total number of patients admitted to the hospital and emergency department was 755.535 and 175.086, respectively. In other words, 5.85% of the total number of patients admitted to the hospital and 11.02% of the patients who applied to the emergency department were evaluated by CT.

According to the body region examined, total and emergency service-originated CT numbers, respectively it was like this: 9.527 and 6.846 for brain CT (%71.86), 8.525 and 2.505 for thorax CT (%29,38), 6.623 and 1.482 for upper abdominal CT(%22,38) , 6.498 and 1.488 for lower abdominal CT (%22.90), 3.325 and 2.718 for vertebral CT (%81.74), 2.320 and 1.789 for extremity CT (%77,11), 1.074 and 903 for maxillofacial CT (%84,08).

TABLE 1

Name of CT	Number	Rate (%)
Brain	9.527	21,54
Thorax	8.525	19,27
Upper abdomen	6.623	14,97
Lower abdomen	6.498	14,69
Vertebra	3.325	7,52
Extremity	2.320	5,25
Angiography	1.965	4,44
Other	1.881	4,25
Paranasal	1.126	2,55
Maxillofacial	1.074	2,43
Temporal bone HRCT	869	1,96
Orbit	499	1,13
TOTAL	**44.232**	**100**

TABLE 2

Name of CT	Number	Rate (%)
Brain	6.846	35,48
Vertebra	2.718	14,09
Thorax	2.505	12,98
Extremity	1.789	9,27
Lower abdomen	1.488	7,71
Upper abdomen	1.482	7,68
Maxillofacial	903	4,68
Other	818	4,24
Orbit	455	2,36
Angiography	179	0,93
Temporal bone HRCT	66	0,34
Paranasal	47	0,24
TOTAL	**19.296**	**100**

4. Discussion and Conclusions

Excessive use of diagnostic imaging in emergency medicine is highly debated. Use of advanced diagnostic imaging has increased in the United States exponentially since the advent of computed tomography and magnetic resonance imaging. CT use has increased from 3 million scans in 1980 to greater than 60 million in 2005, and is still increasing (2). Overall, CT use during ED visits increased 330%, from 3.2% of encounters in 1996 to 13.9% in 2007 (3).

The results we obtained in our study were as follows: Although the number of patients admitted to the emergency department within a 1-year period is 23.17% of the total number of patients, the number of emergency CT examinations constitutes 44.63% of all CT examinations. CT examination was performed in 5.85% of all patients and 11.02% of those who applied to the emergency department. The increase in the number of CT examinations from the emergency department can be attributed to the high number of trauma-related patients, workload, agitation of patients and their relatives, doctor's experience, desire to avoid risk and forensic concerns.

The rate which was 13.9% in the USA in 2007 is now 11.02% in our hospital. However, I think it would be more accurate to compare with current data. Nevertheless, I would like to think and say that our doctors are more experienced, knowledgeable and dedicated.

References

[1] Kumaş A. (2009) *Radyasyon Fiziği ve Tıbbi Uygulamaları* (2. Baskı). Palme Yayıncılık.
[2] Brenner DJ, Hall EJ.(2007). Computed tomography-an increasing source of radiation exposure. *N Engl J Med,* 2 (357), 2277-2284.
[3] Kocher KE, Meurer WJ, Fazel R, et al. (2011) National trends in use of computed tomography in the emergency department. *Ann Emerg Med,* (58), 452-462.

A Hybrid Metaheuristic Approach for Solving Vehicle Routing Problem with Mixed Pickup and Delivery

A. Hamzadayı[1], A. Baykasoğlu[2], Ş. Akpınar[3]

[1] *Department of Industrial Engineering, Van Yüzüncü Yil University, Van, Turkey, alperhamzadayi@yyu.edu.tr*
[2] *Department of Industrial Engineering, Dokuz Eylül University, İzmir, Turkey, adil.baykasoglu@deu.edu.tr*
[3] *Department of Industrial Engineering, Dokuz Eylül University, İzmir, Turkey, sener.akpinar@deu.edu.tr*

Abstract

One extend of the classical vehicle routing problem (VRP) is the vehicle routing problem with mixed pickup and delivery (VRPMPD). In the VRPMPD, customers may have either collect or deliver the goods. Two important aspects must be considered when designing an efficient metaheuristic: diversification and intensification. In general, basic population based metaheuristics (PBMs) tend to be more diversification oriented, whereas basic single solution based metaheuristics (SSBMs) tend to be more intensification oriented. In this article, for being able to solve the problems of VRPMPD effectively, the algorithms Simulated Annealing (SA), Great deluge (GD), Threshold Accepting (TA), Greedy (GS) and Random Search (RS) that are SSBM are hybridized with the Generic algorithm (GA) that is PBM. In the proposed hybrid algorithm (called SSBMs/GA), SSBMs aim at providing intensification, while GA aim at providing diversification. A new neighborhood structure adding a kind of memory characteristic to the proposed algorithm is proposed by providing the information sharing between the individuals in the population. In this study, a new crossover operator is also proposed. The proposed hybrid algorithm has been elaborately tested through set of VRPMPD instances that are taken from the literature. The results of the computational experiments show that the SSBMs work within the GA very concordantly; and the SSBMs/GA is an effective algorithm in solving VRPMPD.

1. Introduction

There are three variants of this problem in the current literature [1]: (1) the vehicle routing problem with backhauls (VRPB) in that all customers that have delivery demand have to be served, and the vehicle would be empty before the customers that have pickup demand are serviced. (2) The vehicle routing problem with mixed deliveries and pickups (VRPMDP) in that customers having delivery and pickups demands are served in any order on the vehicle route. (3) The vehicle routing problem with simultaneous deliveries and pickups (VRPSDP) in that customers may simultaneously be served a delivery and a pickup.

The VRPMDP is known to be NP-hard because it is a variant of the classical Vehicle Routing Problem (VRP) which is a well-known NP-hard problem [2]. In the current literature, only [3] uses the genetic algorithm (GA) for solving the VRPSPD. To the best knowledge of the authors, this is the first study that uses genetic algorithm for hybridizing to solve VRPMPD.

```
Initialize the algorithm parameters;
Population (,,) = Randomly generate NA number of initial solutions by using initial solution construction procedure
F (.) = Evaluate fitness values of the initial solutions;
Best_Population (,,) =Population (,,);
BF (.) = F (.);
character_matrix = zeros (NA, 5);
move_structure = zeros (NA, 1);
for i = 1: NA
move_structure (i, 1) = Randomly select a neighborhood structure between move 1 and move 11;
character_matrix (i, 1) = Select a SSBM algorithm; //Randomly assigning a different algorithm to each individual
endfor
for i = 1: NA
Call Function CHAR_ASSIGN (i, Parameters of algorithms); //Parameter assignment for the individual i
endfor
crossover_decision = zeros (NA, 1);
parameter_count = zeros (NA, 2); //The parameters of SSBMs are controlled via this matrix
iteration=0;
While iteration<= MAX_ITERATION
                    for i = 1: NA
    if crossover_decision (i, 1) == 0
                                        chromosome (:) = Population (i, :);
                                        ALGORITHM = character_matrix (i, 1);
                                        switch ALGORITHM
                                        case {1} run great deluge algorithm ();
                                        case {2} run threshold accepting algorithm ();
                                        case {3} run greedy search algorithm ();
                                        case {4} run random search algorithm ();
                                        case {5} run simulated annealing algorithm ();
                                        endswitch
                                        Population (i, :) = chr_c (1, :);
                                        F (i) = obj_c;
                                        if obj_c < BF (i)
                                        BF (i) = obj_c;
                                        Best_Population (i, :) = chr_c (1, :);
                                        endif
```

```
36.                                    else
37.                                        if Is there any individual (j) (except i) whose
crossover_decision equals to 1 in crossover_decision (:, 1) array?
38.                                        [child1, child2] = Apply crossover operator to obtain two new individuals
from the individuals i and j;
39.                                            crossover_decision (i, 1) = 0;
40.                                            crossover_decision (j, 1) = 0;
41.                                            f_value = Calculate the fitness value of
child1;
42.                                            if f_value < BF (i)
43.                                            BF (i) = f_value;
44.                                            Best_Population (i, :) = child1 (1, :);
45.                                            endif
46.                                            f_value = Calculate the fitness value of
child2;
47.                                            if f_value < BF (j)
48.                                            BF (j) = f_value;
49.                                            Best_Population (j, :) = child2 (1,:);
50.                                            endif
51.                                        Call Function CHAR_ASSIGN (i, Parameters of algorithms); //New
parameter assignment for the individual i
52.                    move_structure (i, 1) = Increase the move number (if move number is 11, it will start again from move 1)
53.                                        Call Function CHAR_ASSIGN (j, Parameters of algorithms); //New
parameter assignment for the individual j
54.                    move_structure (j, 1) = Increase the move number (if move number is 11, it will start again from move 1)
55.                                    end if
56.                                    endif
57.                            endfor
58.                            Population (:, :) = Best_Population (:, :);
59.                            F (:) = BF (:)
60.                    endwhile
61.                    Find BestSolution and BestFitness, and report them
```

Fig. 2. Main steps of the algorithm.

In this paper, for being able to solve the problems of VRPMPD effectively, a new hybrid algorithm (SSBMs/GA) is proposed by hybridizing the general framework of the population based genetic algorithm (GA) with single solution based search metaheuristics (SSBMs) such as great deluge (GD) [4], threshold-accepting (TA) [5], greedy search (GS), random search (RS), and simulated annealing (SA). In the proposed SSBMs/GA, the crossover operator of the GA enables the search process to explore different search regions in the search space while SSBMs enable high-quality solutions from the examined regions. Moreover, a new neighborhood structure is proposed to share information between the individuals in the population. Proposed neighborhood structure adds a kind of memory characteristic to the proposed algorithm by providing the information flow between the individuals in the population. SSBMs in the proposed SSBMs/GA change their neighborhood structures whenever they are converged to a certain point. This characteristic of the proposed algorithm resembles to the systematic neighborhood variation of the Variable Neighborhood Search (VNS) [6] that is used frequently in the solution of these and similar problems in the literature. This feature increases the probability of escaping local optimum and gives an advantage of providing intensive local. The proposed neighborhood structure adds a kind of memory characteristic to the proposed algorithm by providing the information flow between the individuals in the population. In this study, a crossover operator is also proposed. The proposed algorithm requires the optimization of only a single parameter in real terms. In terms of this characteristic, it makes its implementation to the optimization problems much easier. The performance of the SSBMs/GA is tested on well-known VRPMPD benchmark instances from the literature. The computational results indicate that the proposed algorithm generates competitive results with the most powerful algorithms developed so far, for the VRPMPD.

The remainder of this paper is organized as follows: In Section 2, the proposed hybrid SSBMs/GA algorithm is presented. The computational results are presented in Section 3. Finally, concluding remarks are given in Section 4.

2. The Proposed Hybrid SSBMs/GA Algorithm

In this paper, in order to be able to solve the problems of VRPMPD effectively, the algorithms of SA, GD, TA, GS, and RS that are SSBM are hybridized with the GA that is PBM. The main steps of the algorithm are given in Fig. 1.

Table 1. Computational results for the VRPMPD instances labelled H of [2].

N	WNA	GKA	ALS	SSBMs/GA		
				Best	Avg.	GAP%
50	598.27	**567.20**	**567.20**	**567.20**	567.20	0.0000
50	624.86	**610.66**	**610.66**	**610.66**	610.66	0.0000
50	611.78	**584.52**	**584.52**	**584.52**	584.52	0.0000
50	650.03	**608.15**	**608.15**	**608.15**	608.15	0.0000
50	624.77	**597.35**	**597.35**	**597.35**	597.35	0.0000
50	602.57	**585.09**	**585.09**	**585.09**	585.09	0.0000
50	605.00	**574.25**	**574.25**	**574.25**	574.25	0.0000
50	579.64	**575.99**	**575.99**	**575.99**	575.99	0.0000
50	676.44	**635.84**	**635.84**	**635.84**	635.84	0.0000
50	645.69	**600.92**	**600.92**	**600.92**	600.92	0.0000
50	724.51	**714.86**	**714.86**	**714.86**	714.86	0.0000
50	787.66	**780.77**	**780.77**	**780.77**	780.77	0.0000
N	**WNA**	**GKA**	**ALS**	**SSBMs/GA**		

50	**739.93**	740.55	740.55	740.55	740.55	**0.0008**
50	764.52	**761.83**	**761.83**	**761.83**	761.83	0.0000
50	819.39	**794.69**	**794.69**	**794.69**	794.69	0.0000
50	788.42	**783.04**	**783.04**	**783.04**	783.04	0.0000
50	731.41	**723.21**	**723.21**	**723.21**	723.21	0.0000
50	735.98	**725.79**	**725.79**	**725.79**	725.79	0.0000
50	856.20	850.91	**849.20**	**849.20**	849.20	0.0000
50	803.18	**771.96**	**771.96**	**771.96**	771.96	0.0000
50	604.65	**579.67**	**579.67**	**579.67**	579.67	0.0000
50	537.88	**513.46**	**513.46**	**513.46**	513.46	0.0000
50	502.23	**482.74**	**482.74**	**482.74**	482.74	0.0000
50	**549.68**	**549.68**	**549.68**	**549.68**	549.68	0.0000
50	568.28	**551.92**	**551.92**	**551.92**	551.92	0.0000
50	534.22	**529.93**	**529.93**	**529.93**	529.93	0.0000
50	480.00	**466.61**	**466.61**	**466.61**	466.61	0.0000
50	551.45	**527.98**	**527.98**	**527.98**	527.98	0.0000
50	500.77	**475.77**	**475.77**	**475.77**	475.77	0.0000
50	559.57	**536.89**	**536.89**	**536.89**	536.89	0.0000
50	716.41	**697.34**	**697.34**	**697.34**	697.34	0.0000
50	611.86	**610.23**	**610.23**	**610.23**	610.23	0.0000
50	601.73	**590.43**	**590.43**	**590.43**	590.43	0.0000
50	663.75	**653.81**	**653.81**	**653.81**	653.81	0.0000
50	659.69	**641.59**	**641.59**	**641.59**	641.59	0.0000
50	623.46	**608.59**	**608.59**	**608.59**	608.59	0.0000
50	557.52	**546.06**	**546.06**	**546.06**	546.06	0.0000
50	682.53	**652.70**	**652.70**	**652.70**	652.70	0.0000
50	591.18	**595.60**	**595.60**	**595.60**	595.60	0.0000
50	631.23	**624.03**	**624.03**	**624.03**	624.03	0.0000

3. Computational Results

The proposed hybrid algorithm is run ten times for each benchmark problem. The hybrid algorithm is compared with the algorithms that yield the best solutions for the VRPMPD benchmark instances in the current literature. To the best of our knowledge, the best known upper bounds for [2] data set were found by the following algorithms: WNA: Reactive tabu search [1], GKA: Hybrid discrete particle swarm optimization [7], ALS: Adaptive local search algorithm [8].

For the problems labelled as T, SSBMs/GA, ALS and GKA reach 34 best solutions over 40 instances while WNA produces only 10 best solutions. For the problems labelled as Q, SSBMs/GA, ALS and GKA obtain 36 best solutions out of 40 instances while WNA reaches 7 best solutions. As seen from the Table 1, SSBMs/GA and ALS reach 39 best solutions over 40 instances labelled as H. This result indicates that the developed algorithm is effective in solving the VRPMPD in reasonable computation time.

4. Conclusions

The performance of SSBMs/GA is tested by using well-known benchmark instances for the VRPMPD. The computational results indicate that the SSBMs/GA generates high-quality solutions for the instances and perform as well as the most sophisticated methods proposed for the VRPMPD up to date.

References

[1] Wassan, N., and Wassan, A.H., Nagy, G. (2008). A reactive tabu search algorithm for the vehicle routing problem with simultaneous pickups and deliveries. *Journal of Combinatorial Optimization*, 15, 368–386.

[2] Dethloff, J. (2001). Vehicle routing and reverse logistics: The vehicle routing problem with simultaneous delivery and pick-up. *OR-Spektrum*, 23, 79–96.

[3] Tasan, A.S., and Gen, M. (2012). A genetic algorithm based approach to vehicle routing problem with simultaneous pick-up and deliveries. *Computers & Industrial Engineering*, 62, 755–761.

[4] Dueck, G. (1993). New optimization heuristics: The great deluge algorithm and the record-to record travel. *Journal of Computational Physics*, 104 (1), 86–92.

[5] Dueck, G., and Scheuer, T. (1990). Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics*, 90, 161–175.

[6] Mladenović, N., and Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, 24, 1097–1100.

[7] Goksal, F.P., Karaoglan, I., and Altiparmak, F. (2013). A hybrid discrete particle swarm optimization for vehicle routing problem with simultaneous pickup and delivery. *Computers & Industrial Engineering*, 65, 39–53.

[8] Avci, M., and Topaloglu, S. (2015). An adaptive local search algorithm for vehicle routing problem with simultaneous and mixed pickups and deliveries. *Computers & Industrial Engineering*, 83, 15–29.

Data-Driven Multi-Criteria Decision Making in Decision Engineering

Adil Baykasoğlu

Dokuz Eylül University, Department of Industrial Engineering, İzmir, Turkey, baykasoglu@gmail.com

Abstract

A new data-driven Multi-Criteria Decision Making (MADM) model is introduced in this invited talk. The presented approach makes use of metaheuristic algorithms (Jaya algorithm) for training Fuzzy Cognitive Maps (FCMs) in order to enable learning from past data in a dynamic multi-criteria decision making scenario. Trained FCMs are used to predict future performance scores of decision alternatives by incorporating present subjective evaluations. The application of the proposed approach is also presented trough an example.

1. Introduction

Decision engineers have a degree in engineering or a strongly similar field, and have a profound understanding of systems analysis, uncertainty handling and decision theory (Barrager, 2016). Therefore, decision engineers have ability to understand complex systems that are characterized by dynamism and uncertainty. The task of decision engineers is to design computer-assisted mathematical models for real-life complex decision problems. A decision engineer brings processes, tools, facilitation and project management phases into the decision making projects (Barrager, 2012). The decision making situation is considered as a process, beginning from problem framing and alternative generation to the synthesis of the overall information. On the other hand, a decision engineer utilizes a set of computer-assisted mathematical models for the analysis. The field of MADM is one of the most significant components of decision engineering tools (see Fig. 1). A vast array of MADM approaches have been proposed in the literature in order to cope with real-life decision making problems. However, the proposed approaches have some drawbacks. One of the major tasks of a decision engineer is to manage the complexity of the problem. Complexity is mainly emerged from the interactions among the system components. These interactions cannot be quantified with exact terms, rather, they are characterized by linguistic terms which are inherently uncertain. Therefore, a decision engineer is expected to collect and process uncertain information while keeping the complexity of the problem at a manageable level. On the other hand, decision engineer is forward looking and a sound decision making requires incorporation of future-oriented thinking and being flexible to include changes in the preferences of decision makers (Barrager, 2016; Saaty, 2007). Therefore, learning from the past data is of crucial importance for tracking possible changes in the future preferences of decision makers. Data analytics have much to offer in this field by providing decision makers with a plethora of machine learning tools to extract valuable information from the available data. Although the availability of information has been increased dramatically over the past decade, historical records regarding past decision making practices involve imprecision and uncertainty. Therefore, fuzzy system models and system identification tools come into play. A plethora of approaches have been proposed to solve decision engineering problems in the literature. These approaches can be mainly categorized as expert-driven and data-driven approaches. In the expert-driven decision making, all of the required data is obtained from decision makers at the time of decision making. On the other hand, data-driven decision making models make use of historical records related to problem in hand and experts are elucidated about the temporal performances of the alternatives. These problems are usually called as data-driven, dynamic or multi-period MADM in the literature.

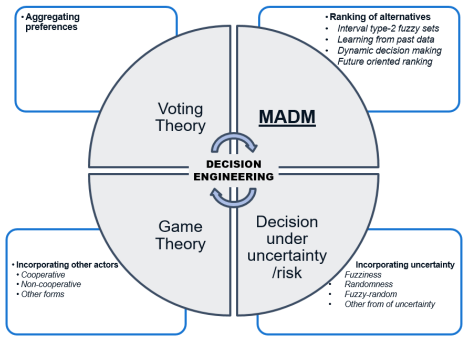


Fig.1. MADM within decision engineering (inspired from De Smet, 2019)

2. Data-Driven MADM

The increasing competition of the rapidly changing business environment entails effective use of resources such as people, material, capital, information in order to attain organizational goals. To carry out business functions effectively, managers are continuously engaged in a process of making decisions. These decisions involve consideration of diverse set of factors such as environmental, social, economic and etc., which are characterized by their multiple and conflicting nature. Decision making is becoming more complicated and difficult due to the dramatic increase in the availability of information obtained from diverse resources. Therefore, MADM offers a set of sophisticated techniques to help decision makers to select the best alternative by considering multiple, conflicting and incommensurate criteria.

Recently, with the rapid advances in the information technologies and digital transformations, data is becoming an indispensable part of decision making practices, and the field of MADM undergo a paradigm shift towards data-driven and dynamic decision making. Data-driven MADM problems have grabbed considerable attention in the literature recently. In the data-driven MADM problems, at least two-period decision making information is considered. In other words, in addition to attributes and alternatives, time dimension is considered in data-driven MADM problems.

3. Analyses of Data-Driven MADM Approaches

The literature of the field of data-driven MADM can be investigated under three categories:

- Aggregating all of the past decision making matrices and then applying a MADM technique (aggregating beforehand)
- Aggregating results of different periods (aggregating afterwards)
- Other methods (scenario analysis, time-dependent eigenvectors, system dynamics, hybrid techniques and etc.)

Aggregating all of the past decision making matrices beforehand: Xu and Yager (2008) investigated data-driven MADM problems under intuitionistic fuzzy environment. Two new aggregation operators were developed; namely dynamic intuitionistic fuzzy weighted averaging and uncertain dynamic intuitionistic fuzzy weighted averaging. Dynamic aggregation operators were used to aggregate all the intuitionistic fuzzy decision matrices pertaining to past periods, and then the classical methods are used to rank the alternatives. Bali et al. (2015) integrated AHP and dynamic intuitionistic fuzzy weighted averaging operator in dynamic personnel promotion problem in the same manner with Xu and Yager approach. Park et al. (2013) proposed dynamic intuitionistic fuzzy weighted geometric and uncertain dynamic intuitionistic fuzzy weighted geometric operators for dynamic MADM problems. After aggregation past decision matrices by using the proposed aggregation operators, VIKOR method was used to obtain rankings of alternatives. Chen and Li (2011) proposed a dynamic MADM model based on triangular intuitionistic fuzzy numbers and a new distance measure is suggested. In this paper, decision matrices at the different time periods were aggregated by using weighted arithmetic averaging operator for triangular intuitionistic fuzzy numbers, and then the ranking orders of alternatives were obtained by using closeness coefficients. There are similar aggregation operator-based studies in the literature which utilize intuitionistic fuzzy numbers (Wei, 2009; Zhang, 2016), 2-tuple linguistic representation (Ai & Yang, 2014; H. Liu et al., 2018; Y. Liu, 2014; Z. Xu, 2009; Zulueta-Veliz & Sanchez, 2018), grey numbers (Cui et al., 2011; Shen et al., 2015), and etc. For more information about dynamic aggregation operators, we refer to a review paper (Mardani et al., 2018).

Aggregating results of different periods: Xu (2009) proposed a dynamic weighted geometric aggregation operator for dynamic MADM problems. Three-period investment decision making was used to show the methodological steps. Crisp, interval, and fuzzy number representations were used to compute ranking orders. The aggregation was performed based on the closeness coefficients of the different time periods. Lin et al. (2008) employed grey numbers and Minkowski distance to deal with multi-period and uncertain decision matrices. The TOPSIS method was used as a main MADM method. Rather than aggregating all of the decision matrices at the beginning, the proposed model calculated period-weighted distances to ideal and anti-ideal solutions.

Other studies: Saaty (2007) studied time-dependent eigenvectors and approximating functional forms of relative priorities in data-driven (dynamic) MADM. Hashemkhani Zolfani et al. (2016) emphasized the relevance and necessity of future studies in MADM problems. In the paper, scenario-based MADM papers were reviewed and analyzed. Orji and Wei (2015) integrated fuzzy logic and system dynamics simulation to sustainable supplier selection problem. In the system dynamics simulation, future time horizons were considered.

What is lacking in the literature?

There are two major concerns that the current methods are not able to satisfy: Learning from the past data is missing. As ever-more data pour through the networks of organizations, considerable effort has been made in order to cultivate valuable information from these resources. As the notions of big data, digital revolution or internet of things pervade most organizations, firms steadily recognize the importance of data, and mining of firms' data to identify patterns and trends is about to become standard business practice. However, learning models that make use of past data have not gained attention in the data-driven (dynamic) MADM literature.

Future performance scores of alternatives are not considered. It is evident that sound decision making not only considers past data but also looks ahead as a process of thinking and planning (Saaty, 2007). Past, current, and future performance scores should be taken into consideration simultaneously. In parallel with these efforts, the culture of decision making is changing as well. Style of decision making evolving from intuition-based to data-driven. Learning of past data helps predict future performance scores of alternatives.

Our contributions

FCM learning is employed to capture patterns and trends in historical decision matrices first time in the literature. Jaya algorithm, a simple and effective metaheuristic optimization algorithm, is used to determine degree of interconnections and activation function parameters in the FCM. FCMs are able to model interactions among attributes while learning of the underlying structure, and thus performance patterns of alternatives are captured by means of multiple-attributes simultaneously. The present study is the first work which uses FCM in the domain of data-driven (dynamic) MADM (Baykasoğlu, Gölcük, 2019). Moreover, first time in the literature, a learning algorithm is used for data-driven MADM (Baykasoğlu, Gölcük, 2019). Considering different decision horizons, future decision matrices are generated via trained FCMs. First time in the literature, future decision making matrices generated by means of an intelligent algorithm are analyzed in decision making (Baykasoğlu, Gölcük, 2019). Short-, medium-, and long-term rankings of alternatives are provided to decision makers by utilizing past, current, and future information about alternatives. A real-life supplier performance evaluation problem is used to demonstrate practicality of the proposed model (Baykasoğlu, Gölcük, 2019).

4. Conclusions

A new hybrid dynamic MADM model, which combines Jaya algorithm, FCMs, time series weights and closeness coefficients, is proposed. The proposed model exhibits desirable properties that help overcome the drawbacks of the traditional dynamic MADM approaches.

References

References can be obtained from the author. They are not provided here due to space limitations.

Analysis of Earthquake Data in Eastern Anatolia Region Using Data Mining Techniques

A. Cengiz [1], D. Avcı[2]

[1]*Firat University, Elazig, Turkey, aayscengiz@gmail.com*
[2]*Firat University, Elazig, Turkey, davci@firat.edu.tr*

Abstract

In the last 10 years, earthquake data of the provinces in the Eastern Anatolia Region were obtained and inferences were made by applying Data Mining methods to this data stack. In the Weka program, K nearest neighbor algorithm, C4.5 algorithm and Random Forest algorithms were applied to the data stack. Accuracy values and complexity matrices are compared.

1. Introduction

In our country, located at the intersection point of the fault lines, high-intensity earthquakes occur in which life and property losses occur at uncertain intervals. Leading earthquakes and aftershocks occur in addition to high-intensity main earthquakes. The main reason for this is that the fault lines are active. The Eastern Anatolia Region is a risky earthquake zone with first and second-degree active fault lines.

The earthquake data used were obtained from the Kandilli Observatory Earthquake Research Institute (KRDAE). With the help of various classification methods to be applied to the earthquake data, it is aimed to create a model by analyzing the earthquake data which is restricted to the Eastern Anatolia Region.

2. Earthquake Data

Earthquakes occur with the vibration of the energy generated by the breaking of the rock called earth fault under high pressure. The magnitude of the earthquake is a measure that indicates the magnitude of the fractured surface and thus the level of energy generated. Earthquake intensity measurement is done by certain methods [1].

Md

Measures how long the earthquake generates vibration on the seismometer. It is measured by the distance of the earthquake center. This method is used for earthquakes with a magnitude less than 5.0 and close to the surface and depth less than 300 km.

Ml

It is the first method used. The sound waves generated in the earth's crust are recorded and the maximum amplitude value recorded in the sound recording is measured by distance and the magnitude is determined. This method is used for earthquakes with magnitudes below 6.0 and depths less than 700 km.

Ms

This method is used to measure large earthquakes with magnitudes above 6.0. It is scaled by measuring the amplitude of the seismic waves occurring on the earth due to the movements of the earth's crust.

Mb

In this method, it is calculated by recording the object waves (sound waves and shear waves) that go deep into the ground. The currents emitted by the recorded object waves are scaled.

Mw

It is scaled by making a mathematical model of earthquake formation. Calculated only for earthquakes with a magnitude over 4.0

3. Data Collecting

Data were obtained from the Kandilli Observatory Earthquake Research Institute (KRDAE) page in Excel format [2]. In order to be used in Weka, Cvs (comma separated) format is converted to the required definitions on this file, and the file .arff Extension after the Weka program is loaded. Between 2010-2019, 3338 earthquake data with a magnitude greater than 3.0 and a depth of less than 300 km were selected in the Eastern Anatolia region. These data are classified by column data such as date, time, latitude and longitude information, depth of occurrence, occurrence places and magnitudes of earthquakes. "The Miss Replace Missing Values" module is used to prevent missing data from causing problems during the classification process. With this method, the lost values are replaced by the mean or mode of the other values of the classes to which they belong. In the Filters section, "String to Nominal" filter is used for Location data.

KNN (K Nearest Neighbour - IBK) Algorithm

KNN algorithm is a lazy learning algorithm since it is not a learning-based algorithm. A distance measure is used to determine which of the K samples in the training data set most closely resembles a new input. [3]. If KNN is used in classification, the output K can be calculated as the class with the highest frequency from the most similar samples. In fact, each data instance is calculated for its own classes and assigned data to the class with the highest accuracy.

C4.5 (J48) Algorithm

Decision trees try to classify test data using a tree structure. The nodes in the tree indicate the property, the branches indicate the property value, and the leaf nodes indicate the class label. The reason for the widespread use of decision trees is a simple and simple definition of tree structures. The rules learned in this way are easily transferred. The most widely used decision tree algorithm is the C4.5 algorithm, which is an improved version of Quinlan's ID3 algorithm [4]. It is also known as boosting trees because it uses the boosting algorithm to improve accuracy.

Random Forest Algorithm

Random forest, a collection of classifiers, was first proposed by Leo Breiman. Base learners are decision trees [5]. In the training phase, 63% of the sample uses the Bagging algorithm. However, attribute selection is performed randomly. In the testing phase, the decision-making process is based on democracy. The result is obtained by combining the decisions of each major learner. In our study, we use the C4.5 algorithm as the basic classifier in Random Forest. The default number of trees used in random forests is 10 trees.

4. Classification of Data

Looking at the detailed accuracy values of the classes, it is seen that J48 Algorithm makes 97.077% of the data stack, IBK Algorithm 92.1592%, Random Forest Algorithm 96.695%.

Confusion Matrix

Calculating a confusion matrix is useful to see the points where the classification model correctly draws out and at which points an error occurs. The J48 algorithm graded with the highest accuracy and misplaced only 98 of the 3338 data. The IBK algorithm incorrectly inserted 252 of the 3338 data. The RF algorithm incorrectly inserted 100 of the 3338 data.

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
 174    0    0    3    0    2    0    0    0    0    0    0    0    1    0 |    a = ELAZIG
   0 2136    0    0    0    0    0    0    4    0    2    1    0    0    0 |    b = VAN
   0    0  120    0    0    2    1    4    1    0    0    0    1    0    0 |    c = ERZURUM
   3    0    0   84    0    0    2    0    0    0    0    0    0    0    0 |    d = MALATYA
   0    0    0    0   31    0    0    0    0    0    0    0    1    0    0 |    e = ARDAHAN
   3    0    0    0    0  102    2    2    0    0    0    0    0    5    0 |    f = BINGOL
   0    0    1    0    0    1  135    0    0    0    0    0    0    1    0 |    g = ERZINCAN
   0    0    4    0    0    3    0   89    0    0    3    0    0    0    0 |    h = MUS
   0    1    1    0    0    0    0    0   53    0    1    0    0    0    1 |    i = AGRI
   0    2    0    0    0    0    0    0    0   57    0    0    0    0    0 |    j = HAKKARI
   0   13    0    0    0    0    0    1    0    0   96    0    0    0    0 |    k = BITLIS
   0    0    0    0    0    0    0    0    0    0    0   88    0    0    0 |    l = SIRNAK
   0    0    3    0    2    0    0    0    0    0    0    0   19    0    0 |    m = KARS
   3    0    0    1    0    3    6    0    0    0    0    0    0   35    0 |    n = TUNCELI
   0    0    0    0    0    0    0    0    6    0    0    0    0    0    0 |    o = IGDIR
```

Figure 3: J48 Confusion Matrix

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
 167    0    0    7    0    5    0    0    0    0    0    0    0    1    0 |    a = ELAZIG
   0 2112    0    0    0    0    0    0    0    5   22    4    0    0    0 |    b = VAN
   0    0  107    0    2    4    3    5    4    0    0    0    4    0    0 |    c = ERZURUM
  16    0    0   71    0    0    1    0    0    0    0    0    0    1    0 |    d = MALATYA
   0    0    0    0   27    0    0    0    0    0    1    0    4    0    0 |    e = ARDAHAN
   8    0    4    0    0   92    2    2    0    0    0    0    0    6    0 |    f = BINGOL
   1    0    1    0    0    3  125    0    0    0    0    0    0    8    0 |    g = ERZINCAN
   0    0    7    0    0    6    0   80    0    0    6    0    0    0    0 |    h = MUS
   0    6    5    0    0    0    0    2   38    0    1    0    5    0    0 |    i = AGRI
   0    5    0    0    0    0    0    0    0   51    0    3    0    0    0 |    j = HAKKARI
   0   41    0    0    0    0    0    8    1    0   60    0    0    0    0 |    k = BITLIS
   0    4    0    0    0    0    0    0    0    1    1   82    0    0    0 |    l = SIRNAK
   0    0    4    0    5    0    0    0    3    0    0    0   11    0    1 |    m = KARS
   0    0    1    0    0    8    7    0    0    0    0    0    0   32    0 |    n = TUNCELI
   0    0    0    0    0    0    0    0    5    0    0    0    0    0    1 |    o = IGDIR
```

Figure 4: IBK Confusion Matrix

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k    l    m    n    o   <-- classified as
 175    0    0    3    0    1    0    0    0    0    0    0    0    1    0 |    a = ELAZIG
   0 2139    0    0    0    0    0    0    1    0    3    0    0    0    0 |    b = VAN
   0    0  123    0    0    1    1    2    1    0    0    0    1    0    0 |    c = ERZURUM
   3    0    0   85    0    0    1    0    0    0    0    0    0    0    0 |    d = MALATYA
   0    0    1    0   29    0    0    0    0    0    0    0    2    0    0 |    e = ARDAHAN
   4    0    1    0    0  104    0    2    0    0    1    0    0    2    0 |    f = BINGOL
   0    0    3    0    0    0  132    0    0    0    0    0    0    3    0 |    g = ERZINCAN
   0    0    0    0    0    3    0   91    0    0    5    0    0    0    0 |    h = MUS
   0    2    1    0    0    0    0    0   53    0    1    0    0    0    0 |    i = AGRI
   0    3    0    0    0    0    0    0    0   56    0    0    0    0    0 |    j = HAKKARI
   0   18    0    0    0    0    0    2    0    0   90    0    0    0    0 |    k = BITLIS
   0    1    0    0    0    0    0    0    0    1    0   86    0    0    0 |    l = SIRNAK
   0    0    4    0    3    0    0    0    2    0    0    0   15    0    0 |    m = KARS
   1    0    0    3    0    3    4    0    0    0    0    0    0   37    0 |    n = TUNCELI
   0    0    0    0    0    0    0    0    6    0    0    0    0    0    0 |    o = IGDIR
```

Figure 5: RF Confusion Matrix

5. Result

For the 15 provinces in the Eastern Anatolia region, data with 3338 data intensity greater than 3.0 and depth less than 300 km Between 2010 and 2019 were selected for the classification study and the results were compared. The data were classified by applying the J48 algorithm based on the C4.5 decision tree, IBK algorithm based on the KNN classification algorithm and Random Forest algorithms. Correctly Classified Instances were compared with accuracy and sensitivity values. The algorithm that makes the most accurate classification for this data stack is found to be J48 Algorithm based on C4.5 decision tree with an accuracy rate of 97.077%. The lowest success rate was found to be the IBK algorithm based on the KNN classification algorithm, which made the wrong classification with the rate of 7.8408%.

References

[1] Özlem K. *Çanakkale 18 Mart University Faculty of Engineering and Architecture Department of Geophysical Engineering - Lecture Notes Section 6/7*

[2] Kandilli Observatory Earthquake Research Institute, http://www.koeri.boun.edu.tr/sismo/2/deprem-verileri/sayisal-veriler/.

[3] K Nearest Neighbor Algorithm, http://www.mademir.com/2010/12/k-nn-algoritmas.html.

[4] Quinlan, J. R. (1993) C4.5: Programs for Machine *Learning. Morgan Kaufmann Publishers, USA.*

[5] L. Breiman (2001) Random forests, *Machine Learning, vol. 45, no.1, pp.5 -32.*

[6] www.bilgisayarkavramlari.sadievrenseker.com.

Relationships between Land Use Types and Soil Development in the Narman-Alabalik Micro-catchment

E. Çomaklı[1], M. Özgül[2], T Öztaş[3]

[1]Atatürk University, Erzurum, Turkey. emrecomakli@atauni.edu.tr
[2] Atatürk University, Erzurum, Turkey. mozgul@atauni.edu.tr
[3] Atatürk University, Erzurum, Turkey. toztas@atauni.edu.tr

Abstract

The need for determining the effects of land use types and changes on soil quality is essential because of the increasing pressure on natural resources. The objective of this study was to evaluate the effects of land use changes on soil development in the Narman-Alabalik micro-catchment. Soil formation and erosion conditions of soils developed on different physiographic land groups (terraces, foot-slopes and back-slopes) in forestland, pastureland and cultivated land were evaluated. Top soil (Ap and A) development and thickness, horizons and layers, and other profile characteristics were obtained in situ. Physical and chemical properties of soil samples collected from different horizons and layers were analyzed for comparing and evaluating soils under different land use types. Due to the improper management of agricultural lands without considering topography and soil characteristics, early and overgrazing of pasturelands and misuse of forestlands have led to deterioration of soil properties and weakening of pasture vegetation and forests. Erosion caused by the deterioration of forest vegetation and pasturelands lead severe sediment accumulation in terraces. Land use methods and measures for land degradation and erosion prevention have been proposed by evaluating the socioeconomic structure of the micro catchment.

1. Introduction

Sustainable practices can halt the loss of forests, pastureland and agriculture areas. Anthropogenic impacts have important effects on land degradation (Baude et al., 2019). Misuse of lands has generally been considered a local problem, but it is day-by-day becoming a force of global importance. Land-use and land-cover changes are significantly affect soil behavior. The ecosystems is often modified by land use type, production practices and the land use intensity (Kasperson et al., 1995; Lambin et al., 2001). Such changes in land use have enabled humans to appropriate an increasing share of the land resources, but they also potentially undermine the capacity of ecosystems to sustain food production, maintain forest resources, regulate climate and air quality. Land use has also caused declines in biodiversity through the loss, modification of habitats and degradation of soil and water resources (Foley et al., 2005).

The objective of this study was to assess and compare soil development in farmland, pastureland and forestlands in the Alabalik micro catchment regarding to land use changes.

2. Methodology

2.1 Description of the study area

This study was conducted in Alabalik micro catchment about 110 km north of Erzurum (40°22′–40°27′N and 41°49′–41°58′E), Turkey. Total area of the study catchment is 2.895 ha. Annual rainfall averages 340.75 mm and is concentrated during the spring and summer (May-June). The dry season extends during winter (December–March). Mean temperature of the warmest month (August) and coldest month (January) are 22.7 and -4.2°C, respectively. The soils in the study site are shallow, sandy loam textured, colluvial and alluvial material originated without salinity and sodicity problems. Soils are classified commonly in Inceptisol. Average slope is 39%, very steep. Minimal slope (3%) was found in agricultural lands and maximum slope (66%) was found in forestlands. Intense human activities and topographic factors have hampered the regeneration of existing residual vegetation on degraded reforested lands and pasturelands. The forests have been extravagant exploited to meet the increasing demands for firewood and timber materials. Dominant tree species Pinus sylvestris L. in the forestlands. The pastureland in research area are very degraded by overgrazing. Dominant species Festuca ovina and Astragalus sp. in the pasturelands.

2.2 Soil sampling and vegetation analysis

Fifty surface soil samples (0–30 cm depth) were collected from 4 different land use/land cover types (forest, plantation area, pastureland and cultivated fields). Each land use type were randomly sampled and mixed to obtain a composite sample that was sealed in a plastic bag. Soil samples were air dried for 24 h, sieved through a 2 mm mesh. For each soil sample, the following soil characteristics were measured: organic matter (SOM) by the wet method of Smith-Weldon (Nelson and Sommer, 1982), soil pH was determined in 1:2.5 soil–water suspension, using a combination glass electrode (McLean, 1982), soil particle size analysis was done by the Bouyoucos hydrometer method (Gee and Bauder, 1986), $CaCO_3$ contents by the Scheibler calcimeter methot (Kacar, 1994) and soil EC was determined in saturation paste extraction, using a electrical conductivity instrument (Rhoades, 1982).

3. Results and Conclusion

The pH values of the forest, plantation area, pastureland and cultivated soils changed between 7.6 to 7.9 (Table 1). Forest, grassland and cultivated sites did not show alkalinity problem. Soil pH and EC were highest in cultivated lands (7.9 and 0.72 dS m-1 respectively), and lowest in forest areas (7.6 and 0.50 dS m-1 respectively) within the catchment. Soil pH and EC were likely more regulated in the more conventional cultivated land uses from fertilizer additions. Brye and Pirani (2005), similarly concluded that soil pH and EC were generally greater under tilled cultivated than under native land use. On the average soil organic matter content was 2.9 times higher in forestlands than that of cultivated land (Table 1). Similarly soil organic matter content was 4 times higher in forestlands as compared that of pastureland (Table 1). The differences in SOM between land uses indicated that in the cultivated land use in general, the SOM pool is less diverse with other soil nutrients than in the forestland. The reason for the lowest SOM in the pasturelands, is the over grazing of sloping areas. Soil tillage, overgrazing and land use change greatly influence on SOM content. $CaCO_3$ content of the forest, plantation area, pastureland and cultivated lands varied significantly from 3.7 to 12.4%. $CaCO_3$ content were highest in cultivated sites (12.4%), and lowest in the forestlands (3.7%) (Table 1). $CaCO_3$ content was high at the cultivated lands. Soils on limestone material are A-C horizon. Land use changes and management type, especially cultivation of deforested land rapidly diminish soil quality effects of misuse agricultural practices and overgrazing to pasture land. Land use changing resulted in deterioration of soil properties soils under forest and pastureland. As a result, degradation in soil quality may lead to a permanent reduce of land productivity. Overgrazing in the micro-catchment is very detrimental to soil fertility, resulting in a marked decrease in SOM contents. In the forest area, soils are systematically deeper in the pastureland and cropland positions compared. Despite differing types of managed land use type the micro catchment physiographic conditions region clearly had a greater influence. Results of this study demonstrate how the change of land use type, which constitutes the major differences between physiographic conditions that were investigated in this study, can have. Especially mismanagement can result in surface erosion at varying levels or quantities. Micro catchment are used for agriculture and livestock. Therefore, the local population needs to be developed economically to restore the natural balance in the micro-basin. To reduce erosion and sedimentation in micro catchment, management model that collaborates with native people should be using.

Table 1. Some average physical and chemical analysis results of soil sample

Soil properties	Cultivated land	Pastureland	Forestland
Clay,%	13	10	6
Loam, %	23	19	20
Sandy,%	64	71	74
Texture	Sandy loam	Sandy loam	Sandy loam
pH	7,9	7,8	7,6
EC, dS m^{-1}	0,72	0,65	0,50
$CaCO_3$, %	12,4	9,7	3,7
Organic matter, %	3,0	2,1	8,6

Acknowledgments

The authors are grateful to transport support was obtained from the Erzurum Regional Directorate of Forestry and Eastern Anatolia Forestry Research Institute.

References

Baude, M., Meyer, B. C., and Schindewolf, M. (2019). Land use change in an agricultural landscape causing degradation of soil based ecosystem services. Science of the Total Environment, 659, 1526-1536.

Brye KR, Pirani AL. Native soil quality and the effects of tillage in the Grand Prairie region of eastern Arkansas. The American Midland Naturalist. 2005;154:28-41

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., and Helkowski, J. H. (2005). Global consequences of land use. Science, 309(5734), 570-574.

G.W. Gee, J.W. Bauder Particle-size analysis A. Klute (Ed.), Methods of Soil Analysis, Part 1: Physical and Mineralogical Methods, Agronomy Monograph No. 9 (2nd ed.), American Society of Agronomy, Madison (1986), pp. 383-411

Kacar, B., 1994. Bitki ve Toprağın Kimyasal Analizleri III. Toprak Analizleri. Ankara Üniv. Ziraat Fak. Eğitim Araş. ve Gel.Vakfı Yay., No:3 Ankara.

Kasperson, J.X., Kasperson, R.E., Turner, B.L. II (Eds.), 1995. Regions at Risk: Comparisons of Threatened Environments. United Nations Univ. Press, Tokyo.

Lambin, E. F., Turner, B. L., Geist, H. J., Agbola, S. B., Angelsen, A., Bruce, J. W.,and George, P. (2001). The causes of land-use and land-cover change: moving beyond the myths. Global environmental change, 11(4), 261-269.

McLean, E. (1982). Soil pH and lime requirement. Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties, 199-224.

Nelson, D.W., Sommer, L.E., 1982. Total carbon, organic carbon and organic matter. In: Page, A.L., Miller, R.H., Keeney, D.R. (Eds.), Methods of Soil Analysis. Part 2. Agronomy, Vol. 9. Am. Soc. Agron., S.S.S. Am., Madison, WI, pp. 539–579.

Rhoades, J. (1982). Cation Exchange Capacity 1. Methods of soil analysis. Part 2. Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties, 149-157.

Using Deep Learning Models in Problem Solving

İ. Türkoğlu[1]

[1]Fırat University, Elazığ, Turkey, iturkoglu@firat.edu.tr

Abstract

The importance of artificial intelligence has been increased since the developing technology and digital era. The development of autonomous systems which can make decisions on its own serves both humanity and provides new job opportunities. Artificial intelligence can be defined as software and hardware systems that exhibit human behaviors, conduct numerical logic, and have many abilities including movement, speech and voice recognition. In general, AI can be divided into two groups: Machine learning and deep learning. Machine learning represents an algorithmic structure which learns from examples of qualitative information extracted from the data. Yet, deep learning systems learn from the data without training.

In this study, problem solving approaches are given by applying artificial intelligence techniques and deep learning models used the applications. Leaf classification, colon cancer detection and epileptic seizure recognition are presented to solve problems with artificial intelligence techniques.

1. Introduction

A new period is began with increasing technological developments in education, security, health, industry, etc. With this period, new business models, changing systems, various working and communication methods are emerged. Also, these developments start to reduce the need of human labor, and cause artificial intelligence based systems are more effective. Artificial intelligence can be defined as a model which can imitate the human behaviors, skills, and actions and can be used to process in computer based systems. It is able to solve problems, generate solutions, make signification and generalization, and learn from experiments. Artificial intelligence is a developing science with various applications and various subjects. In today, it is widely used in robotics, speech and image recognition, and health applications. Yet, in daily life, people need too many information and need to use that information to make a living. Most of this information is subjective and intuitive, thus it is difficult to comprehend and express. In artificial intelligence based systems, this information is need to be uploaded to the computer to make the systems process. Therefore, one of the most important difficulties encountered in this filed is how to transfer this information into the computer [1]. Artificial intelligence can be divided into two categories: machine learning and deep learning. Machine learning represents a structures that learns information from extracted data labels. In machine learning, information are obtained after pre-processing phase which are called features. However, obtaining key features require a specialization and it is a time consuming process. In order to solve this problem, deep learning has emerged. Deep learning allows computers to build more complex concepts and information. It is able to learn from raw data. Although, it is seen as a new concept, deep learning is a discipline which has existed since 1940s. In Figure 1, the history of artificial intelligence, machine learning and deep learning is given. Increasing size and number of data, and development of powerful computers make deep learning is more popular nowadays. Many commercial companies use deep learning process in their applications.

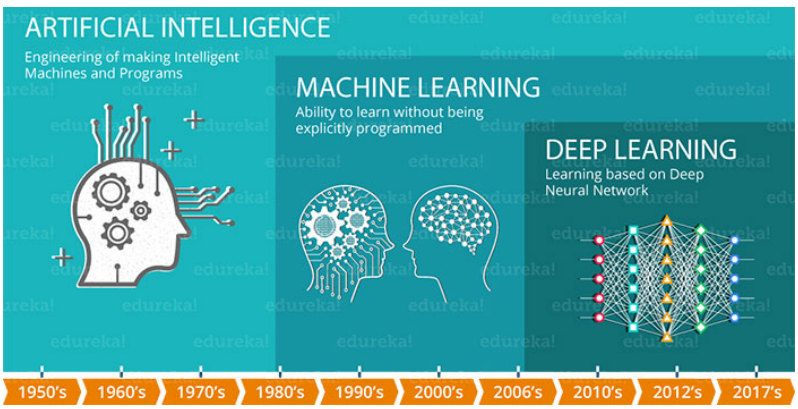


Figure 1: The historical development of artificial intelligence, machine learning, and deep learning [2].

2. Problem Solving Process

Deep learning process has begun at 2006. Prior to this year, feature extraction was the key process. Yet, with the development of powerful computers, feature extraction process was included in the artificial intelligence model. In this artificial intelligence model, the specialization need for high classification accuracy is decreased with increasing number of data. The components of problem solving process of deep learning are given below:

1. *Identification of the Problem:* Evaluation the suitability of the problem for deep learning.
2. *Preparation of the Data:* Formation of data set and analyzation of the data.
3. *Choosing a Deep Learning Model:* Determination of deep learning model based on the problem.
4. *Testing the Classification Accuracy:* Evaluation of the selected deep learning model based on the selected criteria.

3. Applications

In this section, problem solving applications are mentioned based on deep learning. These applications are; the classification of two-dimensional leaf images, determination of key features with deep learning to recognize colon cancer based on FTIR signals, and epilepsy recognition based on one-dimensional EEG signals. In the study of leaf classification, classification accuracies of deep learning models are compared and evaluated. To do that, AlexNet, GoogleNet, VGG-16, VGG-19, and ResNet50 are applied. In machine learning based image recognition studies, generally, preprocessing, feature extraction and classification steps are included. On the other hand, in deep learning applications, feature extraction and preprocessing steps are determined by deep learning. In the study, there are 32 leaf classes exist and each class includes 240 number of leaf images. Workflow of the proposed is given in Figure 2 below.

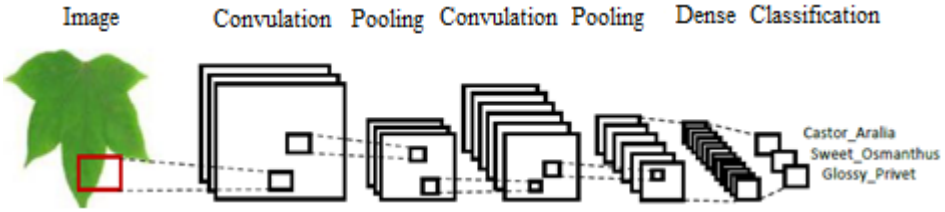


Figure 2: Workflow of the proposed method [3].

In the classification process, 80% of the data is used for training, and 20% of the data is used for testing. In the end of the study, best classification accuracy is obtained from AlexNet model with 99%. In the literature, classification accuracies differ from 88% to 97% with machine learning. Performing machine learning techniques is difficult both in terms of time and workload. With deep learning, accuracy is increased and classification performance differs from 97% to 99%. The detailed information can be seen in [3].

In the second study, deep learning is applied in order to detect colon cancer from FTIR signals and features are obtained with deep learning model. One of the difficulty encountered in cancer diagnosis based on blood samples is the similarity of FTIR (Fourier Transform Infrared) signals between normal and diseased patients. In machine learning applications, the difference between FTIR signals are determined based on the peak values. To solve this problem, CNN (Convolutional Neural Network) based method is proposed. In the first part of the study, spectrogram images are extracted from FTIR signals. The study is conducted on 70 data with 30 patients and 40 healthy subjects. AlexNet model is used as a deep learning model. Workflow of the proposed method is given in Figure 3.

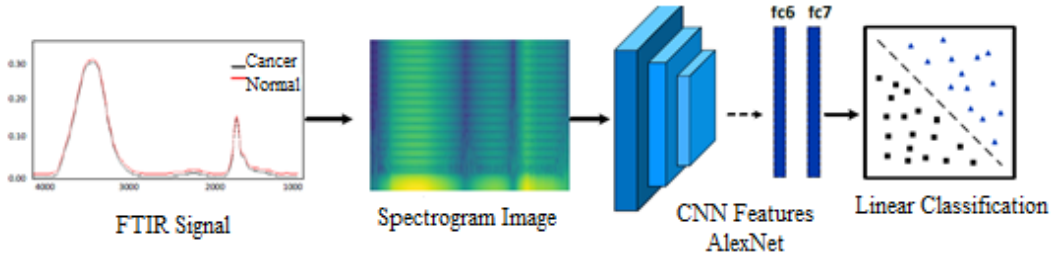


Figure 3: Workflow of the proposed method. [4].

Deep learning model is used to extract features from images. Firstly, key features are obtained with CNN model. Then, features are used as an input for SVM (Support Vector Machines) for classification process. Accuracy, sensitivity, and specificity are calculated the results are calculated as 86%, 85%, and 88% respectively. The detailed information can be seen in [4].

In the last study, deep learning model is used to determine the epileptic seizures. The main idea of the study is use the deep learning models on one dimensional EEG signals and determine the performance of these models. Brain signals (EEG) are non-stationary signals and include long time series. Interpretation and analysis of brain signals require an expert opinion and take some time. In this study, AlexNet, and VGG-16 models are used. During the study, 7500 number of data were used which include 3750 epileptic and 3750 non-epileptic signals. The workflow of the proposed is given in Figure 4.

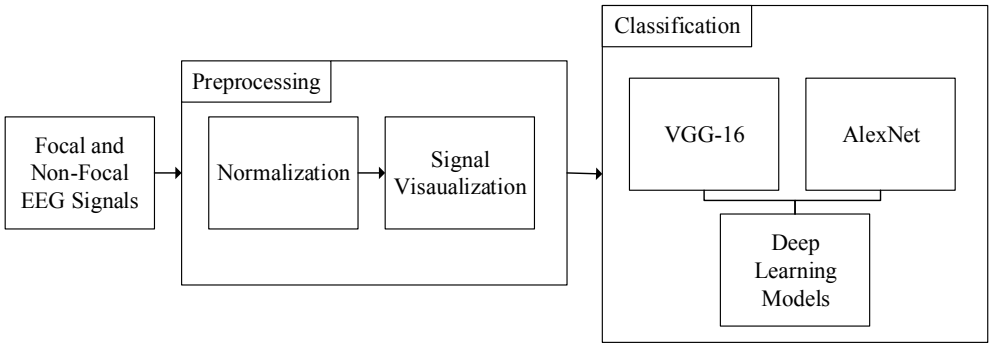


Figure 4: Workflow of the proposed method. [5].

EEG signals are converted into two-dimensional images and classified with deep learning models. Classification accuracies are obtained as 91%, and 86% for AlexNet, and VGG-16 respectively. Thus, it is shown that successful classification results are obtained with deep learning modes without extracting features from EEG signals. The detailed information can be seen in [5].

4. Result

In this study, the usage and the classification performance of deep learning models are evaluated in two-dimensional images and one-dimensional signals. It has also been shown that, in the case of feature extraction processes which requires specialization, deep learning can be used successfully for feature extraction without expert opinion.

References

[1] Goodfellow, I., Bengio, Y., and Courville, A., (2016). Deep Learning, *The MIT Press.*
[2] URL: What is Deep Learning?, https://mc.ai/, May 2019.
[3] Doğan, F., Türkoğlu, I., (2018). The Comparison of Leaf Classification Performance of Deep Learning Algorithms, *Sakarya Univ. J. of Computer and Information Sci.*, 1, 10-21.
[4] Toraman, S., Türkoğlu, I., (2018). Determination of Colon Cancer Risk from FTIR Signals by Deep Learning, *Science and Eng. Journal of Fırat Univ.,* 30(3), 115-120.
[5] Alakuş, T.B., Türkoğlu, I., (2019). Detection of Epileptic Seizures with Focal and Non-Focal Brain Signals by Using Deep Learning, *2nd International Congress on Applied Sciences*, Diyarbakir, 458 – 469, 2019.

Sentiment Analysis Of Tweets Using Machine Learning

M. Pek[1], M. Turan [2]

[1]University of Istanbul Commerce, Istanbul, Turkey, mesutpek@gmail.com
[2] University of Istanbul Commerce, Istanbul, Turkey, mturan@ticaret.edu.tr

Abstract

Today, with the improvements in technology, comments and sharing of people increased. In the past, before widespread internet usage, creating reports were easier. But now with advancements in internet and technology reporting got harder. Analyzing and processing data from this sharing platforms where millions of people are registered and expressing their opinions is easier. Important works on this field includes emotion analysis. In this study emotion analysis is done using machine learning and compared to other applications tried to improve success rate with improvements on feature selection. While using any other set than Naïve Bayes classification learning set this study has achieved %80 success rate.

1. Introduction

With the development of Web 2.0 on the internet, one of the new communication technologies, user-based "social media offers many feedback opportunities to individual users such as instant notifications, comments and complaints. Facebook, Twitter, LinkedIn, Instagram, micro blogs, also web-based news sites and even shopping sites, including many other platforms, instant messaging and comments to activate the user. Platforms can analyze the stored data in-house or through different companies. Politics is one of the areas in which data analysis is used. Kassraie et al. conducted an analysis with a data set containing more than 370,000 tweets in the 2016 US Elections and used Gauss regression model in their study. Estimation of the study of tweet data had a %1variance on the election result.[6] With emotion analysis, users sharing on products or subjects can be grouped as positive or negative. Akgül et al. in their study wrote an application to acquire tweet data from the twitter API.[2] They tagged tweets from Twitter API as positive, negative and neutral. After clearing the data, they removed the word frequencies and the program they wrote found to have a success rate of 70% with n gram method. Bari and Saatçioğlu conducted studies on data sets belonging to Amazon, Cornell, IMDB, Twitter, Yelp, Reviews and Kaggle in his study on Emotion Analysis.[7].Textblob and Stanford NLP gave good results. Opinion Finder gave the worst result.

Onan preferred used the zemberek library for working with Turkish tweet data.[3] At the end of the study, the results were compared. The best predictor of the algorithm was Naive Bayes with 76%. After analyzing emotions with Tweet data, Şeker and Yeşilyurt conducted a study on expressing the emotion of the person by turning them into emojis.[4] In the data mining program called Rapid Miner, many algorithms have been studied by using machine learning. The most successful algorithm was Bayes with 52%. Gamal et al. examined 4 different data sets with machine learning algorithms in his study.[1] Naive Bayes algorithm has found a 74% success rate in this study on twitter data set. Naive Bayes algorithm gives the best results in different data compared to other algorithms. A study covering different languages is done by Hartmann et al.[5] That study compared the performance of ten different approaches (five dictionary-based and five machine learning algorithms) in 41 social media datasets covering various sample sizes and languages from different social media platforms. In that study, it is found that Random Forest (RF) and Naive Bayes (NB) algorithms gave the best results.

2. Data Set And Data Cleaning

Kaggle is a platform where competitions are held with data sets used in machine learning. In the data set; there are two main information in the form of emotion and tweet.

Data cleaning is an important process for the model to learn well. By removing unnecessary words and expressions in the data set, the model can be improved. Data can be edited by converting URL (www., http, https), username, hashtag, signs (- and /) and converting to lowercase for cleaning with regular expressions. Example: if the model is trained without clearing data, passing @Mesut in the tweet, there will be a possibility of incorrect classification of the model and the success rate will decrease.

3. Machine Learning

The concept of machine learning has become widespread in recent years. Machine learning, searches for some patterns in the data with various algorithms and methods and learns by looking at the labels corresponding to these patterns, then, when faced with a similar situation as they have learned, it creates systems capable of extrapolating from past experiences. This possibility is provided by many algorithms using various mathematical and statistical methods. One or more of these methods and algorithms are used together to create a model and this model aims to predict what is to be estimated in the most efficient, most precise and fastest way.[8]

3.1 Naive Bayes Algorithm

It is a classification algorithm that classifies data by calculating with probability principles. Simply put, a Naive Bayes classifier assumes that the existence of a particular property in a class does not depend on the existence of any other property. It is one of the most preferred algorithms in machine learning. There are many studies in the literature (Wikipedia, 2019). Formula of Naive Bayes algorithm:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(A|B) ; is the probability of A happening in case B happens
P(B|A) ; is the probability of B happening in case A
P(A) ve P(B) ; is the prior probabilities of A and B cases

4. Creating A Model

The structure diagram that we will use for machine emotion analysis with tweets is shown in Figure 1. For the words from the two data sets, the data is cleaned as a pre-stage. With data cleanup, URL, name, tag parsing and lowercase conversion are applied. The purpose of converting to lowercase is because we want these words to take up a single place in the list. If the data cleaning phase is not applied, the success rate of the system decreases, and the runtime increases because the model is not properly trained. When the ineffective words and pronouns are removed, the success rate increases.By finding similar words, the word list is normalized. Thus, the list is processed with fewer and correct words. The algorithm calculates the probability of each case for each word due to the way it works and classifies it according to the highest probability value. After training the model, tests can be performed according to any data set.

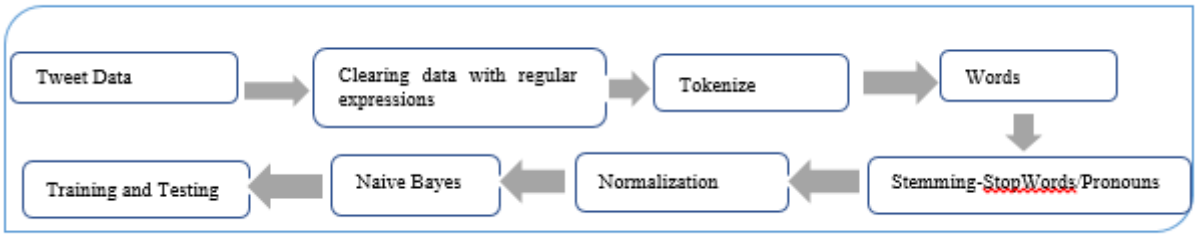


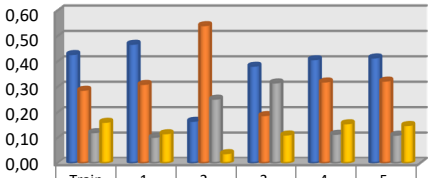
Figure 1: Flow chart of the proposed model

5. Conclusion

In the first stage of the study, transactions were performed with a single data set without using pronouns. The first data set is reserved for 75% training and 25% testing. By writing the codes of the Naive Bayes algorithm from scratch, it was possible to increase performance and observe the operations performed. 73% success rate was achieved during the development phase. While the model was being trained, the effect of pronouns and nouns were removed, and the code improved to increase the success rate. Figure 1 shows the list of names 80% success rate was achieved in a second independent data set. In the confusion matrix, the values of the first and second tests were compared by normalizing. You can see the comparison numerically and graphically in Table 1 As shown in Table 2, an increase in the success rate was observed. When we examined the accuracy value in the success rate, for the first data set it is found approximately 72% while for the second dataset it is found %78. When we look at it, we can see that the accuracy value increases by 6%. When the results were examined, the learning data set had a prediction rate of 73%, and the value reached 80% when tested with the second data set. There was an increase of about 7. The data set consisting of tweets for the airline we used for the test found 70% accuracy and 81% consistency values. In the

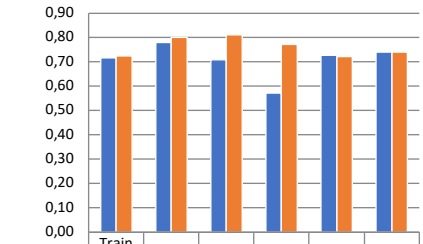
results, 57% accuracy rate was lower than other data sets. The success rate for accurate prediction is 77%. In fourth test; The model was trained with the first 50,000 tweets included in the 1st data set used for testing. When the complexity matrix and success rate were examined, it was observed that the values were close to the first test with the educational data. In order to increase the success rate, the number of tweets used in the training was updated to 100,000 and the 5th test was conducted. Accuracy and estimated value were observed to be 74%. There was a 2% increase in success with the fourth test.

In the study conducted with Naive Bayes, it was observed that the success rate was increased by improving the model. Compared to other studies, Akgül et al.[2] 70% success rate in his study, Onan in his study 76%, Şeker and Yeşilyurt his study 52%, Gamal et al. 74% success rate was found.[1][3][4] As a result of the success in our model, we achieved a higher success rate than other studies. In order to achieve a high success rate, it is recommended to perform a preprocessing on the data set to increase the success rate. In subsequent studies, support vector machines or random forest studies from other machine learning algorithms can be performed and success rates can be compared.



	Train Datase t	1. Testing	2. Testing	3. Testing	4. Testing	5. Testing
TP	0,429	0,469	0,166	0,383	0,408	0,415
TN	0,288	0,311	0,542	0,188	0,321	0,324
FN	0,121	0,103	0,253	0,317	0,115	0,111
FP	0,162	0,117	0,039	0,112	0,156	0,149

Table 1: Confusion Matrix



	Train Datase t	1. Testing	2. Testing	3. Testing	4. Testing	5. Testing
Accuracy	0,716	0,78	0,708	0,571	0,728	0,74
Precision	0,7257	0,8003	0,8109	0,773	0,723	0,74

Table 2: Success Rate

References

[1] Gamal D., Alfonse M., El-Horbaty E. M., Salem A.B.M, (2018), Analysis Of Machine Learning Algorithms For Opinion Mining İn Different Domains, *Machine Learning And Knowledge Extraction* — Open Access Journal.

[2] Akgül E.S., Ertano C., Diri B.,(2016), Sentiment analysis with Twitter, *Pamukkale University Journal of Engineering Sciences* (Vol. 22, No. 2, pp 106-110).

[3] Onan A., (2017), Emotion Analysis Based on Machine Learning Methods on Twitter Messages, *Journal of Management Information Systems* (Vol. 3 No. 2).

[4] Şeker, Ş. E., Yeşilyurt A.,(2017), Twitter Sentiment Analysis using Text Mining Methods, *YBS Ansiklopedisi* (Vol. 4, No. 2).

[5] Hartmann J., Huppertz J., Schamp C., Heitmann M.,(2019), Comparing automated text classification methods, *International Journal of Research in Marketing.*

[6] Kassraie P., Modirshanechi A., Aghajan H.,(2017), Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls, *In Proceedings of the 6th International Conference on Data Science.*

[7] Bari A., Saatçioğlu G., (2018), Emotion Artificial Intelligence Derived from Ensemble Learning, *Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering.*

[8] Machine Learning, Date of Access: 01.05.2019, https://tr.wikipedia.org/wiki/Makine_%C3%B6%C4%9Frenimi.

Empiric Findings For The Relationship Between Information Communication Technologies And Economic Growth

V. Inal[1], M. Torusdag[2]

[1] *Sakarya University, Faculty of Political Sciences, Department of Finance, veyselinal@sakarya.edu.tr*

[2] *Van Yüzüncü Yıl University, Faculty of Economics and Administrative Sciences, Department of Economics, mustafatorusdag@yyu.edu.tr*

Abstract

The aim of this study is to investigate the relationship between information and communication technologies and economic growth. Information communication technologies, which have a significant proportion in the world economy market, make it possible to attain economic markets more easily due to the developing infrastructure opportunities. Therefore, the importance given to the studies on information communication technologies is increasing day by day. In our study, the relationship between information communication technologies and economic growth for 15 OECD countries was analyzed by panel causality analysis and Smith et al. (2004) panel unit root methods are investigated. As a result of the analyzes, there is a causal relationship between the information communication indicators and economic growth, and their economic performances will contribute positively if these countries fulfil the deficiencies in the infrastructure investments related to information and communication technology and adopt and choose policies towards this.

1. Introduction

The developments seen in information and communication technologies in today's world, which is called knowledge economy, affect every aspect of life as well as economic life in real terms. The production, consumption, distribution composition and market structure in the economy are influenced by knowledge-based changes, leading to intense competition at the international level. Economic activities have become a global structure with the importance of information technologies and developments in this field. The main actor in the formation of this structure is the integration of the Internet into all areas of life and the exponential growth of telecommunication investments. It has been shown in many studies that there is a positive relationship between information communication technologies and economic growth. Especially developed countries are in a better position in this sense. Developing countries, on the other hand, make the necessary infrastructure investments and pursue development-oriented policies and try to close the gap between developed countries.

Information communication technologies form the basis of the concept of "new economy" which is frequently encountered in the literature. This economy is defined as "new economic relations based on information technologies, new business areas and the reshaping of existing business areas using new communication environments" (Baily and Lawrence, 2001: 8). Today, world economies are going through a rapid transformation from industrial society to information society and the main reason for this transformation is the speed of development and expansion of new technologies and the adaptation of countries to these technologies. While some sectors have lost their importance in today's new economic structure, the fact that new sectors with high profit margins gained importance and spread rapidly have changed the dimensions of both the economy and competition. Together with the new economic conditions based on information and technology, the realization of economic activities in a virtual environment, a digital revolution in communication, and the ability to carry out transactions with the state without going to government offices also reveal the extent of the developments. In this new environment, the transnational borders have disappeared and new economic developments have increased the importance of information in the economic field, and the "knowledge based economy" and/or "knowledge economy" highlighted (Yeloglu, 2004: 177-179).

While the introduction of information and communication technologies in our lives dates back to the 1950s, it began to gain momentum in the 1980-1990s. With the widespread use of computers, the process accelerated. In order to solve the economic crisis especially in the western world in the 1970s, the application of neo-liberal economic policies in the early 1980s and the idea that the multiplying effect of information technology-based industrial policies and information investments would have positive effects on employment and growth were adopted by many countries (Kevük, 2006: 320). In 1995, information and communication technologies became the "World Wide Web (www)' application that enabled the masses to use the Internet. In addition, the decrease in hardware costs in ICT has led to significant increases in the number of beneficiaries. This increase in demand has brought the ICT industry to the forefront (OECD Publication, 2000: 10).

2. Literature Summary

Dewan and Kraemer (2000), examined the impact of ECO (Economic Cooperation Organization) on economic growth. Panel data analysis was conducted for 36 countries consisting of DC (Developed Countries) and DC (Developing Countires) using the annual data of 1985-1993 period. In the study, ECO investments were positive and statistically significant for the EMUs; It was determined that it was not significant and meaningful for the developing countries. The effects of ECO capital and non-ECO capital on output (the output elasticities of the variables) are 0.057 and 0.160, respectively. Calculations for developing countries showed that the output elasticity coefficient of non-ECO capital was 0.593, and the output elasticity of ECO capital was statistically non-zero. The differences in the findings of the study according to the country groups are explained by the fact that developed and developing have different physical capital accumulation and that there is a sufficient and appropriate infrastructure. As a result, it was emphasized that the development required for the transition to the information age cannot be achieved only through ECO investments, and that it should be at a certain level of development.

Roller et al. (2001), in their study on 21 OECD countries, examined the extent to which telecommunication investments affect the economic growth rate in the 1980-2000 period by establishing different growth models. As a result of the study, they stated that telecommunication investments had a positive and statistically significant effect on growth, but the degree of this effect differs according to the models used. Based on the estimation results, the researchers stated that telecommunication investments contributed positively to economic growth but that ignoring the individual country effects would produce biased results.

Yamak and Bozkurt (2003), in the period of 1996-2000 in 47 countries, the effects of ECO (Economic Cooperation Organization) investments on economic growth for developed and developing countries are investigated by OLS (ordinary least squares) method. In addition to the investment expenditures of ECO, the production function consisting of the rate of increase in money supply, inflation rate and rate of increase in export revenues was used. In the study in which data belonging to 1996-2000 period were analyzed with panel data technique, the data set was classified according to the development level of the countries. In addition, the effects of ECO investment expenditures on economic growth in G-7 countries (USA, Japan, Germany, United Kingdom, France, Italy, Canada) were investigated. According to the findings, the effect of ECO investments on economic growth is negative or zero for developed and developing countries and this effect is positive for G-7 countries.

Seki (2008), tested the performance of ICT (information and communications technology, or technologies) sectors in selected OECD countries by using data envelopment analysis method based on 1980-2003 period. It calculated the level of technical efficiency, changes in technical efficiency, technological change and changes in total factor productivity for selected OECD countries. Luxembourg was selected as the reference country according to the technical productivity index. Accordingly, South Korea was the worst performing country. Technological progress has been observed in all countries included in the analysis. All countries except Mexico had a positive change in total factor productivity. According to the total factor productivity and technical efficiency index remained below the OECD average of Turkey. Japan has the best performance in terms of technological change index, while Norway has been the most successful country in terms of technical efficiency.

Petals and Scott (2010), studies have tested the relationship between the 1980-2008 period, using time series analysis of ICT and economic growth in Turkey. According to the analysis results, economic growth in the short and long term is positively affected by ICT. It was also agreed that the contribution of ICT to economic growth in Turkey are lower than the other factors of production. Granger causality tests showed a two-way causality between economic growth and knowledge.

Gulmez and Akpolat (2014), Turkey and 15 EU countries, R & D activities in the study, they investigated the relationship between innovation and long-term economic growth. In this study, R&D expenditures per capita for R&D activities, number of patents for innovation and income per capita for growth are used. As a result of their panel data analysis, they showed that R&D expenditures are 4 times more effective on economic growth than patents. As a result, they stated that there is a positive and significant relationship from R&D expenditures and patent numbers to economic growth in the long run. When the studies are examined in general, it is seen that information communication technologies contribute positively to economic growth.

3. Empirical Analysis

In this section where the econometric analysis of the relationship between ICT and economic growth is made, the effects of ICT on economic growth are examined empirically. Turkey, Austria, Belgium, France, Germany,

Hungary, Ireland, Italy, Japan, Netherlands, Poland, Portugal, Spain, the United States and consisting of Slovenia 15 OECD countries in the 1990-2016 period of years in the analysis made in the study, which uses data generated models, panels data analysis techniques were used. In this study, internet users data (INT) per 100 people was used as an indicator of information communication technology. The GDP per capita representing economic growth was included in the analysis.

In the study, the existence of the dependence among the cross-sections (countries) forming the panel was determined by the cross-section dependence tests developed by Breusch-Pagan (1980) and Pesaran (2004). Unit root tests to determine the stability of the series vary according to the homogeneous and heterogeneous curve parameters. In the study, Delta test developed by Peseran and Yamagata (2008) was applied before unit root tests and curve parameters were found to be heterogeneous. In accordance with this feature, unit root test developed by Smith et al. (2004), which is one of the second generation unit root tests, was applied and the degree of stationary was determined as I (1). The causality relationship between the series was investigated by Dumitrescu and Hurlin (2012) panel causality test. While there is a cross-sectional dependence between the series, taking into consideration this fact has an important effect on the results to be obtained Breusch and Pagan, 1980; Pesaran, 2004. Because when selecting methods for unit root and cointegration tests, this situation should be taken into consideration.

Table.1: Cross Section Dependency Test Results of INT Variable

Test	Statistic	Prob.
Breusch-Pagan LM	4412.695	0.00
Pesaran scaled LM	215.5937	0.00
Bias-corrected scaled LM	215.1770	0.00
Pesaran CD	66.39112	0.00

According to the results in the tables; Since the probability values are less than 0.10, H_0 hypothesis (cross-section dependence) is strongly rejected. In dynamic panel data analysis, firstly, whether the curve coefficients are homogeneous or heterogeneous for each country should be determined by the delta tests developed by Pesaran and Yamagata (2008). Homogeneous or heterogeneous curve coefficients change the form of unit root and causality tests to be applied. In this study, it is examined whether the curve coefficients are homogeneous with the help of delta test.

Table 2: Delta Test Results

Test	Test istatistiği	Prob.
$\tilde{\Delta}$	2.850	0.000***
$\tilde{\Delta}_{adj}$	3.389	0.000***

*** Indicates significance at the level of 1%.

According to the results in the table, the slope coefficients, which are the null hypothesis, are homogeneous and the hypothesis is rejected. That is, curve coefficients are heterogeneous. In this study, since the cross-sectional dependence was determined among the countries forming the panel, the stationarity of the series and the unit root test developed by Smith et al. (2004), one of the second generation unit root tests, were used. Smith et al. (2004) $\overline{LM}$, $\bar{t}$, $\overline{Min}$, $\overline{Max}$ ve $\overline{WS}$ and named 5 bootstrap panel unit root test is existed.

Table 3: Smith et al. (2004) Panel Unit Root Test Results

TESTS	INT (Fixed Model)		INT(Fixed-Trend Model)	
	Level	Differenced	Level	Differenced
$\bar{t}$	-1.337 (0.456)	-2.383*** (0.000)	-0.253 (0.998)	-2.912*** (0.000)
$Ma\bar{x}$	-1.055 (0.066)	-2.200*** (0.000)	-0.075 (0.977)	-2.782*** (0.000)
$L\bar{M}$	2.455 (0.882)	5.615*** (0.000)	1.873 (0.974)	6.901*** (0.000)
$Mi\bar{n}$	1.884 (0.519)	5.047*** (0.000)	1.518 (0.691)	6.478*** (0.000)
$W\bar{S}$	-1.089 (0.598)	-2.389*** (0.000)	-0.218 (0.999)	-3.007*** (0.000)

Note: The model includes fixed and trend terms. *** Indicates that unit root zero hypothesis is rejected at 1%, ** 5% and * 10% significance levels. Probability values are derived from 5000 bootstrap cycles. The block volume and the maximum delay length are 100 and 4, respectively.

While all variables have unit root in level values, they become stationary in their first difference. Therefore, it is determined that the stationary level of the variables are I (1). Dumitrescu & Hurlin (2012) panel causality test, which takes into account the cross-sectional dependence, is a prerequisite for the variables to be stationary at the same level. For the purpose of the study, the method developed by Dumitrescu and Hurlin (2012) was used to test the causality relationship between the series. The advantages of this method are; it can take into account both the horizontal cross-section dependence and heterogeneity between the countries that make up the panel, can be used when the time dimension is smaller than the horizontal cross-section dimension (N) and can produce effective results in unstable panel data sets (Dumitrescu and Hurlin, 2012). Another feature of the Dumitrescu and Hurlin test is that it can analyze both in the presence and absence of cointegrated relationship.

Table 4: Dumitrescu ve Hurlin (2012) Panel Causality Test Result

Direction of Causality	W stat	Z-bar stat	Significance Value	Decision based on W stat.
INT → LKGSYH	4.34532	3.02038	0.0000***	INT ↔ KGSYH
LKGSYH → INT	7.84310	5.42695	0.0000***	There is a bi-directional causality relationship between the variables.

Note: *** 1%, ** 5% and * 10% means the level of significance. The lag length (K): 3 was determined. ↔: indicates bidirectional causality. There is no causal relationship. ↛: shows one-way causality relationship.

4. Result

As a result of the analyzes, there is a causal relationship between the information communication indicators and economic growth, and their economic performances will contribute positively if these countries fulfil the deficiencies in the infrastructure investments related to information and communication technology and adopt and choose policies towards this. As a result of panel causality test, it was revealed that there is a two-way causality relationship between economic growth and information communication technology. A consistent result was obtained with the studies in the literature. The reciprocal causal relationship between ICT representation variable internet usage rates and economic growth emphasizes the importance of e-commerce in economic life. The importance of information and communication technologies has emerged once again for people who are distant from the globalizing economic markets. In this sense, the validity of the necessary infrastructure investment policies of the OECD countries in the sample group has been demonstrated and found to be viable policies. It is obvious that the countries that have not completed their development or developing countries can close the gap between developed countries if they give importance to information communication technologies. Empirical results from our study may be presented as proof of this.

References

[1] BAILY M. N. and R. Lawrence (2001), "Do we have a new e-conomy?" National Bureau of Economic Research.
[2] YELOGLU O. (2004), "Knowledge Economy and Variables: Comparison of Turkey and OECD Countries". 3rd National Congress on Knowledge, Economy and Management, 25-26 November, 177-185.
[3] KEVÜK S. (2006), "Knowledge Economics", *Journal of Yaşar University,* C: 1 No: 4 ss. 319-350.
[3] OECD (2010), Communications outlook:, www.oecd.org/sti/ict/outlook, Date of Access: 10.05.2019.
[4] DEWAN S. and K. L. KRAEMER (2000), "Information technology and productivity: evidence from country-level data", *Management Science*, 46.4: 548-562.
[5] ROLLER L. H., and Waverman L. "Telecommunications Infrastructure and Economic Development: A Simultaneous Approach," *American Economic Review* (91:4), 2001, pp. 909-924.
[6] YAMAK R. and H. BOZKURT (2003), "Knowledge Technology and Economic Growth: Panel Data Analysis". II. Knowledge, Economy and Management Congress, Kocaeli University, Faculty of Economics and Administrative Sciences, Kocaeli.
[7] SEKI I. (2008), "*The importance of ICT for the knowledge economy: A total factor productivity analysis for selected OECD countries*" , In: Papers of the Annual IUE-SUNY Cortland Conference in Economics, *İzmir University of Economics*, p. 72-90.
[8] GÜLMEZ A. and A. G. AKPOLAT (2014), "R & D, Innovation and Economic Growth: A Panel Data Analysis for Dynamic Case of Turkey and the EU", *Journal of Abant Izzet Baysal University Institute of Social Sciences*, C.14(2), ss. 1-17.
[9] SMITH, L. Vanessa, et al(2004), "*More powerful panel data unit root tests with an application to mean reversion in real exchange rates*", *Journal of Applied Econometrics*, 19.2: 147-170.
[10] BREUSCH, T.S and A.R. PAGAN (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification Tests in Econometrics*"*, *Review of Economic Studies*, 47, 239-53.

[11] PESARAN, M. H., A. ULLAH and T. YAMAGATA (2008), "A bias - adjusted LM test of error cross section independence", *The Econometrics Journal,* 11.1: 105-127.
[12] PESARAN M. H. and T. YAMAGATA (2008), "Testing slope homogeneity in large panels", *Journal of Econometrics*, 142.1: 50-93.
[13] SMITH, L. Vanessa, et al(2004), "*More powerful panel data unit root tests with an application to mean reversion in real exchange rates*", *Journal of Applied Econometrics*, 19.2: 147-170.
[14] DUMITRESCU E. I. and C. Hurlin (2012), "Testing for Granger non-causality in heterogeneous panels", *Economic Modelling*, 29.4: 1450-1460.

Real Time Application of Self Organizing Maps (SOM) and Neural Network (ANN) Algorithms for Credit Score Card Production

A. Tunc[1]

[1]*Kuveyt Türk Participation Bank R&D Center, Konya, Turkey, ali.tunc@kuveytturk.com.tr*

Abstract

Credit score card production and development work has a very important place in the finance sector. These scorecards developed in line with the customers' information are taken as basis in any financial transaction requested. And according to this information, the decision is made whether the customer can be allocated the necessary loans and funds. In this context, the most accurate results are obtained by analyzing the data available by the financial institutions in the best way. Existing customers, as well as the status of newly acquired customers need to be determined. In other words, there is no need for any financial transactions of the customer and there is also a need for the financing and score card of this customer. In the proposed study, SOM (self-organizing maps) and ANN (Artificial Neural Network) algorithms were used to analyze the data obtained from a special financial institution. Then, when an unknown customer arrives, the new customer can use the credit or not use information is estimated by including instantaneous customers upon the learned data. The proposed method has been run in a number of iterations, the performance and success rates of the method are presented in real time.

Introduction

Computer technologies are widely used in daily life. Most of the information in our daily life is kept on computers. The amount of data stored on the computer is constantly increasing. Many algorithms and methods have been developed to make these data meaningful and to use in various fields. These methods and methods are commonly referred to as data mining. The concept of data mining is to create meaningful information by evaluating the available data [1]. It has been successfully used in many fields such as medicine, engineering and finance. Data mining studies have achieved significant results, especially in many medical studies [2]. The basic method is to convert hidden information and relationships in the raw data into estimated information [3] [4]. By means of these methods and methods developed, the relations between the data were determined and the results based on these relations were tried to be put forward. In order to establish these relationships correctly, the data should be passed through the preprocess techniques and the necessary statistical and learning algorithms [5].

In this study, a basic data set with 1000 lines of record and credit result information from 35% negative and 65% positive results are investigated. In the a random set of data row in the data set, the result column value was deleted to simulate the real time learning environment. For the records that come in random order via this data set, the data up to that record is trained by the Self Organization Map (SOM) and Artificial Neural Network (ANN) methods based on the training the user can used the credit or not as a result of the credit . In this study, the real-time learning outcomes of SOM and ANN algorithm were evaluated separately and the results were compared and the success percentages were determined.

When the literature scans are performed, SOM algorithms are used for the analysis of high dimensional data [6], [7] for the determination of properties [7,8], for map applications [9], for initial determination [10], for prioritization selection studies [11], for classification, optimization, accelerating methods and it is observed that it is used in many areas such as knowledge discovery [12]. Using the ANN algorithm, classification [13,14] is used in many areas such as detection of diseases [15], estimation [16,17], modeling, simulation and resource management.

Material and Method

Within the scope of the proposed article, a real-time machine learning method based on SOM (self-organizing maps) and ANN (Artificial Neural Network) algorithms was tried to be put forward. In order to provide a real-time information flow, the process of converting raw data to meaningful information can be reached very quickly and the data must be correctly analyzed and organized. The online training method has been adopted so that the inputs can be updated in real time as the inputs become available. In this method, the input samples are given to the model in random order and are passed through the input set up to that time [18].

The SOM method is a typical artificial intelligence technique used in clustering analysis. One of the most important reasons why neural networks-based techniques are preferred to statistical modeling techniques is that they do not require assumptions about the distribution of data. SOM provides a map in which high-sized data is depicted in lower dimensions and this map displays the cluster to data mining researchers. The SOM method does not require assumptions about the number of clusters, such as some statistical clustering methods, the probability distributions of variables and the independence of variables [19]. The primary goal of SOM is to convert pattern signals of any size to one or two dimensional finite maps. To be Traning and Mapping as name works in two stages. Generally used for classifying. The main feature of these networks is that they do not need a teacher to learn about the events. The signals (input values) sent to SOM networks pass through some operations (transmission-weighting) to the map layer. This layer consists of 1 or 2 dimensional nerve cells. The structure of the self organizing maps is given in Figure 1.

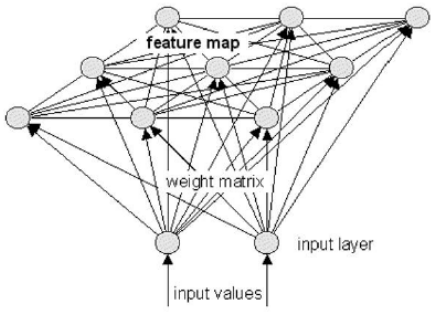


Figure 1 - SOM (Self Organizing Maps)

Type	- Feed Back / Feed Forward
Layers	- 1 Input Layer, 1 Maps Layer
Types of Input Value	- Binary, Real
Activation Function	- Sigmoid
Learning Methods	-
Learning Algorithm	- Self Organization
Usage Areas	- Pattern Classification - Optimization Problem - Simulation

The working structure of the SOM algorithm is as follows.

1. The weight values of the neurons in our network are started randomly.
2. Receive input vectors. (Target vectors in the system)
3. All values on the map are navigated. The distance between the input vector and the measured map value is calculated as the euclide distance.
4. The node with the shortest distance is taken (called the best matching unit, BMU)
5. All nodes adjacent to the selected optimal node are updated and approached to the input vector. Equation (1) uses the formula for calculation.
6. Return to step 2 as long as $t < \lambda$ is repeated.

$$Wv(t+1) = Wv(t) + \Theta(t)\alpha(t)(D(t) - Wv(t)) \quad (1)$$

t = current step , λ = time limit on the step , Wv = weight vector , D = target input value , Θ(t) = neighborhood function , α(t) = time-dependent learning limit

The concept of Artificial Neural Network was introduced with the idea of mimicking the working principles of the brain on digital computers, and the first studies focused on the mathematical modeling of the biological cells forming the brain, or the name of the neurons in the literature. Artificial Neural Network can be used to assemble many cells in a certain order and by using appropriate learning algorithms, neural networks can be established and these networks can perform very complex tasks successfully [20]. The working structure of artificial neural networks is as follows:

- Determination of model
- Determining the topology of the network
- Determination of the number of inputs and outputs
- Determining the learning parameters of the network
- Determination of learning coefficients and constants
- Assigning the initial values of the network
- Epoch count
- Display to sample network
- Calculation of the error
- Update weights according to the found error
- Calculation of the total error of the system

ANN is the structure of artificial nerve cells connected to each other. But the combination of nerve cells is not random. The structure of the artificial neural network is given in Figure 2.

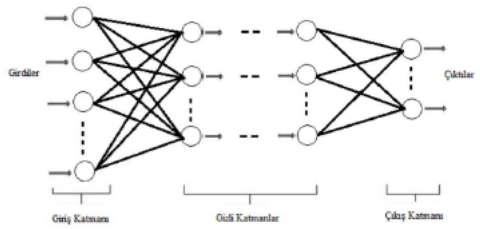


Figure 2 – Artificial Neural Network Structure

Input Layer is the layer where the input from the outside world to the artificial neural network. This layer contains as many cells as the number of inputs, and the inputs are passed to the hidden layer without any processing.

Hidden Layer: Processes the information from the input layer to the next layer. The number of hidden layers and the number of cells in the hidden layer can change from network to network. Cell numbers in hidden layers are independent of input and output numbers.

Output Layer: Processes the information to the hidden layer and sends the output produced in accordance with the input from the input layer to the outside world. The number of cells in the output layer may be greater than one. Each output cell has one output. Each cell is connected to all cells in the previous layer.

Application

In the study, 1000 records of the loan applications of a private bank were considered as learning data. Of these records, (65%) consisted of 650 of which are positive (approved) loan applications and and (35%) 350 of them are negative (rejected) applications. In data mining applications, It is possible to use open source applications such as 'Clementine' of SPSS, 'MineSet' of Silicon Graphics, IBM's Intelligent and SAS Institute's Enterprise Miner or WEKA [21]. The application is written specifically for the proposed article. The operating software for the study was developed using the C # programming language in the Visual Studio 2015 development environment. Machine learning approach was adopted in the proposed structure and SOM and ANN Algorithms were used together in the production process. These algorithms include; It was included in the program developed using the functions of the DLL set for the C # programming language of the WEKA program and used these algorithms to produce the required results.

In the developed software, an iteration number and an interface that can have been taken from outside is prepared. Afterwards, the random line numbers of the 1000 records are selected for the test and the success rate of the proposed method has been determined by checking the results of the records in the selected row through the trained system. Recordings up to the selected random order are written to a separate file and used as data training data. After the necessary training on this training data, the result of the selected recording was tried to be estimated according to this learning. Then, the result obtained in the estimation is included in the training set to update the training data. This process; 5, 10 and 20 random records were repeated on the results were tried to be compared. The procedures were tested in different iterations on both the SOM algorithm and the ANN algorithm. The results were recorded and compared with other methods.

The tables shown below show the correct or incorrectly estimated results of the study records. The result of the randomly selected recording is recalculated according to the existing training set. If the result of the record according to the learning data set and the result before entering the learning data is the same, it is written in the correct number field. If the result of the learning data set is not the same as the result, the wrong number field is written. The result of the recording is updated as a result of the learning data set.

In the developed software; The results of the accepted online records on the SOM algorithm show Table 1, Table 2 and Table 3. Table 1 shows the results of the five selected randomly selected records on the SOM algorithm according to the iterations. Table 2 shows the results obtained from the iterations of 10 randomly selected records on the SOM algorithm of the training data. Table 3 shows the results of 20 randomly selected records on the SOM algorithm according to the iterations. According to these results, the results of SOM algorithm according to the

success of online learning and the number of noisy recordings were observed. The success and values obtained in the conclusion part are mentioned.

Similarly, the results on the ANN algorithm which are accepted as online are shown in Table 4, Table 5 and Table 6. Table 4 shows the results of the 5 randomly selected recordings on the ANN algorithm of the training data according to the iterations. Table 5 shows the results obtained from the iterations of 10 randomly selected records on the ANN algorithm of the training data.

Table 6 shows the results of the 20 records selected randomly on the ANN algorithm of the training data according to the iterations. According to these results, the success of the ANN algorithm in online learning and the results of the selected noisy recordings were observed. The success and values obtained in the conclusion part are mentioned.

Table 1. Results obtained for 5 iterations of 5 randomly selected recordings over 1000 recordings with SOM Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	4	1	5	%80
2. Iteration	3	2	5	%60
3. Iteration	3	2	5	%60
4. Iteration	5	0	5	%100
5. Iteration	4	1	5	%80

Table 2. Results obtained for 5 iterations of 10 randomly selected recordings over 1000 recordings with SOM Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	8	2	10	%80
2. Iteration	6	4	10	%60
3. Iteration	7	3	10	%70
4. Iteration	10	0	10	%100
5. Iteration	6	4	10	%60

Table 3. Results obtained for 5 iterations of 20 randomly selected recordings over 1000 recordings with SOM Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	17	3	20	%85
2. Iteration	10	10	20	%50
3. Iteration	15	5	20	%75
4. Iteration	13	7	20	%65
5. Iteration	16	4	20	%80

Table 4. Results obtained for 5 iterations of 5 randomly selected recordings over 1000 recordings with ANN Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	5	0	5	%100
2. Iteration	4	1	5	%80
3. Iteration	4	1	5	%80
4. Iteration	3	2	5	%60
5. Iteration	3	2	5	%60

Table 5. Results obtained for 5 iterations of 10 randomly selected recordings over 1000 recordings with ANN Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	6	4	10	%60
2. Iteration	6	4	10	%60
3. Iteration	7	3	10	%70
4. Iteration	8	2	10	%80
5. Iteration	8	2	10	%80

Table 6. Results obtained for 5 iterations of 20 randomly selected recordings over 1000 recordings with ANN Algorithm

Iterations	Accurate Number	Incorrect Number	Total Count	Success Rate
1.Iteration	17	3	20	%85
2. Iteration	16	4	20	%80
3. Iteration	13	7	20	%65
4. Iteration	16	4	20	%80
5. Iteration	15	5	20	%75

Results

The results obtained in the study are shown in Table 1-6. According to the results, the success results of randomly selected records in a set of 1000 records are shown in percent (%) in Table 7. In the light of the data obtained, it was observed that SOM algorithm gave very close results according to ANN algorithm for real time and the success rate of both algorithms was approximately 74%.

Table 7. Successful results according to randomly selected record numbers of 1000 records on algorithms

	Success results by number of records			
Learning Algorithms	Success rate according to 5 records	Success rate according to 10 records	Success rate according to 20 records	Average Success Rate
SOM	%76	%74	%72	%74
ANN	%76	%70	%77	%74

Table 8. Results of 5 iteration run times based on randomly selected record numbers of algorithms on 1000 records

	Work-Time results by number of records			
Learning Algorithms	Work-Time (second) according to 5 records	Work-time (second) according to 10 records	Work-time (second) according to 15 records	Work-time (second) according to 20 records
SOM	2	4	5	7
ANN	170	325	472	610

If these algorithms are to be compared as the run times, the run time of the SOM algorithm is much shorter than the ANN algorithm. In case of use of the SOM algorithm in the studies that have time constraints on in-line data, it is determined that the result of the lending will result in a shorter time compared to ANN algorithm. Table 8 shows the table of working time.

In future studies, more records and different kinds of test data will be examined and the performance and success rates of the algorithms will be tested. In addition, SVM, Navie Bayes and Dynamic Bayesian algorithms will be applied to the online learning set and the results will be compared with the results of SOM and ANN algorithms.

References

[1] Murray J. Mackinnon ve Ned Glick (1999), *'Data Mining and Knowledge Discovery in Databases- An Overview', J.Statists.* (Vol.41). No.3, s.260.
[2] A.Kusiak, K.H. Kernstine, J.A.Kern (2000). *K.A.McLaughlin and T.L.Tseng: Medical and Engineering Case Studies*.
[3] Yan, L. and Miller, J. and Mozer, M. and Wolniewicz R (2001). *Improving Prediction of Customer Behaviour in Nonstationary Environments, Proceedings of International Joint Conferance on Neural Networks*. 2258-2263.
[4] Rud, O.P. (2001). *Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management*, John Wiley & Sons, Inc. Canada.
[5] Özkan, M. & Boran, L. (2014). *Veri Madenciliğinin Finansal Kararlarda Kullanımı. Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 4(1). 59-82.
[6] Kohonen, Teuvo; Honkela, Timo (2007). *"Kohonen Network". Scholarpedia*
[7] Ultsch, Alfred (2007). *"Emergence in Self-Organizing Feature Maps". In Ritter, H.; Haschke, R. Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM '07)*. Bielefeld, Germany: Neuroinformatics Group. ISBN 978-3-00-022473-7.
[8] Liu, Yonggang; Weisberg, Robert H.; and Mooers, Christopher N. K. (2006). *Performance Evaluation of the Self-Organizing Map for Feature Extraction, Journal of Geophysical Research*, 111, C05018, doi:10.1029/2005jc003117.
[9] Liu, Y.,and R.H. Weisberg (2011). *A review of self-organizing map applications in meteorology and oceanography. In: Self-Organizing Maps-Applications and Novel Algorithm Design*, 253-272.
[10] A.A. Akinduko, E.M. Mirkes, A.N. Gorban (2016). *SOM: Stochastic initialization versus principal components, Information Sciences*. http://dx.doi.org/10.1016/j.ins.2015.10.013
[11] Zheng, G. and Vaishnavi, V. (2011*). "A Multidimensional Perceptual Map Approach to Project Prioritization and Selection," AIS Transactions on Human-Computer Interaction.* (3) 2, pp. 82-103
[12] Saadatdoost, Robab, Alex Tze Hiang Sim, and Jafarkarimi, Hosein (2011). *"Application of self organizing map for knowledge discovery based in higher education data." Research and Innovation in Information Systems (ICRIIS), 2011 International Conference on. IEEE*.
[13] D. C. Ciresan, U. Meier, J. Masci, J. Schmidhuber (2012). *Multi-Column Deep Neural Network for Traffic Sign Classification. Neural Networks*.
[14] D. C. Ciresan, U. Meier, J. Schmidhuber (2012). *Multi-column Deep Neural Networks for Image Classification. IEEE Conf. on Computer Vision and Pattern Recognition CVPR.*
[15] Hentrich, Michael (2015). *"Methodology and Coronary Artery Disease Cure".*
[16] Tahmasebi; Hezarkhani (2012). *"A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation". Computers & Geosciences* 42: 18–27. doi:10.1016/j.cageo.2012.02.004.
[17] M. Forouzanfar; H. R. Dajani; V. Z. Groza; M. Bolic & S. Rajan (2010). *Comparison of Feed-Forward Neural Network Training Algorithms for Oscillometric Blood Pressure Estimation (PDF). 4th Int. Workshop Soft Computing Applications. Arad, Romania: IEEE*.
[18] Tunca C. Toplana E. Işıklab S (2010). *Yapay Sinir Ağları ile WiFi Tabanlı İç Mekan Konumlandırma*
[19] Dunham Margaret H. (2003). *Data Mining Introductory and Advanced Topics, Prentice Hall, USA*
[20] Efe, Ö., Kaynak, O (2000). *Yapay Sinir Ağları ve Uygulamaları, Boğaziçi Üniversitesi*.
[21] Cios, K.J., Pedrycz, W., Swiniarski, R.W. ve Kurgan, L.A (2007). *Data Mining: A Knowledge Discovery Approach. USA: Springer Science Business Media, LLC*.

Use Case Study: Data Science Application for Microsoft Malware Prediction Competition on Kaggle

Çayır[1], U. Ünal[1], I. Yenidoğan[1], H. Dağ[1]

[1]*Kadir Has University, Istanbul, Turkey,(aykut.cayir, ugur.unal, hasan.dag, isil.yenidogan)@khas.edu.tr,*

Abstract

Malware prediction is the most prominent area of cybersecurity domain. Malware predictionapplications are leaned to be empowered by machine learning due to rapidly emergingintrusion attacks. In this perspective, defense systems aim to conjoin data science and cybersecurity. There are many platforms which provide public datasets and organize competitions for sector specific problems. For instance, Netflix has organized a competition with $1 Million prize to develop a new recommendation system. Kaggle organizes many competitions for research, business and educational purposes. Microsoft has sponsored two important malware prediction competitions on Kaggle in 2015 and 2019. In this paper, we present our solution as a use case which is placed at 5th among 2,426 teams on Microsoft Malware Prediction 2019 dataset.

Introduction

Cybersecurity is one of the significant example of domains that make use of widespread machine and deep learning methods. Malware prediction problem can be divided into two basic parts. Predicting whether a software is a malware or not and type of malware within a given malicious software. Recently, competitions point out applicability of these methods on malware prediction [1]. Preparing a data science competition via allowing usage of company-owned datasets is certainly preferable way to do. For instance, Netflix announced to award $1M to a team in 2009 for a recommendation model that was better than company's recommendation engine [2]. This is a showpiece of data science competition to help improve business objectives. Similarly, Kaggle is the biggest and a well known competition platform that incorporates businesses with research [3, 4].

Microsoft has sponsored for two data science competitions in cybersecurity domain at Kaggle platform. The first competition titled Microsoft Malware Challenge 2015 was announced in 2015. The second competition titled Microsoft Malware Prediction has been completed in 2019. In this paper, our solution, which is placed at 5th among 2,426 teams, is presented for Microsoft Malware Prediction competition. This paper, as a use case of a data science application in cybersecurity domain, aims to show that a general data science pipeline might be more useful than many complex models.

This paper is organized as follows. Our solution is described as a pipeline, which can be applied to any data science competition, in Section 2 and Section 3 concludes the paper with competition results.

Pipeline for a Malware Prediction Competition

Microsoft Malware Prediction dataset has a target column titled "HasDetections" and the target is binary. There are 83 columns in the original training set and the training set has approximately 8.9M samples. The original test set has 82 columns and approximately 7.8M samples. Machine identifier column has been dropped because the column is not informative for any model. The training set is balanced in terms of class distribution.

Proposed pipeline is applicable for other data science problems in different domains. General pipeline contains three main parts such as model selection, feature engineering, Bayesian hyper-parameter tuning and submission creation as shown in Fig. 2.

The first part of the general pipeline is feature engineering (Fig. 3a). In the feature engineering, first of all, columns have been dropped if they have missing values greater than or equal 70%. If the missing value percentage is less than 70% then we can apply missing value imputation by feature type. If the feature type is categorical then the missing values are filled by mode value of the column, otherwise the missing values are imputed by mean. The second part of the feature engineering is categorical feature engineering. One of the most important categorical feature encoding methods is one-hot encoding. However, the dataset has a lot of categorical features, which have high cardinality. High cardinality increases the number of dimensions when one-hot encoding is applied. To avoid curse of dimensionality, we have used target and count-frequency encoding. Target encoding has caused overfitting, so we

have preferred to use count-frequency. At the end of the early test steps, we have noticed that making the model robust against overfitting, requires applying rare encoding with count-frequency method. Rare encoding makes a group called ”rare” from categorical features occurring less than a specific threshold value. In our case the threshold is 0.005. For the feature engineering part, we have used FeatEngine tool [5]. The last feature engineering step is to select top-k features using LOFO [6]. LOFO uses tree based models to calculate feature importance. Fig. 1 shows importance values after creating feature matrix. From the results, ”SmartScreen” can be seen as the most important feature. Top 30 features were selected to be used in our model.

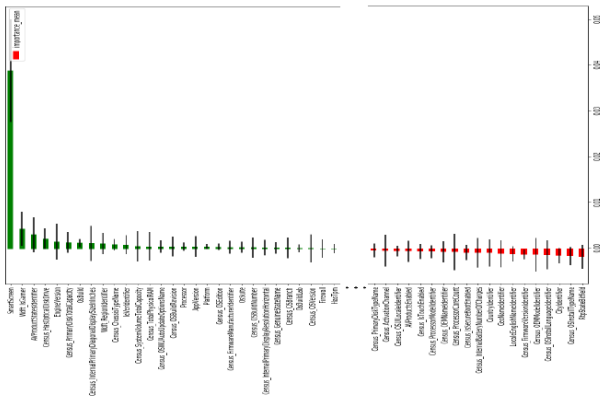


Figure 1: Feature Importance by LOFO

The second part of the general pipeline is model selection (Fig. 3b). The task type is a binary classification. Therefore, we have used received operating curve - area under curve evaluation metric. This metric has been also used by Kaggle to evaluate submissions for the competition. For this competition, we focused on tree based models. Especially, gradient boosting trees are

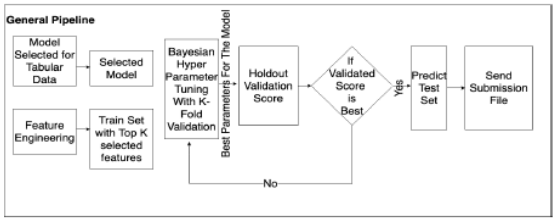
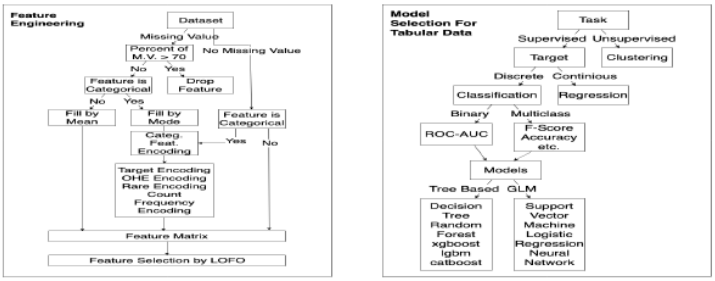


Figure 2: Proposed General Pipeline

useful for tabular data because they can consider feature interaction automatically. We have tried XGBoost [7], Lgbm [8] and Catboost [9]. We have noticed that all models have been highly overfitting data except XGBoost. Thus, XGBoost algorithm is selected as a predictive model for the competition.



(a) Feature Engineering for the Dataset (b) Model Selection for the Task

Figure 3: Details of Feature Engineering and Model Selection Parts

Third part of the our proposed pipeline is to tune hyper-parameters of the predictive algorithm.Bergstra et al. [10] propose a Bayesian optimization framework. We define a hyperparameter list and specify the upper and lower values of these parameters. According to third part of the general pipeline, if the ROC-AUC score of holdout validation set is worse than the previous score, we tweak the values of ranges otherwise we create a submission file using the original test set as shown in Fig. 2. Hyper-parameter tuning phase takes approximately 12 hours in this pipeline. Kaggle uses public and private leaderboards for submission evaluation. This approach aims to show whether the model is overfit or not. For this competition, 37% of the original test dataset is dedicated to calculate private leaderboard score and the rank is the final result for the competition. Tables 1 and 2 show public and private

leaderboard ranks in the competition respectively. According to result tables, our public and private scores are close to each other. This closeness shows that our model generalizes well in range of our scores.

Conclusion

This paper introduces our solution for Microsoft Malware Prediction data science competition in Kaggle. We propose a simple general pipeline, which is placed at 5th in the competition.

Table 1: Public Leaderboard

Rank	Team Name	ROC-AUC
1	Sashimi P.	0.7144
2	APTX4869 P.	0.7116
3	Tofu P.	0.7113
4	J. Serrano P.	0.7098
1539	khas_ccip	0.6781

Table 2: Private Leaderboard

Rank	Team Name	ROC-AUC
1	abuurista	0.6758
2	Confiniti	0.6653
3	ken10ML	0.6652
4	John DiMarco	0.6647
5	khas_ccip	0.6640

This pipeline generates the simplest model with 30 features and a single fine tuned XGBoost predictive model. The pipeline has three main parts such as model selection, general feature engineering and hyper-parameter tuning. For these steps, well known libraries have been used effectively. The most important of the proposed model is to make the predictive model robust to overfitting. These types of competitions with real business data researchers aim to test their algorithms and validate these results against those around world. Hence, it is a good way to ensure the validity of test results.

References

[1] Royi Ronen, Marian Radu, Corina Feuerstein, Elad Yom-Tov, and Mansour Ahmadi. Microsoft malware classification challenge. arXiv preprint arXiv:1802.10135, 2018.

[2] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS), 6(4):13, 2016.

[3] Kaggle: Your home for data science. https://www.kaggle.com/. (Accessed on 05/03/2019).

[4] Souhaib Ben Taieb and Rob J Hyndman. A gradient boosting approach to the kaggle load forecasting competition. International journal of forecasting, 30(2):382–394, 2014.

[5] Feature engine. https://feature-engine.readthedocs.io/en/latest/. Accessed: 2019-03-21.

[6] aerdem4/lofo-importance: Leave one feature out importance. https://github.com/ aerdem4/lofo-importance, Feb 2019. (Accessed on 05/03/2019).

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM, 2016.

[8] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, pages 3146–3154, 2017.

[9] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems, pages 6638–6648, 2018.

[10] James Bergstra, Dan Yamins, and David D Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in science conference, pages 13–20. Citeseer, 2013.

Comparison of Feature Selection Methods for Detection of Alert Sounds in Traffic

C. Akyürek Anacur[1], R. Saraçoğlu[1]

[1]Van Yüzüncü Yıl University, Van, Turkey, cansuakyurekanacur@gmail.com, ridvansaracoglu@yyu.edu.tr

Abstract

People with hearing loss may have difficulty hearing the sound of the siren and alert signals while driving in traffic. In this study, it is aimed to distinguish the warning sounds of the vehicles such as police, fire brigade and ambulance from other voices easily and with high accuracy in order to facilitate the use of these people in traffic. For this reason, it is necessary to extract the features of the sounds in traffic and to reduce the number of features. Feature selection was used for this dimension reduction process, and various feature selection algorithms were compared and tested. By this way, the rate of recognition of the sounds is increased. This study is important for comparing the methods of feature selection on a real-world problem.

Introduction

According to Turkey Health Survey conducted by the Turkish Statistical Institute (TUIK) in 2016, the 12.9% of the population aged between 15 and 64 is composed of people with hearing loss [1]. In order to increase the number of people with hearing loss in traffic and to connect more people with this impairment to the life, the Directorates of Traffic Registration and Auditing that are parts of Provincial Police Departments started Sign Language and Traffic Training programs for people with hearing loss in 2016-2017 academic year [2]. People with hearing loss may be late making way for vehicles with pass priority such as police, fire brigade and ambulance vehicles while driving due to that they cannot hear sirens and warning signals. Especially for ambulance even one minute is vital. Thus, it is aimed to give visual warnings to the people with hearing loss in traffic to inform them about the direction that police, fire brigade and ambulance vehicles come from by detecting and distinguishing the siren sound from other sounds to help them drive more comfortably. In this study, feature selection methods are tested to support the developments for detection of the sound to support people with hearing loss to drive more comfortably in traffic during the implementation of such systems by selecting the most meaningful and the most representative values from the whole data to recognize the sound signal. Feature selection methods are compared on their effectiveness in the classification of sound.

Literature Review

Extraction of unnecessary and non-data attributes in feature selection data sets is a strategy that has been applied frequently by many researchers as it increases the proportion of meaningful data and success rate by reducing size of data in detection and classification systems. Many researchers have used feature selection methods in their recognition systems to reach the most recognizable data.

Kaynar et al. compared the success rate of non-size-reduced data with size-reduced data in feature selection methods by using numerous feature selection algorithms together with various classification algorithms for intrusion detection [3]. Eskidere experimented with six different feature selection methods in a data set constituted by biomedical sound measurements for diagnosis of Parkinson's disease. As a result of this comparison, the SVM-RFE quality selection method for the Parkinson's patient data set was found to be the feature selection method which gave the best recognition result with the correct classification rate of 95.13%. [4]. In his study, Yakut proposed a method of classifying heartbeats with automatic arrhythmias. For the classification of arrhythmias in the ECG signal, he formed the most meaningful subsets by doing feature selection after the feature detection with use of four different feature extraction methods. The classification method proposed by use of these subsets and known classification methods were compared after they were classified. The classification was more accurate with the feature selection methods which selected the most representative data [5]. Vogt et al. worked on the recognition of emotions online from the voice. To be able to extract the emotional state of the online user from the voice information; first they extracted the feature vectors of the sound signal and then classified the data after creating ideal data set by feature selecting the data that best represent the data in this extracted data [6].

Material and Method

For a voice recognition system to recognize the sound, the audio signal must be correctly expressed. Some pre-treatments have been applied to enable the data to be feature selected. The following part describes the structure of the system from raw data collection to recognition:

In this system, the warning sounds of the vehicles with the priority of traffic in the traffic are considered as positive sounds, while the other sounds recorded outside these sounds are considered as negative voices. Sounds, which are in ".wav" format on the computer, are called from the coding environment and transforms into a matrix with numerical values.

Digitized audio data gets subject to feature extraction operation. Feature extraction is one of the important steps of the recognition system because it affects the performance of the classification [7]. In this study, feature vectors are extracted by using Linear Predictive Coding (LPC) method as a feature extraction method. The steps of feature extraction are shown in Figure 1.

Figure 1. Feature Extraction Steps [7]

After feature extraction process, feature selection is applied to ensure high accuracy before classification. Feature selection methods are the methods by which some features are eliminated and the remaining ones are used without modification [3]. The feature selection methods that are applied to the data set respectively are: Fast Feature Selection Approach Based on Extreme Learning Machine and Coefficient of Variation, ReliefF, Fisher Score, Infinite Latent Feature Selection (ILFS), Feature Selection Via Concave Minimization (FSV), Recursive Feature Elimination (RFE), and Least Absolute Shrinkage and Selection Operator (LASSO). The data set is classified after the determined feature selection method is applied to the data set and the features that represents the data most are selected. Support Vector Machine (SVM) are used to make classification.

Application (Findings)

Each two-second sound converted into a numerical data transformed into an 88200x1 matrix. Each audio signal that was expressed by the value of 88200 was transformed into a single-line 1x5500 matrix, which can be expressed by 5500 values after feature extraction. The data set, which consisted of a total of 94 sounds, finally became a matrix of 94x5500. Then, a total of 7 different feature selection methods have been tried. They were then classified by SVM and a performance comparison was made.

First, "Fast Feature Selection Approach Based on Extreme Learning Machine and Coefficient of Variation" was tried. With this method, feature ranking was done and 5 column data from the data set were selected as the best features. Finally, the 94x5500 matrix obtained as a result of feature extraction has been transformed into a 94x5 matrix. Classification success rate was 72.34% in this feature selection method when classified with SVM. The data were then subjected to the ReliefF feature selection method. The data set, expressed by a matrix of 94x5500 by feature extraction, has been transformed into a 94x1006 matrix as a result of the ReliefF feature selection method. Then it was subjected to classification with SVM after the application of this feature selection method and the selection of the values that represented the data best. As a result of this classification, success rate was 78,72%. As the third method, Fisher Score feature selection method was tested. With the application of the Fisher Score method, the 94x5500 matrix obtained as a result of feature extraction was transformed into a 94x1000 matrix. When the features selected by this feature selection method were classified by SVM, a success of 56.38% was observed. Then, the feature selection was carried out by ILFS method. The sound signal, which was expressed with the matrix of 94x5500 after feature extraction, was transformed into a 94x1006 matrix after the feature selection process with the ILFS method. After feature selection was completed, the rate of success in the data set classified by SVM was 77.65%. Then, the feature selection process was carried out by FSV method. The feature vectors, which were expressed with a matrix of 94x5500 after feature extraction, have been transformed into a 94x1006 matrix after the selection of the most significant features. After feature selection was completed, it was classified with SVM. The classification success of the data set resulted from this feature selection method was 78,72%. As the sixth method, RFE was used to select features in the data set. The sounds that were taken in the computer environment and digitized had been expressed with a matrix of 94x5500 after feature extraction process, and this value transformed into a 94x1011 matrix after the RFE method was applied. As with the other methods, classification was made with SVM

after the selection of features and the success rate was found to be 53.19%. Lastly, feature selection was carried out with LASSO method. Before the LASSO feature selection method was applied, the feature vectors had been expressed with a matrix of 94x5500, and after the feature selection process, it turned into a 94x1000 matrix. When the data set was classified with SVM, the success rate was found to be 79.78%.

Conclusion

As a result of the study, feature selection algorithms are compared for the classification of the warning sounds of vehicles with pass priority such as police, fire brigade and ambulance vehicles and the impacts of these algorithms on the success of the classification is determined.

After the characteristics of the received audio signal have been transformed into the vectors, seven different feature selection methods were compared on their success in the data set which included the audio signals in the traffic. As a result of these operations, the most successful feature selection method for the data set containing the warning sounds in the traffic was found to be the LASSO with a success rate of 79.78%; and the least successful feature selection method was RFE with a success rate of 53.19%.

As a result, the success rates of these tested feature selection methods for the data set consisting of warning sounds in traffic have been demonstrated.

References

[1] TÜİK, 2016. *http://www.tuik.gov.tr/PreTablo.do?alt_id=1017*. Turkish Statistical Institute, Ankara. Date of access: 30.04.2019.

[2] Harmancı, S., (2016). *http://www.hurriyet.com.tr/isitme-engelli-ogrencilere-isaret-diliyle-trafi-40274838*. DHA, Amasya. Date of access: 30.04.2019.

[3] Kaynar, O., Arslan, H., Görmez, Y. and Işık, Y. E., (2018). Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti. *Journal of Information Technologies, 11*(2), 175-185.

[4] Eskidere, Ö., (2012). A Comparison Of Feature Selection Methods For Diagnosis Of Parkinson's Disease From Vocal Measurements. *Sigma*, 30, 402-414.

[5] Yakut, Ö., (2018*). Ekg İşaretindeki Aritmilerin Yumuşak Hesaplama Algoritmaları Kullanılarak Sınıflandırılması* (PhD Thesis). KÜ, Institute of Science, Kocaeli.

[6] Vogt, T., André, E., and Bee, N., (2008). EmoVoice—A framework for online recognition of emotions from voice. *In International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems* (pp. 188-199). Springer, Berlin, Heidelberg.

[7] Uğuz, H., Arslan, A., (2010). A new approach based on discrete hidden Markov model using Rocchio algorithm for the diagnosis of the brain diseases. *Digital Signal Processing*, 3(20), 923-934.

Mobile Application Price Estimation Using Deep Learning Techniques

Ö. Arslan[1], M. Yıldırım[1], S. A. Uymaz[1]

[3]*Konya Technical University, Konya, Turkey, omer_arslan@kuveytturk.com.tr, mehmet-yildirim@kuveytturk.com.tr, sauymaz@ktun.edu.tr*

Abstract

Today, mobile applications have become an inseparable part of the people. They can reach their phones whenever they want and so that people started to satisfy all their needs using phones. So, this has increased the popularity of mobile applications. Correspondingly, count of the mobile application developers increased and thousands of applications took place in the market. The main aim of the developers while doing this is, arouse the interest of customers and generate an income. The proposed application aims to give a piece advice about the state of the current application by reviewing all other applications. So, the system will be trained with the information about the applications in AppStore. After training, the system will ask information as above about current application. The system will give advice whether the application should be paid or not as a result.

Introduction

The Last step of the application development process is sending the application to the application store and let customers reach applications. As much as the development process, appraising and deployment of the application is also a hard process. Because the influence of the pricing strategy is varied customer to customer and serious pricing strategy is needed for better marketing. An expensive application may lose some customers but with a correct strategy, it also can reach it's target customers. When it comes to free applications, they can be separated into two groups as free and paid. Because today "In-App Purchases" term has emerged and thus event though come applications can be downloaded free, after downloading they may request some prices in behalf of some in-app services.

In aggregate, there are some aims of the applications when they are paid, completely free or has in-app purchases Some of these aims are; generating an income, publishing advertisement via application and gathering customer data with application.

Generally, paid applications are the applications that received appreciation with previous applications by developers of companies. Applications that offering in app purchases are not requesting price for whole application. They aims to receive appreciation with a portion of the application and then aims to request price for the remain part of the application. Completely free applications don't request price from customers directly but they aim to introduce themselves by their free application and then request price for their future applications. Even though these applications are completely free, still they can generate an income to their developers by the advertisements that published via certain application. Besides, if we take into account that are free aplications are downloaded more than paid applications, we can conclude that free applications are more beneficial factor for he Big Data field because, by them, developers or companies can generate vast customer data.

Effect of the free and paid applications on the customers was already discussed before. In this case, another point we should make is "Most Favored Customer Clause" that also known as MFC. This clause, predicts a regulation on the pricing strategy of the company and guarantees that the prices that given by the company to its customer are the best prices among the prices that given to other customers by the same company[1]. Here Apple and Amazon companies can be the examples. Because these companies are following the MFC rules hence they guarantee that price of the books and journals are not more expensive than books and journals that offered from similar vendors. Applying these rules to the platforms lets customer reach products with a balanced pricing policy[2]. Applying these rules to mobile application market may give confidence to the customers. Because customers are perplexed because of the variable prices offered from different developers and companies.

The Proposed application aims to give a new impulse to pricing strategies that aforementioned. Thus, while predicting the price of the new application, firstly similar applications will be compared and a more fair price scale will be designated.

Artificial Neural Network

Artificial neural network is a network that accepts the brain of the human as a model. Shortly, this model produces an output in behalf of values that were sent to input layer. Then by propagation of the output layer error on the weights, network stars to be able to generalization. As shown in the figure 1; we can roughly say that a basic neural network is consisted from five elements that are inputs, weights, sum function, activation function and output neurons determined by the user to present their own class label[3].

Artificial neural networks were inspired by single neurons. Every neurons are connected to the neuron in the other layer by a weighted connected. Neurons are also known as process elements. Actually, single neurons can also make simple calculations by themselves but real power of the neural networks comes from the weighted connection structure of the neuron. Established neural network architectures can rarely involve hundreds or thousands of neurons while human brain is consisted from about 100 million neurons. As shown in figure 2, after the producting input values with their weights, they sent to sum function. Activation function is consisted from weighted inputs. Sigmoid function is commonly used as a transfer function[4].

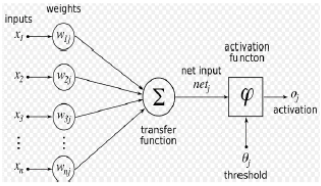


Figure 1: Basic Artificial Neural Network Architecture

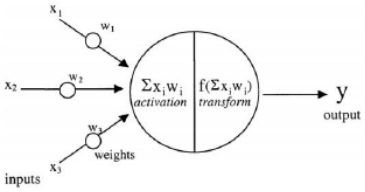


Figure 2: Activation and Transfer Functions

Maybe the best known feuture weighting method is the multi layer perceptron rule. By this way, based on the error of the output, some changes made on the weights depending on the thresold value. For backpropagation and linear elements, least squares mean algorithm, makes updates on the weights and reduces the error then increases the generalization ability of the network[5].

We have already discussed that first layer of the neural networks is the input layer. Before sending values to the input layer, a proper dataset should be generated. After the generation of the dataset, next step is seperating the dataset as test and training data. Because network is trained by training dataset and tested with test data. Hence the accuracy of the network is measured by using the test dataset. To separate dataset into two groups, two methods commonly being used. First method is called as Hold-Out method and by this method, dataset is separated into training and test dataset by a rate that determined by the user. Commonly used rate is 70% and this means 70% portion of the dataset will be used as training set and remained portion will be used as test set. This method eases the process of selecting training and test dataset but it also has a disadvantage that this method chooses test and training dataset randomly. This randomly selection brings some unwanted results. By this way, same data can be used in both training and test steps and this may affect the accuracy negatively. To overcome these unwanted results, cross validation method being used.

Cross Validation

This method defends that selection of the test and training dataset should be done by following some rules. There are three types of this method determined in the literature. These types are; Leave One Out, K-Fold Cross Validation and Monte Carlo Cross Validation method[6]. Leave one out method runs in loops as much as count of dataset element. During every loop, one element chose as test data and remained elements chose as training data.
Multiple validation method or by known name K-Fold Cross Validation method separates dataset into K parts. K is determined by the user and after the determination of the K value, as Leave One Out method, algorithm runs in K loops. In every loop one group chose as test data and remained groups chose as training data[7]. Schematic working of the K-Fold Cross Validation was shown in the Figure 3.

K-Fold Cross Validation method separates dataset into K parts, and as it might be expected, it takes more time to be completed. But as a result of controlled selection process of test and training dataset, it has more generalization ability and accuracy than randomly selection methods. Thus this naturally increases the classification accuracy[8]. By the raising of the K value in the Cross Validation method, calculation time will naturally increased and hence calculation cost will be also increased. Recommended K values are 5 and 10. Because by selecting these two values, bias will be less than K=2 value[9]. In this direction, a schema of the K=3 Cross Validation architecture was given in the Figure 4.

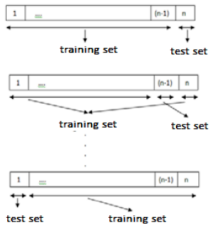


Figure 3: Schematic View Of K-Fold Cross Validation

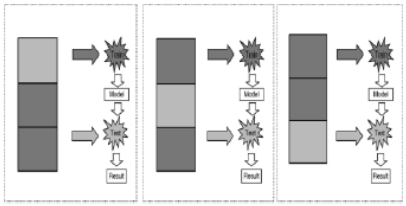


Figure 4: Cross Validation Schema For K=3 Value

Normalization

It is also an important step to put every feature into same form. Some features in the dataset range between 0 and 100 while some features range between 0 and 1. This produces unwanted results during updating the weights. Normalization is a data preprocessing technique that depending on mapping old values into recently determined range. Some of commonly used normalization methods are Min-Max and Z-Score normalization[10]. Normalization process also enhances the performance of the network. A scientific study on diabetic patients made for predicting waiting times of the patients in the queue. As the result of the this scientific study, Min-Max normalization was chosen as the most successful normalization type. Here most important condition to be a successful method is, predicting the truth more than other methods[11].

Min-Max Normalization

After the finding of against values in the dataset, by using Min-Max normalization, all values can be mapped into 0 and 1 range. In the beginning of the this method, biggest and smallest values of the every feature were found. Then as shown in the figüre 5, Min-Max normalization was applied[12].

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure 5: Applying Of Min-Max Normalization

Here X* value represents the value that produced after the normalization method. X is the value before the calculation. Xmax is the biggest value in the feature map and Xmin is the smallest value in the feature map.

K-Means Clustering Algorithm

K – Means is a commonly used clustering algorithm to cluster values into proper clusters. This algorithm is an unsupervised algorithm and clusters values by their properties and distances to each other. Basic aim of this algorithm is clustering close values into same cluster and clustering far values into different clusters. Here close and far mean the amount of the difference between values. K value within this algorithm is determined by the user and it designates the count of the clusters that our data will be clustered. Hence, determined K values represent the centers of the each clusters. Distances between the features are commonly calculated using Euclidean distance as shown in the figure 6[13].

$$distance(x, y) = \left\{ \sum_{i=1}^{n} (x_i - y_i) \right\}^{1/2}$$

Figure 6 : Distance Calculation In The K-Means Algorithm

As most of other algorithms, K-Means algorithm also has some disadvantages. Major deficient of this algorithm is selection method of the K value. Finding proper value of the K presents challenge for users. Because, to find proper value for K, by trial and error method, many K values should be tried[14]. Placement of the K values in the beginning of the algorithm is a randomly proceeded step[15].

Naive Bayes Algorithm

A clustering algorithm named Naive Bayes is not an iterative algorithm and thus it is more simple and presents a better working than other complicated algorithms[16]. A and B and random probabilities and when the probability of the B is known, probability occurrence of the A is represented as P(A\B) and calculated as shown in the figure 7. Here P(A) value represents the independent probabilty of the event A while P(B) represents the independent probability of the event B. Finally P(B\A) represents the probability of the event B when its known that event A occured.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Figure 7: Calculation Of The Bayes Theorem

Application Versioning

After the updates that made on the application, to inform customers about the recent version, a version number is given to the application. Semantic Versioning is the commonly used versioning method. This versioning method is consisted from three parts. First part is Major, second part is Minor and last part is called as Patch part. When unconformable Api changes were made, Major part is increased, when new function was added to application then Minor part is increased. Patch part is about bugs. When bugs were fixed then Patch part is increased[17]. When this process comes complex to the developers, they can also use dual versioning type that contains only Major and Minor parts[18]. Apart from these, Github platform can also be used for this versioning process. By this way, a new changeset id is given to each changes, thus, this can ease the following of the code changes[19].

Material And Method

Proposed application aims to make a connection between its price and other features. Here, a dataset that contains the some features of the old application was used as aforementioned. Proposed application roughly contains following steps; choosing proper features from raw dataset, a prerocessing step for features to make them more suitable for training, clustering applications by their prices and finally mapping application to a proper class by its trained features.

Purpose

There are two core purposes of the application. First purpose is determining whether application should be free or paid. Another purpose is predicting the price range of paid applications. In this step only paid application dataset were used to partly overcome the imbalanced dataset problem. Because, free applications were much more than paid applications.

Clustering Application As Free Or Paid

To cluster the applications as free or paid, firstly little changes were made on the dataset. Dataset was separated into three groups. First group was consisted from all applications raw data. Second dataset contained same number of free and paid applications. Last dataset was consisted from all applications except against values.
To determine the proper prices for every group, K-Means algorithm was used. Firstly a clustering transaction was made on the dataset that contains all applications and result of the clustering was shown in the Figure 8.

As it can be seen in the figure 9, most of the applications were accumulated between 0 and 100 values. Thusi values that bigger than 100 were accepted as against values and they were removed from dataset. After that, a new clustering was made and result of the new clustering was shown in the figure 9.

Difference of the element count of the each groups has a negative effect on the training and this shortly known as imbalanced data. Finally a dataset that contains same number of elements from each classes was created and then a new clustering process was done on this dataset. Result of the K-Means clustering on this dataset was shown in the figure 10.

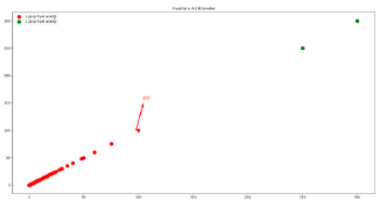


Figure 8: Price Clustering On The All Applications DataSet

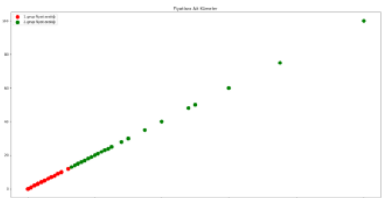


Figure 9: Clustering Except Against Values

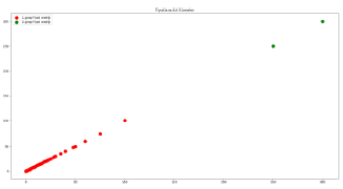


Figure 10: Clustering On The Dataset That Contains Same Element Count From Each Classes

Predicting Price Range Of The Paid Applications

There was a serious count difference between the count of the free and paid applications in the dataset. Thus, to determine the prices of the paid applications, only paid applications were used as the dataset. To measure better

price ranges, dataset were separated into 2, 3, 4 and 5 groups. These ranges also found via K-Means algorithm. For instance, after the removing of the against values, a clustering was done on the 2-Parts group. Clustering result was shown in the figure 11.

Here that can be seen that graphic is consisted from two colors and every color is representing one price group. In this group, 12.5 was used as distinction point and it also found with K-Means algorithm. So, application will predict that certain application should be more expensive or less expensive than determined price as result.
Part counts of the every group and determined prices that separated groups are shown in the figure 12.

	1st Range(USD)	2nd Range(USD)	3rd Range(USD)	4th Range(USD)
2 Parts	12.5			
3 Parts	4.5	16.5		
4 Parts	3.5	8.5	18.5	
5 Parts	3.5	8.3	17.5	32.5

Figure 12: Price Ranges Based On Part Counts

Based on the figure shown above, we can make following interpretations for the 4-Parts group. If the certain application is less expensive or equals to 3.5$ then it will be mapped to the first class. And also if it more expensive than 8.5$ and less than 18.5$ then it will be mapped to the third class and finally if it more expensive than 18.5% then it will be mapped to the fourth class. K-Means algorithm was run on this 4-Part group was shown in the figure 13.

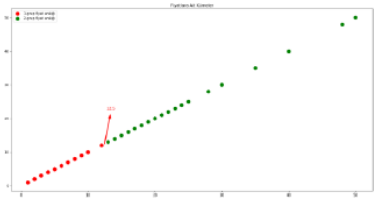


Figure 11: Seperating Paid Applications Into 2 Parts

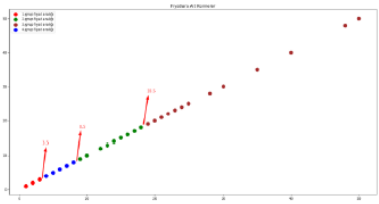


Figure 13: Partitioning Paid Applications Into 4 Groups

Results And Discussion

By the proposed project, artificial neural network and naive bayes algorithms were compared on the mobile application dataset. Only Weka framework was used to run naive bayes algorithm while artificial neural network was run on two softwares. Artificial neural network architecture was run on both Weka and the neural network that developed by us. Developed artificial neural network used feature count as the input neuron count. And class counts of each datasets were used as output neuron count. For instance, 4 output neurons were used for 4-Parts dataset. Because this dataset has 4 subclasses inside.

In our developed network, we used single layer hidden neurons and our hidden neuron count was 25. Network was run for 1000 epochs and momentum factor was determined as 0.3. Finally our bias and learning rate factors were chose as 1 and 0.1 respectively. Every feature firstly sent to normalization process then used. As aforementioned, Min-Max normalization was used during the tests. We also used versions of the applications to make a connection between version numbers and the development of the application. To compare neural network and naive bayes algorithm, an application was developed and developed application was shown in the figure 14.

Figure 14: Application That Developed On .Net

Developed neural network, neural network and naive bayes algorithm that come from Weka were tested via developed application. Comparison of the neural network of the Weka and our neural network was shown in the figure 15. In the test results that shown above, Hold-Out method was used. Neural network was also run by Cross Validation method and K value was chosen as 4. Results of the 4-Fold Cross Validation on the dataset was shown in the figure 16. Finally, naive bayes algorithm based on the Weka software was tested on our mobile application dataset. Results of the naive bayes algorithm on our datasets were shown in the figure 17.

App Type	DataSet Info	Element Count	Weka (ANN)	Developed (ANN)
All values	All values	7195	59%	57%
All values	Except against values	7199	56%	60%
All values	Groups are equal	6282	52%	50%
Paid Applications	2 Parts	3142	97%	96%
Paid Applications	3 Parts	3142	77%	78%
Paid Applications	4 Parts	3142	70%	68%
Paid Applications	5 Parts	3142	68%	70%

Figure 15: Comparison Of The Neural Network Of The Weka And Developed Neural Network

App Type	DataSet Info	Element Count	Weka (ANN) K-Fold = 4
All values	All values	7195	60%
All values	Except against values	7199	70%
All values	Groups are equal	6282	76%
Paid Applications	2 Parts	3142	96%
Paid Applications	3 Parts	3142	75%
Paid Applications	4 Parts	3142	65%
Paid Applications	5 Parts	3142	63%

Figure 16: Results Of 4-Fold Cross Validation

App Type	DataSet Info	Element Count	Weka (Naive Bayes)
All values	All values	7195	60%
All values	Except against values	7199	56%
All values	Groups are equal	6282	56%
Paid Applications	2 Parts	3142	20%
Paid Applications	3 Parts	3142	38%
Paid Applications	4 Parts	3142	28%
Paid Applications	5 Parts	3142	24%

Figure 17: Results Of The Naive Bayes

Related Work

A website that called EstimateMyApp is trying to appraise applications by their features. It requests size of the app, level of the UI of the application, billing information etc. to predict the price range of the application[20].

Future Work

Proposed project aims to predict the price of the application. But there are some other predictable features. In the other steps we will work on already appraised applications and try to predict some key features about them. How many customers will download our application if we keep request this price. When should we revise the price of the application. We will try to answer these unanswered questions in our future project.

References

[1] https://govcontractassoc.com/most-favored-customer-clause/
[2] Gans, J. (2012). Mobile application pricing. Information Economics and Policy, 24(1), pp.52-59.
[3] Mustafa Furkan Keskenler1, Eyüp Fahri Keskenler2(2017), Geçmişten Günümüze Yapay Sinir Ağları ve Tarihçesi, pp.8-18.
[4] Agatonovic-Kustrin, S. and Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis, 22(5), pp.717-727.
[5] Blum, A. and Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2), pp.245-271.
[6] Ömer Faruk Ertuğrul1, Mehmet Emin Tağluk2, Yılmaz Kaya3 (2017), A Basic and Brief Scheme of an Application of a Machine Learning Process
[7] Gülhan OREKEKİCİ TEMEL1, Semra ERDOĞAN1*, Handan ANKARALI2, Sınıflandırma Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı
[8] Aran, O., Yıldız, O. and Alpaydın, E. (2009). An Incremental Framework Based On Cross-Valıdatıon For Estımatıng The Archıtecture Of A Multılayer Perceptron. International Journal of Pattern Recognition and Artificial Intelligence, 23(02), pp.159-190.
[9] Agatonovic-Kustrin, S. and Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis, 22(5), pp.717-727.
[10] Patro, S. and sahu, K. (2015). Normalization: A Preprocessing Stage. IARJSET, pp.20-22.
[11] Selahattin YAVUZ1, Muhammet DEVECİ2 (2013), İstatistiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi, pp.167-187
[12] Saylı, A., Akbulut, C. And Kosuta, K. (2018). Multıple Regressıon Analysıs System In Machıne Learnıng And Estımatıng Effects Of Data Transformatıon&Normalızatıon. Journal of Engineering Technology and Applied Sciences.
[13] Hüseyin Eriştı1, Vedat Tümen2(2012), K-means Kümeleme Yaklaşımı Kullanarak Elektrik Dağıtım Sistemlerindeki Harmoniklerin Zamansal Değişimlerinin İncelenmesi
[14]Turgay Tugay BİLGİN1, Yılmaz ÇAMURCU2(2005), DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması, pp.139-145
[15] Güncel SARIMAN(2011), Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması,pp.192-202
[16]Necati Arslan (2018), Özelleştirilmiş Naive Bayes Algoritması, Döviz Alış-Satış Önerisi Sistemi Uygulaması
[17] https://semver.org/
[18]https://dzone.com/articles/how-to-version-your-software
[19] https://medium.com/@jameshamann/a-brief-guide-to-semantic-versioning-c6055d87c90e
[20] https://estimatemyapp.com/

Environmental Sound Recognition With Varıous Feature Extraction And Classification Techniques

Y. Arpacı[1], H. Canbolat[1]

[1]*Ankara Yıldırım Beyazıt University, Ankara, Turkey, arpacyasemin@gmail.com, hcanbolat@ybu.edu.tr*

Abstract

This thesis proposes "An Environmental Sound Recognition with various feature extraction and classification techniques" for environmental sound recognition. Study in Environmental Sound Recognition (ESR) has taken attention in recent years. In the past decade, research on the Environmental Sound Recognition (ESR) area has accelerated. ESR has important role on intelligent computer systems and robots for the purpose of identification, recognition and discrimination. In this survey, I will put forward a survey on which various feature extraction and classification techniques is better to recognize environmental sounds. Survey includes these parts: environmental sound recognition system processing, feature extraction techniques, classification techniques, and performance comparison of selected techniques. At long last, finishing up comments and future innovative work slants in the ESR field will be given.

Introduction

Artificial intelligence has started to take place in our lives for about ten years with the latest technological developments. Artificial intelligence related studies have accelerated and environmental sound recognition has been one of them. Although there is a lot of research about environmental sound recognition, there is limited research which comparing feature extraction and classification methods to make the most effective environmental sound recognition. Considering the constantly developing field of artificial intelligence, in this research, it is aimed to make environmental sound recognition in the most efficient way by using various feature extraction and classification techniques and making comparisons with these techniques. There are many articles in the literature on environmental sound recognition [1] [2] [3] [4] [5]. In these articles, a database containing various sounds and information about the development of a sound voice recognition system are given.

Environmental Sound Recognition System Processing

Environmental Sound Recognition System Design starts with database selection. After, select the right environmental sound database firstly, feature extraction techniques are applied the environmental sounds. Then, these data which are obtained from applying feature extraction techniques use for learning. Before the learning phase we divide the data train and test. For learning, train data are used and then, learning techniques are applied these train data. System learns from train data and system aims to recognize environmental sounds with test data. Each try, system learn better to recognize environmental sounds.

Database and Experimental Setup

In the research, http://wiki.cnbc.cmu.edu/Sound_Databases is used as a database. Various environmental sounds have been selected from this database. Anaconda plugin Spyder, a program developed for Python, is used to create the sound recognition system to work. The reason why Python language is preferred is that its libraries are suitable for both feature extraction and classification techniques. Another reason is that it is efficient in such applications. The majority of the sounds are used for learning data and the rest are used as test data (learning data 2/3, test data 1/3). Audio files with. wav extension selected as learning data and loaded into the sound recognition system using the Python library which name is librosa. Different feature extractions (Features Extraction section) are applied to the audio signals which are obtained from these environmental sounds. The features are obtained from each feature extraction techniques and combined together to be used in learning techniques. It is aimed to learn the system by applying various classification techniques (Classification section). After the system learning, the sounds that are separated as test data are tested. For each combination, accuracy percentages are obtained from these test results and best combinations are determined.

Feature Extractİon

Every audio signal consists of many features. However, we must extract the characteristics that are relevant to the problem we are trying to solve. The process of extracting features to use them for analysis is called feature

extraction [6]. Feature extractions is the principle part of the environmental sound recognition system. The work of this is to extract those features from the input signal that help the system in identifying the sound. Feature extraction compresses the magnitude of the input signal without causing any harm to the power of sound signal.

Sound Feature Extraction Techniques and Application

There are many feature extraction techniques. In this thesis, few of the features are used like a Zero-Crossing Rate, Spectral Centroid, Mel-Frequency Cepstral Coefficients, Chroma Frequencies, Continuous Wavelet Transform. Firstly, the sound is loaded as an audio. wav file with python code in Spyder which is environment for python. Then, a few libraries are used to extract the signal features. Librosa and signal libraries are used for feature extraction properties.

Zero Crossing Rate
The Zero Crossing Rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval [6].

Spectral Centroid
The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of mass" of the spectrum is located. Perceptually, it has a robust connection with the impression of "brightness" of a sound [34].
It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights: [8]

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

where x(n) represents the weighted frequency value, or magnitude, of bin number n, and f(n) represents the center frequency of that bin.

Mel-Frequency Cepstral Coefficients (MFCC)
In audio processing, Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the short-term power spectrum of a sound based on a linear cosine transformation of a log power spectrum on a non-linear frequency scale.

Chroma
Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave [6].

Continuous Wavelet Transform (CWT)
The conventional method of producing a time-frequency map using the short time Fourier transform (STFT) limits time-frequency resolution by a predefined window length. In contrast, the Continuous Wavelet Transform (CWT) method does not require preselecting a window length and does not have a fixed time-frequency resolution over the time-frequency space. CWT uses dilation and translation of a wavelet to produce a time-scale map. A single scale encompasses a frequency band and is inversely proportional to the time support of the dilated wavelet [7].

Classification

The concept of classification is simply to distribute the data between the various classes defined on a data set. Classification algorithms learn this distribution from the given training set and then try to classify them correctly when the test data is not specified.

Classification Techniques and Application

There are many feature extraction techniques. In this thesis, we use few of the classification techniques like a K nearest neighborhood (KNN), Support Vector Machine (SVM), Decision Tree Classifier.
Firstly, the sound is loaded as an audio .wav file with python code in Spyder which is environment for python. Then, a few libraries are used to extract the signal features. Librosa, signal libraries are used for feature extraction

properties. After that, these features are used for learning and classification techniques. System learn from these data and try to recognize and classify test data which environmental sound type are they.

K nearest neighborhood (KNN)
According to this algorithm used in the classification, the characteristics removed during the classification are looking at the similarity of the new individual from previous individuals to k.

Support Vector Machine (SVM)
It is one of the most effective and simple methods used in classification. For classification, it is possible to separate two groups by drawing a border between two groups in a plane. The place where this limit will be drawn should be the most distant from the members of both groups. Here SVM determines how this limit is drawn.

Decision Tree Classifier
In the decision tree learning, a tree structure is formed and the class labels on the leaf level of the tree and the handles that go to these leaves and with the arms coming from the beginning are expressed.

Results

Comparisons are made from the results and the best combinations are determined. These are MFCC-SVM 95%, CWT-SVM 95%, CWT-DecisionTreeClassifier 95% as shown in Table 5.1. Mel-Frequency Cepstral Coefficients (MFCC) and Continuous Wavelet Transform (CWT) are the most effective feature extraction techniques. As a classification technique, the most effective techniques are Support Vector Machine (SVM) and Decision Tree Classification.

Combinations	KNN	SVM	Decision Tree Classifier
MFCC	%75	%95	%75
CWT	%50	%95	%95
Chroma	%50	%30	%50
Zero Crossing Rate	%40	%60	%70
Spectral Centroid	%75	%65	%80

Table 5.1: Performance comparison result

Conclusion

The aim of this paper is to make environmental sound recognition in the most efficient way. It is aimed to find the combination of the most successful feature extraction and classification for the most efficient environmental sound recognition. In this, various learning techniques are applied to the features which are obtained after time-frequency measured feature extraction techniques and for using different feature extraction techniques and learning techniques together various combinations are obtained to find best sound recognizing system. Comparisons are obtained from the test data result and in this way, the best combinations are determined. According to the comparison result MFCC-SVM 95%, CWT-SVM 95%, CWT-DecisionTreeClassifier 95% rate is obtained. Further elaboration on these techniques can be achieved by obtaining better classification rates.
The field of artificial intelligence, which is constantly evolving, enables the testing of new types of feature extraction and classification techniques, and paves the way for environmental sound recognition studies.

References

[1] K. Lupatka, P. Zwan, A. Czyzewski, (2010) *Dangerous sound event recognition using support vector machine classifiers*, Advances in Multimedia and Network Information System Technologies of the series Advances in Intelligent and Soft Computing (Vol. 80, pp. 49-57)
[2] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, (2010) *Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring*, Pattern Recognition Letters (Vol. 31, No. 12, pp. 1524–1534)
[3] S. Chu, S. Narayanan, and C.-C. J. Kuo, (2009) *Environmental sound recognition with time– frequency audio features,* IEEE Trans. Audio, Speech, Lang. Process. (Vol. 17, No. 6, pp. 1142–1158)
[4] M. Lojka, M. Pleva, E. Kiktova, J. Juhar, A. Cizmar, (2016) *Efficient acoustic detector of gunshots and glass breaking,* Multimedia Tools and Applications, 75(17):10441–10469.

[5] Music Genre Classification with Python, (2018). *A Guide to analysing Audio/Music signals in Python* (https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8) Accessed 12 May 2019.
[6] *https://towardsdatascience.com/music-genre-classification-with-python*
[7] Satish Sinha, Partha S. Routh, Phil D. Anno, and John P. Castagna. *Spectral decomposition of seismic data with continuous-wavelet transform*. Geophysics Volume 70, Issue 6.
[8] *A Large Set of Audio Features for Sound Description,* (2003) - technical report published by IRCAM. Section 6.1.1 describes the spectral centroid

Statistical Analysis Used in Scientific Studies in Nursing

M. Yildiz[1], E. Senturk[1]

[1]Agri Ibrahim cecen University School of Health, Agri, Turkey, myildiz@agri.edu.tr, ecavusoglu@agri.edu.tr

Abstract

There is a need for advanced statistical methods in order to distinguish successful, practical and methods from nurses. Evidence-based nursing increases the effectiveness of all kinds of medical process to prove itself, to show the validity of statistics and analysis tools provided by the statistics. Cumulative growth of knowledge on nursing, solutions to new problems or better solutions to old problems is possible through scientific research. Correct analysis of nursing increases the effectiveness of the study. In the studies conducted in the field of nursing, statistical analyzes such as Descriptive Statistics, Compare Means, Nonparametric Test, Correlation Analysis are used.

Introduction

Statistics is a branch of science that plays an important role in objective decision making by using the data obtained from the data and ultimately dealing with the development of observations which are the basis of positive sciences, the collection, analysis and interpretation of the data. The use of statistics in biology, medicine and other health sciences is defined as diğer biostatistics İst. The first use of statistics in the field of health M.S. 720 started with the necessity of live birth, death and marriage records in Japan; however, the collection of health statistics in our country coincides with after 1910.

The fields of use of biostatistics in health sciences are generally classified as service planning, diagnosis and treatment procedures, examination of social changes, protective services, identification of biological, morphological and physiological characteristics, scientific studies and measurement of service.[1]

Descriptive Statistics

In the studies, frequency tables, descriptive statistics, scatter measurements, calculations and graphs of the data are drawn from the ‘’Descriptive Statistics’’ menu with Frequencies. Descriptives are used to determine the statistics of the data and the skewnes (kurtosis) and asymmetric distribution dimensions. With Exploration, the descriptive statistics of all units or units in each group are calculated and scatter plots are drawn. One variable has statistics according to the other variable. With Crosstabs, two or more variable diagrams are arranged. In addition, tests are done in the prepared tables and summary statistics are found and numerical variables are divided into a small number of groups by coding and cross tables are arranged. In the analyzes, Kurtosis and skewness coefficients are analyzed for the distribution of the data.

Compare Means

Normal distributions use the 'Compare Means’ menu. It includes methods to test the difference between two or more averages in dependent and independent samples. Means: Calculates the statistics of subgroups according to the variables alone or by other independent variables. One Sample T test: One sample T test. Independent Samples T Test: Two independent samples T test. Paired Samples T Test: Makes two samples T Test. One Way ANOVA: One-way analysis of variance is used to test the significance of the two independent samples.

Nonparametric Test

Nonparametric Test menu is used in non-normal distributions. The Mann - Whitney U Test as a non - parametric alternative for independent variables t-test for scale variables; The Wilcoxon Test is used as a non-parametric alternative to the t test. One-way analysis of variance (ANOVA) for non-parametric variables is the non-parametric Kruskal Wallis Test. Correlation analysis is a statistical analysis which shows the relationship between two or more variables, if there is a relationship, the severity of this relationship. ’In the ate Correlate’ menu, the bivariate relationship between two or more variables; direction, size and importance are determined. Also partial (partial) correlation analysis is performed. According to at least one variable, similarities are used to determine the similarities or dissimilarities of the units. As an alternative to the parametric Pearson Correlation test, the sample size is 51 and above, while Spearman, 50, and the Kendall test are used for gold.

Conclusion

By analyzing the data with statistical tests, p value is obtained. The p value is the probability that the results of the research will be obtained depending on the chance if the null hypothesis is correct. If P is less than the value α, the H0 hypothesis is rejected in favor of the alternative hypothesis; this is indicated by definitions such as $p < 0.05$ or $p < 0.01$. In summarizing the results of the research, the following points should be given importance.
In the research, which statistical method is used to obtain the result, if more than one method is used,
• What is compared in group comparisons (averages, percentages, medians),
• According to the p-value found with the analysis result, it should be noted whether the hypotheses established are rejected and what it means. (For example, H0 hypothesis is rejected because $p < 0.05$).
In conclusion, the reliability and validity of the research in the field of nursing depends on the evaluation of the data by appropriate statistical methods. Therefore, it is important for the researchers to prepare their projects in consultation with a statistical expert at the planning stage of the studies and to perform the data collection / analysis procedures correctly.

References

[1] Sümbüloğlu K., Sümbüloğlu ., Biyoistatistik, 9. Baskı, Ankara, 2000, 1-6.
[2] Lang T. Twenty statistical errors even you can find in biomedical research articles. Croat Med J 2004;45:361-70.
[3] Biyoistatistik, Kadir Sümbüloğlu, Vildan Sümbüloğlu, 12. Basım, 2012, Hatipoğlu yayın evi, Ankara.
[4] Fundamantals of Biostatistics, Bernard Rosner, 7. Basım, 2011, 2006 Brooks/Cole, Cengage Learning, USA.
[5] Basic and Clinical Biostatistics, Beth Davson, 4. Basım, 2004. McGraw-Hill Professional Publishing, USA.
[6] Akgül A. Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri, Ankara, 1997, 236-237.
[7] Bluman A.G., Elementary Statistics, McGraw-Hill ,2004, 365-424.
[8] Glover T., Mitchell K., An Introduction to Biostatistics,McGraw-Hill, 2002, 113-127.
[9] Çelik Y. Biyoistatistik Araştırma İlkeleri, Diyarbakır, 1. Baskı, 1999, 125-127.
[10] Bernstein S., Bernstein R., Elements of Statistics II: Inferential Statistics, McGraw-Hill, 1999, 379-394.
[11] Aksakoğlu, G., Sağlıkta Araştırma Teknikleri ve Analiz Yöntemleri, İzmir, 2001, 308.

Software Used for Drug Design and Development

M. Senturk[1]

[1]Agri Ibrahim Cecen University, Pharmacy Faculty, Agri, Turkey, msenturk@agri.edu.tr

Abstract

The discovery of a drug is a process involving the design and development of the drug. Software-based drug discovery and development methods have played a major role in the development of compounds that have demonstrated biological activity over the last 35 years. Methods such as software-based molecular modeling, structure-based drug design, structure-based virtual screening, ligand interaction and molecular dynamics have been developed. This software is thought to be a very powerful tool that helps to investigate the pharmacokinetic and pharmacodynamic properties of the drug and the structural activity relationship between ligand and drug. Computational approaches, such as target molecule placement, allow the calculation of the interaction of small molecules with structural macromolecules. This makes it easier to identify the target and optimization of the precursor. These softwares help determine experimental findings and mechanisms of action more quickly. In addition, the correct use of these software provides a reduction in drug design and development costs. Nowadays in biomedical sciences, these software plays a mandatory role in different stages of discovery of the drug. Today, over 50 docking software has been developed and used for various purposes. In this study, we have tried to illustrate how appropriate docking programs are compared with experimental data.

Introduction

The development of model and software-based tools has significantly contributed to drug discovery and development for the pharmaceutical industry. These software have become a component of drug discovery and have a very important role in determining the potential of new bioactive molecules to become drugs. The accurate and effective use of these software and computer-based modeling methods in our day reduces the costs in the process of drug discovery and accelerates the development of new drugs. Molecular modeling, structure-based drug discovery, structure activity relationship (SAR), ligand-based modeling and molecular chemistry methods such as molecular dynamics, pharmacokinetic and pharmacodynamic properties are widely used in determining a tool. In addition, one of the most important findings is used as a very powerful tool for determining the relationship between structure and ligand-target activity (Table 1) [1,2]. It provides a powerful paradigm for modernizing experimental design design and analysis combining modeling applications such as ligand-based computer-aided drug design (CADD) and simulations [3,4]. These software-aided techniques make researches advantageous over time and expense. These advantages; The need for materials used in research reduces the number of animals needed and the preclinical stages of drug discoveries. It can also help generate solutions for handling huge data and can improve accuracy of study results [5,6].

Table 1. Some computer and software based programs used in new drug discovery and design.

	Software name	Major use	References
Ligand interactions and molecular dynamic	AutoDock	Ligand-protein interaction	[7]
	Schrodinger	Ligand-receptor docking	[8]
	GOLD	Protein-ligand docking	[9]
	BioSuite	Genome and sequence analyzing	[10]
Molecular modeling and structural activity relationship (SAR)	Maestro	Molecular modeling analysis	[11]
	ArgusLab	Molecular docking calculations and molecular modeling package	[12]
	GRAMM	Protein-protein docking and protein-ligand docking	[13]
	SYBYL-X Suite	Molecular modeling and ligand based design	[14]
	PASS	Create and analysis of SAR models	[15]
Data analysis	GeneSpring	Identify variation across set of sample and for correction method in samples	[16]
	QSARPro	Protein-protein interaction study	[17]
	REST 2009 Software	Analysis of gene expression data	49,50

Today, on average, one billion dollars is spent on a single drug and about 12 years is needed to achieve a successful product. High cost, long time need, high level of risk, the results are both uncertain and highly complex procedures are the main challenges in discovery of a new drug. Today, in order to overcome these problems, new and cheaper drug discovery and design methods such as software and computer based drug design and molecular coupling are used [5,7]. This paper highlights the commonly used software for the discovery of new drugs as well as potentially used drugs [19].

Conclusion

In this study, we have briefly mentioned some software-based programs that play a very important role in drug design and discovery recently. Today, successful implementation of software and computer-based techniques helps identify bioactive agents in vitro. It also provides an opportunity without prejudice to well-known companies or potential customers. New methods, such as deployment, help solve a wide variety of mechanisms underlying the complex target ligand interaction. Significant advances in the field of pharmacokinetics and pharmacodynamics and the continued application of new software make use of the drug discovery process. This reduces candidate drug discovery and the cost constraints of various biochemical industries. Some previous drug samples such as indinavir, HIV protease inhibitor, Software and software-based approaches can be used to assist in expensive, complex and highly challenging drug design and discovery operations.

References

[1] Ferreira, L.G., Santos, R.N., Oliva, G., and Andricopulo, A.D. (2015) Molecular docking and structure-based drug design strategies. *Molecules*, 20, 13384-13421.
[2] Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W. (2014). Computational methods in drug discovery, *Pharmacol. Review*. 66, 334-395.
[3] Kimko, H., and Pinheiro, J. (2014). Model-based clinical drug development in the past, present and future: a commentary, *Br. J. Clin. Pharmacol.* 79, 108-116.
[4] Gill, S.K., Christopher, A.F., Gupta, V., and Bansal, P. (2016). Emerging role of bioinformatics tools and software in evolution of clinical research, *Perspect. Clin. Res.* 7, 115-122.
[5] Hodos, R.A., Kidd, B.A., Shameer, K., Readhead, B.P., and Dudley, J.T. (2016). *In silico* methods for drug repurposing and pharmacology, computational approaches to drug repurposing and pharmacology, *WIREs Syst. Biol. Med.* 186.
[6] Liu, T., Lu, D., Zhang, H., Zheng, M., and Yang, H. (2016). Applying high-performance computing in drug discovery and molecular simulation, *Natl. Sci. Rev.* 3, 49-63.
[7] Rizvi, S.M.D., Shakil, S., and Haneef, M. (2013). A simple click by click protocol to perform docking: Autodock 4.2 made easy for non- Bioinformaticians, *Excli. J.* 12, 831-857.
[8] Dineshkumar, B., Kumar, P.V., Bhuvaneshwaran, S.P., and Mitra, A. (2010). Advanced drug designing softwares and their applications in medical research, *Int. J. Pharmacy Pharm. Sci.* 2, 16-18.
[9] Yurieva, E., Agostinoa, M., and Ramsland, P.A. (2011). Challenges and advances in computational docking: 2009 in review, *J. Mol. Recogn.* 24, 149-164.
[10] The NMITLI-BioSuite Team, BioSuite: a comprehensive bioinformatics software package (A unique industry–academia collaboration), *Curr. Sci.* 92 (2007) 29-38.
[11] https://www.schrodinger.com/Maestro (accessed 30.05.2019).
[12] http://www.arguslab.com/arguslab.com/ArgusLab.html (accessed 30.05.2019).
[13] Kundrotas, P.J., and Vakser, I.A. (2010). Accuracy of protein-protein binding sites in high throughput template-based modeling, *PLoS Comput. Biol.* 6, 1-10.
[14] http://www.certara.com/products/molmod/sybyl-x (accessed 30.05.2019).
[15] http://www.genexplain.com/pass (accessed 30.05.2019).
[16] http://www.genomics.agilent.com/en/Microarray-Data-Analysis-Software/Gene Spring GX/?cid=AG-PT-130 & tabId=AG-PR-1061 (accessed 30.05.2019).
[17] http://www.vlifesciences.com/products/QSARPro/Product_QSARpro.php (accessed 30.05.2019).
[18] www.qiagen.com (accessed 30.05.2019).
[19] Jamkhande, P.G., Ghante, M.H., and Ajgunde, B.R. (2017). Software based approaches for drug designing and development: A systematic review on commonly used software and its applications. *Bulletin of Faculty of Pharmacy, Cairo University* 55, 203-210.

In Vitro and In Silico Calculation of Melatonin and Dopamine on Acetylcholinesterase and Butyrylcholinesterase

M. Senturk[1], E. Senturk[1], H. Ustundag[2]

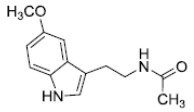
[1]*Agri Ibrahim Cecen University, Faculty of Pharmacy, Agri, Turkey, msenturk@agri.edu.tr, ecavusoglu@agri.edu.t r*
[2]*Artvin Coruh University, Faculty of Health Sciences, Artvin, Turkey, hurcar@artvin.edu.tr*

Abstract

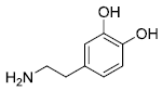
Melatonin (N-acetyl-5-methoxytryptamine) is the main product of the pineal gland, and remains an intriguing indoleamin. This compound is widely distributed among living organisms and has been detected in all living species [1,2]. Dopamine (3,4-dihydroxyphenethylamine) is an organic chemical of the catecholamine and phenethylamine families. It functions both as a hormone and a neurotransmitter, and plays several important roles in the brain and body [3]. Agents that inhibit acetylcholinesterase (AChE) and Butyrylcholinesterase (BuChE) enzymes are compounds which can be used for different therapeutic applications in which Alzheimer's and Parkinson's disease is involved [4]. In this study, we investigated the inhibition of these cholinesterase enzymes with melatonin and dopamine. The inhibition mechanism of these enzymes has been found to be caused by blockage of the active site entrance by the inhibitor scaffold. Molecular insertion and molecular dynamics (MD) simulations have shown significant interactions between the ligands and amino acid residues studied in different regions of the active regions of AChE and BuChE. A variety of classical AChE and BuChE inhibitors can be used as a guiding molecule for the design of the investigated substances, the novel therapeutically effective enzyme inhibitors.

Introduction

The hormone melatonin (N-acetyl-5-methoxytryptamine) regulates the sleep-wake cycle. It is firstly synthesized by the pineal gland and left in circulation [1]. Dopamine (3,4-dihydroxyphenylamine) is a member of the catecholamine group compound family. This substance acts as both a hormone and a neurotransmitter. Dopamine is known to play several important roles in the brain and the body. This substance is an amine synthesized by removing a carboxyl group from L-DOPA synthesized in the brain and kidneys. Dopamine is also a natural molecule synthesized in most animals and plants. In the brain, dopamine acts as a neurotransmitter, a chemical released by neurons to signal to other nerve cells [5].



Melatonin



Dopamine

Melatonin is a well-established endogenous free radical scavenger and an effective antioxidant [6]. It also regulates selected antioxidant enzymes such as glutathione peroxidase (GPx) and superoxide dismutase (SOD) [7]. Melatonin has been thought to be a good clinical candidate for the treatment of a potent acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) inhibitor (takrin) as a neuroprotective agent, a dual activity drug and Alzheimer's disease, and recently some promising results have been reported [8], it was determined that the activity of both drugs, melatonin-takrin hybrid compounds were effective inhibitors of these enzymes [8,10]. In a study by Klagaris *et al.*, dopamine was determined to be 21 µg / mL for the AChE enzyme and 56 µg / mL for the BChE enzyme [9].

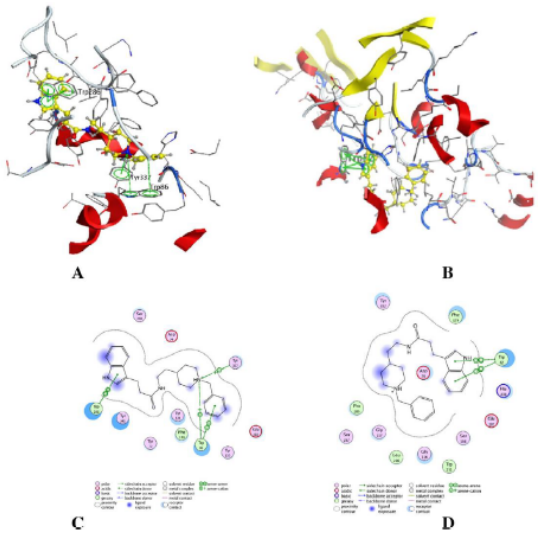


Figure 1. (A) 3D docking model of melatonin derivative with hAChE. The dashed lines represent the interactions between the protein and the ligand. (B) 3D docking model of melatonin derivatives with hBuChE. (C) 2D schematic diagram of docking model of melatonin derivative with hAChE. (D) 2D schematic diagram of docking model of melatonin derivative with hBuChE [10].

Conclusion

In this study, we determined that they are powerful inhibitors of melatonin and dopamine or their derivatives and AChE and BChE enzymes. These derivatives are potent cholinesaterase inhibitors and have been supported by *in vitro*, *in silico* and *in vivo* studies that have potential to be used in the treatment of diseases such as Alzheimer's disease, Parkinson's disease, Myestenia gravis.

References

[1] Reiter, R.J., Tan, D.X., Rosales-Corral, S., and Manchester, L.C. (2013). The universal nature, unequal distribution and antioxidant functions of melatonin and its derivatives. *Mini Review of Medicinal Chemistry*. 13, 373-384.

[2] Calvo, J.R., Gonzales-Yanes, C., and Maldonado, M.D. (2013). The role of melatonin in the cells of the innate immunity: A review. *Journal of Pineal Research.* 55, 103-120.

[3] Berridge, K.C., Robinson, T.E., and Aldridge, J.W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*. 9, 65-73.

[4] Cavdar, H., Senturk, M., Guney, M., Durdagi, S., Kayik, G., Supuran, C.T., and Ekinci, D. (2019). Kinetic and *In Silico* Studies of Some Uracil Derivatives on Acetylcholinesterase and Butyrylcholinesterase Enzymes. *Journal of Enzyme Inhibition and Medicinal Chemistry*. 34, 429-437.

[5] Berridge, K.C., Robinson, T.E., and Aldridge, J.W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*. 9, 65-73.

[6] Reiter, R.J., Tan, D-X., Mayo, JC. (2003). Melatonin as an antioxidant: biochemical mechanisms and pathophysiological implications in humans. *Acta Biochim Pol.* 50, 1129-1146.

[7] Reiter, R.J., Tan D-X, Cabrera J et al. (1999). The oxidant/antioxidant network: role of melatonin. *Biol Signals Recept*. 8, 56-63.

[8] Rodriguez-Franco, M.I., Fernandez-Bachiller, M.I., Perez, C. et al. (2006). Novel tacrine-melatonin hybrids as dual-acting drugs for Alzheimer disease with improved acetylocholinesterase inhibitory and antioxidant properties. *J Med Chem*. 49, 459-462.

[9] Klegeris, A., Korkina, L.G., and Greenfield, S.A. (1995). A Possible Interaction Between Acetylcholinesterase and Dopamine Molecules During Autoxidation of the Amine. *Free Radical Biology & Medicine*, 18, 223-230.

[10] Wang, J., Wang, Z-M., Li, X-M., Li, F., Wu, J.J., Kong, L.J., and Wang X. B. (2016). Synthesis and evaluation of multi-target-directed ligands for the treatment of Alzheimer’s disease based on the fusion of donepezil and melatonin. *Bioorganic & Medicinal Chemistry*. 24, 4324-4338.

AHP – TOPSIS Hybrid Approach for Research Assistant Selection

A. Aktas[1], M. A. Yerlikaya[1], M. Kabak[1], B. Ecer[2]

[1]*Gazi University, Ankara, Turkey, aaktas@gazi.edu.tr , akifyerlikaya@gazi.edu.tr , mkabak@gazi.edu.tr*
[2]*Ankara Yıldırım Beyazıt University, Ankara, Turkey, becer@ybu.edu.tr*

Abstract

Academic staff, who are working in Turkish universities can be classified into three classes as faculty members, lecturers and research assistants. Each class of these academicians have some various responsibilities in education given and researches conducted by the institution. Therefore, different criteria are needed to be considered in the personnel selection process of each class of academicians. There may be more than one applicants for a job advertisement and the number of criteria and alternatives needed to be thought brings the personnel selection decision in to a complex situation. Multi criteria decision making techniques are used for solving decision problems with several criteria and a number of decision alternatives. These techniques determine the best alternative by making some calculations on the decision matrix of the problem. They are very useful for obtaining a compromise solution in complex decision problems. The main aim of this study is to propose an analytic model for research assistant selection decision of universities. In this regard, a hybrid decision making model based on Analytic Hierarchy Process and Technique of Order Preference by Similarity to Ideal Solution (TOPSIS) is developed. Graduate Record Examination Score (C1), Grade Point Average (C2), Language Exam Score (C3), University Entrance Exam Rank (C4) and Research Experience of Applicant (C5) are considered as research assistant selection criteria within the model. A case study in Gazi University is presented in the study to demonstrate the applicability of the proposed model. Results of the current selection procedure and proposed model are compared.

Introduction

Faculty members, lecturers and research assistants are the academic personnel in universities. Each types of academicians have different responsibilities in educational processes and research activities. Faculty members are professors, associate professors and assistant professors. They are responsible for conducting academic research, teaching courses and administrative duties in university. They may serve as rector, dean or department chair as an administrative duty. Publishing articles and conference papers, supervising master and doctoral theses and developing project proposals are academic responsibilities of faculty members. Also, they are expected to teach courses to students. On the other hand, lecturers are only responsible for teaching courses. Research assistants are expected to assist faculty members and lecturers for their responsibilities. So, research assistants have different responsibilities for administrative, academic and educational processes in universities. Hiring process of each type of academic staff have different steps. For example, faculty members are evaluated according to their previous academic, educational and administrative experiences. Lecturers are being hired after a two-stage evaluation process. In the first step, their exam scores and graduate point average is considered, then they take an exam for evaluation of their capabilities and professional experience on a desired subject. Employment of research assistants is a two-stage process, too. But, in the first stage they are evaluated by graduate record exam and language exam scores, next they are having an exam for assessment of general basic knowledge of their area. In this study, research assistant selection problem is considered. Determination of the best research assistant to hire requires considering different criteria simultaneously and this makes the selection of the best staff to be a complex decision. Moreover, a number of applicants are needed to be evaluated, and one applicant may have different advantages over the other applicants. In such complex decisions, which contains different criteria and alternatives, multiple criteria decision making approaches are very useful to obtain a compromise decision [1].

Using multiple criteria decision making approaches for personnel selection is very common in the literature. Çelik et al. [2] used triangular fuzzy sets, fuzzy AHP and fuzzy TOPSIS methods for academic personnel selection. Aksakal and Dağdeviren [3] solved personnel selection problem with interdependent criteria by using a hybrid model based on DEMATEL and ANP. Application of fuzzy TOPSIS method on personnel selection problem is presented in Kelemenis and Askounis's study [4].

The main aim in this study is to propose an analytic decision making approach based on AHP and TOPSIS methods for choosing the best research assistant among a number of applicants. A case study in Gazi University is given in this study to demonstrate the applicability of the proposed approach. The rest of the paper organized as follows: the proposed methodology for research assistant selection is given in the second part. Next, research assistant selection application is presented in the third part. The paper is concluded in the fourth part by giving suggestions for further applications.

Methodology

Research assistant selection is made by an integrated AHP and TOPSIS methodology in this study. AHP is used to determine criteria weights and TOPSIS is used for evaluation of alternatives.
Analytic Hierarchy Process (AHP) method is proposed by Thomas L. Saaty in 1980 [5]. The basic idea of the method is to derive ratios from pairwise comparisons.

TOPSIS method is firstly proposed by Hwang and Yoon [6] to find solutions in MCDM problems. TOPSIS method aims to find the shortest alternative to positive ideal solution and the farthest alternative to negative ideal solution. Readers may refer to Saaty [5] for detailed expression of AHP and Hwang and Yoon [6] for TOPSIS.

AHP – TOPSIS Application for Research Assistant Selection

In this part, application of the proposed approach is presented. Seven applicants for research assistant position in Gazi University Industrial Engineering Department are evaluated by AHP – TOPSIS approach in views of five criteria including Graduate Record Examination Score (C1), Grade Point Average (C2), Language Exam Score (C3), University Entrance Exam Rank (C4) and Research Experience of Applicant (C5).

Pairwise comparison matrix for criteria evaluation is constructed and presented in Table 1. Criteria weights are obtained by using this matrix and consistency ratio is checked. University entrance exam rank is seemed to be the most important criterion and consistency of evaluations is satisfied.

Table 1. Pairwise comparison of criteria

	C1	C2	C3	C4	C5	Priority
C1	1	5	3	3	1	0.3201
C2	1/5	1	1/3	1	1/5	0.0674
C3	1/3	3	1	1/3	1/5	0.1013
C4	1/3	1	3	1	1/5	0.1200
C5	1	5	5	5	1	0.3912
Consistency Ratio: 0.094						

Obtained criteria weights are used for obtaining weighted normalized decision matrix in TOPSIS. Original decision matrix of the problem is presented in Table 2, normalized decision matrix, weighted normalized decision matrix and ideal solution sets are not given due to paper limitations.

Table 2. Decision matrix of the problem

Alternative	C1	C2	C3	C4	C5
A1	89.167	85.00	73.40	1	0.9547
A2	87.423	90.00	91.36	3	0.9766
A3	80.731	85.00	75.03	2	0.9514
A4	86.417	83.75	75.96	1	0.9603
A5	88.252	97.50	80.86	3	0.9514
A6	87.328	96.25	72.00	2	0.9494
A7	84.813	87.50	75.96	1	0.9930

By calculating normalized and weighted normalized decision matrices, ideal solution sets are determined. Then, distances of alternatives to positive and negative ideal solutions are calculated by Euclidian distance. Rank of alternatives is determined by rank of relative closeness value, which is calculated by using distances to positive and negative ideal solutions. Distance to ideal solutions, relative closeness values and rank of alternatives are presented in Table 3, as follows:

Table 3. Distance to ideal solutions and ranking of alternatives

Alternative	Distance to Positive Ideal Solution	Distance to Negative Ideal Solution	Relative Closeness to Ideal Solution	Rank
A1	0.045939	0.011873	0.205372	5
A2	0.004106	0.046736	0.919248	1
A3	0.027454	0.022334	0.448578	4
A4	0.045803	0.008369	0.154495	7
A5	0.008304	0.046161	0.847539	2
A6	0.025262	0.024382	0.491140	3
A7	0.045693	0.009078	0.165742	6

According to results obtained by proposed approach, A2 seems as the best candidate for this position. This applicant is followed by A5, A6, A3, A1, A7 and A4, respectively. We can see that A5 has greater points in C1 and C2, which are generally considered as important aspects for graduate studies, but A2 has better scores in C3, C4 and C5. The importance of criteria weighting can be seen from these results.

Conclusion

Research assistants have important responsibilities in educational institutions. So, selection of the best research assistant is an important decision. In this study, AHP and TOPSIS methods are integrated to determine the best research assistant among possible candidates. These methods are integrated, because AHP is commonly used for weighting selection criteria and TOPSIS is a suitable method for evaluation among alternatives. A case study in Gazi University is presented to demonstrate the applicability of proposed methodology. In further studies, effects of different criteria or results obtained by different MCDM methods can be analyzed.

References

[1] Kabak M., Sağlam F., and Aktas A. (2017). Usability analysis of different distance measures on TOPSIS, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 32 (1), 35-43.
[2] Celik M., Kandakoglu A., and Er I.D. (2009). Structuring fuzzy integrated multi-stages evaluation model on academic personnel recruitment in MET institutions, *Expert Systems with Applications*, 36 (3), 6918-6927.
[3] Aksakal, E., Dağdeviren M., (2010). An Integrated Approach for Personel Selection with DEMATEL and ANP Methods, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 25, 905-913.
[4] Kelemenis A., and Askounis D. (2010). A new TOPSIS-based multi-criteria approach to personnel selection, *Expert Systems with Applications*, 37, 4999-5008.
[5] Saaty, T.L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill.
[6] Hwang C.L., and Yoon K. (1981). *Multiple Attribute Decision Making: Methods and Applications*, Springer-Verlag, New York, USA

Mathematical Model of Flow Shop Scheduling Problems and Solution with Metaheuristic Methods

S.Kaya[1], A. Çelik[1], İ. H. Karaçizmeli[1], İ. B. Aydilek[1], A. Gümüşçü[1], M.E.Tenekeci[1]

[1]Harran University, Şanlıurfa, Turkey, serkankaya@harran.edu.tr , aysecelik@harran.edu.tr , hkaracizmeli@harran.edu.tr, berkanaydilek@harran.edu.tr, agumuscu@harran.edu.tr, etenekeci@harran.edu.tr

Abstract

Flow shop scheduling is the type of scheduling that occurs when n jobs are processed in the same order on m machine. In the beginning, the problem was to schedule n jobs on two machines. With the increase in the number of jobs and machinery, the problem has entered in the NP-Hard scope. In this study, mathematical model is presented for the solution of flow shop scheduling problems. Small sized problems are solved by using mathematical model and a model based on particle swarm optimization algorithm is proposed for the solution of medium and large sized problems. The results showed that particle swarm optimization algorithm yields effective results in solving such problems.

Introduction

One of the most commonly used types of scheduling is flow shop scheduling, which generally aims to minimize makespan. The problems that have different *m* machines and *n* jobs and each job is done in the same order are defined as flow shop scheduling problems (FSSP). Mathematical model is enough to reach optimal solution for small size problems. However, with the increase in the number of machines and jobs, the problem falls within the scope of NP-hard and it becomes very difficult to reach an optimal solution [1].

Ruiz et al. [2], Chen et al. [3] and Chen et al. [4] proposed genetic algorithms for FSSP, Liu and Liu [5] presented hybrid artificial bee colony algorithm, Fernandez-Viagas and Framinan [6] proposed NEH based heuristic algorithm, Liu et al. [7] improved NEH heuristic algorithm, Tayeb et al. [8] proposed a hybrid algorithm from genetic and artificial immune algorithms for FSSP. Wang et al. [9] proposed cuckoo search algorithm, Liu et al. [10] presented a hybrid algorithm of memetic and particle swarm optimization algorithms, Yang and Deb [11] presented a particle swarm optimization algorithm for FSSP. Abedinnia et al. [12] presented a NEH based local search algorithm for the solution of the same problem for total flow time. Taşgetiren et al. [13] presented the discrete bee colonies algorithm and Li and Yin [14] bee colonies algorithm for FSSP.

In this study, a proposal was made to minimize the objective of makespan in FSSP. While it is shown that the solution can be reached by using the mathematical model for small size problems, it has been shown that with the growth of the problem size, the mathematical model is inadequate and a value close to the target can be reached by meta heuristic methods. Particle Swarm Optimization (PSO) algorithm was used as the meta-heuristic method in the study. In the second part of the study, the mathematical model of FSSP, in the third part, PSO, in the fourth part experimental results, in the fifth part conclusion are given.

Mathematical Model

The mathematical model which aims to minimize the makepan value of the FSSP is as follows.

Indices

i: jobs ($i = 1, 2, \ldots, I$)

j: machines ($j = 1, 2, \ldots, J$)

k: sequence position ($k = 1, 2, \ldots, K$)

Parameters

t_{ij} : processing time of job i on machine j $(\forall i \in I, \forall j \in J)$

Decision Variables

x_{ik}: if job i is performed k. position 1, otherwise 0 $(\forall i \in I, \forall k \in K)$

C_{kj}: Completion time of job k on machine j $(\square k \in K, \square j \in J)$

P_{kj}: Processing time of job k on machine j $(\forall k \in K, \forall j \in J)$

C_{max}: Time for the last job to leave the last machine

Mathematical Model

$$Min\ C_{max} \quad (1)$$

$$\sum_{k=K} x_{ik} = 1 \quad ,\forall i \in I \quad (2)$$

$$\sum_{i=I} x_{ik} = 1 \quad ,\forall k \in K \quad (3)$$

$$P_{kj} = \sum_{i=I} x_{ik} t_{ij} \quad ,\forall k \in K, \forall j \in J \quad (4)$$

$$C_{k1} = C_{(k-1)1} + P_{k1} \quad ,\forall k \in K / \{1\} \quad (5)$$

$$C_{kj} \geq C_{k(j-1)} + P_{kj} \quad ,\forall j \in J / \{1\}, \forall k \in K \quad (6)$$

$$C_{kj} \geq C_{(k-1)j} + P_{kj} \quad ,\forall j \in J / \{1\}, \forall k \in K / \{1\} \quad (7)$$

$$C_{11} = P_{11} \quad (8)$$

$$x_{ik} \in \{0,1\} \quad ,\forall i \in I, \forall k \in K \quad (9)$$

$$C_{kj}, P_{kj} \geq 0 \quad ,\forall k \in K, \forall j \in J \quad (10)$$

In the model, equation (1) aims at minimizing the makespan. Equation (2) shows that each job can only be assigned to one position, and equation (3) indicates that only one job can be assigned to each position. Equation (4) k. means that the processing time of the job performed in machine j is equal to the processing time of that job on that machine. Equation (5) k. the time to complete the next job on the first machine, equation (6) and equation (7) k. shows the time when the next job is completed on other machines. Equation (8) states that the completion time of the first job on the first machine is equal to the processing time of that job on that machine. Equation (9) and equation (10) are the sign constraints of decision variables.

Particle Swarm Optimization

One of the meta-heuristic methods used to solve FSSP is the PSO algorithm. PSO is an optimization technique developed by Kennedy and Eberhart [15] in 1995 based on the movements of flocks of birds and fish during food search [16]. Compared to other evolutionary and mathematically based algorithms, PSO is an algorithm that does not require much memory, has effective computational capabilities, is easy to implement and has fast convergence features [17].

Experimental Results

Mathematical model and PSO algorithm for the solution of FSSP were solved with 6 job 6 machine problem from Benavides and Ritt [18] studies and the first 3 problems of Taillard's [19] 20 job 5 machine data set. The problem [18] is solved primarily with the mathematical model and PSO algorithm for the minimization of makespan and the results obtained are compared with the results (Benavides and Ritt, 2018). According to the result obtained with the mathematical model, the job sequence of the problem was realized as *{J5-J6-J1-J3-J4-J2}* and *Cmax* was *41*. The job sequence obtained by Benavides and Ritt (2018) was obtained as *{J5-J4-J6-J2-J1-J3}* and *Cmax* was *43*.

The data of the first 3 problems were solved with the mathematical model and developed PSO algorithm from Taillard's [19] 20 job 5 machine problem sets. In the PSO algorithm, the best solutions obtained after 3 trials were considered. The comparison was made according to the makespan values obtained in Table 1.

Table 1. Comparative makespan results.

Problem Nr.	Taillard Results [19]	Mathematical Model Results	PSO Best Results		
1	1278	1278	1312	1297	1278
2	1359	1359	1359	1359	1360
3	1081	1081	1089	1098	1098

When Table 1 is examined, it is seen that the mathematical model can find Taillard's optimal result for all three problems. The developed PSO algorithm was able to find the optimal result for 2 of 3 problems.

Conclusion

In this study, a mathematical model and PSO algorithm are presented for the makespan solution of FSSP. The solution of the small size problems obtained from the literature was realized with mathematical model and PSO algorithm. It has been seen that the mathematical model yields optimal results for small size problems. Since the mathematical model could not provide the optimal solution as the problem size increased, PSO algorithm was developed for the solution of these problems. The developed PSO algorithm was tested on 3 problems solved by mathematical model. When the results obtained in the first 3 trials for these problems were examined, it was seen that the optimal solution for the 2 problems could be achieved.

Acknowledgements

This study is a part of the project which is supported by The Scientific and Technological Research Council of Turkey (TUBITAK). The authors thank the TUBITAK for financial support of this work under the grant number 118E355.

References

[1] Baker, K. R., and Trietsch D. (2009). Principles of sequencing and scheduling. John Wiley&Sons.
[2] Ruiz, R., Maroto, C., Alcaraz, J. (2006). Two new robust genetic algorithms for the flowshop scheduling problem. Omega 34 (5), 461–476.
[3] Chen, S. H., Chang, P. C., Cheng, T. C. E. and Zhang, Q. (July, 2012). A Self-guided Genetic Algorithm for permutation flowshop scheduling problems. Computers & Operations Research (Vol 39, Issue 7, Pages 1450-1457).
[4] Chen, Y.M., Chen, M.C., Chang, P.C. and Chen, S.H. (2012). Extended artificial chromosomes genetic algorithm for permutation flowshop scheduling problems. Computers & Industrial Engineering (62: 536-545).
[5] Liu, Y.F. and Liu, S.Y. (2013). A hybrid discrete artificial bee colony algorithm for permutation flowshop scheduling problem. Applied Soft Computing (13: 1459-1463).
[6] Fernandez-Viagas, V. and Framinan, J.M. (2014). On insertion tie-breaking rules in heuristics for the permutation flowshop scheduling problem. Comput. Oper. Res. (45, 60–67).
[7] Liu, W., Jin Y. and Price, M. (2017). A new improved NEH heuristic for permutation flowshop scheduling problems. Int. Journal of Production Economics (193: 21-30).
[8] Tayeb, F. B., Bessedik, M., Benbouzid, M., Cheurfi, H. and Blizak, A. (2017). Research on Permutation Flow-shop Scheduling Problem based on Improved Genetic Immune Algorithm with vaccinated offspring. Procedia Computer Science(112,427-436).
[9] Wang H., Wang W., Sun H., Cui Z., Rahnamayan S. and Zeng S. (2017). A new cuckoo search algorithm with hybrid strategies for flow shop scheduling problems. Soft Computing (21:4297–4307).
[10] Liu B., Wang L. and Jin Y. H. (2007). An effective pso-based memetic algorithm for flow shop scheduling. IEEE Trans Sys Man Cybern Part B Cybern 37(1):18–27.
[11] Yang X. S. and Deb S. (2009). Cuckoo search via Lévy flights. In: World congress on nature and biologically inspired computing IEEE, pp 210–214.
[12] Abedinnia, H., Glock, C. H. and Brill, A. (2016). New simple constructive heuristic algorithms for minimizing total flow-time in the permutation flowshop scheduling problem. Computers & Operations Research (Vol 74, Pages 165-174).
[13] Taşgetiren M.F., Liang Y. C., Şevkli M., Gençyılmaz G. (2007). A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem. Eur J Oper Res 177(3):1930–1947.
[14] Li X. and Yin M. (2012). A discrete artificial bee colony algorithm with composite mutation strategies for permutation flow shop scheduling problem. Scientia Iranica 19(6):1921–1935.
[15] Kennedy J. and Eberhart R. C. (1995). Particle swarm optimization. International Neural Networks (95, 1942-1948).
[16] Dahal K.P., Tan K. C. and Cowling, P.I. (2007). Evolutionary Scheduling. Springer-Verlag, New York.
[17] Kaya S. and Fığlalı N. (2018). Çok Amaçlı Esnek Atölye Tipi Çizelgeleme Problemlerinin Çözümünde Meta Sezgisel Yöntemlerin Kullanımı. Harran Üniversitesi Mühendislik Dergisi, 3(3): 222-233.
[18] Benavides, A. J. and Ritt, M. (2018). Fast heuristics for minimizing the makespan in non- permutation flow shops. Computers and Operations Research (100, 230–243).
[19] Taillard, E. (2004). Best known lower and upper bounds of the PFSSP for Taillard's instances. http://mistic.heig-vd.ch/taillard/problemes.dir/ordonnancement

Evaluating Solution Performance of Hybrid Firefly and Particle Swarm Optimization Algorithm in Flow Shop Scheduling Problems

İ.H. Karaçizmeli[1], İ.B. Aydilek[1], A. Gümüşçü[1], M.E. Tenekeci[1], S. Kaya[1]

[1]*Harran University, Şanlıurfa, Turkey, hkaracizmeli@harran.edu.tr, berkanaydilek@harran.edu.tr, agumuscu@harran.edu.tr, etenekeci@harran.edu.tr, serkankaya@harran.edu.tr*

Abstract

In this study, the effect of swarm size and number of iteration parameters on the hybrid firefly and particle swarm optimization algorithm are investigated in flow shop scheduling problems. There are sequential *m* machines and *n* jobs following the same route in the permutation flow shop scheduling. Different swarm size and number of iteration parameters can be preferred in the hybrid firefly and particle swarm optimization algorithm. Therefore 8 levels for swarm sizes and 5 levels for number of iterations are tested for the better makespan value which is used as an objective function of the scheduling problem in this study. It is observed that the parameters examined can change the performance of the algorithm depending on the problem type.

Introduction

In the production facilities, the type of layout in which the machines are placed sequentially is the flow shop layout and scheduling problems in such types of facilities are called flow shop scheduling (FSS) problems [1]. The FSS in which machines have the same job sequences, is the permutation flow shop schedule (PFSS) and the FSS in which the job sequences can be different, is called non-permutation flow shop schedule [1]. In production layout of PFSS, there are sequential *m* machines and *n* jobs following the same route. All jobs are processed in the same order in machines.

One of the most preferred objective functions in solving PFSS problems is to minimize the makespan. This problem is a NP-Complete problem [2] and it is not possible to get an optimal solution in polynomial time in big problems even if the optimal result is reached in small problems. Therefore, many heuristic and metaheuristic algorithms are used to solve the PFSS problem such as genetic algorithm, particle swarm optimization (PSO), tabu search, ant colonies optimization, scatter search, simulated annealing algorithms and so on. In the past studies, it is seen that the most preferred algorithm is genetic algorithm [3]. On the other hand, hybrid methods are also preferred to increase the success of the algorithms. For example, Lin et al. 2015; Lin and Ying, 2009; Tseng and Lin, 2009; Xie et al. 2014, Zobolas et al. 2009 have proposed different hybrid algorithms [4 - 8]. However, the success of the algorithms may change according to the problem characteristics. PSO algorithm is one of the most commonly used algorithms among metaheuristics. The PSO algorithm is a very successful algorithm due to the speed parameter and storing past best value information feature. Therefore PSO has taken place in the literature as an effective method in global search. However, sometimes it may be delayed to achieve results due to oscillations in local search. Therefore, local search capability is slightly less successful than global search capability.
The firefly algorithm (FA) does not include the speed parameter and storing past best value information feature. Therefore, FA is less affected by local search and can find a more suitable solution. In this way, FA performs a successful local search. However, sometimes it can reach a global result in a longer period than PSO.
The hybrid firefly and particle swarm optimization algorithm (HFPSO) proposed by Aydilek was created by combining firefly and particle swarm optimization (PSO) algorithms [9]. It has been shown that obtain a more successful algorithm by the author. HFPSO algorithm is used by combining search capabilities of FA and PSO algorithms.

Different swarm sizes and number of iterations can be preferred in scheduling studies, using PSO algorithm. In this study, the effect of swarm size and number of iteration parameters on the HFPSO algorithm were investigated. In the next section, the research methodology is explained. Then the results are discussed. At the end of the study conclusions and future comments are shared.

Research Methodology

9 different problems consisting of "20; 50; 100" jobs and "5; 10; 20" machines combination, from Taillard's FSS problems were used as data set [10]. The swarm size and number of iteration parameters were tested at the levels of "20; 40; 60; 80; 100; 120; 150; 200" particles and "500; 1,000; 5,000; 10,000; 25,000" iterations respectively. Each trial was run on a Intel Core i5-4570 pc (4 GB Ram) with 10 replicates.

The average relative deviation (ARD) value of each problem from the upper bound of the Taillard problem was calculated for the evaluation of the results [11]. At the Eq. (1) *CCd* is the calculated makespan value; *CUB* is the best known solution or the upper bound of the corresponding Taillard problem.

$$ARD = \frac{(C^{Cd} - C^{UB})}{C^{UB}} \times 100 \qquad \text{Eq. (1)}$$

Results and Discussions

The mean ARD value decreased as the swarm size increased, compared to the average of repeated trials at different levels of parameters in nine different problems (Figure 1). Similarly, when the number of iterations increased, the mean ARD value decreased. The rate of decrease, due to the number of iterations is higher than the swarm sizes. Especially when the number of iterations increased from 1,000 to 5,000 the ARD value decreased significantly.

In Figure 2, the interaction between the parameters of swarm size and number of iteration according to the ARD values can be seen. It can be said that there is no interaction between these two parameters. Almost similar conditions occurred at different levels of swarm size and mean ARD values decreased as the number of iterations increased. Another result that can be interpreted from this figure is that the contribution to the result in cases where the swarm size is greater than 100 is relatively limited. Larger improvements are achieved when increasing from 20 particles to 100 particles.

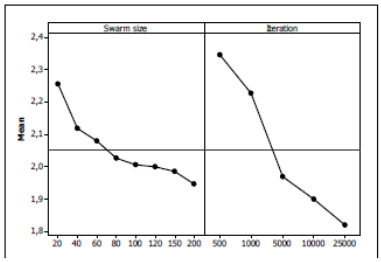


Figure 1. Main effects plot for ARD

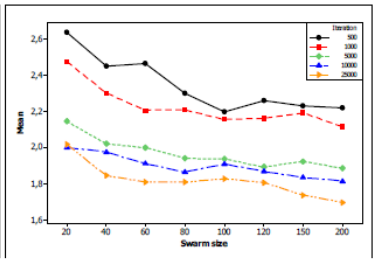


Figure 2. Interaction plot for ARD

Table 1 shows the maximum and minimum ARD values obtained in nine problem types. Especially as the number of machines increases, ARD values increase because the problem becomes more complicated. Although the effect of different parameter levels varies according to the problem, there are differences up to about 146% between maximum and minimum ARD values.

Table 1. Maximum and minimum ARD values for problem type

Problem	Job	Machine	Max ARD	Min ARD
1	20	5	1,47	0,00
2	20	10	2,65	1,53
3	20	20	2,71	1,10
4	50	5	0,08	0,00
5	50	10	5,14	3,61
6	50	20	6,98	5,03
7	100	5	0,36	0,34
8	100	10	1,17	0,99
9	100	20	3,34	3,18

Conclusions

As a result, it is observed that the parameters examined can change the performance of HFPSO algorithm up to 146% depending on the problem type. The number of iteration is more effective than the swarm size. In the trials, even in the highest number of iterations cpu time did not exceed 30 min. at the largest FSS problem. These cpu times are acceptable as solution times of FSS problems. On the other hand, for the swarm size parameter, the improvement rate of ARD value is higher, up to 100 particles. Similarly, the improvement rate of ARD value is very high, up to 5,000 iterations but then remains low for the number of iterations parameter. In the future research, adaptive variable swarm size and number of iteration parameters can be studied depending on the problem size.

Acknowledgements

This work is a part of the project which is supported by The Scientific and Technological Research Council of Turkey (TUBITAK). The authors thank the TUBITAK for financial support of this work under the grant number 118E355.

References

[1] Pinedo, M. (2002). *Scheduling: Theory, Algorithms and Systems*. Prentice Hall, New Jersey, second edition.
[2] Rinnooy Kan, A. H. G. (1976). *Machine Scheduling Problems: Classification, Complexity and Computations*. Martinus Nijhoff, The Hague, The Netherlands.
[3] Rossit, D.A., Tohmeb, F. and Frutos, M. (2018). The non-permutation flow-shop scheduling problem: A literature review". *Omega*, 77, 143–153.
[4] Lin, Q., Gao, L., Li, X. and Zhang, C. (2015). A hybrid backtracking search algorithm for permutation flow-shop scheduling problem. *Comput. Ind. Eng.*, 85, 437–446.
[5] Lin, S.-W. and Ying, K.-C. (2009). Applying a hybrid simulated annealing and tabu search approach to non-permutation flowshop scheduling problems. *Int. J. Prod. Res.,* 47 (5), 1411–1424.
[6] Tseng, L.-Y. and Lin, Y.-T. (2009). A hybrid genetic local search algorithm for the permutation flowshop scheduling problem. *Eur. J. Oper. Res.,* 198 (1), 84–92.
[7] Xie, Z., Zhang, C., Shao, X., Lin, W. and Zhu, H. (2014). An effective hybrid teaching-learning-based optimization algorithm for permutation flow shop scheduling problem. *Adv. Eng. Softw.,* 77, 35–47.
[8] Zobolas, G.I., Tarantilis, C.D. and Ioannou, G. (2009). Minimizing makespan in permutation flow shop scheduling problems using a hybrid metaheuristic algorithm. *Comput. Oper. Res.*, 36 (4), 1249–1267.
[9] Aydilek, İ. B. (2018). A hybrid firefly and particle swarm optimization algorithm for computationally expensive numerical problems. *Applied Soft Computing*, 66, 232-249.
[10] Taillard E. "Benchmarks For Basic Scheduling Problems", http://mistic.heig-vd.ch/taillard/problemes.dir/ordonnancement.dir/ordonnancement.html (04.02.2019).
[11] Wang, H., Wang, W., Sun, H., Cui, Z., Rahnamayan, S. and Zeng, S. (2017). A new cuckoo search algorithm with hybrid strategies for flow shop scheduling problems". *Soft Computing*, 21, 4297–4307.

Solution of Exam Supervisor Assignment Problem to Final Exams by Goal Programming

A. Çelik[1], H. N. Alp[1], S. Kaya[1], İ. H. Karaçizmeli[1]

[1]*Harran University, Şanlıurfa, Turkey, aysecelik@harran.edu.tr, handenuralp@harran.edu.tr, serkankaya@harran.edu.tr, hkaracizmeli@harran.edu.tr*

Abstract

In this study, 34 exam supervisors from 6 different departments of Harran University Faculty of Engineering were assigned to 170 final exams. Since this problem has more than one objective, goal programming is used as solution method. When performing these assignments, taking into account the special requests of the exam supervisors, the supervisor in each department has been assigned in an equal number of exams in its own department and in such a way that it will not be assign one after another as much as possible.

Introduction

Productivity is one of the most important requirements of our era and it is the concept that we use continuously in many sectors. It is very important to use time efficiently and to perform the work done in the most efficient way. Scheduling works for personnel in business life are also aimed at this. Scheduling is a method of decision making to keep productivity at the top when it is done in a fair and correct manner, considering the wishes of the personnel.

In the second part of the study, Goal Programming, in the third part, literature review, in the fourth part, application and in the fifth and last part, results and suggestions are given.

Goal Programming

In the studies that have been done and are being done to date, the complexity of the event makes the solution of the problem as difficult as possible. In this direction, multi-criteria decision making method is used as the most appropriate method. In solving the curret situation, it may be desirable to realize more than one goal at the same time. In this case, one of the most important multi-criteria decision-making methods, which is the most important goal programming method is preferred.

Literature Review

Varlı vd. (2017), have carried out a study with the aim of equal and fair assignment among the supervisors of midterm and final exams in Kırıkkale University Faculty of Engineering [1]. Varlı ve Eren (2017), established a goal programming model to ensure that nurses are balanced and fairly appointed as needed for shifts in a hospital in Kırıkkale [2]. Özçalıcı (2017), has developed an algorithm that prevents supervisors from being assigned to several exams at the same time and has an equal number of assignments between them. This algorithm is designed to solve even large-scale problems [3].

Application

In this study, the supervisors of the Faculty of Engineering at Harran University were assigned to the final exams for Autumn Period. Scope of the study, a total of 6 departments, 34 supervisors and 170 final exams were modeled and solved with goal programming.

It is considered that each supervisor can only take exams in his / her department. In addition, it is aimed that each department should assign an equal number of examiners among their supervisors and not to take 2 consecutive exams as far as possible for each supervisor. The special requests of the supervisors were also taken into consideration. . In the Table 1 it's given the datas about problem.

Table 1. Datas on the problem

Department	Supervisor Code (i)	Exam Code (j)	Required Assignment	Average Assignment for Each Supervisor
Computer	1,2,3,4,5	1,2,……,27	52	4 and 5.supervisors 5, others 14
Environment	6,7,8,9	28,29,……,55	32	8
Electric-Electronics	10,11,12,13,14,15,16	56,57,……,86	60	9
Map	17,18,19	87,88,……,97	20	7
Civil	20,21,22,23,24,25,26	98,99,……,128	115	16
Machine	27,28,29,30,31,32,33,34	129,130,……,170	97	33 ve 34.supervisors 7, others 14

Since the 4th and 5th supervisors in the computer department and the 33rd and 34th supervisors in the machine department are lecturers, they have fewer appointments than the research assistants. When writing the exam codes, numbering was made for each department considering the priority order in the exam program. For example, in the computer department, 1 code is given to the first exam during the exam week and 27 code is given to the last exam.

Indices

i : supervisors *(i=1,2,……..,34)*

j : exams *(j=1,2,……..,170)*

Parameters

$y_{ij} = \begin{cases} 1, \text{i. supervisor can be assigned to the j. exam} \\ 0, \text{i. supervisor can't be assigned to the j. exam} \end{cases}$

t_j : Number of supervisors required for j.exam.

Decision Variables

$x_{ij} = \begin{cases} 1, \text{i. supervisor assigned to the j. exam} \\ 0, \text{other events} \end{cases}$

$d1_{ij}^{+}$: Positive deviation from the goal of assigning an equal number of exams for i.supervisor

$d1_{ij}^{-}$: Negative deviation from the goal of assigning an equal number of exams for i.supervisor

$d2_{ij}^{+}$: Positive deviation from the goal of not taking 2 consecutive exams for i.supervisor

$d2_{ij}^{-}$: Positive deviation from the goal of not taking 2 consecutive exams for i.supervisor

Equation (1) aims to minimize deviation from the specified targets as much as possible.

$$MinZ=\sum_{i\in I}\sum_{j\in J} d1_{ij}^{+} + d1_{ij}^{-} + d2_{ij}^{+} \qquad \text{(Eq. 1)}$$

Goal 1: To conduct an equal number of exams between the supervisors of each department, Equation (2) and (3) computer, equation (4) environment, equation (5) electrical-electronics, equation (6) map, equation (7) construction, equations (8) and (9) targeted total for machine supervisors shows the number of assignments.

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 14 \qquad \text{i=1,2,3} \qquad \text{(Eq. 2)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 5 \qquad \text{i=4,5} \qquad \text{(Eq. 3)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 8 \qquad \text{i=6,7,8,9} \qquad \text{(Eq. 4)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 9 \qquad \text{i=10,11,12,13,14,15,16} \qquad \text{(Eq. 5)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 7 \qquad \text{i=17,18,19} \qquad \text{(Eq. 6)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 16 \qquad \text{i=20,21,22,23,24,25,26} \qquad \text{(Eq. 7)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 17 \qquad \text{i=27,28,29,30,31,32} \qquad \text{(Eq. 8)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 7 \qquad \text{i=33,34} \qquad \text{(Eq. 9)}$$

Goal 2: Supervisors do not take 2 consecutive exams, Equation (10) is the constraint that prevents supervisors from entering 2 consecutive examinations as far as possible. 7. supervisor is exempt because she/he wanted to repeatedly enter some exams.

$$x_{ij} + x_{ij+1} - d2_{ij}^{+} + d2_{ij}^{-} = 1 \qquad \forall i \in I/\{7\}, \forall j \in J/\{170\} \qquad \text{(Eq. 10)}$$

Equation (11) allows each supervisor to be assigned to the exam only in his or her department.

$$x_{ij} \le y_{ij} \qquad \forall i,j \qquad \text{(Eq. 11)}$$

Equation (12) provides the number of supervisors required for each exam.

$$\sum_i x_{ij} = t_j \qquad \forall j \qquad \text{(Eq. 12)}$$

Equation (13) shows that each supervisor can be assigned to a maximum of one more exam from the specified target. Equation (14) shows that no less than one exam assignment can be made.

$$\sum_j d1_{ij}^{+} \le 1 \qquad \forall i \qquad \text{(Eq. 13)}$$

$$\sum_j d1_{ij}^{-} \le 1 \qquad \forall i \qquad \text{(Eq. 14)}$$

Equations (15), (16) and (17) indicate that the 8th supervisor should take the 28th, 31st and 34th exams in the environmental engineering department. Equations (18) and (19) show that the 18th supervisor in the map engineering department cannot take the 90th and 92th exams. Equation (20) shows that at least one of the 11th and 14th supervisors should be assigned to the 61st exam in the department of electrical and electronics engineering. Equation (21) shows that the 15th supervisor should take the 70th exam. Equation (22) shows that at least one of the 12th and 16th supervisors should be assigned to the 71st exam. Equations (23), (24), (25) and (26) indicate that the 20th supervisor in the civil engineering department cannot take the 117, 118, 119 and 120 exams. Equations (27), (28), (29) and (30) indicate that the 7th supervisor in the environmental engineering department wants to take the 47th, 48th, 49th and 50th exams one after the other.

$$x_{11\,61} + x_{14\,61} \geq 1 \quad \text{(Eq. 20)}$$
$$x_{15\,70} = 1 \quad \text{(Eq. 21)}$$
$$x_{12\,71} + x_{16\,71} \geq 1 \quad \text{(Eq. 22)}$$
$$x_{20\,117} = 0 \quad \text{(Eq. 23)}$$
$$x_{20\,118} = 0 \quad \text{(Eq. 24)}$$
$$x_{20\,119} = 0 \quad \text{(Eq. 25)}$$
$$x_{20\,120} = 0 \quad \text{(Eq. 26)}$$
$$x_{7\,47} = 1 \quad \text{(Eq. 27)}$$
$$x_{7\,48} = 1 \quad \text{(Eq. 28)}$$
$$x_{7\,49} = 1 \quad \text{(Eq. 29)}$$
$$x_{7\,50} = 1 \quad \text{(Eq. 30)}$$
$$x_{8\,28} = 1 \quad \text{(Eq. 15)}$$
$$x_{8\,31} = 1 \quad \text{(Eq. 16)}$$
$$x_{8\,34} = 1 \quad \text{(Eq. 17)}$$
$$x_{18\,90} = 0 \quad \text{(Eq. 18)}$$
$$x_{18\,92} = 0 \quad \text{(Eq. 19)}$$

Equation (31) and (32) are sign constraints of decision variables.

$$x_{ij} \in \{1,0\} \qquad \forall i,j \qquad \text{(Eq. 31)}$$
$$d1_{ij}^{+}, d1_{ij}^{-}, d2_{ij}^{+}, d2_{ij}^{-} \geq 0 \qquad \forall i,j \qquad \text{(Eq. 32)}$$

The solution of the problem was realized by coding the mathematical model in GAMS 24.0.2 application on an 8GB Ram Memory computer with Intel (R) Core (TM) i7-6500U CPU @ 2.50 GHz. Table 2 presents the comparison of the current situation with the solution obtained.

Table 2. Comparative results (some part)

	Current State				Recommended Solution			
Supervisor	Total Assignment	1.Positive deviation from target	1.Negative deviation from target	2.Positive deviation from target	Total Assignment	1.Positive deviation from target	1.Negative deviation from target	2.Positive deviation from target
1	14	0	0	11	14	0	0	2
2	14	0	0	9	14	0	0	2
3	14	0	0	5	14	0	0	2
4	3	0	2	1	5	0	0	0
5	7	2	0	0	5	0	0	0

Results And Suggestions

When the optimal solution obtained was compared with the current situation, the positive deviation rate from the target assignment amount for each supervisor was reduced from 22 to 3, while 86% improvement was achieved while the negative deviation rate was reduced from 24 to 5, resulting in approximately 79% improvement. The positive deviation rate of the supervisors from 2 consecutive examinations was reduced from 173 to 54, resulting in an improvement of 69%. In subsequent studies, besides the supervisor assignments can also be made of scheduling the exam or supervisor assignments can be made by considering both midterm and final exams.

References

[1] Varlı, E., Alağaş, H. M., Eren, T., and Özder, E. H. (2017). Sınav görevlisi atama probleminin hedef programlama yöntemiyle çözümü. *Bilge International Journal of Science and Technology Research,* 1(2), 105-118.

[2] Varlı, E., and Eren, T. (2017). Hemşire çizelgeleme problemi ve hastanede bir uygulama. *Academic Platform-Journal of Engineering and Science*, 5(1), 34-40.

[3] Özçalıcı, M. (2017). Sınavlara Gözetmen Atama Probleminin Çözümü İçin Takas Bazlı Bir Algoritma Önerisi. *İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(2), 492-506.

Investigation of Metrics Used to Measure the Distance Between Clusters in Hierarchical Clustering via WEKA

S. Sevimli Deniz[1]

[1]Van Yüzüncü Yıl Üniversitesi, Van, Turkey, sdeniz@yyu.edu.tr

Clustering analysis is an analysis used to group observations. In this analysis using distance and similarity measures, two different methods are used [8]. These; hierarchical clustering and non-hierarchical clustering methods [11]. There are basically two kinds of hierarchical clustering methods. These are referred to as combiner and separator clustering methods [10]. In the combining methods, each observation is initially evaluated as a separate set. In the next steps, the closest clusters are combined. At each step, the number of clusters decreases by one, this process continues until all the data is collected in a cluster. In the case of separating methods, this process works in reverse [9]. At the beginning, there is only one set and different data is separated from each other until each data continues to form a cluster. The maximum number of combinator clustering methods is used. There are different metrics that are used to create clusters. These can be listed as Single, Complete, Average, Mean, Centroid, Ward, AdjComplete and Neighbor Joining [12].

In this study, different metrics used to measure the distance between clusters will be examined in hierarchical clustering. Using WEKA, a data mining tool, the success of different clustering metrics in the same data set will be evaluated [11].

Clustering

Clustering is a data analysis technique aimed at identifying hidden group schemes contained in data sets. Clustering analysis is one of the main data analysis methods that helps to define natural grouping in a range of data elements [8]. Data clustering refers to an unsupervised learning technique that divides a data set into a series of discrete or overlapping groups, providing refined and more abstract images to the clustered structure. Clustering refers to the grouping of objects in data groups so that objects of the same group are similar to objects in other groups. It is divided into three. These; Hierarchical clustering is density-based clustering and division-based clustering [2].

Hierarchical Clustering

In hierarchical clustering, the hierarchy of objects is created. There are two types of hierarchical clustering: Consolidator (bottom-up) and separator (top-down). In combinatorial clustering, we start with a data object and, in the divisive technique, gradually build the cluster; We start with the entire data set, and then divide the data objects into clusters. The combining technique consists of the following steps [2].

Step 1: Assign each data object to a set so that each object is associated with only 1 set. If we have N data objects, N sets are created, each containing 1 data object.
Step 2: Locate the nearest cluster pair and merge them to form a pair so that the number of clusters is N-1.
Step 3: Calculate the distance between the new cluster and each of the old ones.
Step 4: Repeat steps 2 and 3 until all data objects are clustered in N size [1].

When performing step 3, two different methods are used, namely the connection technique and the metric technique. Connection techniques indicate how the distance between two sets is measured, while the metric technique shows how the distance between two data objects is measured [12]. The connection technology is further divided into three broad categories: single connection technology, complete connection technology and average connection technology. The average connection takes into account the average distance between any object of a set and any object of the second set [3].

The second technique that calculates the distance is the metric technique. It can be applied in many ways, but Manhattan distance and Euclidean distance are the most commonly used techniques [6].

$$d = \sum_{i=1}^{n} |x_i - y_i| \qquad \text{(1) Manhattan}$$

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{(2) Öklid}$$

The main methods used to calculate the distance between two clusters are: Single, Complete, avarage, mean, Centroid, Ward method, NeighborJoining and AdjComplete)[7].

Single-link: There is a single link distance and minimum link, which is the closest distance between any element in the first set and any element in the second set [9]. In this technique, the two closest units are placed in the cluster. The next closest distance is then determined and this observation is added to the first cluster or a new two observation sets are created[11].

Complete-link: Finds the maximum link, the full link distance, which is the largest distance between any item in the first set and any item in the second set [3]. This method is also known as the farthest neighborhood. This technique is called full connection technique, because all units in the cluster are connected to each other with maximum distance or minimum proximity [6].

Average-link: Finds the average distance between the elements of two sets. In average connection technique, clusters are connected to each other with small differences [7].

Mean: Calculates the average distance of a combined set [13].

Center (centroid) connection: Finds the distance of the center of the cluster The center of the first cluster (mean vector with p elements) and the distance between the center of the second cluster are calculated [11].

$$U(k_1,k_2)=\frac{1}{|k_1|}\frac{1}{|k_2|}\sum_{x_1\in k_1}\sum_{x_2\in k_2}U(x_1,x_2) \qquad (3)$$

Ward method: Finds the change distance caused by the formation of clusters. In Ward method, the sum of the squares in the group is processed instead of group connections [13]. The method starts with n sets, each with a single unit. Since the first step of each method is a cluster of error squares sum is zero [4]. At each stage, two subsets are combined to form the next level. The Ward method is sensitive to contradictory points [5].

Neighbor joining method: uses the neighbor joining algorithm. This method, which is more complex than other methods, includes the following steps. All pairwise distances are calculated. Select two elements i and j with relative distances of minimum. The rows and columns for the selected elements (i and j) are deleted from the offset matrix. The distance of the new parent element(k) to represent the selected elements is calculated and added to the distance matrix. Repeat until the two elements remain.

Adjusted complete-link: It is obtained by subtracting the distance between the two farthest elements of the two clusters, which is greater than the intra-cluster distance values of the clusters. In other words, it is calculated by subtracting the larger intra-cluster distance values from the intra-cluster distance value [10].

Method

Weka

It is an open source software written in JAVA which is one of the Data Mining tools. Weka is a data mining tool that consists of a series of machine learning algorithms. Preprocessing, classification, regression, clustering, association rules and visualization of data can be performed with Weka.

Data sets

The data used in our experiment are real world data from the UCI repository. During the evaluation, multiple data sizes were used, each data set attribute types, the number of samples stored in the data set, as well as the table, indicate that all selected data sets were used for the clustering task. These data sets were chosen because they have different characteristics and are intended for different fields [14].

Table 1. Description of Databases Used for Trial

Database Name	Number of attributes	Sample number
Diabet	9	768
ionosphere	35	351
iris	4	150
Germancreditcard	21	1000

Results

Table 2. below shows the experimental results obtained when comparing clustering algorithms.

Method	False Clustered Examples			
	Diabet	ionosphere	İris	Germanc.card
Single	267.0 %34.76	125.0 % 35.61	51.0 % 34	301 %30.1
Complete	268.0 %34.89	123.0 % 35.04	18.0 % 12	314 %31.4
Average	267.0 %34.76	125.0 %35.61	**17.0 % 11.33**	**300 %30**
Mean	286.0 %37.23	128.0 %36.46	27.0 %18	458 %45.8
Centroid	267.0 %34.76	125.0 %35.61	51.0 %34	301 %30.1
Ward	**251.0 % 32.68**	**102.0 %29.05**	25.0 %16.6	475 %47.5
AdjComplete	267.0 %34.76	125.0 %35.61	51.0 %34	314 %31.4
Neiphbor-joining	268.0 % 34.89	126.0 %35.89	100.0 %66.6	478 %48

Discussion and Conclusion

Data mining techniques cover every aspect of our lives. Before working on data mining models, it is very important to be familiar with the basic algorithms available. There are many ways to calculate distance in combinatorial hierarchical clustering. If we evaluate the different connection methods used in hierarchical clustering studies according to this research; When we look at the wrong clustering rates of Ward and Avarage methods, it is seen that their performance is better than the others. Again, Mean and Neighbor-joining methods showed lower performance than the others. The selected method affects the result. Therefore, it is necessary to evaluate the data structures well and select the appropriate methods. Groups or clusters can always be created, even if there is no predefined structure in the cluster. The results of clusters should not be generalized.

Referances

[1] Aastha Joshi ve Rajneet Kaur, 2013. "A Review: Comparative Study of Various Clustering Techniques in Data Mining" IJARCSSE.

[2] Andrew Moore: "K-means and Hierarchical Clustering – Tutorial Slides" http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html. Erişim tarihi: 14.05.2019.

[3] Chatfield, C. and Collins, H. 1980. J.A.Introduction to Multivariate Analysis, Chapmanand Hall, p.224

[4] Everitt, B.,1974. Cluster Analysis. Heinmann.London, p.122

[5] Everitt, B.S.,1979.Unresolved problems in cluster analysis, Biometrics, 35, p. 169-181

[6] Green, E.P., 1989. Analysing Multivariate Data, Philadelphia, p.427

[7] Hubert, L.,1974. "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarcihal Clustering Procedures", Journal of the American Statistical Association,69, 698-704.

[8] Jiawei Han M. K., 2006. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier.

[9] Johnson, A.R. ve Wichern, D.W., 1988. Applied Multivariate Statistical Analysis. Prentice-Hall International Editions. New Jersey, p.554.

[10] Kamvar, S.,Klein, D., Manning, C., 2002. "Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach", *19th International Conference on Machine Learning (ICML 2002)*, 8-12 Temmuz 2002, Sydney, Australia, ss. 283-290.

[11] Murtagh, F.,Contreras, P., 2017. "Algorithms for hierarchical clustering: an overview II", *WIREs Data Mining and Knowledge Discovery*, Cilt 7, Sayı 6, ss. 1-16.

[12] Sharma, A.,Jaloree, S., Thakur, R.S., 2018. "Review of Clustering Methods: Toward Phylogenetic Tree Constructions", *International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks and Systems*, Cilt 34, ss. 475-480.

[13] T. Hayasaka, N. Toda, S Usui, K Hagaiwara, 1996. "least square error and prediction square error of function representation with discrete variable basis", Proceedings of the 1996 IEEE Signal Processing Society Workshop.

[14] UCI Machine Learning Repository. https:// archive.ics.uci.edu/ml/datasets.html, ziyaret tarihi: 16 Mayıs 2019.

Analysis of the Results of Artificial Intelligence Perception Research via Tableau Data Visualization Tool

S. Sevimli Deniz[1], N. Genç[1]

[1]Van Yüzüncü Yıl Üniversitesi, Van, Turkey, sdeniz@yyu.edu.tr, nacigenc@yyu.edu.tr

Data visualization is a graphical representation of information and data. Using visual elements, such as graphs and maps, data visualization tools provide an easy way to see and understand trends, patterns and patterns in data. In the Big Data world, data visualization tools and technologies are crucial to analyzing large amounts of information and making data-driven decisions [2]. The purpose of data visualization is to establish flexible, creative, bridges between the human perception system and computer systems. Schematic structures can be abstracted when creating visual representations of data. Visual elements such as tables and graphics can be used to provide a clear flow of information. Thus, cognitive processes such as comparison, interpretation and analysis can be performed more effectively [3]. Thanks to visualization techniques, it is possible to show the bulk of the data in a single area. This allows an analysis to be carried out, taking into account human perception abilities and interpersonal interpretation differences. Data visualization techniques provide a general overview of the data, and it may be possible to discover hidden small patterns that may be important during the analysis. Through data visualization, relationships such as distributions of variables, clusters between variables, correlations can be revealed [4].

The ability of machines to program and display intelligent behavior is known as artificial intelligence and is accepted. The aim of Artificial Intelligence perception research is to determine the point of view of artificial intelligence in various age and education groups [1]. The survey was conducted with 225 people. In this study, the results of this survey were reported with Tableau, a data visualization tool.

Data Visualization

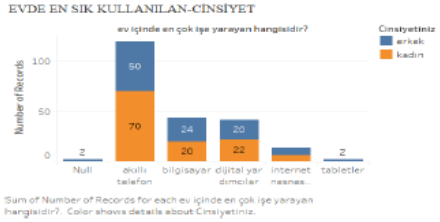
Today, it has become increasingly difficult to analyze high-dimensional data. Data visualization and visual data mining is a tool that helps to deal with big data. There are many information visualization techniques involved in the data mining process. In this study, data visualization and visual data mining techniques are presented [4].
In the Data Visualization process, the next step after completing the data and questions is to decide on visualization. Graphs and charts reduce large amounts of data to simple and easy to understand formats. The purpose of the graphs in the dashboard is to compare data, show relationships, or determine if a trend exists. It is important to select the correct graph from the available options. There are four categories defined for data presentation[1].

Comparison, Consolidation, composition, Deployment, Relationship / Trend

It is known that visual data mining techniques have high potential in exploratory data analysis. Visual data discovery is used especially when little is known about the data and discovery targets are uncertain[3].

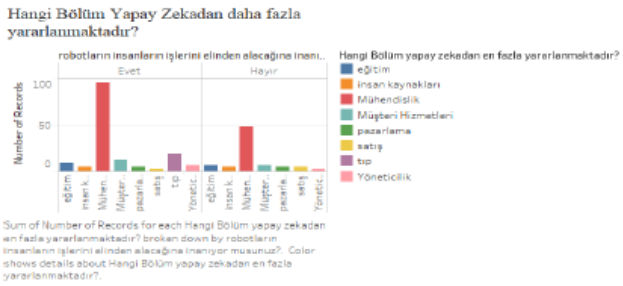
Today data visualization is also used in many techniques. The most commonly used of these techniques are:

1. Bar Graph: A bar graph or bar graph is a graph or graph that shows categorical data with rectangular bars with heights or lengths proportional to the values they represent. The bars can be drawn vertically or horizontally. A vertical bar chart is sometimes called a line chart. A bar chart is a good way to display categorical data.



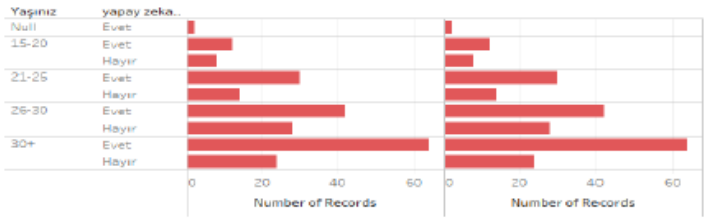
2. Line graphs reveal trends or progressions over time. A good way to visualize a continuous data set or a set of values. It is the most appropriate method to analyze trend-based data and rate of change over time.The

values are drawn on the line chart and the data points are connected to show a trend. Multiple trends can be highlighted and compared by drawing lines of different colors. Used to compare data between categories[5].



3. Pie charts; used to represent a data composition, typically used to represent numbers as a percentage of information or proportions. The sum of all ratios is 100%.

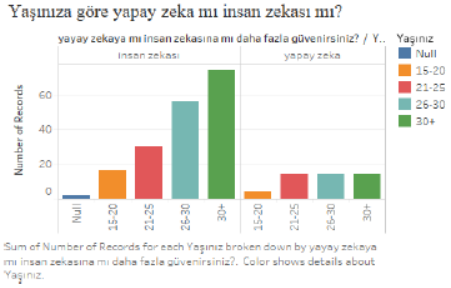
4. Distribution charts; mostly used in correlation and distribution analysis. This is a type of chart that helps determine whether there is a relationship between two variables. It is an effective visual tool to show trends, concentrations and outliers in the distribution of data.



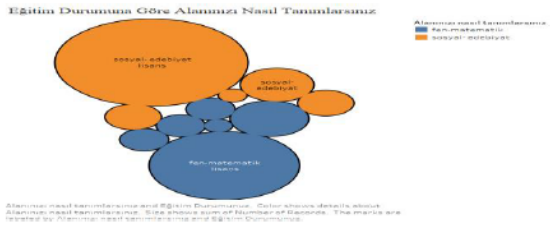
5. Heat maps; mostly used for information comparison. Provides an activity level or rating information (eg High to Low, Strongest and Poorest, Perfect to Poor), all displayed using different colors.

6. The histogram graph is used to see how data is distributed among groups. This is different from a bar graph. Like a bar chart, a histogram consists of columns, but there is no space between columns. Histograms provide continuous data, while the bar graph provides categorical data (data that fits categories).

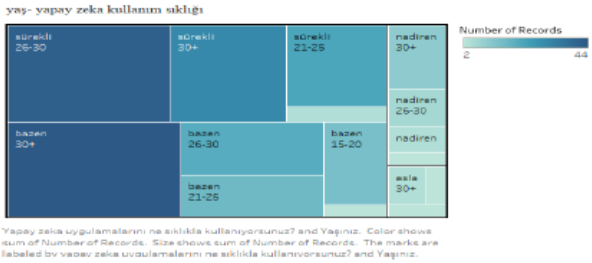
7. Scatter graph: The scatter graph is a two-dimensional visualization of data representing the points representing values obtained for two different variables, one drawn along the x axis and the other drawn along the y axis.



8. Packed bubbles: Use packed balloon graphics to display data in a circle set. Dimensions define individual bubbles, and dimensions define the size and color of individual circles.



9. Treemaps: Use tree maps to display data in nested rectangles. You use dimensions to define the structure of the tree map and measures to define the size or color of individual rectangles. Treemaps is a relatively simple data visualization that can provide insight in a visually appealing format.



Discussion and Conclusion

Data visualization is the presentation of analysis results of abstract information by creating graphs, diagrams, tables, images or animations. The aim of visualization is to make complex data of statistical and variable information presented in classical format easy to understand with easy to understand graphical interfaces. The human perception system is limited to 3 dimensions and higher dimensional data structures go beyond the limits of human perception. Data visualization is a very new and promising field in computer science. Uses computer graphic effects to reveal patterns, trends, and relationships in data sets. Today, thanks to developing technology, large and complex data sets can be analyzed easily. When we used the sampling method when we used to encounter large and complex datasets, we can now process all the data in a simple way.

In order to successfully implement data mining algorithms, visualization and interaction capabilities are crucial factors, as it allows the user to use domain information, guide the data mining process, and better understand the results. The aim of this study is to investigate the results of artificial intelligence perception research with visual data mining method.

References

1. R. Agarwal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining, pages 307–328, 1996.
2. M. Ankerst, M. Breunig, H. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. Proc. ACM SIGMOD ’99, Int. Conf on Management of Data, Philadelphia, PA, pages 49–60, 1999.
3. M. Ankerst, M. Ester, and H. Kriegel. Towards an effective cooperation of the computer and the user for classification. SIGKDD Int. Conf. On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, pages 179–188, 2000.
4. M. Ankerst, D. A. Keim, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In Proc. Visualization ’95, Atlanta, GA, pages 279–286, 1995.
5. M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. In Visualization ’96, Hot Topic Session, San Francisco, CA, 1996.

Bayesian Estimation of the Reduced Kies Distribution Parameters

I . Usta[1], M. Akdede[2]

[1]Eskisehir Technical University, Eskisehir, Turkey, iusta@eskisehir.edu.tr
[2] Usak University, Usak, Turkey, merve.akdede@usak.edu.tr

Abstract

Kies distribution (KD) was first introduced by [1] as a function of the Weibull distribution. The KD has become a good alternative to other extensions of Weibull distribution, since it has a decreasing, increasing and bathtub-shape hazard rate functions [2]. In this study, we propose the Bayesian estimation of the two unknown parameters of the reduced KD which a special case of the KD. Using the independent gamma priors for both parameters, Bayesian estimators of the parameters are obtained under the squared error and linear exponential loss functions. Since the conditional posterior functions cannot be reduced analytically to the well-known distributions, a hybrid Markov chain Monte Carlo (MCMC) method is used for computing the Bayesian estimates of the parameters. Then, the corresponding maximum likelihood estimators (MLEs) of the parameters are obtained and the performances of the estimators are compared through a Monte Carlo simulation study. Finally, a real data set is analyzed for the illustrative purposes.

Introduction

Kies Distribution (KD) was considered by [1] as a functional form of the Weibull distribution. KD has a decreasing, increasing and bathtub-shape hazard rate function like the other extended Weibull models such as the generalized Weibull (GW), the modified Weibull (MW), the beta Weibull (BW) distributions, etc. Due to flexibility feature of the bathtub shaped hazard rate function, it becomes a better alternative than the other extended Weibull distributions for modeling the lifetime data sets [2].Kumar and Dharmaja [2] also studied the properties of a special case of the KD, called as the reduced Kies distribution (RKD), and its applications in the engineering and medical sciences.

In this paper, we consider the parameter estimation for the two-parameter of the RKD, denoted as $RK(\lambda,\beta)$. The probability density function (pdf) of the RKD is given by

$$f(x;\lambda,\beta) = \frac{\lambda\beta x^{\beta-1} exp\left\{-\lambda\left(\frac{x}{1-x}\right)^{\beta}\right\}}{(1-x)^{\beta+1}} \qquad (1)$$

where $0\leq x\leq 1$, $\lambda>0$ and $\beta>0$.

This study aims to obtain the Bayesian estimators and MLEs of the parameters. In the Bayesian analysis, we used the symmetric squared error loss function (SELF) and the asymmetric general entropy loss function (GELF) under the assumption of the informative gamma prior for the unknown parameters. Since the joint posterior density is not in tractable form, we employ a hybrid MCMC method (Gibbs within Metropolis-Hastings algorithm) to compute the approximate Bayes estimates of the parameters.

The rest of the paper is organized as follows: Frequentist and Bayesian estimators of the unknown RKD parameters are obtained in Section 2 and Section 3, respectively. Simulation results are given in the Section 3. Further in Section 4, concluding remarks are provided.

Maximum Likelihood Estimation

Let $X=(X_1,X_2,\ldots,X_n)$ be a random sample of size n from a population following $RK(\lambda,\beta)$, then the likelihood function is obtained as:

$$L(\lambda,\beta;\boldsymbol{x}) = \lambda^n\beta^n\prod_{i=1}^{n} x_i^{\beta-1}\exp\left(-\lambda\sum_{i=1}^{n}\left(\frac{x_i}{1-x_i}\right)^{\beta}\right)\prod_{i=1}^{n}\frac{1}{(1-x_i)^{\beta+1}} \qquad (2)$$

The log-likelihood function is

$$\ell(\lambda,\beta;\boldsymbol{x}) = n\ln\lambda + n\ln\beta + (\beta-1)\sum_{i=1}^{n}\ln x_i - \lambda\sum_{i=1}^{n}\left(\frac{x_i}{1-x_i}\right)^{\beta} - (\beta+1)\sum_{i=1}^{n}\ln(1-x_i) \qquad (3)$$

MLEs of the parameters λ and β can be obtained by solving the following equations with the help of MATLAB.

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} \left(\frac{x_i}{1-x_i}\right)^{\beta} = 0$$
$$\frac{\partial \ell}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \ln(1-x_i) - \lambda \sum_{i=1}^{n} \left(\frac{x_i}{1-x_i}\right)^{\beta} \ln\left(\frac{x_i}{1-x_i}\right) = 0 \quad (4)$$

Bayesian Estimation

In this section, Bayesian estimation of the parameters λ and β are derived. We have assumed the squared error and general entropy loss functions. The priors for the parameters are given as

$$\pi_1(\lambda) \propto \lambda^{a_1-1} e^{-b_1\lambda}, \qquad \pi_2(\beta) \propto \beta^{a_2-1} e^{-b_2\beta} \quad (5)$$

The joint posterior of λ and β is obtained as following

$$\pi(\lambda,\beta|x) = \lambda^{n+a_1-1}\beta^{n+a_2-1} e^{-\lambda\left(b_1+\sum_{i=1}^{n}\left(\frac{x_i}{1-x_i}\right)^{\beta}\right)} e^{-\beta\left(b_2+\sum_{i=1}^{n}[\ln(1-x_i)-\ln x_i]\right)} e^{-\sum_{i=1}^{n}[\ln(1-x_i)+\ln x_i]} \quad (6)$$

Since the posterior density is not in tractable form, the Bayes estimation cannot be performed explicitly. Therefore, we employ the hybrid MCMC (Gibbs within Metropolis-Hastings) method to compute the approximate Bayes estimates of the parameters.

The Gibbs within Metropolis-Hastings algorithm [3] is as follows:

1. Set $t = 0$ and initialize $\lambda^{(0)}$ and $\beta^{(0)}$
2. Set $t = t + 1$
   i. Generate $\lambda^{(t)}|\beta^{(t-1)}$ from $\pi_1\left(\lambda^{(t)}|\beta^{(t-1)}\right)$ as $G\left(n + a_1, b_1 + \sum_{i=1}^{n}\left(\frac{x_i}{1-x_i}\right)^{\beta}\right)$
   ii. Generate $\beta^*|\lambda^{(t)}$ from $\pi_2\left(\beta^{(t-1)}|\lambda^{(t)}\right)$
   - Generate $\beta^* \sim N(\beta^{(t-1)}, V_\beta)$
   - Calculate the acceptance probability:
     $$\alpha = \min\left[1, \frac{\pi_2\left(\beta^*|\lambda^{(t)}\right)}{\pi_2\left(\beta^{(t-1)}|\lambda^{(t)}\right)}\right]$$
   - Draw a random number $u \sim U(0,1)$
     If $u \leq \alpha$ then accept β^* and set $\beta^t = \beta^*$
     else set $\beta^t = \beta^{t-1}$
3. Repeat until $t = N$

Let u be a function of λ and β. Then, the approximate Bayesian estimate under SELF be calculated as follows

$$\hat{u}_{SELF} = \frac{1}{N-m}\sum_{i=m+1}^{N} u(\lambda_i, \beta_i) \quad (7)$$

and the Bayes estimates under GELF are obtained as following

$$\hat{u}_{GELF} = \left(\frac{1}{N-m}\sum_{i=m+1}^{N} [u(\lambda_i, \beta_i)]^{-k}\right)^{-1/k} \quad (8)$$

Simulation

In this section, a Monte Carlo simulation is conducted to compare the performance of the Bayes estimators of the RKD parameters. We consider the samples of size n = 25, 50, 100 from the RKD. The following parameters are used in the simulations: λ = 1.5 and β = 0.75. The comparison between the estimators with respect to the average bias and mean squared errors (MSEs) for 5000 repetitions. We also consider the loss parameter for GELF as and $Gamma(2, 2/\theta)$ as the prior. Burn in time, m, equals to $N * 0.2$ where N = 5000. The results are presented in Table 1.

Table 1. Simulation results for the parameter estimation when λ=1.5 and β=0.75

	$\hat{\lambda}_{MLE}$	$\hat{\beta}_{MLE}$	$\hat{\lambda}_{SELF}$	$\hat{\beta}_{SELF}$	$\hat{\lambda}_{GELF}$	$\hat{\beta}_{GELF}$	$\hat{\lambda}_{GELF}$	$\hat{\beta}_{GELF}$
n=25					k=-0.75		k=0.75	
Mean	1.5980	0.8028	1.5552	0.8071	1.5494	0.8051	1.5139	0.7926
Bias	0.0980	0.0528	0.0552	0.0571	0.0494	0.0551	0.0139	0.0426
MSE	0.1355	0.0210	0.0618	0.0146	0.0615	0.0144	0.0604	0.0132
n=50								
Mean	1.5503	0.7695	1.5500	0.7809	1.5466	0.7798	1.5256	0.7728
Bias	0.0503	0.0195	0.0500	0.0309	0.0466	0.0298	0.0256	0.0228
MSE	0.0596	0.0078	0.0469	0.0077	0.0465	0.0076	0.0449	0.0072
n=100								
Mean	1.5112	0.7618	1.5208	0.7699	1.5189	0.7694	1.5079	0.7659
Bias	0.0112	0.0118	0.0208	0.0199	0.0189	0.0194	0.0079	0.0159
MSE	0.0228	0.0040	0.0232	0.0041	0.0231	0.0041	0.0224	0.0040

From Table 1, Bayes estimators of the both parameters under GELF with k=0.75 performed better than the other estimators in terms of the bias and MSEs for all sample sizes. Furthermore, the Bayes estimators under asymmetric

loss function show better performances than the ones under symmetric loss function to estimate in all cases especially when the loss parameter is k=0.75 which points out that overestimation is more serious than underestimation. Results show that the MSEs decrease as long as the sample size increases.
Furthermore, the different parameter values have also been used in the simulations and the similar results are observed. However, we are not able to show the results because of the page limitation.

Conclusion

Bayesian estimators of the unknown parameters of RKD under SELF and GELF assuming the gamma priors are obtained by using the hybrid MCMC method. We also calculate the MLEs. The comparison between the estimators is made through a Monte Carlo simulation study and the results show that the Bayes estimators under GELF have the minimum bias and the MSEs for all considered cases.

References

[1] Kies, J. A. (2014).*The strength of glass performance*. Washington, D.C., 1958.
[2] Kumar, C.S., and Dharmaja, S.H.S. (2014). On some properties of Kies distribution. *Metron*, 72(1), 97–122.
[3] Tierney, L. (1991). Exploring posterior distributions using Markov chains. *Comput. Sci. Stat. Proc. Symp. Interface Crit. Appl. Sci. Comput.*, 563–570

An Analytic Hierarchy Process Example in Pharmacy Management

M. Arslan[1], N. Tarhan[2]
[1]Van Yüzüncü Yıl University, Van, Turkey, eczmirayarslan@gmail.com
[2]İzmir Katip Çelebi University, İzmir, Turkey, tarhan8840@gmail.com

Abstract

Analytic Hierarchy Process (AHP), evolved by Saaty, is an extensively used mathematical tool to make decision for multiple criteria. There are several studies in the literature using AHP in pharmaceutical sector. The aim of this study is suggesting priorities about influencing factors of pharmacists' pharmaceutical warehouse choice in Van, a city of Turkey. According to the aim of the study, AHP was used for analyzing. Firstly, priority criteria for three pharmaceutical warehouses in the city, which hold an important place in the market share, have been identified for constructing AHP. The criteria were labeled then, pair-wise comparisons of these criteria were made and consistency ratio was checked. Alternatives were lined up; first alternative is seen as the most advantageous for pharmacists. Additionally, the most important criterion was found on this choice. It is thought that these results will not only be helpful for pharmacists for making decision about the issue, but also beneficial for pharmaceutical warehouses to improve their services.

Introduction

Decision-making is a process of selecting one of possible action plans for achieving the goals. When the management functions such as planning, controlling etc. are taken into consideration, it is seen that all of them are actually based on decision making [1]. In the beginning of 1970's Saaty developed a mathematical tool to make decision for multiple criteria named as Analytic Hierarchy Process (AHP) [2,3]. AHP simplifies complex problems and enables both objective and subjective judgments to be included in the decision process. AHP realizes the importance level of decision-making criteria and orders the alternatives with pairwise comparisons [4]. Using such a modern decision support technique will provide a competitive advantage for enterprises [1]. AHP can be used in many areas such as supplier selection [5], performance evaluation [6], job evaluation [7], strategy determination [8], and personnel selection [9].

AHP has 4 axioms as reciprocity, homogeneity, independence and expectations [9]. AHP is based on pairwise comparisons in terms of significance values using a predefined comparison scale [2, 6]. The Saaty's importance scale for making comparisons is given in Table 1.

Table 1. The Saaty's Importance Scale

Importance values	Meanings of the values
1	equally important
3	little important
5	strongly important
7	very strongly important
9	extremely important
2,4,6,8	intermediate values

As in many different sectors, decision making process is very important for pharmaceutical sector. Therefore, to make right selections, AHP will be helpful for researchers. In the literature there are several studies conducted via AHP in pharmaceutical sector [10,11,12].

In this regard, the aim of this study is suggesting priorities about influencing factors of pharmacists' pharmaceutical warehouse choice in Van, a city of Turkey. To the best of authors' knowledge, this is the first study evaluating this selection process via AHP in Turkey.

Material and Method

According to the aim of the study, AHP was used for analyzing. Firstly, priority criteria for three pharmaceutical warehouses in the city (A1, A2, A3), which hold an important place in the market share, have been identified for constructing AHP. For this, primarily getting connect with warehouses and some basic information about their services were get such as product range, number of daily services, etc. Accordingly, community pharmacists' judgments were taken to evaluate priority criteria for selection process. As a result, five priority criteria

were determined and named as: communication (C1), service quality (C2), product range (C3), number of daily services (C4), and service at out of shift (C5).

After this, pair-wise comparisons of these criteria were made by taking opinions of community pharmacists and pair-wise comparison matrix of the selection criteria (Table 2) was created according to Saaty's importance scale by following rules (i=1,2,3 ; j=1,2,3,4,5) [13]:

i. $a_{i,i} = 1$, for all i values
ii. If $a_{i,j} = x$, then $a_{j,i} = 1/x$, $x \neq 0$
iii. If C_i and C_j were determined as having equal importance, $a_{i,j} = 1$ and $a_{j,i} = 1$

Lastly, for representing the relative importance of the criteria, Weighted SuperMatrix was created and for getting solution limiting matrix was derived from it by taking sensitivity of the final outcome into account.

Results

The limiting matrix was given in Table 2, which shows that the best alternative is *Alternative 1*.

Table 2. The Limiting Matrix. (W: warehouse)

	Communication	Product range	Quality of service	Number of daily services	Services at out of shift	
W	0,508	0,095	0,211	0,143	0,043	
W_1	0,600	0,649	0,747	0,655	0,655	0,646
W_2	0,200	0,279	0,119	0,290	0,290	0,207
W_3	0,200	0,072	0,123	0,055	0,055	0,147

According to Table 2, the most important criterion was found as *Communication* and *Service Quality* was followed it. The consistency ratio of the result was calculated as 0,079, which shows the solution is consistent [2].

Conclusion

The result of this study shows that communication with pharmaceutical warehouses was the most important criterion from the perspective of pharmacists. In other words, warehouses establishing a good communication network with pharmacists are more advantageous. The second important criterion was found as quality of service, and number of daily services, product range, and services at out of shift were followed it respectively. According to these results, W1 seems the most appropriate option for the pharmacists in Van. It is thought that these results will not only be helpful for pharmacists for making decision about the issue, but also beneficial for pharmaceutical warehouses to improve their services.

References

[1] Kuruüzüm, A., & Atsan, N. (2001). Analitik hiyerarşi yöntemi ve işletmecilik alanindaki uygulamalari. *Akdeniz University Faculty of Economics & Administrative Sciences Faculty Journal*, 1(1):83-105.
[2] Saaty, T. L. (1980). The Analytic Hierarchy Process. New York: McGraw-Hill.
[3] Elahi, F., Muqtadir, A., Anam, S., & Mustafiz, K. (2017). Pharmaceutical Product Selection: Application of AHP. *International Journal of Business and Management*, 12(8): 193-200.
[4] Ömürbek, N., & Şimşek, A. (2014). Analitik hiyerarşi süreci ve analitik ağ süreci yöntemleri ile online alişveriş site seçimi. *Yönetim ve Ekonomi Araştırmaları Dergisi*, 12(22): 306-327.
[5] Dağdeviren, M., & Eren, T. (2001). Tedarikçi firma seçiminde analitik hiyerarşi prosesi ve 0-1 hedef programlama yöntemlerinin kullanılması. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 16(2):41-52.
[6] Yaralıoğlu, K. (2001). Performans Değerlendirmede Analitik Hiyerarşi Proses. D.E.Ü.İ.İ.B.F. Dergisi, 16(1): 129-142.
[7] Dağdeviren, M., Diyar, Akay., & Mustafa, Kurt (2004). İş Değerlendirme Sürecinde Analitik Hiyerarşi Prosesi Ve Uygulamasi. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 19(2): 131-138.
[8] Yüksel, İ., Akın, A. (2011). Analitik hiyerarşi proses yöntemiyle işletmelerde strateji belirleme. Doğuş Üniversitesi Dergisi, 7(2): 254-268.
[9] Ünal, Ö. F. (2011). Analitik Hiyerarşi Prosesi Ve Personel Seçimi Alaninda Uygulamalari. *Journal Of Alanya Faculty Of Business/Alanya İsletme Fakültesi Dergisi*, 3(2):18-38.

[10] Jaberidoost, M., Olfat, L., Hosseini, A., Kebriaeezadeh, A., Abdollahi, M., Alaeddini, M., & Dinarvand, R. (2015). Pharmaceutical supply chain risk assessment in Iran using analytic hierarchy process (AHP) and simple additive weighting (SAW) methods. Journal of Pharmaceutical Policy and Practice, 8, 9.
[11] Velmurugan, R., & Selvamuthukumar, S. (2012). The analytic network process for the pharmaceutical sector: Multi criteria decision making to select the suitable method for the preparation of nanoparticles. DARU Journal of Pharmaceutical Sciences, 20(1), 59.
[12] Bahmani N, Blumberg H. (1987). Consumer Preference and Reactive Adaptation to a Corporate Solution of the Over-The-Counter Medication Dilemma-An Analytic Hierarchy Process Analysis. Math. Modelling. 9 (6): 293-298.
[13] Koklu O., Jakubowski E, Servi T, Huang J. (2016). Identifying and Interpreting Subjective Weights for Cognitive and Performance Characteristics of Mathematical Learning Disability: An Application of a Relative Measurement Method. The Journal of International Lingual, Social and Educational Sciences, 2(2):136-151.

Maximum Likelihood Estimation of the Parameters of the Exponentiated Reduced Kies Distribution via Artificial Bee Colony Algorithm

I .Usta[1], V. Yesildere[1]

[1]Eskisehir Technical University, Eskisehir, Turkey, iusta@eskisehir.edu.tr,volkanyesildere@eskisehir.edu.tr

Abstract

In recent years, the exponentiated reduced Kies distribution (ERKD), which is described [1], has been used in the modeling of data sets with monotonic and non-monotonic hazard rate functions. However, the maximum likelihood (ML) estimation for the three parameters of ERKD is difficult to obtain. In this study, artificial bee colony (ABC) algorithm is used to compute the ML estimates of the ERKD parameters. The ML estimates are also obtained by the numerical methods. Then, an extensive Monte Carlo simulation study is conducted to compare the performances of the ABC algorithm. The results show that the ABC algorithm perform better than the other considered methods.

Introduction

The Weibull distribution (WD) is widely-used distribution in reliability and lifetime data analysis. However, its hazard rate function cannot exhibit nonmonotonic as the bathtub shape or unimodal shape. Therefore, many researchers have studied various modified forms of the WD in order to achieve nonmonotonic shapes. One of the modified forms of the WD was introduced as Kies distribution (KD) by [2]. For further details about the KD, the readers can refer to [3]. The reduced Kies distribution (RKD), which is a special case of the KD, was considered by [4]. In addition, Kumar and Dharmaja [1] proposed a new distribution with two shape parameters, called the exponentiated reduced Kies distribution (ERKD), which is an exponentiated version of the RKD. They studied the properties of the ERKD and stated that the ERKD shows a good modeling performance compared with the commonly used statistical distributions.

In this paper, we consider the maximum likelihood (ML) estimation for the three-parameter version of the ERKD, denoted by $ERKD(\lambda,\beta,\delta)$, with two shape parameters and one scale parameter. It should be noted that the ML estimators of the unknown three parameters are not obtained in closed forms. Hence, numerical methods, such as the Newton–Raphson, are required to compute the ML estimates. However, numerical methods may have some problems such as convergence to the wrong root, non-convergence of iterations and suchlike. In order to overcome these problems, in this paper, artificial bee colony (ABC) algorithm, which is a swarm based meta-heuristic approach, is used to obtained the ML estimates of the parameters of $ERKD(\lambda,\beta,\delta)$. The ML estimates are also obtained by the numerical methods. Then, an extensive Monte Carlo simulation study is conducted to compare the performances of the ABC algorithm.

The rest of this paper is organized as follows. In Section 2, we give a brief description about $ERKD(\lambda,\beta,\delta)$ and the ML estimators of the parameters of $ERKD(\lambda,\beta,\delta)$. Section 3 presents general information about ABC algorithm. A Monte-Carlo simulation study and its results are given in Section 4.

Background for the research

Exponentiated reduced Kies distribution

Let X be a random variable from $ERKD(\lambda,\beta,\delta)$. The cumulative distribution function (cdf) and probability density function (pdf) of X are defined, respectively, as follows:

$$F(x) = \left[1 - e^{-\lambda\left(\frac{x}{1-x}\right)^{\beta}}\right]^{\delta}, \quad 0 < x < 1, \quad \lambda,\beta,\delta > 0, \qquad (1)$$

$$f(x) = \lambda\beta\delta \frac{x^{\beta-1}}{(1-x)^{\beta+1}} e^{-\lambda\left(\frac{x}{1-x}\right)^{\beta}} \left[1 - e^{-\lambda\left(\frac{x}{1-x}\right)^{\beta}}\right]^{\delta-1} \qquad (2)$$

where λ is the scale parameter, β and δ are the shape parameters. The ERKD reduces to RKD, introduced by [4], when λ=1 and δ=1.

Maximum likelihood estimators

Let x=(x_1,x_2,… x_n) be a random sample of size n from E $ERKD(\lambda,\beta,\delta)$. Then, the log-likelihood function (ℓ) is as follows:

$$\ell = n\ln(\lambda\beta\delta) + (\beta-1)\sum_{i=1}^{n}\ln x_i - (\beta+1)\sum_{i=1}^{n}\ln(1-x_i) - \lambda\sum_{i=1}^{n} z_i^{\beta} + (\delta-1)\sum_{i=1}^{n}\ln\left(1-e^{-\lambda z_i^{\beta}}\right). \quad (3)$$

where $z_i = \left(\frac{x_i}{1-x_i}\right)$. After taking the first derivatives of ℓ with respect to λ, β and δ, the likelihood equations for the parameters are obtained by

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}\ln(z_i) + (\delta-1)\sum_{i=1}^{n}\frac{z_i^{\beta}e^{-\lambda z_i^{\beta}}}{1-e^{-\lambda z_i^{\beta}}} = 0, \quad (4)$$

$$\frac{\partial \ell}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n}\ln(z_i) - \lambda\sum_{i=1}^{n} z_i^{\beta}\ln(z_i) + (\delta-1)\sum_{i=1}^{n}\frac{\lambda z_i^{\beta}\ln(z_i)\,e^{-\lambda z_i^{\beta}}}{1-e^{-\lambda z_i^{\beta}}} = 0, \quad (5)$$

$$\frac{\partial \ell}{\partial \delta} = \frac{n}{\delta} + \sum_{i=1}^{n}\ln\left(1-e^{-\lambda z_i^{\beta}}\right) = 0. \quad (6)$$

It is obvious that equations (4−6) cannot be solved in closed form. Hence, numerical methods, such as the Newton–Raphson (NR), are required to compute the ML estimators of λ,β and δ.

Artificial Bee Colony Algorithm

The ABC algorithm is a swarm based meta-heuristic approach which was first introduced by [5] for the solution of multi-dimensional and multi-modal optimization problems. The ABC algorithm is inspired by the intelligent foraging behavior of honeybee swarms. The ABC algorithm contains three types of bees including employed bees, onlooker bees, and scout bees. Employed bees are responsible for finding new sources of food and providing information to onlooker bees. Onlooker bees wait in the nest, evaluate the information coming from employed bees, and decide which food source go to. Scout bees emerge lastly and find new food resources that the algorithm could not have previously discovered, i.e. a set of possible solutions with high suitability.

The ABC algorithm has some assumptions. For example, the nectar of the food source (solution) is collected only by a bee. In other words, the number of food sources should be equal to the number of employed bees. Another assumption of the algorithm is that the number of onlooker bees should be equal to the number of employed bees. The flow chart of the ABC algorithm is given in Figure 1.

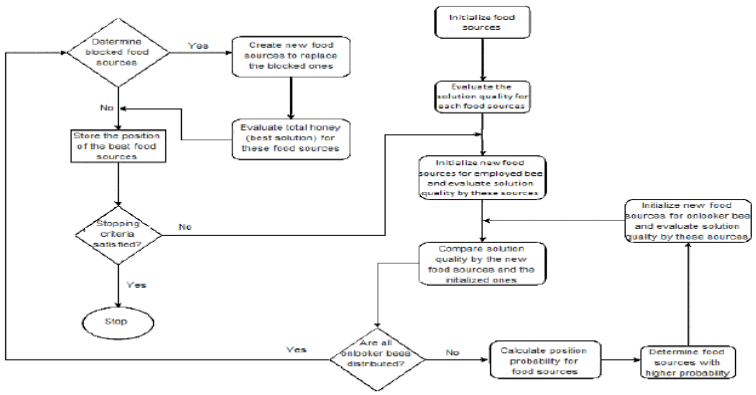


Figure 1. Flow chart of the ABC algorithm.

Simulation Study

In this section, a Monte Carlo simulation study has been conducted to compare the performance of the ML estimates of the ERKD parameters obtained by the NR method and the ABC algorithm. We consider the samples of size n=25,50 ,100 from the $ERKD(\lambda,\beta,\delta)$ with the parameters λ=1,β=0.5,2 and δ=0.5,2. For 100000n⁄ repetitions, the performance of the ML estimates is measured with different criteria such as mean, mean square error (MSE). In the ABC algorithm, the number of iterations is taken as 10000. All simulation programs were written by MATLAB (2019a). The simulation results are presented in Table 1.

Table 1. Simulated mean and MSE (in brackets) values for $\lambda = 1, \beta = 0.5, 2$ and $\delta = 0.5, 2$.

n	P	MLE_{Num}	MLE_{ABC}	MLE_{Num}	MLE_{ABC}	MLE_{Num}	MLE_{ABC}	MLE_{Num}	MLE_{ABC}
		$\beta = 0.5, \delta = 0.5$		$\beta = 0.5, \delta = 2$		$\beta = 2, \delta = 0.5$		$\beta = 2, \delta = 2$	
25	λ	0.9601	0.9709	0.8074	0.8083	1.0949	1.1030	0.9344	0.9262
		(0.6176)	(0.5998)	(0.4499)	(0.4277)	(0.8318)	(0.8098)	(0.4910)	(0.4573)
	β	0.9487	0.9268	1.0216	0.8773	3.3324	3.2674	3.0119	2.8964
		(0.7470)	(0.6858)	(1.5203)	(0.6986)	(7.8183)	(7.2801)	(5.7349)	(4.1951)
	δ	0.5594	0.5610	1.9546	1.9388	0.7349	0.7320	2.4416	2.3628
		(0.3053)	(0.2993)	(3.0681)	(2.7596)	(0.8831)	(0.8435)	(5.6358)	(4.6595)
50	λ	0.9915	0.9956	0.9601	0.9580	1.0215	1.0264	1.0198	1.0153
		(0.3712)	(0.3644)	(0.3084)	(0.2983)	(0.4276)	(0.4197)	(0.3676)	(0.3522)
	β	0.6775	0.6697	0.6347	0.6186	2.7397	2.7040	2.4306	2.4025
		(0.1981)	(0.1811)	(0.1746)	(0.1042)	(3.4748)	(3.1869)	(1.8412)	(1.4677)
	δ	0.5505	0.5517	2.2614	2.2425	0.5926	0.5937	2.5210	2.4819
		(0.1631)	(0.1617)	(2.6416)	(2.4620)	(0.3021)	(0.2979)	(4.4228)	(3.9707)
100	λ	1.0198	1.0213	0.9961	0.9947	0.9790	0.9799	1.0471	1.0446
		(0.2059)	(0.2038)	(0.1900)	(0.1867)	(0.1819)	(0.1804)	(0.2494)	(0.2435)
	β	0.5676	0.5656	0.5481	0.5479	2.3139	2.3053	2.1638	2.1637
		(0.0570)	(0.0539)	(0.0271)	(0.0266)	(0.9628)	(0.9053)	(0.5086)	(0.5009)
	δ	0.5415	0.5420	2.2228	2.2139	0.5181	0.5184	2.4247	2.4076
		(0.0818)	(0.0815)	(1.6947)	(1.6307)	(0.0755)	(0.0750)	(2.8610)	(2.6888)

The simulation results show that the ABC algorithm outperforms the corresponding NR method in terms of MSE for all the cases. Thus, we could conclude that the ML estimates via ABC algorithm is a good approach for estimating the parameters of the ERKD.

References

[1] Kumar, C. S., Dharmaja, S. H. S. (2017). The exponentiated reduced Kies distribution: Properties and applications. Communications in Statistics – Theory and Methods, 46(17), 8778-90.

[2] Kies, J. A. (1958). The strength of glass performance. Naval Research Lab Report No. 5093, Washington, D.C.

[3] Kumar, C. S., Dharmaja, S. H. S. (2014). On some properties of Kies distribution. Metron, 72, 97-122.

[4] Kumar, C. S., Dharmaja, S. H. S. (2013). On reduced Kies distribution. In: Kumar, C.S., Chacko, M., Sathar, E.I.A., eds. Collection of Recent Statistical Methods and Applications, (pp. 111–123). Trivandrum: Department of Statistics, University of Kerala Publishers.

[5] D. Karaboga. (2005). An idea based on honey bee swarm for numerical optimization, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department.

An Application on Comparison of Classical and Artificial Neural Network Approaches on Type II Regression Analysis

B. Tunca[1], S. Saraçlı[1], Ç. H. Aladağ[2], İ. Kılıç[3]

[1]*Afyon Kocatepe University, Afyonkarahisar, Turkey ,berkalp0606@gmail.com, ssaracli@aku.edu.tr*
[2]*Hacettepe University, Ankara, Turkey, chaladag@gmail.com*
[3]*Afyon Kocatepe University, Afyonkarahisar, Turkey, ibrahimkilic@hotmail.com*

Abstract

The aim of this study is to compare the performances of OLS-Bisector and Neural Network Bisector Regression models. With this aim, the data set of an early study on health workers, about the effect of organizational stress on burnout in Afyonkarahisar city is used.
In this study, a new, artificial neural network approach is applied to calculate the bisector regression line. The performance comparisons of ANN-Bisector and OLS-Bisector regression models are based on the MSE, RMSE and AIC criteria. The results of the study indicate that performance of ANN-Bisector is better than the performance of OLS-Bisector regression. Detailed results are given in related tables and figures.

Section-I

Artificial Neural Networks (ANN) or Neural Networks (NN) are parallel and distributed information processing structures which are developed by inspiring from the human brain and composed of process elements connected by means of weighted connections. The most important feature is the ability to learn by experience. Artificial neural networks have been developed in order to realize the skills such as the ability to derive new information, to create new knowledge and to discover by means of learning which are features of the human brain without any help. Artificial neural networks have the ability to create relationships between knowledge as well as learning [1].

The ANN method does not require various model assumptions like classical methods [2,3]. ANN can provide the necessary modeling without the need for any prior knowledge and assumption between input and output variables. Therefore, ANN has more advantages than other methods and can produce more successful results as a forecasting tool [4].

It is used successfully in the fields such as mathematics, statistics, physics, engineering and computer sciences with its advantages such as good results and ease of applicability produced by ANN [5,3].
The use of classical regression techniques that are always known in the study of the compatibility between the two methods is not accurate. Considering the measurements made with the two techniques considered, when the technique is considered as dependent and the other as independent variable, it is assumed that the independent variable does not contain any measurement error according to the classical regression assumption, the resulting errors are caused by the dependent variable. However, both of the measurements made with these two techniques actually contain some error. Considering this idea, regression models are considered as Type II regression models in the literature considering that both observation variables contain errors. Some of the Type II regression techniques are as follows; OLS Bisector, Orthogonal Major Axis, Reduced Major Axis, Deming, Passing-Bablok and York Regression Technique [6].

One of the Type II regression techniques, the OLS Bisector technique tries to minimize the distance of the observation points from the predicted regression line by taking into account the line-up of the line-of-line OLS(Y|X) and OLS(X|Y) regression line [7,8]. Here, $\beta 1(Y|X)$ is the inclination coefficient of the OLS(Y|X) regression equation calculated as the independent variable of X in the classical regression technique, and $\beta 1(Y|X)$ is the same as an independent variable is calculated as a result of the OLS(X|Y) is the slope coefficient of the regression equation [8].

Section-II

In this study, a questionnaire was applied to 411 individuals between September and November 2018 in order to determine the opinions of health workers in public hospitals in Afyonkarahisar on organizational stress and vocational burnout. Organizational stress and occupational burnout levels of the participants were measured with this questionnaire.

First of all, data processing methods were used to increase the performance and efficiency of ANN education. Considering the advantage of the "Min-Max Normalization" method, the normalization of the data has been deemed appropriate for the normalization of the data.
Afterwards, 70% of the data were allocated as training data and 30% as test data. While the data set was reserved for education, 90%, 80% and 70% were divided into two groups. Since the lowest error was obtained at 70%, the data was divided into 70% training and 30% testing.

These calculated outputs were then transferred to the activation function. In order to determine the most suitable function, the trial and lapse method was applied and "Sigmoid Function, Purelin Function, Step Function and Tansig Function" were tried. As a result of the experiments, the function which gives the most appropriate output to the data structure is determined as "Tansig Function".
Levenberg-Marquardt algorithm was used for the learning algorithm of the network because it works faster than other learning algorithms.

For the network architecture, the trial and error path was again made and for the second hidden layer, one to one hundred trials were performed. As a result of these experiments, one cell was used in the first hidden layer and ten cells were used in the second hidden layer. Since the learning rate yielded different results with different data sets and sizes and there is no definite information about the best learning rate, the trial and error method was applied as 0,01, 0,05, 0,001, 0,005 respectively. The best result was calculated as 0.005, so the learning rate was 0.005.

In order to use the ANN-Bisector (ANNBIS) technique, the input variable dependent variable was taken as the target variable independent variable. Then the input variable independent variable target variable was taken as dependent variable. In this way, ANN(Y|X) (Y dependent variable, X independent variable) equation and ANN(X|Y) (X dependent variable, Y independent variable) equations were calculated. While calculating these equations by using MATLAB program, the curve structure of ANN is transformed to a linear structure by applying a transformation with observation values as a result of the established neural network. In this way, the regression equation of that linear structure was obtained. The MSE, RMSE and AIC criteria were calculated to measure and compare the performance of these two models.

Section-III

The calculated slope coefficients and constant coefficients for the models are given in Table 1 below.

Table 1. Slope coefficients and constant coefficients for models (S: organizational stress, T: vocational burnout)

OLS	Constant Coefficient (β_0)	Slope Coefficient (β_1)
OLS(T|S)	1,343	0,546
OLS(S|T)	0,765	0,692
ANN		
ANN(T|S)	0,053	0,460
ANN(S|T)	-0,024	0,440
Type II Regression (Bisector)		
OLS-Bisector	0,977	0,617
ANN-Bisector	1,449	0,450

When Table 1 is examined, OLS(T|S) model's is constant coefficient 1,343, slope coefficient 0,546, OLS(S|T) model's is constant coefficient 0,765, slope coefficient 0,692, ANN(T|S) model's is constant coefficient 0,053, slope coefficient 0,460, ANN(S|T) model's is constant coefficient -0,024, slope coefficient 0,440,OLS-Bisector model's is constant coefficient 0,977, slope coefficient 0,617 and ANN-Bisector model's is constant coefficient 1,449, slope coefficient 0,450.

The calculated MSE, RMSE, AIC values for the OLS-Bisector and ANN-Bisector techniques' are given in Table 2 below.

Table 2. OLS-Bisector and ANN-Bisector techniques' related MSE, RMSE, AIC values

n=411	MSE	RMSE	AIC
OLS-Bisector	0,4457	0,6676	4,7065
ANN-Bisector	0,4434	0,6658	4,7023

When Table 2 is examined, OLS-Bisector techniques' is MSE value 0,4457, RMSE value 0,6676, AIC value 4,7065 and ANN-Bisector techniques' is MSE value 0,4434, RMSE value 0,6658, AIC value 4,7023. As a result, according to, MSE, RMSE and AIC criteria, ANN-Bisector technique gives better results with lower error than OLS-Bisector technique.

Section-IV

If a prediction is to be made for any problem and its assumptions provide appropriate conditions, the first model that comes to mind is classical regression modeling. However, if there is a measurement error for this problem, it would be a more correct decision to use Type II Regression techniques. This is because Type I Regression techniques are a disadvantage for Type I Regression due to the fact that errors caused by independent variables are not included in the model. The use of Type II Regression techniques to overcome this disadvantage is a very correct decision.

In this study, ANN-Bisector technique is presented by combining the advantage of the Type II Regression technique to estimate the errors in the independent variable and the superiority of the Artificial Neural Network in estimating. With this technique, organizational stress and vocational burnout levels of individuals were modeled by ANN-Bisector and OLS-Bisector techniques.

In this study, as in the ANN comparisons in this literature, ANN-based bisector technique yielded better results than OLS-based bisector technique. At the same time, it is foreseen that this ANN-Bisector technique can make a more successful estimation compared to classical regression methods since it deals with the error in the independent variables. The performance of Type II Regression and Artificial Neural Networks in the estimation was measured and it was tested whether the Artificial Neural Networks model was superior to Type II Regression as in the other comparisons in the literature. According to the findings, the artificial neural network model was also successful in the Type II bisector model.

For this study, no studies have been conducted on the optimization of the Artificial Neural Network. The parameters such as neuron number and activation function of the hidden variable in the network were calculated according to the best results obtained from the experiments. In another study, it is predicted that different network parameters can be obtained and better results can be obtained. In this study, it has been shown that Artificial Neural Network technique can be used as a very successful alternative technique by giving better results than OLS technique.

References

[1] Uğur, A., Kınacı, A. C. (2006). Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması . *"Türkiye'de İnternet" Konferans Bildirileri*. Ankara: TOBB Ekonomi ve Teknoloji Üniversitesi, 345-349.

[2] Gutierrez, R.S., Solis, A.O. and Mukhopadhyay S. (2008). Lumpy demand forecasting using neural networks, *International Journal of Production Economics*, 111, 409-420.

[3] Aladağ, Ç.H. (2009). Yapay Sinir Ağlarının Mimari Seçimi için Tabu Arama Algoritması. Doktora Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü. Ankara.

[4] Ataseven, Satış Öngörü Modellemesi Tekniği Olarak Yapay Sinir Ağlarının Kullanımı: "PETKİM'de Uygulanması", Yüksek Lisans Tezi, Celal Bayar Üniversitesi, Sosyal Bilimler Enstitüsü, Manisa.

[5] Günay, S., Eğrioğlu, E. ve Aladağ, Ç.H. (2007) Tek değişkenli zaman serileri analizine giriş. Hacettepe Üniversitesi Yayınları, Ankara: 230s.

[6] Saraçlı, S. (2008). Ölçüm Hatalı Modellerde Doğrusal Regresyon Tekniklerinin Karşılaştırılması: Monte-Carlo Simülasyon Çalışması Eskişehir Osmangazi Üniversitesi. Eskişehir.

[7] Isobe, T., Feigelson, E.D., Akritas, M.G. and Babu, G.J. (1990). Linear Regression in Astronomy I. *The Astrophysical Journal*, 364, 104-113.

[8] Saraçlı, S. (2011). Tip II Regresyon Tekniklerinin Monte-Carlo Simülasyonu ile Karşılaştırılması. *e-Journal of New World Sciences Academy, 6* (2), 26-35.

Using Dimensionality Reduction Techniques to Determine Student Success Factors

O. Doguc[1], Z. N. Canbolat[1]

1Medipol University, Istanbul, Turkey, odoguc@medipol.edu.tr, zncanbolat@medipol.edu.tr

Abstract

This study aims to determine the factors that contribute to student success by analyzing data from Gazi University. To collect data, around 6000 students were asked to rate the courses and instructors' performances with 28 course-specific questions, including the course content coverage, textbook, quizzes, instructor preparedness, attendance etc. We analyzed the questions and student responses, and found out which ones have more profound effect on student success in the courses. We used data mining techniques to associate students' responses to their successes, calculate correlations between them, and determine the top measurement factors (dimensions) that correlate most with student successes. focusing on those key factors, Also we used the k-means clustering algorithm to detect patterns in students' responses, and determine if certain groups of students tend to have similar success rates.

Introduction

What are the driving factors to success for students? This is a question that every educator and education professional is interested to learn. There are numerous factors that can impact students' success in class, and each student has different success factors. That being said, some factors are more common among students and some are more important for students' success and failure. Data mining is the process of analyzing data and transforming it into insight and rules. Data mining algorithms and applications enables organizations to analyze data from several sources in order to detect patterns. One important application area of data mining is called dimensionality reduction. While most data mining applications provide more accurate results when more data points available, performance suffer a lot with too many dimensions (columns). Therefore, dataheavy data mining applications have to reduce the number of dimensions before searching through the data for patterns and rules. Dimensionality reduction is done based on the correlations between dimensions and output; and also between the dimensions themselves. Dimensions that have lowest correlations and highest impact on the overall output are picked, and remaining ones are reduced. Data mining has application areas in the education sector, where large amount of data from students and classes are available. However, these applications are focused on specific areas in the sector such as enrollment and grade evaluations.

This study aims to provide insights to the educators by identifying student success factors through data with the help of dimensionality reduction algorithms. We used survey results from Gazi University in Ankara, Turkey; with around 6,000 participating students. [1] The dataset contains 33 dimensions, including instructor and class information and a number of survey questions about instructors and courses. This study uses the students' survey answers to discuss the important factors in courses and instructors that make them more successful.

Dimensionality Reduction

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Data Set

The data set used in this study contains 5820 evaluation scores provided by students from Gazi University in Ankara,Turkey. [1]In the survey, students were asked to evaluate 13 courses, taught by 3 different instructors. Along with the 28 course specific questions, students were also asked about their attendance levels, if they were repeating the course and finally how difficult the course is. Answers to the questions are captured in Likert scale between 1 and 5; where 1 denotes 'strongly disagree' and 5 denotes 'strongly agree'.

Evaluation

In this study, we used the MDR (Multi-factor dimensionality reduction) software to evaluate the dataset and infer most important attributes (dimensions) for students' successes. MDR requires at least one attribute to be picked as the goal or the result; which is used for calculating the correlations against. This study focuses on the factors that impact students' successes; and among the attributes that are available in the dataset, 'difficulty' is the one that represents the goal best. Therefore, we picked the 'difficulty' attribute as the goal for MDR. As mentioned in the Data Set section, difficulty shows how students perceive the course difficulty. This section will present what factors influence students' perception on course difficulty; which would provide educators insights about what drives success for students.

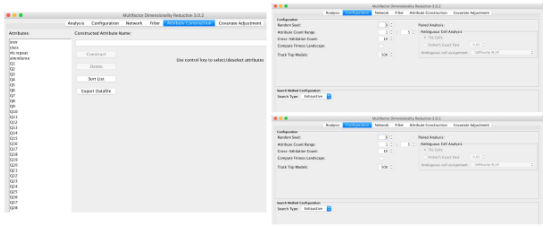


Figure 1. MDR setup parameters

Figure 1 shows our setup for the MDR software. MDR can compute correlations between dimensions and the goal (difficulty) one by one, or can group some of the dimensions before computing the correlations. MDR then picks the dimension and/or groups that present the best results in a few categories. For this study, we limit the group size to 5, as we want to isolate a few factors that have the highest impact on success. Given these settings and using the data set discussed earlier, following correlations are discovered by MDR:

Model	T-statistic CV Training	T-statistic CV Testing	CV Consistency
attendance	42.24	44.53	10/10
nb.repeat, attendance	48.5	51.13	10/10
class, nb.repeat, attendance, Q9, Q23	58.42	37.59	2/10

Table 1. Best models for MDR

Table 1 shows that when a single attribute is considered for success, it is the 'attendance' of the student. Because that model is composed of only 1 attribute, its TA is relatively low, but it is very consistent factor. Figure 2 shows how attendance is correlated with success. No attendance causes very severe negative impact on success. Also, it can be inferred that regardless of the class and the instructor; attendance is the number one factor that drives students to success.

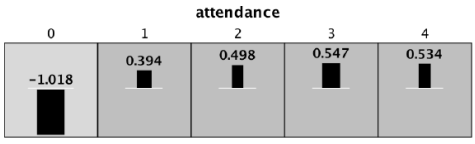


Figure 2. MDR output for 'attendance' model

Figure 3 shows TA values for other attributes and survey questions. It can be seen that 'attendance' has by far the highest value when considered alone.

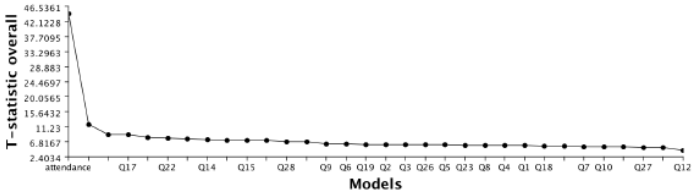


Figure 3. TA values for attributes and survey questions

When multiple attribute groups are considered, 'attendance' still remains an important factor, in addition to number of repeats and class itself; as well as Q9 and Q23 in the survey. This group represents a very high TA value (0.58) that makes it significant over the success result. Q9 and Q23 of the survey are as follows:

Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
Q23: The Instructor encouraged participation in the course.

Interestingly, both questions are about in-class participation. Students who can actively participate in the lectures; who are encouraged by the instructors find the class less difficult. So, leaving the class and the instructor aside, inclass attendance and participation are most important factors for success.

Conclusion

This study aims to provide the educators insights about the factors that drive success in classes, by using data mining. Because there can be too many factors that impact students and instructors, this study uses dimensionality reduction techniques to evaluate these factors and isolate the ones that have highest impact. For this purpose, the study uses survey results from Gazi University in Ankara Turkey, where about 6,000 students from various classes participated. With the help of MDR software, this study showed that attendance and in-class participation are the most important driving factors for students to success.

References

[1] G. Gunduz and E. Fokoue, "Turkiye Student Evaluation Data Set," 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/turkiye+student+evaluation.

Performance of Type II Regression in Time Series Analysis

E. Özgören[1],S. Saraçlı[1],B. Tunca[1]

[1]Afyon Kocatepe University, Afyonkarahisar, Turkey, ecem_ozgoren@hotmail.com, ssaracli@aku.edu.tr, berkalp0606@gmail.com

Abstract

The aim of this study is to examine the performance of Type II Regression in time series analysis. With this purpose, current exchange rates of Euro, Dollar and GBP between January 2007 - December 2018 are used. The data set is obtained from the web site of Turkish Central Bank. Forecast of January-May 2019 are compared with the real values. Earlier studies emphasize that performance of OLS-Bisector regression is the best among all Type II regression techniques. This technique simply bisects the two regression lines, which are obtained by considering each variable as dependent and independent respectively. To obtain this regression line, exchange rates of Euro- Dollar, Euro-GBP and Dollar-GBP are taken into consideration and the forecasts are compared with the forecasts of trend line and performances of these two techniques are compared according to MSE criteria.
The results of the study indicate that OLS-Bisector Technique's forecast are much closer than the forecast of Trend analysis to the real values, which also support that the MSE of OLS-Bisector Technique are much lower than the trend technique.

Section I

Time series analysis plays an important role in a large variety of application fields, such as economics, medicine, astronomy, geology and many others [1,2]. A time series can be defined as a sequence of observations, usually collected over regular time intervals with the temporal order preserved [3].
The real exchange rate in the literature is defined as the relative national price levels between two economies with the corresponding nominal exchange rate being an auxiliary to convert the unit of account such that two price levels are measured in a single currency. Its subject is not the currency or exchange rate, but the relative cost of living between two economies [4].

In the literature several econometric techniques are used to study exchange rate pass-through (ERPT). The most frequently applied methods include standard single-equation regression estimation, VAR (vector autoregression model), introduced by [5] or VEC (vector error correction model allowing for cointegration between variables) introduced by [6]. The latter two methods are the most widely used during the last two decades [7].

The aim of this study is to forecast the exchange rates of Dollar, Euro and GBP via Trend analysis and OLS-Bisector Regression technique and to compare the performances of these two techniques.

Section II

To obtain the forecasts of a given time series via Trend analysis, the least-squares method is frequently used to calculate the slope and intercept of the best line through a set of data points. However, least-squares regression slopes and intercepts may be incorrect if the underlying assumptions of the least-squares model are not met [8]. OLS assumes an error-free x variable and a constant analytical imprecision (sa) of the y variable (also called “homoscedastic” variance), both of which are seldom met in practice [9]. When both the dependent and the independent variables include some measurements errors, Type II Regression techniques must be used to obtain the correct parameters [10]. Earlier studies indicate that within the Type II regression Techniques, Performance of OLS-Bisector technique is better than the others.

The OLS-Bisector regression technique simply defines the line that mathematically bisects the OLS(Y|X) and the OLS (X|Y) lines [11]. As Isobe et al. mentioned, there is not any study in the literature regarding the merits or deficiencies of the OLS-Bisector line. When attempting to determine the underlying relationship between two uncensored variables, an OLS-bisector fit is likely to be the most reliable fitting method [12].

By OLS-Bisector technique, the slope can be calculated as in equation 1.

$$\hat{\beta}_{Biz} = \left(\hat{\beta}_1 + \hat{\beta}_2\right)^{-1}\left[\hat{\beta}_1\hat{\beta}_2 - 1 + \sqrt{\left(1+\hat{\beta}_1^2\right)\left(1+\hat{\beta}_2^2\right)}\right] \quad (1)$$

Here $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ is the slope of OLS(X|Y) regression and $\hat{\beta}_2 = \frac{S_{yy}}{S_{xy}}$ is the slope of OLS(Y|X) regression.

Section III

In the application part of this study, first of all exchange rates for Dollar, Euro and GBP are forecasted by linear Trend analysis, then these forecasts are obtained via OLS Bisector Technique. To obtain the regression model via OLS-Bisector technique for forecasts of Dollar exchange rate, exchange rates for Euro and GBP are used. Same procedure is applied while forecasting the exchange rates of Euro and GBP. Performances of Trend and OLS-Bisector Regression are compared via MSE criteria and how close do they forecast the real exchange rates.

To obtain the forecast via these two techniques, exchange rates of Euro, Dollar and GBP between January 2007 - December 2018 are used. Related data set is obtained from the web site of Turkish Central Bank. Forecasts of these exchange rates between January-May 2019 are compared with the real values and performances of these techniques are compared by MSE criteria.

Tend and Bisector regression models for the exchange rates are obtained as in equation 2 respectively.

$$\begin{aligned} &\text{Dollar}=0{,}6400+0{,}02276*t \qquad \text{Bisector}_{E/GBP}=2{,}7608+0{,}0241*t \\ &\text{Euro}=1{,}1864+0{,}02286*t \qquad \text{Bisector}_{D/S}=2{,}2075+0{,}0240*t \\ &\text{Sterling}=1{,}6070+0{,}02533*t \qquad \text{Bisector}_{D/E}=2{,}2254+0{,}0228*t \end{aligned} \tag{2}$$

Related with these equations, forecasts of Trend and OLS-Bisector techniques and the real exchange rate values are given in Table 1. MSE values as the performance criteria of these two techniques are given in Table 2.

Table 1. Trend, Bisector Forecasts and Real values of GBP, Dollar and Euro.

	GBP			Dollar (D)			Euro (E)		
Year/Month	Real	Trend	Bisector D/E	Real	Trend	Bisector GBP/E	Real	Trend	Bisector D/GBP
2019/01	6,93	5,28	5,53	5,38	3,94	6,25	6,14	4,50	5,68
2019/02	6,86	5,30	5,55	5,27	3,96	6,28	5,99	4,52	5,71
2019/03	7,20	5,33	5,58	5,45	3,98	6,30	6,17	4,55	5,73
2019/04	7,51	5,36	5,60	5,75	4,01	6,33	6,46	4,57	5,76
2019/05	7,89	5,38	5,62	6,07	4,03	6,35	6,80	4,59	5,78

As it can be seen from Table 1, forecasts for GBP, Dollar and Euro of OLS-Bisector for the first five months of 2019 are much closer to the real exchange rates than the forecasts of Trend analysis.

Table 2. MSE values of OLS-Bisector and Trend methods for the forecasts of GBP, Dollar and Euro.

	MSE		
Method	**GBP**	**Dollar**	**Euro**
Trend	3,91	2,63	3,19
OLS-Bisector	3,02	0,58	0,40

Supportively the results given in Table 1, the MSE values of OLS-Bisector technique are much lower than the MSE values of Trend Analysis as given in Table 2.

Section IV

There are many statistical techniques in time series analysis to obtain the forecasts. Some of them may be listed as Trend analysis, Box-Jenkins Models, Support Vector Machines, Neural Networks and etc.. Because of its nature, a time series may be affected from various factors. It is very important to determine these factors and eliminate them to obtain much closer forecasts with smallest error terms.

Both for the literature and the upcoming studies, it is very important for a new suggested method to obtain better forecasts then the current ones. Considering the advantages and the disadvantages of this new method, it must also be tested statistically and interpretations must be done considering the assumptions of each method.
The originality of this study is: OLS-Bisector technique which is a kind of Type II regression analysis and use in the method comparison studies is used first time in the literature to obtain the forecasts of a time series. One of the disadvantages of this technique is, even this technique considers the error terms of both dependent and independent variables in the model, it requires two similar variables to obtain the model and the model has a linear form. Because of this situation, performance of this technique is compared with the performance of linear Trend analysis.

One of the most important differences of OLS-Bisector technique from the classical time series analysis is; it uses the past values of other two exchange rates to obtain the model and makes the forecasts of related time series.

Results of the study indicate that to forecast the exchange rates, OLS-Bisector technique gives much better results (much closer to real values) and MSE values are lower than the Trend analysis. Related with this result, considering the other effects of the related time series, OLS-Bisector technique may be suggested to obtain the forecasts in time series analysis.

References

[1] Rani, S., et al. (2014). Review on time series databases and recent research trends in Time Series Mining. In 2014 5th international conference on confluence the next generation information technology summit, 109-115.
[2] Schlüter, T., and Conrad, S. (2011). About the analysis of time series with temporal association rule mining. In 2011 IEEE symposium on computational intelligence and data mining, 325-332.
[3] De Gooijer, J. G., and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22 (3), 443–473.
[4] Yang B.Z. and Zeng T. (2014) A Note on the Real Currency Exchange Rate: Definitions and Implications. *Journal of International Business and Economics,* 2 (4), 45-55.
[5] Sims, C. A. (1972). Money, income, and causality. American Economic Review, 62 (4), 540-552.
[6] Engle R.F and Granger C.W.J. (1987) Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55 (2), 251-276.
[7] Bodrug N. (2011) Estimating Exchange Rate Pass-Through In The Republic Of Moldova, MS. Thesis, Kyiv School of Economics.
[8] Cornbleet PJ. and Gochman N. (1979). Incorrect Least-squares Regression Coefficients in Method-Comparison Analysis. *Clinical Chemistry*, 25 (3), 432-438.
[9] Stöckl D., Dewitte K, and Thienpont LM. (1998). Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? *Clinical Chemistry*, 44, 2340-2346.
[10] Saraçlı S. and Çelik H.E. (2011) Performance of OLS-Bisector Regression in Method Comparison Studies, *World Applied Science Journal* 12 (10), 1860-1865.
[11] Saylor RD, Edgerton ES. and Hartsell BE. (2006) Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation. *Atmospheric Environment*, 40, 7546-7556.
[12] Isobe T., Feigelson ED., Akritas MG. and Babu GJ. (1990) Linear Regression in Astronomy I. *The Astrophysical Journal*, 364, 104-113

Examining Effective Factors on Emotional and Cognitive Demands by Copenhagen Psychosocial Questionnaire (COPSOQ) via Statistical Modeling

İ. Berk[1], S. Saraçlı[1], G. Boca[2], B. Tunca[1]

[1]*Afyon Kocatepe University, Afyonkarahisar, Turkey, ihsanberk1993@gmail.com, ssaracli@aku.edu.tr, berkalp0606@gmail.com*
[2]*Technical University, Cluj Napoca, Romania, bocagratiela@cunbm.utcluj.ro*

Abstract

Purpose of this study is to examine the effective factors on emotional and cognitive demands at work by Copenhagen Psychosocial Questionnaire (COPSOQ) via Structural Equation Modeling (SEM). Different from the earlier studies, there has been seven meaningful and significant factors, named as Quality of Leadership and Social Support, Mean of Work Satisfaction and Commitment, Social Support and Sense of Community, Role Conflict, Insecurity at Work, Emotional Demand and Cognitive Demand. By the help of SEM, the effects of other sub-factors on Emotional Demand and Cognitive Demand are statistically modeled. The results are given in related tables and figures.

Section I

International studies show that exposures to psychosocial risks are important characteristics of work. There is clear evidence that aspects and levels of job strain vary between countries, branches and professions. COPSOQ is a well proven instrument to gather valid and reliable information about main risk factors [1]
The Copenhagen Psychosocial Questionnaire (COPSOQ I) was developed in 1997 to satisfy the need of Danish work environment professionals and researchers for a standardized and validated questionnaire that covered a broad range of psychosocial factors [2,3,4]

Section II

In this study, to collect the data set, Copenhagen Psychosocial Questionnaire (COPSOQ) which is Likert type scale questionnaire, ranging from 1 ‘strongly disagree’ to 5 ‘strongly agree’ translated and applied to 338 workers at Afyon Kocatepe University between the dates 1-30 April 2019. To analyze the data, SPSS and LISREL software are used to perform the Explanatory Factor Analysis (EFA) and Structural Equation Modeling (SEM).

Section III

Results of EFA and SEM are given in Table 1, Figure 1 Figure 2 and Table 2

Table 1 Results of Explanatory Factor Analysis

Factors		Loadings	Eigenvalues	% of Variance
QLSS	**Quality of Leadership and Social Support**			
QLSS1	To what extent would you say that your immediate superior is good at solving conflicts?	,834		
QLSS2	To what extent would you say that your immediate superior is good at work planning?	,832		
QLSS3	To what extent would you say that your immediate superior makes sure that the individual member of staff has good development opportunities?	,830	8,460	13,479
QLSS4	To what extent would you say that your immediate superior gives high priority to job satisfaction?	,831		
QLSS5	How often is your immediate superior willing to listen to your work related problems?	,727		
QLSS6	How often do you get help and support from your immediate superior?	,661		

MWSC	**Mean of Work Satisfaction and Commitment**			
MWSC 1	Do you feel that the work you do is important?	,779		
MWSC 2	Is your work meaningful?	,771		
MWSC 3	Do you feel motivated and involved in your work?	,758		
MWSC 4	Do you feel that your place of work is of great personal importance to you?	,722	3,580	12,462
MWSC 5	Would you like to stay at your current place of work for the rest of your working life?	,662		
MWSC 6	How pleased are you with your job as a whole, everything taken into consideration	,578		
MWSC 7	How pleased are you with the way your abilities are used	,520		
SSSC	**Social Support and Sense of Community**			
SSSC1	How often are your colleagues willing to listen to your work related problems?	,770		
SSSC2	Is there good co-operation between the colleagues at work?	,745	2,645	9,425
SSSC3	How often do you get help and support from your colleagues?	,705		
SSSC4	Is there a good atmosphere between you and your colleagues?	,722		
SSSC5	Do you feel part of a community at your place of work?	,512		
RC	**Role Conflict**			
RC1	Do you sometimes have to do things, which ought to have been done in a different way?	,829		
RC2	Do you sometimes have to do things, which seem to you to be unnecessary?	,795	2,187	8,098
RC3	Are contradictory demands placed on you at work?	,772		
RC4	Do you do things at work, which are accepted by some people but not by others?	,686		
IW	**Insecurity at Work**			
IW1	Are you worried about it being difficult for you to find another job if you became unemployed?	,867		
IW2	Are you worried about becoming unemployed?	,852	1,813	8,030
IW3	Are you worried about new technology making you redundant?	,733		
IW4	Are you worried about being transferred to another job against your will?	,722		
ED	**Emotional Demand**			
ED1	Does your work require that you have very clear and precise eyesight?	,780		
ED2	Does your work demand your constant attention?	,734	1,501	7,668
ED3	Does your work require that you have to control your movements, e.g. your arms and hands consciously?	,716		
ED4	Does your work require a high level of precision?	,697		
CD	**Cognitive Demand**		1,266	5,873
CD1	Does your work demand that you are good at coming up with new ideas?	,796		
CD2	Does your work require you to make difficult decisions?	,780		
CD3	Does your work require that you remember a lot of things?	,674		

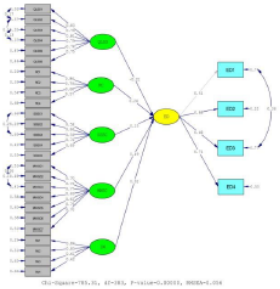


Figure 1. Structural Model for CD.

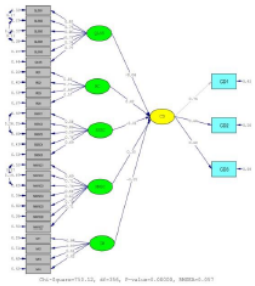


Figure 2. Structural Model for ED.

Table 2. The Goodness of Fit Indices for the Structural Models.

Criteria	Perfect Fitness	Acceptable Fitness	CD Model	ED Model
RMSEA	0<RMSEA<0.05	0.05 ≤ RMSEA ≤ 0.10	0.057	0.059
NFI	0.95 ≤ NFI ≤1	0.90≤ NFI ≤ 0.95	0.94	0.94
NNFI	0.97 ≤ NNFI ≤1	0.95≤ NNFI ≤ 0.97	0.96	0.96
CFI	0.97 ≤ CFI ≤1	0.95≤ CFI ≤ 0.97	0.97	0.97
GFI	0.95 ≤ GFI ≤1	0.90≤ GFI ≤ 0.95	0.87	0.87
AGFI	0.90 ≤ AGFI ≤1	0.85≤ AGFI ≤ 0.90	0.84	0.84

(Source: [5])(RMSEA: Root Mean Square Error of Approximation, NFI: Normed Fit Index, NNFI: Non-Normed Fit Index, CFI: Comparative Fit Index, GFI: Goodness of Fit Index, AGFI: Adjusted Goodness of Fit Index)

Table 2 indicates that the Structural Model for CD and Structural Model for ED statistically significant. Results of Standardized parameter estimate values, t values and hypotheses for the models given Figure 1 and Figure 2 are also given in Table 3 and Table 4.

Table 3. Standardized parameter estimate values, t values and hypotheses

Hypotheses	Paths	Standardized parameter estimates	*t* values	Results
H_1	(QLSS)→(ED)	-0.21	-2.31	Confirmed
H_2	(MWSC)→(ED)	0.49	4.27	Confirmed
H_3	(SSSC)→(ED)	0.15	1.70	Not Confirmed
H_4	(RC)→(ED)	0.24	3.26	Confirmed
H_5	(IW)→(ED)	0.02	0.39	Not Confirmed
Structural Equations				
ED = -0.21*QLSS + 0.49*MWSC + 0.15*SSSC + 0.24*RC + 0.024*IW				(R^2=0.22)

Hypotheses	Paths	Standardized parameter estimates	*t* values	Results
H_1	(QLSS)→(CD)	-0.04	-0.34	Not Confirmed
H_2	(MWSC)→(CD)	0.33	2.91	Confirmed
H_3	(SSSC)→(CD)	-0.01	-0.13	Not Confirmed
H_4	(RC)→(CD)	0.42	5.30	Confirmed
H_5	(IW)→(CD)	-0.05	-0.75	Not Confirmed
Structural Equations				
CD = -0.04*QLSS + 0.33*MWSC - 0.01*SSSC + 0.42*RC - 0.05*IW				(R^2=0.17)

Section IV

The results of the study indicate that, the most effective factor on the Emotional Demands of the workers is Mean of Work Satisfaction and Commitment. On the other hand the most effective factor on the Cognitive Demands of the workers is Role Conflict. Besides most of the workers are on the opinion of they do the right work which they can.

Similar with Emotional Demands, the Quality of Leadership and Social Support has a negative effect on the Cognitive Demands of the workers but this effect on Cognitive Demands is not statistically significant.
Improvements these effective factors on the workers Emotional Demands and Cognitive Demands may also improve the quality of work.

References

[1] (https://www.copsoq-network.org/) Date of access: 10.06.2019.
[2] Kristensen TS, Hannerz H, Høgh A, Borg V. (2005). The Copenhagen Psychosocial Questionnaire – a tool for the assessment and improvement of the psychosocial work environment. *Scand J Work Environ Health*, 31, 438-49.
[3] Kristensen TS, Bjorner JB, Christensen KB, Borg V. (2004). The distinction between work pace and working hours in the measurement of quantitative demands at work. *Work & Stress*, 18, 305-22.
[4] Pejtersen J.H., Krıstensen T.S., Vılhelm Borg V., and Bjorner J.B. (2010). The second version of the Copenhagen Psychosocial Questionnaire. *Scandinavian Journal of Public Health*, 38 (3), 8-24.
[5] Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8* (2), 23-74.

A Special Case of the p-Hub Center Network Problem

E. Engür[1], B. Soylu[1]

[1]*Erciyes University, Kayseri, Turkey, engur.enver@gmail.com, bsoylu@erciyes.edu.tr*

Abstract

The p-hub center problem aims to minimize the maximum travel time/distance in a network in order to increase the customer service level. We consider the special case of the uncapacitated p-hub center problem, where the weight/flow matrix includes 0 entities. In some real life networks, such as airline networks, cargo networks etc., the zero flow might exist between certain demand points. Typically in airline networks, nobody travels from city i to city i. In the literature, the general assumption is that all entries of the weight/flow matrix are positive. In this study, we discuss the effects of zero flows on the optimal solution and modified the current formulation of the uncapacitated p-hub center problem in order to handle this case. We present our computational results.
Keywords: hub center, optimization, linear programming.

Introduction

Hubs are special facilities managing the switching, sorting or consolidation functions in a network. The p-hub center problem determines the location of p-hubs in a network in order to minimize the maximum travel time/distance. This also helps to increase the customer service level. This problem is important for on-time delivery or emergency services systems. There are two types of allocation strategies as single and multiple. In this study, we consider the single allocation p-hub center problem, which assigns a non-hub node to a single hub only. We assume that the number of hubs is fixed to p and there is no capacity restriction.

The hub center problem was addressed by [1] and [2]. [2] formulated the uncapacitated single allocation p-hub center problem (USApHCP) as a quadratic integer problem. [2] also provides the linearization of this model, which requires huge number of binary variables as a disadvantage. [3] proved that the USApHCP is NP-hard and proposed an improved model. [4] considered the p-hub center allocation problem and showed that many versions of the problem are NP-hard. They also presented polynomial time algorithms for some special cases. Recently, [5] provided the novel formulation for the USApHCP based on the hub radius.
In the literature, the general assumption is that a positive flow is associated with every (i,j) origin-destination pair, i.e. $wij>0\ \forall(i,j)$. In some real life network structures such as airline networks, logistics networks etc., the zero flow might exist between certain demand points. The typical example is that in an airline network nobody travels from city i to city i, i.e. $wii=0\ \forall i$. The CAB data set [6] also includes zero flow on the diagonal of the flow matrix. In some logistics networks, the inner-city flow is directly delivered without needing to send it to hub. [2] also mentioned that the zero-flow might exist in the network and this can be handled by restricting (i,j) pairs such that $wij>0$ in the proposed four-index formulation. In this study, we modify the radius formulation of [5].

Problem Definition

Let $G=(N,E)$ be a connected network with the node set $N=\{1,2,...,n\}$, which corresponds to origins and destinations, and the arc set E, which corresponds to pathways. Any pair of nodes $i,j\in N$ is connected with an arc (i,j) and tij is the length/travel time of arc (i,j). We assume that the distance $\boldsymbol{d}$ satisfies the triangle inequality and $dii=0\ \forall i\in N$ and $dij>0\ \forall i,j\in N,i\neq j$. There exists a flow wij over arc(i,j). Assume that there is no innercity flow and the flow exists only if to other cities (nodes), i.e. $wii=0\ \forall i\in N$ and $wij>0\ \forall i,j\in N,i\neq j$.

The flow is routed through hub(s) associated with nodes i and j. Let (i,k,m,j) be a path from origin i to destination j via hubs k and m. The distance of this path, denoted Dij, is $Dij=dik+\alpha dkm+dmj$ where α is a discount parameter for the inter-hub transfer and in general $\alpha<1$, which indicates that inter-hub transfer is cheaper than node-hub transfer due to the economies of scale.

In the USApHCP, the aim is to minimize the longest path. [5] present the radius formulation of the USApHCP as follows.

Decision variables

$$x_{ik} = \begin{cases} 1 & \text{if node } i \text{ is allocated to hub } k \\ 0 & \text{otherwise} \end{cases} \quad \forall i,k \in N$$

If $x_{kk} = 1$, then node k is a hub.

r_k is the maximum distance (radius) to hub k among nodes allocated to it.

$$minimize \ D_{max} \tag{1}$$

Subject to

$$\sum_{k=1}^{n} x_{ik} = 1 \qquad \forall i = 1,2,\dots,n \tag{2}$$

$$x_{ik} \le x_{kk} \qquad \forall i,k = 1,2,\dots,n \tag{3}$$

$$\sum_{k=1}^{n} x_{kk} = p \tag{4}$$

$$r_k \ge d_{ik}x_{ik} \qquad \forall i,k = 1,2,\dots,n \tag{5}$$

$$D_{max} \ge r_k + r_m + \alpha d_{km} \qquad \forall k \le m = 1,2,\dots,n \tag{6}$$

$$x_{ik} \in \{0,1\} \qquad \forall i,k = 1,2,\dots,n \tag{7}$$

$$r_k \ge 0 \qquad \forall k = 1,2,\dots,n \tag{8}$$

Where the objective function is to minimize the longest distance $Dmax$. The constraint set (2) ensures that each node is allocated to exactly one-hub (single allocation restriction). The constraint set (3) requires that a node i can only be allocated to node k if node k is a hub. The constraint (4) states that exactly p hubs to be established. The constraint set (5) determines the radius of a hub k. The constraint set (6) ensures that the objective value should be greater than all travel times via hub(s).

The above formulations are suitable if inner-city transportation is applicable in reality, i.e. $wij>0 \ \forall i,j \in N$. Otherwise, the optimum solution might correspond to the path i-hub(i)-i, which is not applicable in some real life networks. In order to handle this case, we modify the [5] formulation as follows.

$$D_{max} \ge r_k + r_m + \alpha d_{km}(x_{kk} + x_{mm} - 1) \qquad \forall k \ne m = 1,2,\dots,n \tag{6'}$$

$$D_{max} \ge \sum_{k} (d_{ik}x_{ik} + d_{kj}x_{jk}) \qquad \forall i \ne j = 1,2,\dots,n \tag{10}$$

(1)-(5), (7), (8)

The constraint set (6) is modified as (6'), which evaluates the longest distance via two hubs. On the other hand, the new constraint set (10) evaluates the travel time from node i to node j if both assigned to hub k.

Computational Results

We performed experiments on the CAB data set [6] since innercity flow is not applicable. Each problem instance is coded as n×p×α, where n is the number of nodes in the network, p is the number of hubs will be installed and α is the discount factor for the inter-hub transportation cost. We used the Cplex 12.6.1 solver. We tested the modified radius formulation for the USApHCP and also compared its results with the well-known results from the literature. The following measures are reported.

Obj. radius: the optimum objective value of the radius formulation presented by Ernst et al. (2009).
Obj.modified: the optimum objective value of the modified formulation.
CPU_MRF: the CPU time required for solving the modified radius formulation.
%Gap: the gap between the optimum objective function values of the mixed-integer problem and the relaxed LP problem, i.e. ($obj-relaxed\ objobj*100$).

Computational results on the USApHCP are presented in Table 1. Accordingly, the optimum solutions of 6 instances have changed as a result of using modified formulations. These instances are marked with a "*" sign. We also observe that all CPU times are very low with a maximum 0.42 sec.

Table 1. Computational results on the modified formulations of the USApHCP for CAB instances

Instance	Obj.radius	Obj.modified	CPU_MRF	%gap
10x2x.2	1425.58	1425.58	0.11	35.6
10x2x.4	1627.52	1627.52	0.16	43.6
10x2x.6	1759.13	1671.49*	0.11	45.0
10x2x.8	1759.13	1743.86*	0.08	47.3
10x2x1	1839.65	1839.65	0.08	50.1
10x3x.2	1119.54	1115.83*	0.09	46.7
10x3x.4	1185.07	1185.07	0.09	49.9
10x3x.6	1387	1387	0.11	57.2
10x3x.8	1588.94	1588.94	0.08	62.6
10x3x1	1790.55	1790.55	0.09	66.8
10x4x.2	830.25	809.37*	0.14	51.8
10x4x.4	968.2	968.2	0.09	59.7
10x4x.6	1146.19	1146.19	0.09	65.9
10x4x.8	1454.44	1454.44	0.06	73.1
10x4x1	1764.79	1764.79	0.06	77.6
15x2x.2	2005.02	2005.02	0.31	33.0
15x2x.4	2160.75	2017.41*	0.31	33.4
15x2x.6	2214.09	2102.18*	0.25	36.1
15x2x.8	2423.8	2423.8	0.25	44.5
15x2x1	2609.18	2609.18	0.17	48.5
15x3x.2	1749.04	1641.7*	0.33	43.8
15x3x.4	1760.15	1738.32*	0.42	46.9
15x3x.6	1844.92	1844.92	0.38	50.0
15x3x.8	2166.54	2166.54	0.17	57.4
15x3x1	2600.08	2600.08	0.16	64.5
15x4x.2	1340.96	1322.92*	0.36	48.2
15x4x.4	1434.38	1434.38	0.19	52.3
15x4x.6	1754.51	1754.51	0.16	61.0
15x4x.8	2080.06	2080.06	0.16	67.1
15x4x1	2600.08	2600.08	0.17	73.7

* instances of which optimum solutions change in the modified formulation

References

[1] O'Kelly, M. E., & Miller, H. J. (1991). Solution strategies for the single facility minimax hub location problem. *Papers in Regional Science*, 70(4), 367-380.

[2] Campbell, J. F. (1994). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research,* 72(2), 387-405.

[3] Kara, B. Y., & Tansel, B. C. (2000). On the single-assignment p-hub center problem. *European Journal of Operational Research*, 125(3), 648-655.

[4] Campbell, A. M., Lowe, T. J., & Zhang, L. (2007). The p-hub center allocation problem. *European Journal of Operational Research,* 176(2), 819-835.

[5] Ernst, A. T., Hamacher, H., Jiang, H., Krishnamoorthy, M., & Woeginger, G. (2009). Uncapacitated single and multiple allocation p-hub center problems.*Computers & Operations Research,* 36(7), 2230-2241.

[6] O'Kelly, M. E. (1986). The location of interacting hub facilities. *Transportation science,* 20(2), 92-106.

Time Series Analysis of the Monthly Forecast of Crude Oil Exports in Iraq

M. Qais[1], M. Jafar[1]

[1] *Duhok Polytechnic University, Duhok, Iraq, qais.mustafa@dpu.edu.krd, jafarsindi@yahoo.com*

Abstract

In this paper, the procedure of Box-Jenkins of Autoregressive Integrated Moving Average ARIMA has been applied for analyzing and forecasting the exports of crude oil in Iraq by taking (72) observations of the monthly exports from 2013 to 2018. The data were divided into two parts, the first (60) observations was for training and the next (12) observations prepared for making a comparison between forecast and real values. Several suitable models of time series have been obtained and built and some of the performance measures have been used for the purpose of comparison between models. Results of the analysis concluded that the ARIMA(0,1,1) model is adequate to be used to forecast the monthly exports of crude oil in Iraq. The forecasted and real values of the year 2018 were close to each other. The forecasting process has done for the next 24 months. During the period 2019 to 2020, the exports of crude oil will reach (148281000) barrel per month.

Introduction

The world oil prices have witnessed a significant decline in the middle of 2015, which negatively affected the economies of many oils producing and exporting countries, and Iraq was among the most affected countries to dependence and a large proportion of imports of oil export to cover its internal expenses. The use of time series analysis on crude oil data has been widely applied like [1,2,3] and [4] but for Iraq, its application is limited due to limited research. The author in [5], used the Box-Jenkins approach to predict the amount of crude oil export in Iraq for the time period 2016-2017. The results showed that a suitable model for representing data is ARIMA(0,1,1). The objectives of this study are to fit an adequate model for the crude oil export and estimate the parameters in the model and also to forecast for the coming 24 months. The rest of this paper is prepared as follows: in section II. the methodology of Box-Jenkins is introduced. In section III. the application of real data is presented. Finally, in section IV. conclusions are stated.

Methodology Using ARIMA Models

Forecasting using time series represents the process of using a model to obtain forecasts for future events depending on past data. The Box-Jenkins method or ARIMA is a univariate technique representing self-projecting time series forecasting method. It became known by George E. and Gwilym M. Jenkins in 1970 [6]. ARIMA model, described as ARIMA (p, d, q), which can be formulated as the form:

$$y_t = \alpha + \emptyset_i Y_{t-1} + \cdots + \emptyset_p Y_{t-p} + \theta_i \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (1)$$

where α represents the constant term, $\emptyset i$ refers to the i-th autoregressive parameter, θj is the j-th moving average parameter, εt is the error term at time t, and yt is the value of the crude oil export observed at the time t. In this paper, the procedure of obtaining a suitable ARIMA model is described as follows. Firstly, the stationarity of the data concerning mean, variance, and trends will be tested by checking the plots of Auto-Correlation Function ACF and Partial Auto-Correlation Function PACF [7]. The two measurments such as root mean square error RMSE, and Akaike Information Criterion AIC are applied to evaluate the prediction accuracy of the approaches. Finally, to test the white noise or randomness of the series, the study will depend on the Box-Pierce test [8].

Application on Real Data

The data used in this study represents the monthly data of crude oil export in Iraq (in thousands) by taking a sample size (72) observations from January 2013 to December 2018 as shown in figure1. In fitting the appropriate model, the observations were divided into two parts which represent training and test set. Observations from January 2013 to December 2017 were used as the training set, and the data from January 2018 to December 2018 were prepared as the test set and was used to assess the predictability accuracy of the fit. Finally, the forecasting model was extended from January 2019 to December 2020.

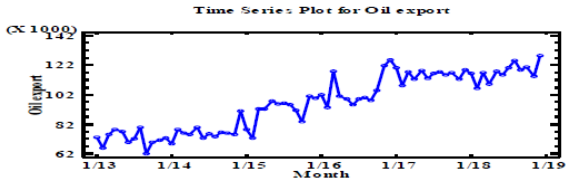


Figure1. Monthly crude oil export of Iraq during the period 2013-2018

Figure1 shows upward increasing and pattern to trend can be seen. This suggests that the given time series is non-stationary. Figure2 presents the ACF and PACF plots. The values of the ACF are gradually decreasing from the beginning to the end. The computed Portmanteau test of Box-Pierce with 24 lags takes a value of 521.609 (p-value = 0.00), which is highly significant, confirming the features of autocorrelation . The PACF shows a large peak at lag 1 with decline thereafter, which indicates the highly persistent autoregressive pattern in the series. Depending on the two functions ACF, and PACF, and also on the randomness test one concludes that the series is non-stationery and some necessary transformations should be done for the current series.

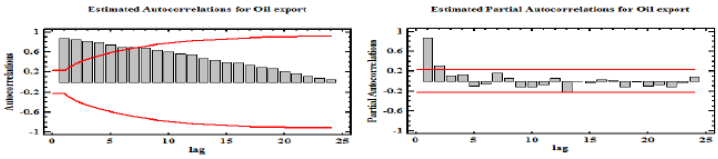


Figure2. Plots of ACF and PACF for crude oil export of Iraq during the period 2013-2018

After many trials, the study concluded that the series needs to be transformed so as to be random and stationary in mean and variance and this is by taking the log transformation with a first non-seasonal difference.

Fitting ARIMA Model

The Box-Jenkins methodology represents an iterative scheme for building a suitable model involving identification, estimation, the goodness of fit and model forecasting. A close examination was done on the ACF and PACF functions. Some adequate models were chosen depending on some performance criteria such as the RMSE and AIC. Table1 shows some adequate ARIMA models and the estimated criteria values.

Table1. Summary of models fitted adequately

Models	RMSE	AIC
ARIMA(0,1,1) with constant	**6849.05**	**17.7193**
ARIMA(1,1,1) with constant	6874.41	17.7545
ARIMA(0,1,2) with constant	6877.38	17.7553
ARIMA(0,1,1)	7076.62	17.7569
ARIMA(2,1,1) with constant	6907.90	17.7920

Depending on the selection criteria RMSE, and AIC, the above table shows that ARIMA (0, 1, 1) with constant was selected to be the best model. Also, the computed Box-Pierce test with 24 lags takes a value of 33.293 (p-value = 0.08) confirming the randomness of the fitted model. Furthermore, the parameters of the model were all significant at 5% level of significance.

Forecasting

Using the selected model ARIMA (0,1,1), we forecast future quantities of monthly crude oil export in Iraq for 24 months from January 2019 to December 2020; the first 12 actual values of the year 2018 compared with the forecasted values as shown in table2. The forecast time series values for monthly oil export is shown in table3. The data behavior of the forecasted values shows to follow the same as the original data. All forecasted values lie between the upper and lower boundaries of the 95% confidence intervals, concluding that the forecasting was accurate. In December 2020, the export of crude oil in Iraq will reach (148281000) barrels.

Table2. The actual and forecast values of the year 2018

Residuals	Forecast	Actual (x1000)	Period	Residuals	Forecast	Actual (x1000)	Period
2238.57	117917	120156	7/2018	-2927.17	119394	116467	1/2018
5514.48	119447	124961	8/2018	-13394.1	119514	106120	2/2018
-3275.95	121866	118590	9/2018	384.584	116609	116994	3/2018
-1348.74	121908	120559	10/2018	-8543.6	117624	109080	4/2018
-8216.17	122486	114270	11/2018	1762.02	116131	117893	5/2018
7304.27	121129	128433	12/2018	-1866.92	117517	115650	6/2018

Table3. Forecasts data from January 2019 to December 2020 using ARIMA(0,1,1) model

Forecast	Period	Forecast	Period	Forecast	Period	Forecast	Period
142636	7/2020	136145	1/2020	129949	7/2019	124035	1/2019
143748	8/2020	137206	2/2020	130961	8/2019	125001	2/2019
144868	9/2020	138275	3/2020	131982	9/2019	125975	3/2019
145997	10/2020	139353	4/2020	133011	10/2019	126957	4/2019
147135	11/2020	140439	5/2020	134047	11/2019	127947	5/2019
148281	12/2020	141533	6/2020	135092	12/2019	128944	6/2019

Conclusions

Statistical tests showed that the time series of monthly crude oil export in Iraq involves non-stationarity with the general trend. The log transformation with a first non-seasonal difference was used to achieve stationarity in the series. The most adequate model for the data of the monthly crude oil export in Iraq was found to be ARIMA (0 , 1 , 1). This model was chosen depending on the smallest values of the RMSE and AIC criteria, as well as the value of Box-pierce test. Using the final model, monthly crude oil export in Iraq was forecasted for 24 months. The forecast values for 2018 were in keeping with the original series values. The increasing trend of the data showed in the forecast values. At the end of December 2020, the crude oil export is expected to reach (148281000) barrel.

References

[1] Suleiman, S., Alabi, M. A., Suleman, I., Usman, U., and Adamu, U. (2015). Modeling and forecasting the crude oil price in Nigeria. *International Journal of Novel Research in Marketing Management and Economics*, 2(1), 1-13.
[2] Omekara, C.O., Okereke, O. E., Ire, K.I., and Okamgba, C. O. (2015). ARIMA modeling of Nigeria crude oil production. *Journal of Energy Technologies and Policy*, 5 (10), 1-5.
[3] Kayode A., and Habib A. (2013). Modeling and forecasting Nigerian crude oil exportation: seasonal autoregressive integrated moving average approach. *International Journal of Science and Research,* 2(12), 245-249.
[4] Obinna A., and Jolayemi E. T. (2015). Estimating the Impact on the Nigeria Crude Oil Export from 2002 to 2013. (An Arima-Intervention Analysis). *International Journal of Scientific & Engineering Research*, 10(6), 878-886.
[5] Ammar, SH. A. (2017). The use of time series models to predict the quantity of Iraq's oil exports for the years 2016 – 2017. Journal of *Baghdad College of Economic Sciences University*, (52), 415-428.
[6] George E., Jenkins G. M., and Reinsel G. C. (2008). Time series analysis: Forecasting and control, (4th ed.). John Wiley & Sons, INC.
[7] Karel D. B., Walter E., and Jared L. (2010). Applied econometric time series, 4th edition, WILLEY.
[8] Qais M. A. (2016). Time series forecasting using ARIMA methodology with an application on census data in Iraq. *Journal of the University of Zakho***,** 4(2), 258-268**.**

Evaluation of Factors Effective On the Performance of Surgical Teams in Operating Rooms

Şeyda Gür[1], Tamer Eren[2]

[1]*Harran University Sanliurfa Vocational School of Technical Sciences, Şanlıurfa, Turkey, seydagur@harran.edu.tr*

[2] *Kırıkkale University Department of Industrial Engineering, Kırıkkale, Turkey, tamereren@gmail.com*

Abstract

Hospital managers are developing multiple strategic goals to meet increasing demand and manage hospital costs effectively. Efforts are being made to increase efficiency in operating rooms, which are considered the most important units by hospital managers. The key concepts that have an impact on operating room efficiency are focused. These key concepts include the performance of surgical teams. The higher the performance within the surgical team, this affects the satisfaction of the personnel and the efficiency of the operating rooms. For this reason, in this study, the performance evaluation problem of surgical teams was discussed. In this problem, which directly affects the efficiency of the operating room and the efficiency of the hospital, an evaluation is made using the analytical network process method which is one of the multi-criteria decision-making methods. Criteria that have an impact on surgical team performance were determined and evaluated in consultation with experts. The factors that affect the personnel responsible for the welfare and safety of the patient undergoing surgical operation while serving the patients were examined. Thus, with effective evaluation methods, hospital managers benefit from performance monitoring and improvement.

Introduction

Operating rooms, which are one of the biggest expenditure items of hospitals, are mentioned among the most critical units. It is considered as one of the most important interaction points for many patients in health systems [1]. Since surgical care provided to patients covers about one third of all health expenditures, many strategies are being developed for the efficiency of these units. There are many factors that cause inefficiencies in operating rooms [2-3]. Unforeseen complexities are encountered in these units, which include many stakeholders, including patients, anesthetists, surgeons, nurses, administrative and facility staff [4-8]. The satisfaction and performance of the surgical team, which has a direct impact on the efficiency of operating rooms, is an important point in reaching the desired levels in the planning of hospital managers [9]. In this context, it can be concluded that in order to be able to talk about performance, there should be a targeted aim and certain criteria to achieve this goal [10]. The evaluation of these criteria by analytical methods based on the opinions of experts increases the accuracy and reliability of the results obtained. Performance evaluation in health systems is defined as the extent to which the hospitals or health institutions can achieve their basic objectives and the extent to which they meet the demands of personnel and patients. In this study, analytical network process method, which is one of the multi criteria decision making methods, has been evaluated for the criteria that affect the surgical teams. Because teamwork in operating rooms is considered as the main point for patient safety and open communication culture.

Evaluation of Factors Effective On the Performance of Surgical Teams in Operating Rooms

Today, hospitals are the most important parts of health systems and operating rooms are the most important parts of hospitals. Health systems work to improve the quality of the services they offer to patients, to increase the satisfaction level of the personnel and to reduce costs. The extent to which these goals are achieved in line with these objectives can be determined by performance measurement. In this study, the problem of evaluation of the factors affecting the performance of surgical teams in operating rooms was discussed. In this study, 7 criteria which were effective on surgical teams were determined. The literature was supported by these criteria. 7 of them are in-team compliance, management support, coping with stress, job satisfaction, physical working environment, task responsibility and motivation. When the effective criterion is considered;

- *In-team compliance:* Improving communication channels in working environments is effective against the uncertainties and unrest caused by misunderstandings. Establishing an effective and constructive communication is one of the most important steps for personnel to share their wishes and thoughts.
- *Management support:* It is examined how hospital managers support the goals and objectives of the organization in order to strengthen organizational performance among personnel.

▪ *Coping with stress:* Stress experienced in the work environment causes the personnel to leave their jobs and reluctant work. As the stress of one employee may affect the other employee, it is among the factors that decrease productivity.

▪ *Job satisfaction:* Another factor affecting productivity is the level of employee satisfaction. In order to be high-performing and efficient, personnel must have a high degree of satisfaction with their work.

▪ *Physical working environment:* The working environment defines the characteristics of the individual's inner environment. Concepts such as temperature, equipment, noise, ventilation, lighting, vibration and ambient cleanliness are accepted among the physical working condition elements. The presence of working environment features that provide physical, mental and social protection of personnel increases the productivity of the work.

▪ *Task responsibility:* It affects the performance of personnel in fulfilling their responsibilities according to their job descriptions.

▪ *Motivation:* In order to keep the work performance of the personnel at the highest level and reduce their absenteeism, institutions should give importance to the factors that increase the motivation of the personnel.

Figure 1 shows the network structure of the criteria used in the problem definition which is the first step of the ANP method.

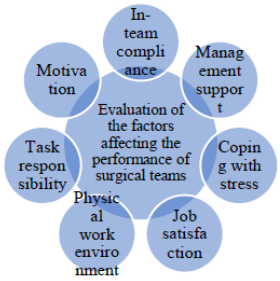


Figure 1. Network structure showing the relationships between criteria

In the comparisons made in the network structure, in team compliance criterion affects the motivation, job satisfaction and task responsibility criteria. Similar interactions were made among the other criteria which are thought to be related between them and the application phase was made. Binary comparison matrices were established based on the network structure and relationships established. The opinions of experts were used in binary comparison matrices. After making these comparison matrices, the importance of the criteria according to each other was obtained. Table 1 shows the weights of the criteria obtained as a result of the ANP method.

Table 1. Criteria weights obtained as a result of ANP method

Sub-Criteria	Sub-Criteria Weight
In-team compliance	0,49015
Task responsibility	0,36692
Management support	0,14292
Motivation	0,25484
Physical work environment	0,16527
Job satisfaction	0,38667
Coping with stress	0,19322

Conclusion

The patient who is decided to undergo a surgical procedure in the hospital is transferred to the operating room personnel. The operating room team, on the other hand, is responsible for the well-being of the patient during the whole operation and recovery process from the moment the patient takes care of the patient. At the same time, the operating room team should both pay attention to the patient's privacy and develop safety measures for the patient. The safety and privacy of the patients in the operating room is protected by the members of the operating room team. The staff working in the operating room; surgery surgeon, surgeon's assistants, anesthesiologist and nurse. Each member of the OR team performs a specific function in coordination with each other to create an atmosphere that best benefits the patient.

In this study, the problem of evaluating the performance of the operating room teams was discussed. Seven criteria were determined in the light of literature review and expert opinions. When there are many factors that affect

the problem, there are multi-criteria decision-making methods that facilitate the evaluation process. In this study, in order to evaluate the criteria that affect the performance evaluation problem, analytical network process method which is one of the multi-criteria decision-making methods was used. When the results obtained are analyzed, the in-team compliance with the value of 49% has come to the fore. The first way to see the increasing effect on the performance of the personnel is through the harmony between the personnel they work within the working environment. There will be an increase in the quality of service provided by the personnel who establish their communication in each team and complete each other in every aspect. Constructive and effective communication can prevent potential negativities. Personnel who establish effective communication within the team increase their satisfaction with the work they do. In this case, it increases the level of satisfaction with their jobs. Personnel who are satisfied with their jobs will indirectly establish positive relationships with other team members and will pass on this positive atmosphere to their teammates as well as patients. The patient's level of satisfaction will increase with positive interest from the staff. It is also important that the physical working environment is at a level to meet the needs and to be supported by the management.

References

[1] Lee, D. J., Ding, J., and Guzzo, T. J. (2019). Improving Operating Room Efficiency. *Current urology reports*, *20*(6), 28.

[2] Tyler D.C., Caroline A.P. and Chun-Hung C., (2003). Determining Optimum Operating Room Utilization. *Anesthesia & Analgesia*, 96(4), 1114-1121.

[3] Persson M.J. and Persson J.A., (2010). Analysing Management Policies for Operating Room Planning Using Simulation. *Health care management science*, 13(2), 182-191.

[4] Gür, Ş., Uslu, B., Eren, T., Akca, N., Yilmaz, A., and Sönmez, S. (2018). Evaluation Of Operating Room Performance In Hospitals By Using Analytic Network Process. *Gazi Journal of Health Sciences*, *3*(3), 10-25.

[5] Lin Q.L., Liu L, Liu H.C. and Wang D.J., (2013). Integrating Hierarchical Balanced Scorecard with Fuzzy Linguistic for Evaluating Operating Room Performance in Hospitals., *Expert Systems with Applications*, 40(6), 1917-1924.

[6] Eren, T. and Gür, Ş. (2018). Evaluation of The Factors Affecting The Performance of Operating Room By Fuzzy AHP. *Harran University Journal of Engineering*, *3*(3), 197-204.

[7] Leu J.D., Lee L.J.H. and Huang Y.W., (2016). A Regression-Based Approach to Identifying Factors Affecting Operational Efficiency in Surgical Rooms. In Industrial Engineering and Engineering Management (IEEM), IEEE International Conference on IEEM, 311-315.

[8] Gómez-Ríos, M. A., Abad-Gurumeta, A., Casans-Francés, R., and Calvo-Vecino, J. M. (2019). Keys to Optimize the Operating Room Efficiency. Revista Española de Anestesiología y Reanimación *(*English Edition*)*, *66*(2), 104-112.

[9] Ateş H. and Kırılmaz H., (2015). The Effects of Personal Factors on Health Staff's views of Performance Management., *Journal of Amme İdaresi*, 48(4), 97-128.

[10] Aksoy, E., Ömürbek, N. and Karaatlı, M. (2015). Use Of Ahp-Based Multimoora And Copras Methods For Evaluating The Performance Of Turkish Coal Enterprises. *Hacettepe University Journal of Faculty of Economics and Administrative Sciences*, 33(4).

Goal Programming Method for Shift Scheduling: Application for A Private Hospital Staff Nurse

A . Cürebal[1], T. Eren[1], H. M. Alakaş[1]

[1]University of Kirikkale, Kirikkale, Turkey, ahmet.crbl@gmail.com, teren@ kku.edu.tr, hmalagas@kku.edu.tr

Abstract

The shift scheduling problem is a problem that needs to be worked on continuously for all companies in both production and service sectors. Especially, in a area like health sector where service quality and customer pleasure is expected, scheduling problem is a problem that needs to be improved. The quality service of a hospital depends on many factors. Nurses are the most important factors for hospital quality. The quality of the nurses affects the quality of the hospital. The most important factor for nurses to make a quality study is the shift situation. The fact that the number of shift assigned to each nurse is as equal as possible will have a great effect on nurses to make a better quality of work. In this study, shift scheduling was tuned for 4 sections of a private hospital. Scheduling is modeled by using goal programming. The model is solved by ilog cplex studio ide program.

Introduction

The basis of the service sectors is the service to the customer. The important thing is to ensure customer satisfaction. Ensuring customer satisfaction; largely through careful planning of the labor force requirements of the service workers. The health sector is one of the most critical service sectors. Customer service quality is very important in this sector. In health sector; nurses are always in close contact with the customer. Therefore, the service provided by nurses directly affects the service provided by the organization to a great extent. Nurses also to be able to offer a quality service; work planning should be physically and mentally efficient.

In this study, monthly shift scheduling of nurses working in neurology, internal medicine, infection and ENT departments of a special hospital was done by setting goal programming model. The aim of the study was to ensure a fair distribution of shifts and to ensure that employees are not physically and mentally tired. Goal programming is one of the models developed to measure multi-purpose decision-making methods. This model is based on taking into account a number of objectives while making the decision maker find the best solution from a group of possible solution areas.[1] There are many studies on personnel scheduling and shift scheduling in the literature. Some of those; Eren and Ünal [2] in their studies using the target programming staff in the state institution has made the seizure scheduling. Varli and Eren [3] in their work on the nursing scheduling problem. Ergişi et al [4] conducted a study of the specific constrained nurse scheduling problem with the goal programming approach.

Monthly Shift Schedule of Nurses Working in Some Departments of A Private Hospital

A total of 52 nurses work in the departments of the private hospital subject to the study. The hospital has two shifts, morning and evening. Each shift is 12 hours. It is aimed that the monthly work plans of the nurses are arranged as equally and as balanced as possible. For this purpose, the study rules determined by the hospital were also taken into consideration. The conclusion of the study; In addition to the objectives, the study rules determined by the hospital were also taken into consideration and goal programming model was established. The operating rules determined by the hospital are shown below

- Working rules determined by the hospital:
- Each nurse should work at most 4 days a week.
- Each nurse should only work one shift per day.
- A nurse working on the evening shift on a given day should not work on the morning shift of the next day.
- The need for nurses in departments should be met. (This needs are shown in the Table-1)

Objectives of the study:

- The total number of appointments for each nurse should be as equal as possible.
- The number of appointments for each nurse at the weekend should be as equal as possible.
- The types of shifts that nurses serve should be as equal as possible.

The following table shows the morning and evening shift nurse needs of all departments of the hospital on weekdays and weekends.

Tablo 1. Nurse needs of related departments

Departments	Shifts	Weekdays	Weekend
neurology	Morning	5	2
	Evening	2	1
internal medicine	Morning	5	2
	Evening	2	1
infection	Morning	4	2
	Evening	2	1
ENT	Morning	4	2
	Evening	2	1

Weekends are as follows: 4,5,11,12,18,19,25 and 26th days. In addition, the need for nurses on May 1, which is a public holiday, is planned as in the weekend.

Parameters of the Problem

i: Nurse index — $i=1,2,3\ldots\ldots,52.$
j: day index — $j=1,2,3\ldots\ldots,31.$
k: Shift index — $k=1,2.$
l: Department index — $l=1,2,3,4.$
w_k: Shift time of k
a_{jkl}: Nurse needs of day of j, shift of k, department of l
K: Maximum number of appointments for each nurse on weekend days (K=3)
C: Minimum number of appointments for each nurse on weekend days (C=2)
S: Number of nurses assigned to morning shift (S=9)
L: Total number of appointments per nurse on all shifts (L=13)

$$X_{ijkl}: \begin{cases} 1, & \text{if empoyee } i \text{ is assigned to shift } k, \text{on } j \text{ day, in } l \text{ department} \\ 0, & \text{otherwise} \end{cases}$$

d_{3i}^{-}: 3. Negative deviation of target — $i=1,2,\ldots,52.$
d_{3i}^{+}: 3. Pozitive deviation of target — $i=1,2,\ldots,52.$
d_{2i}^{+}: 2. Negative deviation of target — $i=1,2,\ldots,52.$
d_{2i}^{+}: 2. Pozitive deviation of target — $i=1,2,\ldots,52.$
d_{1i}^{-}: 1. Negative deviation of target — $i=1,2,\ldots,52.$
d_{1i}^{+}: 1. Pozitive deviation of target — $i=1,2,\ldots,52.$

Mathematical Model

1. Restriction: Nurses should not work more than 48 hours per week.

$$\sum_{l=1}^{4}\sum_{j=1}^{7}(w_1x_{ij1l}+w_2x_{ij2l}) \le 48 \qquad i=1,2,\ldots,52.$$
$$\sum_{l=1}^{4}\sum_{j=8}^{14}(w_1x_{ij1l}+w_2x_{ij2l}) \le 48 \qquad i=1,2,\ldots,52.$$
$$\sum_{l=1}^{4}\sum_{j=9}^{15}(w_1x_{ij1l}+w_2x_{ij2l}) \le 48 \qquad i=1,2,\ldots,52.$$
$$\sum_{l=1}^{4}\sum_{j=16}^{22}(w_1x_{ij1l}+w_2x_{ij2l}) \le 48 \qquad i=1,2,\ldots,52.$$

2. Restriction: The nurse needs in each shift should be met for the relevant departments

$$\sum_{i=1}^{52} x_{ijk1} = a_{jk1} \qquad j=1,2\ldots,31 \quad k=1,2$$
$$\sum_{i=1}^{52} x_{ijk2} = a_{jk2} \qquad j=1,2\ldots,31 \quad k=1,2$$
$$\sum_{i=1}^{52} x_{ijk3} = a_{jk3} \qquad j=1,2\ldots,31 \quad k=1,2$$
$$\sum_{i=1}^{52} x_{ijk4} = a_{jk4} \qquad j=1,2\ldots,31 \quad k=1,2$$

3. Restriction: Any nurse should work in only one shift per day.

$$\sum_{k=1}^{2}\sum_{l=1}^{4} x_{ijkl} \le 1 \qquad i=1,2,\ldots,52. \quad j=1,2,\ldots,31$$

4. Restruction: A nurse working the night shift on any day should not work on the day shift of the next day.

$$x_{ij2l} + x_{i(j+1)1l} \le 1 \qquad i=1,2,\ldots,52. \quad j=1,2,\ldots,30. \quad l=1,2,3,4.$$

5. Restruction: Each nurse cannot work more than the maximum number of appointments on weekend days.

$$\sum_{k=1}^{2}\sum_{l=1}^{4} x_{i1kl} + x_{i4kl} + x_{i5kl} + x_{i11kl} + x_{i12kl} + x_{i18kl} + x_{i19kl} + x_{i25kl} + x_{i26kl} \le K \qquad i=1,2,\ldots,52.$$

Goal 1: The number of appointments for each nurse on the weekend should be as equal as possible.

$$\sum_{k=1}^{2}\sum_{l=1}^{4}(x_{i1kl} + x_{i4kl} + x_{i5kl} + x_{i11kl} + x_{i12kl} + x_{i18kl} + x_{i19kl} + x_{i25kl} + x_{i26kl}) + d_{1i}^{-} - d_{1i}^{+} = C \qquad i=1,2,\ldots,52.$$

Goal 2: The total number of appointments for each nurse should be as equal as possible.

$$\sum_{j=1}^{31}\sum_{l=1}^{4} x_{ij1l} + x_{ij2l} + d_{2i}^{-} - d_{2i}^{+} = L \qquad i=1,2,\ldots,52.$$

Goal 3: The total number of morning shift assignments for each nurse should be as equal as possible.

$$\sum_{j=1}^{31}\sum_{l=1}^{4} x_{ij1l} + d_{3i}^{-} - d_{3i}^{+} = S \qquad i=1,2,\ldots,52$$

Objective Function:

$$Min\ Z = \sum_{i=1}^{52} d_{1i}^{-} - d_{1i}^{+} + d_{2i}^{-} - d_{2i}^{+} + d_{3i}^{-} - d_{3i}^{+}$$

Result

The model was solved with IBM ILOG CPLEX Optimization Studio. Target deviation values were very small compared to such a model. As a result, a successful scheduling was made in which nurses could work more fit and effectively. Equal distribution of work was observed in the total assignments made according to the assignment types specified in the targets. These results are shown in Table-2.

Table 2. Total number of appointments of nurses

N	WD	WE	TA	TE	T	N	WD	WE	TA	TE	T	N	WD	WE	TA	TE	T	N	WD	WE	TA	TE	T
1	11	2	9	4	**13**	**14**	11	2	9	4	**13**	**27**	11	2	9	4	**13**	**40**	12	2	9	5	**14**
2	11	2	9	4	**13**	**15**	11	2	9	4	**13**	**28**	11	2	9	4	**13**	**41**	12	2	9	5	**14**
3	11	2	9	4	**13**	**16**	12	2	9	5	**14**	**29**	11	2	9	4	**13**	**42**	11	2	9	4	**13**
4	11	2	9	4	**13**	**17**	11	2	9	4	**13**	**30**	10	3	9	4	**13**	**43**	11	2	9	4	**13**
5	11	2	9	4	**13**	**18**	11	2	9	4	**13**	**31**	11	2	9	4	**13**	**44**	11	2	9	4	**13**
6	11	2	9	4	**13**	**19**	11	2	9	4	**13**	**32**	11	2	9	4	**13**	**45**	11	2	9	4	**13**
7	11	2	9	4	**13**	**20**	10	3	9	4	**13**	**33**	11	2	9	4	**13**	**46**	11	2	9	4	**13**
8	11	2	9	4	**13**	**21**	10	3	9	4	**13**	**34**	11	2	9	4	**13**	**47**	11	2	9	4	**13**
9	11	2	9	4	**13**	**22**	12	2	9	5	**14**	**35**	11	2	9	4	**13**	**48**	11	2	9	4	**13**
10	11	2	9	4	**13**	**23**	11	2	9	4	**13**	**36**	11	2	9	4	**13**	**49**	11	2	9	4	**13**
11	11	2	9	4	**13**	**24**	11	2	9	4	**13**	**37**	11	2	9	4	**13**	**50**	11	2	9	4	**13**
12	11	2	9	4	**13**	**25**	11	2	9	4	**13**	**38**	10	3	9	4	**13**	**51**	11	2	9	4	**13**
13	11	2	9	4	**13**	**26**	11	2	9	4	**13**	**39**	11	2	9	4	**13**	**52**	11	2	9	4	**13**

N: Nurse Number, **WD:** Total Number of Weekdays Assignments, **WE:** Total Number of Weekends Assignments **TM:** Morning Shift Total Assignments, **TE:** Evening Shift Total Assignments, **T:** Total Assignments.

The goal deviation values of the model are shown in Table-3. A successful planning was carried out according to the deviation values.

Table 3. Deviation values of goal parameters

Goal Parameter	Deviation Value
d_{1i}^{+}	4
d_{1i}^{-}	0
d_{2i}^{+}	4
d_{2i}^{-}	0
d_{3i}^{+}	0
d_{3i}^{-}	0

References

[1] Dağdeviren, M., Eren, T. (2001) Analytical Hierarchy Process and Use of 0-1 Goal Programming Methods in Selecting Supplier Firm, *Gazi Üniv. Müh. Mim. Fak. Der. 16,2, 41-52.*

[2] Ünal F.M., and Eren, T. (2016). The Solution of Shift Scheduling Problem by Using Goal Programming. *Akademik Platform Mühendislik ve Fen Bilimleri Dergisi*, 4,1, 28-37.

[3] Varlı E., and Eren, T. (2017). Nurse Scheduling Problems and An Application in Hospital, . *Akademik Platform Mühendislik ve Fen Bilimleri Dergisi*, 5,1, 34-40.

[4] Varlı, E., Ergişi, B., Eren, T. (2017). Özel Kısıtlı Hemşire Çizelgeleme Problemi: Hedef Programlama Yaklaşımı. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 0 (49), 189-206.

Solutions of Technology Manager Selection Problem with ANP and PROMETHEE Methods

A . Cürebal[1], T. Eren[1], E. Özcan[2]

[1]University of Kirikkale, Kirikkale, Turkey, ahmet.crbl@gmail.com, teren@ kku.edu.tr

[2]University of Kirikkale, Kirikkale, Turkey, evrencan.ozcan@kku.edu.tr

Abstract

The development of information technologies has increased the competition in the sector. Businesses need to follow the current changes. For this reason, enterprises need staff who can follow the rate of change and carry out original studies. Personnel selection problem is important for businesses, especially in the field of technology personnel selection problem is critical. Technology, nowadays it is very important for all sectors. Corporations have to make a successful management of technology in order to achieve success. Businesses should pay special attention when choosing personnel to be employed in technology departments and use scientific methods. In this study, the staff selection problem for the technology manager staff of a company working in the field of web design and development is discussed.. In solving the problem, ANP and PROMETHEE were used. There are 5 alternative candidates in the problem. 5 criteria were determined by an expert. These criterias are experience, exam score, paying attention to trend, responsibility and social life.

Introduction

The competitive environment accelerated with the development of technology has necessitated continuous renewal of enterprises. No matter how much an enterprise invests in a brand, the value of that enterprise determines the personnel it employs. Choosing personnel who can meet the requirements of the enterprise in the best way, adapt to the corporate culture and open to change and development are among the primary objectives of human resources. It is not easy for companies to identify the candidates that meet the criteria they are looking for and choose the most appropriate one, which is an important cost item. For this reason, the solution of the decision making problems which are the basis of the selection process should be realized with scientific methods. In this study, the selection of technology manager for a technology firm was made with ANP and PROMETHEE which are multi-criteria decision making methods.

Multi Criteri Decision Making

Multicriteria decision-making involves an analytical selection process that gives the best selection within the scope of the criteria and on the basis of dual comparison based on evaluation criteria identified among alternatives. There are many multi-criteria decision making methods. This methods of AHP, ANP, TOPSIS, VIKOR, PROMETHEE are frequently used.

ANP Methods and Its Stages

Although some problems do not have a hierarchical structure, the criteria, sub-criteria and alternatives discussed in the solution of the problem may interact with each other. The ANP method, which takes into account the relationships between the criteria effective in decision-making, is the generalized version of AHP developed by Thomas L. Saaty. [3]

Steps:

- Determination of the decision-making problem
- Determination of relationships
- Bilateral comparisons between criteria
- Calculation of consistency
- Creation of super matrices sequentially (Unweighted, Weighted, limit)
- Identifying the best alternative

PROMETHEE Methods and Its Stages

The PROMETHEE method is a simple and adaptable method for multiple criteria that can be expressed in real values when compared with other multi-criteria decision making methods in terms of application and scope. [1]

Steps:
• Formation of data matrix
• Defining preference functions for criteria
• Determination of common preference functions
• Determination of preference indices
• Determining positive and negative advantages for alternatives
• Setting partial priorities for alternatives
• Setting full priorities for alternatives

Application

In this study, 5 candidates were evaluated in the light of 5 main and 8 sub-criteria. Criteria weights were determined by ANP method and these weights were used in PROMETHEE method stage. Candidate evaluation in PROMETHEE Method; It was made by using 0-9 scale in Visual PROMETHEE program. Criteria and decision matrix are shown in Table-1.

Main criteria:
• Experience (Ex)
• Exam Score (E.S.)
• Paying Attention to Trend (P.A.T.)
• Responsibility (Re)
• Social Life (S.L.)

Sub-criteria:
• Evaluating Projects Carried out (E.P.C.)
• Year Worked (Y.W.)
• Score in Field English (S.F.W)
• Score in Software Ability (S.F.A.)
• Paying Attention to Technological developments (P.A.T.D.)
• Original Works (O.W.)
• Being Interested in Business (B.I.B)
• Communication Skills (C.S.)

Table 1. Main and sub-criteria decision matrix of candidates

CANDİDATES	Ex.		E.S.		P.A.T	Re.		S.L.
	E.P.C.	Y.W.	S.F.W	S.F.A	P.A.T.D.	O.W.	B.I.B	C.S.
A1	6	4	7	8	7	7	6	8
A2	8	6	8	8	8	8	9	7
A3	3	2	6	7	7	5	6	8
A4	5	6	8	9	9	7	7	6
A5	4	3	7	7	6	6	6	7

ANP Solution

ANP solution of the problem is solved in SuperDecisions program and the result is shown in Figure-1. According to the results of ANP candidate ranking is A2 - A4 - A1 - A3 and A5.

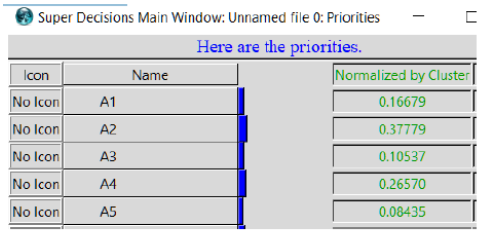


Figure 1. ANP solution screen of the problem

The criterion weights obtained from the ANP solution are shown in Table-2.

Table 2. Main criterion weights obtained from ANP Method

Criteria	Weights
Exam Score	0,28
E.S.	0,24
P.A.T	0,16
Re	0,22
S.L.	0,1

ANP method obtained as a result of the application; The weight of the sub-criteria in the main criterion is shown in Table-3.

Table 3. Sub-Criterion weights obtained from ANP Method

	Ex		E.S.		P.A.T.	Re		S.L.
	E.P.C.	Y.W.	S.F.W.	S.F.A.	P.A.T.D	O.W.	B.I.B	C.S.
Weights	0,75	0,25	0,4	0,6	1	0,7	0,3	1

PROMETHEE Solution

Solution of the problem by PROMETHEE method The solution and solution screen of VISUAL PROMETHEE program are shown in Figure-2. Weights obtained from ANP method were used as criterion weights. The 0-9 decision matrix shown in Table-4 was used as criterion scores. V-Type function is used as the type of function. The order obtained from PROMETHE method is A4 - A2 - A1 - A3 and A5.

Table 4. 0-9 Scale decision matrix

	Ex	E.S.	P.A.T.	Re.	S.L.
A1	6	8	7	8	9
A2	9	8	8	9	8
A3	5	7	7	7	9
A4	7	9	9	9	7
A5	3	7	6	7	8

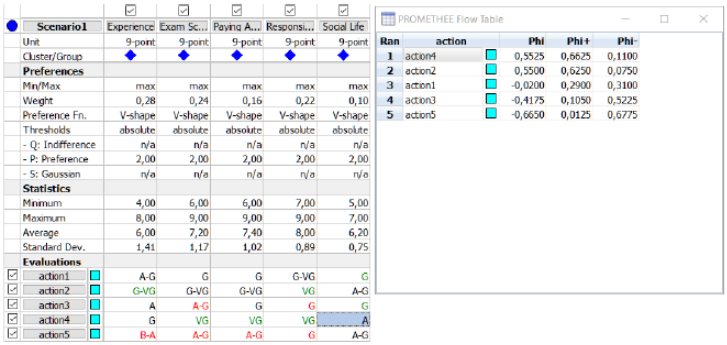


Figure 2. PROMETHEE data and result display

Result

The decision making problem was solved by ANP and PROMETHEE, two of the multi-criteria decision making methods. When we look at the results ranking of the methods, we see that the first two candidates differ and the other three candidates retain their place. The reason for this difference; ANP is one of the multi-criteria decision-making methods, while PROMETHEE is a scoring-based method.

References

[1] Bedir, N., and Eren, T. (2015). *AHP-Promethee* Method with İntegration of Personnel Selection Problem: A Case Study for the Retail Sector, *Social Sciences Research Journal* (Vol.4) 46-58.

[2] Eren, T., and Hamurcu, M. Alağaş, H.M. (2017). Selection of Kirikkale High Speed Train Station Location with Multi Decisions Making Method. *İnternational Symposium on İnnovative Technologies in Engineering and Science*.

[3] Özcan, E.C, and Ünlüsoy, S., and Eren, T., (2017) A Combined Goal Programming AHP Approach Supported with TOPSIS for Maintenance Strategy Selection in Hydroelectric Power Plants. *Renewable and Sustainable Energy Reviews* (Vol. 78), 1410-1423.

Managing Maintenance Cost of a Production System under Vagueness and Uncertainty

M.M. Mozaffari[1], S. H. A. Rahmati[2]

[1] *Imam Khomeini International University, Qazvin, Iran*
[2] *Islamic Azad University, Qazvin, Iran*

Maintenance costs force a vast sum to production systems and their finished product prices. Enormous amounts are waste on breakdown time and on efforts to recover systems suffering from machines malfunction. These costs have two basic features. First of all, the costs are almost entirely probabilistic and uncertain. Moreover, due to computational complexity, they are generally vague and not exact. Therefore, this research focuses on these features and develops a fuzzy model for the maintenance part of a production system. Then, regarding the complexity of the problem, it develops a simulation and evolutionary based meta-heuristic optimization algorithm to solve the problem under investigation. Finally, by implementing different simple and statically tests and figures assess the proposed model and algorithm.

Literature Review

In this research, fuzzy preventive maintenance is supposed to reinforce and equip a popular production problem called flexible job shop scheduling problem (FJSP). It has known as really complex production problem deprived from classical complex job shop scheduling problem (JSP) [1]. In the classical literature of the FJSP, it is usually assumed all machines are available all the time during their working process. It is also assumed that processing times are exact. In real word cases, processing times are usually inaccurate and fuzzy. Therefore, in this paper, this development is under investigation.

Generally, in the rich literature of FJSP and maintenance both of the model development [2-4] and solving method extension [5-6] can be found. Demir and Isleyen [7] do a comprehensive evaluation on the various mathematical models presented for the FJSP. Much more enormous solving methodology researches can be found more rather than mathematical model development [8-10]. In case of concept development, some studies considered the unavailability of machines due to preventive maintenance activities approach. Gao et. al. [11] studied integrated preventive maintenance and FJSP. The developed model, period of maintenance tasks are non-fixed and should be determined during the scheduling procedure. Wang and Yu [12] developed FJSP by maintenance activities that are either flexible in a time window or fixed beforehand. They also considered maintenance resource constraint in their model. Moradi et al. [13] integrated FJSP and preventive maintenance and developed a bi-objective problem that optimizes unavailability and makespan objective functions. Recently, Mokhtari and Dadgar [14] introduced a joint FJSP and PM model that assumes the failure rates are time varying. In their model the duration of PM activities are fixed. New applicable methods are also developed in the literature. However, fuzzy studies are so rarely in this literature. Thus, this research focuses on this gap and tries to fill a part of gap. 2

Problem Description

The Problem under investigation in this paper is the integration of classical FJSP with preventive maintenance. In this problem the processing time of production system are inaccurate. Therefore, they are modeled with fuzzy numbers. In this problem, the setup times between operations are not considered. Moreover, the type of maintenance is scheduled and predetermined.

Biogeography Based Optimization Algorithm

In this research, BBO algorithm is implemented for solving the proposed problem. BBO mimics the migration term of biogeography science. Migration includes two different manners of the species namely emigration and immigration. Each of these manners has a specific rate known as emigration rate and immigration rate. Emigration rate shows how likely a species, i.e. emigrating species, shares its features with other species, i.e. immigrating species. Likewise, immigration rate depicts how likely a species, i.e. immigrating species, accepts features from other species, i.e. emigrating species. Besides, generally features migrate from high-HSI habitats or emigrating habitat to low-HSI habitats or immigrating habitat. In fact, migration rates guide the optimization process to HSI maximization path.

Table 1 The outputs of fuzzy BBO algorithm

Problem	Time1	Fuzzy Cost		
FJSP1	397.11	99.650	101.4	102.521
FJSP2	746.55	158.655	160	161.647
FJSP3	543.23	84.0290	85	85.0208
FJSP4	610.98	134.106	135.6	135.933
FJSP5	1056.43	140.799	142.5	142.790
FJSP6	434.52	63.2097	64.4	65.9654
FJSP7	1127.49	220.586	221.3	222.672
FJSP8	1054.21	557.426	558.7	559.635
FJSP9	2547.36	254.712	255	256.628
FJSP10	2201.74	196.061	198	199.427

Computational Analysis

Table 1 presents the numerical fuzzy costs obtain from the BBO algorithm. In this table the fuzzy cost are based on triangular number on ten test problems [18-19].

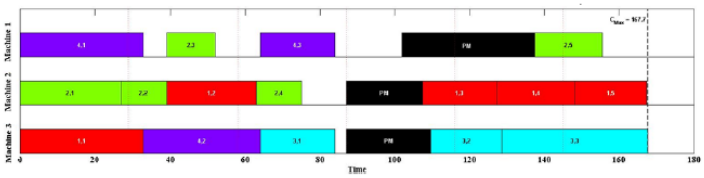


Figure 1 Gant chart of complementation time of preventive maintenance on FJSP

Figure 1 illustrates a sample Gant chart and Figure 2 depicts a sample of a fuzzy convergence plot of the BBO algorithm. Figure 3 presents the increasing processing times on test problems.

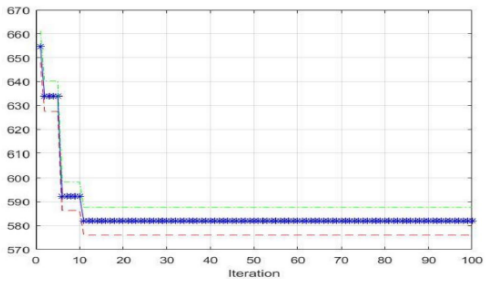


Figure 2 Fuzzy convergence plot obtained form BBO

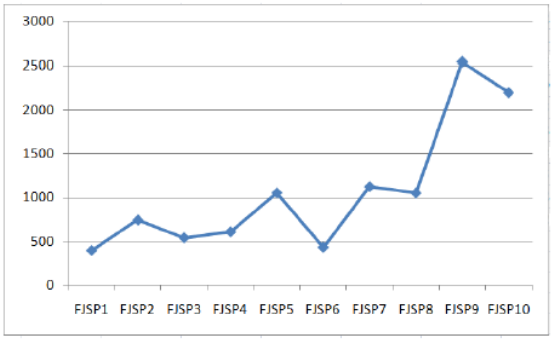


Figure 3 Solving time obtained from BBO on test problems

Conclusion

In this paper, FJSP is developed based on preventive maintenance considering fuzzy costs. The proposed problem is solved with BBO algorithm. The results show the performance of the algorithm.

References

[1] Frutos, M., Olivera, A.C. and Tohmé, F. (2010). A memetic algorithm based on a NSGAII scheme for the flexible job-shop scheduling problem. *Annual Operation Research*, 181, 745-765.

[2] Brucker P. and Schlie R. (1990). Job-shop scheduling with multipurpose machines. *Computing,* 45(4), 369–375.

[3] Choi, I.C. and Choi, D.S. (2002). A local search algorithm for job shop scheduling problems with alternative operations and sequence-dependent setups. *Computers & Industrial Engineering*, 42, 43-58.

[4] Lin, L., and Jia-zhen, H. (2009). Multi-Objective Flexible Job-Shop Scheduling Problem in Steel Tubes production. *Systems Engineering — Theory & Practice*, 29(8), 117–126.

[5] Gao, J., Gen, M., Sun, L.Y. and Zhao, X.H. (2007). A hybrid of genetic algorithm and bottleneck shifting for multiobjective flexible job shop scheduling problems", *Computer and Industrial Engineering*, 53(1), 149–162.

[6] Xing, L.N. and Chen, Y.W. (2009). A Knowledge-based Ant Colony Optimization for Flexible Job Shop Scheduling Problems. *Applied Soft Computing Journal*. doi:10.1016/j.asoc.2009.10.006.

[7] Demir, Y. and Isleyen, S.K. "Evaluation of mathematical models for flexible job-shop scheduling problems", *Applied Mathematical Modelling,* **37**, pp. 977–988 (2013).

[8] Zandieh, M., Khatami, A.R., Rahmati, S.H.A. (2017). Flexible job shop scheduling under condition-based maintenance: Improved version of imperialist competitive algorithm. *Applied Soft Computing* 58 449-464.

[9] Rahmati, S.H.A., Ahmadi, A., Karimi, B., (2018). Developing Simulation Based Optimization Mechanism for Novel Stochastic Reliability Centered Maintenance Problem, Scientia Iranica. *Scientia*.

[10] Rahmati, S.H.A., Ahmadi, A., Govindan, K. (2018). A novel integrated condition-based maintenance and stochastic flexible job shop scheduling problem: simulation-based optimization approach, *Annals of Operation Research* DOI 10.1007/s10479-017-2594-0.

[11] Gao, J., Gen, M. and Sun, L. (2006). Scheduling jobs and maintenance in flexible job shop with a hybrid genetic algorithm. *Journal of Intelligent Manufacturing*. 17, pp. 493-507.

[12] Wang, S. and Yu, J. (2010). An effective heuristic for flexible job-shop scheduling problem with maintenance activities. *Computers & Industrial Engineering*, 59, pp. 436-447.

[13] Moradi, E., FatemiGhomi, S.M.T. and Zandieh, M. (2011). Bi-objective optimization research on integrated fixed time interval preventive maintenance and production for scheduling flexible job-shop problem. *Expert Systems with Applications*, 38, pp. 7169-78.

[14] Mokhtari, H. and Dadgar, M. (2015). Scheduling optimization of a stochastic flexible job-shop system with time-varying machine failure rate. *Computers & Operations Research*, 61, pp. 31–45.

Semi-Supervised Sparse Data Clustering Performance Investigation

S. Sadeq[1], G. Yetkin[1]

[1]*Gaziantep University, Gaziantep,Turkey, salimkochar@gmail.com, golge.ogucu@gmail.com*

Abstract

Technology advancements have influenced the nature of the generated data. Recently, it is quite often to deal with high-dimensional datasets. This type of datasets requires careful attention to alleviate time and space complexities as well as ensuring the desirable outcomes. In machine learning, clustering is used for producing an idea on the underlying knowledge presented in a dataset, which is in many applications in this filed. Essentially, clustering can be obtained using *k*-means; however, deploying high-dimensional datasets imply dealing with the curse of dimensionality and possible noise within the data, which hinder the outliers of the produced clusters. Subspace dimensionality reduction can be applied for improving the robustness and cluster ability of high-dimensional data. In this paper, four soft subspace clustering algorithms were used sparse data. The choice was made based on their mechanism orientation, and their performances were compared with *k*-means in clustering hard-to-cluster datasets. It was found that the performance of *k*-means can be dramatically improved by incorporating soft subspace computing. Particularly, the ReliefF algorithm, in this study, could improve the clustering performance to more than 50% in some cases.

Introductıon

Machine learning applications have been widely used for improving data analysis accuracy and efficiency. Labeling data samples and discovering distinguished patterns within them have been around for quite some time. However, more recently, classifying data patterns and clustering similar data have become possible even without prior knowledge of data labels. This is usually achieved by retrieving the common underlying knowledge that classifies the data. A good example of such an application is data clustering. In data clustering applications, datasets are partitioned into groups referred to by clusters. These clusters are essentially composed based on the similarity of data objects; meaning that similar objects are gathered in one group. The idea of clustering similar objects requires a mechanism for measuring the similarity/dissimilarity of data objects to decide whether they belong to a specific class or to another one.

In the case of high-dimensional or sparse data, prior to applying any measuring process, the main concern is guided towards space or subspace reduction to alleviate the caused problem. Dimensionality reduction can be achieved through the deployment of dimensions themselves as knowledge in the clustering process. Clustering is generally classified as an unsupervised learning method; however, when dimensions or features are used for gaining some insight on dimensionality reduction, the process becomes semi-supervised in its essence [1]. This semi-supervised mechanism has recently been applied to clustering of high-dimensional or sparse data to produce better outcomes. Features deployment can generally be categorized into two main branches: feature extraction and selection. The idea of feature extraction is based on retrieving some features from the dimensions themselves to reduce the dimensionality and perform the clustering process in a more efficient way. On the other hand, it is possible to measure the relevance of feature or a set of features to achieve the required reduction, and in this case, the mechanism is called feature selection. Both methods require some sort of knowledge to obtain satisfying outcomes. In reality, it is nearly impossible to manually label large datasets. Instead, analysts use both labeled and unlabeled datasets in the processing, which refers back to the same idea of semi-supervised learning essence [2].

Related Works

In soft subspace clustering application, it is possible to maximize the performance of the classical *k*-means algorithm by minimizing its objective function. The objective function here refers to the possible error during the updating process of dimensional weights assignment, which continues until convergence . To further elaborate, generally, it is convenient to represent that a dataset X by the number of its samples (N) and the features (F) characterizing these samples. To conduct a soft subspace-clustering, not only similar objects of X are grouped into U clusters associated with their centroids $V = \{v_1, v_2, \ldots, v_i, \ldots, v_U\}$, but also with the set of their corresponding features weights $W = [w_1, w_2, \ldots, w_U]$. Using the listed parameters, a general objective function can be formulated as follow:

$$J(V,U,W) = J_{In}(V,U,W) + \sum_i \gamma_i J_{NC}(V,U,W)$$

Materıal And Methods

The main objective of this research word is to compare the performances of different soft subspace clustering algorithms with the classical *k*-means outcomes deploying datasets that are commonly easy to cluster.

1.Datasets

Four different datasets extracted from the UCI Machine Learning Repository [3-6] were used in this study. The choice of the datasets was made in a way that they should be somehow sparse and, at the same time, they should be uneasy to cluster. Table I.

Table 1. USED DATASETS

Name	Instances #	Features #	Features Type	Field
Fertility	100	10	Real	Health Science
Immunotherapy	90	8	Numeric	Health Science
Cryotherapy	90	7	Numeric	Health Science
Wholesale	440	8	Integer	Business

Deployed Algorithms

In this study, five different algorithms were used for the clustering purpose. The first one is the classical *k*-means clustering algorithm, and the rest is used for providing the semi-supervised nature for the clustering process. These algorithms are listed below:

- **Constraint Feature Selection (CFS):** which is a training-based feature selection algorithm that can be used for semi-supervised sparse data reduction, especially combined with some soft subspace reduction technique [7].
- **Infinite Feature Selection (IFS):** this algorithm is used for incorporating unsupervised soft subspace clustering procedure with feature selection [8].
- **ReliefF:** this algorithm is used as a robust algorithm for overcoming a number of issues such as noisy data presence [9].
- **Unsupervised Discriminative Feature Selection (UDFS):** this algorithm can be used for semi-supervised clustering process. The algorithm is claimed to be powerful in its performance, especially in high-dimensional data, which requires a higher computation cost with many of the available algorithms [10].

Results And Dıscussıon

The selected algorithms were implemented in MATLAB 2018b environment, and a useful Graphical User Interface (GUI) incorporated both the output clustering view, datasets loading and the performance of each algorithm. The developed GUI is shown in Fig. 1.

The summary of the produced results is given in Table II the analysis details for the used datasets and clustering algorithms. The best performance is highlighted by bold font for recognition purpose. It can be easily noticed from Table II that the performances of selected algorithms outweigh *k*-means in all cases except for the *Immunotherapy* dataset. In addition, very good improvements can be achieved considering *Cryotherapy* and *Wholesale* datasets.

Fig. 1. The main view of the developed GUI

TABLE 2. algorithms' performances for the selected datasets

Algorithm / dataset	*k*-means	CFS	IFS	ReliefF	UDFS
Fertility	34	50	59	**64**	52
Immunotherapy	44	44	44	44	29
Cryotherapy	49	59	69	**76**	53
Wholesale	3	**55**	3	53	32

Conclusions and Final Remarks

This study has investigated the performance of some soft subspace clustering algorithms for sparse data. The main idea of the study was guided towards comparing the outcomes of the clustering ability between some common soft subspace reduction algorithms and the classical *k*-means.

The tested algorithms have also shown that, in most cases, ReliefF algorithm could achieve the best outcomes, even though it was not close to optimal. The main issue with not achieving high results is the nature of the dataset, as the main aim was not guided towards producing an optimal solution; rather, it was aimed at enhancing the performance of *k*-means by incorporating these algorithms. Essentially, this aim has been achieved, and generally, above 50% improvement of the performance of *k*-means has been recorded.

One of the main limitations of this study is the lack of extensive testing results for different datasets, especially for clustering purpose. In fact, this can be a direction for future study where the deployed algorithms can be tested using different synthesized and real datasets suitable for clustering purpose. In addition, the selection of algorithms can also be changed or more algorithms can be added for further classification purpose. It is also possible, for future studies, to deploy different levels of soft subspace dimensionality reduction techniques and investigate both the performance and running time. Moreover, recent trends in machine learning, such as deep-learning, swarming behavior as well as kernels can be deployed for exploring better clustering results.

References

[1] D. Zhang, Z. H. Zhou and S. Chen, "Semi-supervised Dimensionality Reduction," Proceedings of the Society for Industrial and Applied Mathematics SIAM International Conference on Data Mining, 2007, pp. 629-634.

[2] S. Sedhai and A. Sun, "Semi-supervised Spam Detection in Twitter Stream. *IEEE Transactions on Computational Social Systems*, *5*(1), 2018, pp. 169-175.

[3] UCI Machine Learning Repository, Fertility Data Set, [Online]. Available at: https://archive.ics.uci.edu/ml/datasets/Fertility (Accessed on 22 December 2018)

[4] UCI Machine Learning Repository, Immunotherapy Data Set, [Online]. Available at: https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset (Accessed on 22 December 2018)

[5] UCI Machine Learning Repository, Fertility Data Set, [Online]. Available at: https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+ (Accessed on 22 December 2018)

[6] UCI Machine Learning Repository, Fertility Data Set, [Online]. Available at: https://archive.ics.uci.edu/ml/datasets/wholesale+customers (Accessed on 22 December 2018)

[7] M. Hindawi, K. Allab and K. Benabdeslem, "Constraint Selection-based Semi-supervised Feature Selection," 11th IEEE Conference on Data Mining (ICDM), 2011, pp. 1080-1085.

[8] G. Roffo, S. Melzi and M. Cristani, "Infinite Feature Selection. Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 4202-4210.

[9] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," European Conference on Machine Learning, 1994, pp. 171-182.

[10] Y. Yang, H. T Shen, Z. Ma, Z. Huang and X. Zhou, "l2, 1-norm regularized discriminative feature selection for unsupervised learning," IJCAI proceedings-international joint conference on artificial intelligence vol. 22(1), 2011, p. 1589

Novel Pattern in Software Architecture Based On Stored Procedure

M. Hoshmandi[1], S.H.A. Rahmati[2]

[1]*Islamic Azad University of Qazvin, Qazvin, Iran, IranErpGroup@gmail.com, sd_rahmati@qiau.ac.ir*

Abstract

Nowadays, introducing of various software architectures have caused the development of software the world over. Obviously, the advantages and disadvantages of each develop architecture determines the popularity and implementation level of that. In this research, a stored procedure called SP-based architecture is introduced to literature. This pattern reduces the large number of the programming lines indeed. Moreover, it can be adapted on most programming areas. Therefore, it is an almost entirely beneficial tool for data scientists and those who are engaged with data processing. Furthermore, it can be used as platform for conducting complex affairs of internet of thing (IOT) applications.

Introduction and Literature Review-I

Software architecture offers a total vision of software systems with dropping low level of complexity implementation [1]. Software architecture is basic phase in system development life cycle (SDLC), just like building basic structure determines final shape and development [2]. The software systems developers select an appropriate software architectural model for the implementation of the system based on business requirements. Change of this choice could be very costly in the future. This selection has led to a wide range of effects in different phases of development and maintenance of software systems. Furthermore, according to the choice of each architecture, developers will face various advantages and disadvantages.
Meanwhile, the proposed architecture of this paper attempts to introduce a flexible structure consisted with most businesses and conditions. It provides the greatest advantage with the least disadvantages. This Architecture very good operates in most area like Data mining, big data and data Science applications. Since the proposed structure operates based on stored procedure, the processing performance is raised considerably. Therefore, It is an effective and fast in tool for processing large data.
This architecture created during an implementation of a project that its business logic was in the stored procedures. This architecture has been further developed by using layered architecture, micro-service, microcontroller and event -driven architecture. This architecture is implemented with relational banks and various languages like Oracle, Sql Server and mySql, and the Node js, .net and php framework.

Famous and Developed Architectures -II
Totality, there are four main architectures known as follows.
- Layer architecture
- Event-driven architecture
- Microkernel architecture
- Microservice architecture

Proposed architecture of this research takes benefit of mentioned popular basic models and presents a practical environment. This integrated structure is called stored procedure architecture. Stored procedure suggests hybrid architecture so similar to event-oriented architecture. So events are defined on html elements that are received from the server in ajax. The type of event to be executed is defined through the various attributes within the element such as data-runsp, data-del and etc for identify of stored procedure and for related data, and the data-in attribute is used. As a result, CRUD operations are defined based on these attributes.
These operations are processed in three stages. The first stage of pre-processing in the client part determines the parameters associated with the operation. The second step is to process the event on the server and execute the related process routine. Finally, the third step is post-processing in the client, which reflects the result of the server in the browser.

This architecture inherits from the layered architecture, the different segments of coding in three layers. However, the final view is created in html format through the stored procedure.
On the other hand, the proposed architecture inherits from the microkernel architecture, the plug-in and development around a central kernel. But, there is far little dependency between the core and the plug-ins in the form of stored procedures. And during the run time, new operations can be added to the system.

The heritage of MicroService architecture to this architecture is the implementation of Run the chain of stored procedure. But unlike MicroService architecture, the standard for connecting and coding the services is defined and specified in the stored procedure. Besides, the control of the access level through the core is dynamically controlled. Moreover, this architecture of event-oriented architecture inherits the creating queues for events and managing requests.

Implementation of proposed model -III

This architecture can be implemented with different languages and databases. On the client and server side, there are generic functions written to run and process the request. These functions prevent the progressive growth of the source code. In this architecture, requests are executed in 5 steps.

First Step: The first stage of the preprocessing function is executed with the occurrence of an event by the user. This function is called to collect the values of the input elements, announce the deletion, fetch the item id, and so on. The preprocessing functions are defined by the on method in jQuery.

Second Step: in this step, the request function is called by the preprocessor functions. In the second step, the executor function is called by the preprocessing functions. For this purpose, the operation identifier is sent along with the handler's address data on the web server.

Third Step: The third step is validation of the request on the web server. After the request arrives to the web server, the requested operation is initially identified. If its execution is limited, it should be validate. This is done only by a simple query. The user ID and the requested operations are in a simple SQL query to check the access. If the result of this query was zero, execution of the operation would be stopped. Otherwise, the operation is performed in the next step.

Fourth Step: The fourth step is the execution of the operation in the form of stored procedures. For this purpose, the connection and stored procedure parameters are created dynamically. Then, the values received from the client are set to input parameters. Now, the stored procedure is ready for execution with definitions and setup. After execution, the stored procedure output is sent to the client.

Fifth Step: The fifth step is processing server output and reflection in the browser. Outputs received from the server are divided into 2 categories:

• The first category is encoded: to display the error message, success, and etc
• The second category is html code which is created directly by stored procedures and these codes are placed in the defined container. The user can interact with that content for future actions.
The following figure shows the steps to execute a request from beginning to end.

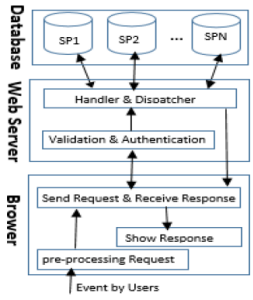


Figure 1 SP-Based architecture

Model Evaluation-IV

In this section, we will analyze the .Net Core object-oriented architecture and the architecture of the stored procedures. An analysis of software architecture is done by a series of parameters. These parameters assess the quality of the selected architecture and determine its usefulness for the stakeholders [5]. In the selected method, key parameters, such as coding volume, development speed, etc., are carefully evaluated. Object-oriented programming was designed to facilitate the implementation of complex and large systems with the principle of division and conquest. The basic drawback of this high-level conflict approach is to detail the implementation of individual classes, methods, and objects. In addition, a bunch of codes should be created by the programmer to implement, call,

and work with the object-oriented model, which has nothing to do with the implementation of business logic. But in a SP-based architecture, the programmer is not involved with objects and details of their implementation, and only focuses on transaction implementation. The approach of this architecture is transaction-centric. As a result, transaction IDs and data associated with an automated transaction mechanism are executed. Transactions in this template are well controlled. Successful and unsuccessful execution is accurately recorded and reported. In this architecture, codes have the least dependency. Operations are encapsulated in stored procedures. Therefore, following advantages are achieved.

• the speed of detection and resolution of errors increases
• the network traffic decreases
• transactions perform faster
• the access control become more accurate and easier to check

The evaluation of the system is limited to about 300 modules and has 1,600 operations. Two approaches to object-oriented .Net Core and the architecture of stored procedures are compared. The following results are achieved due to the approach of placing business logic in stored procedures.

Table 1 Results of an implementation of SP-Based architecture

Title	Explanation	Reduction rate
Number of files	Being a business logic in Database Stored Procedure	98%
Line of code	no need to define actions, models and coding in different layers, as the project volume increases, the number of files remains constant	88%
Number of repeated	No repeat codes in different categories	100%
Execution Costs	low programming with basic language (SQL, html, etc)	82%

Other advantages of this architecture are as follows.

• sharp decrease in coding, the number of project files and coding in basic languages like SQL, html, etc
• greatly cost reduction of implementing the project
• accurate logging of all successful and unsuccessful transactions with associated parameters
• easily publish and backup copies of the project.

Conclusion-V

This research developed an architecture SP-based Architecture which work based on procedures stored in different domains is effective and useful. It is a useful environment for programming in different practical areas like ERP, data mining and data mining systems. Two main features of this architecture make it useful for mass data. First, the encapsulation of executing operations in the form of stored procedures and database functions, which greatly increases the speed of execution of operations. And the second is a light controller layer that allows very wide interaction with a large range of different devices. These are also supported by a experiment.

References

[1] A.Ahmada, M.A Babar, Software architectures for robotic systems: A systematic mapping study, Elsevier 2016
[2] A.S, Manoj Kumarb, S. Agarwalc, A Complete Survey on Software Architectural Styles and Patterns, Elsevier 2015
[3] Software Architecture Patterns - Mark Richards
[4] Event-Driven Architecture Overview 5th Anniversary Edition: February 2, 2011
[5] R.Kazman, L.Bass, M.Klein, T.Lattanze, L.Northrop, A Basis for Analyzing Software Architecture Analysis Methods, Software Quality Journal 2005.

Detection of Pneumonia from X-Ray Images using Convolutional Neural Network

H. S. Omar[1], A. Babalık[1]

[1]Konya Technical University, Konya, Turkey, husham.s.o.89@gmail.com, ababalik@ktun.edu.tr

Abstract

X-ray imaging is one of the important methods of lung diseases detection. The X-ray images of the chest region are evaluated by specialist physicians. However, this evaluation is a time-consuming process and requires specialized expertise. Computer-aided diagnosis (CAD) is a recent research topic in machine learning, and it assists doctors for the analysis of medical images [1]. In the process of analyzing medical images, some of the deep learning techniques such as convolutional neural networks (CNN) are used [2]. In this study, a CAD system based on CNN was proposed for detection of pneumonia using X-ray images of the chest region, taken from the Guangzhou Women's and Children's Medical Center in Guangzhou. The dataset contains 5840 chest X-ray images (anterior-posterior). The training set consists of 5216 images (1341 normal, 3875 pneumonia), and the test set consist of 624 images (234 normal, 390 pneumonia) [3]. According to the experimental studies, the accuracy of the proposed system was achieved as 87.65%.

Introductıon

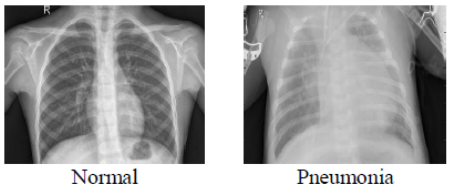
Pneumonia is an inflammation of the lung tissue caused by bacteria and airborne viruses. Pneumonia is the leading cause of child mortality. Every year 1.4 million children die from the disease, and it is about 18% of all deaths [4]. Doctors often use chest X-rays to diagnose chest-related diseases quickly and easily. Other imaging techniques such as computed tomography (CT) or magnetic resonance imaging (MRI) are used to diagnose such diseases [5], but X-ray imaging is cheaper and faster. The number of X-rays has been performed in recent years has increased significantly, and the diagnosis of these X-rays takes a lot of time by the doctors who check them manually. Computer-aided diagnosis (CAD) is a recent research topic in automated learning, and it helps the doctors to analyze medical images. Convolutional neural networks (CNN) are a typical machine learning algorithm used in CAD in recent years. Rajpurkar et al., developed an algorithm to detect 14 disease pathogens using the Chex Net algorithm [6]. Park et al., used an algorithm based on Gabor filter for detection of rib reduction on abnormal tissue obtained by X-ray imaging [7]. Dina A. Ragab used a CAD system to classify benign and malignant tumors in mammography [8]. Sarika C. and Seema B. used an algorithm to detect skin cancer using neural networks [9].
In this study, a CAD model based on CNN that can accurately and quickly detect pneumonia by using chest X-ray images is presented.

Material And Methods

Chest X-ray Dataset: The dataset contains 5840 X-ray images (front and posterior) which were selected retroactively from patients aged from 1 to 5 years. The dataset is compiled by the NIH, and it has been taken from Guangzhou Women's Medical Center in Guangzhou [3][10]. All X-ray images were taken from patients during the examination which was a part of the routine clinical care of patients. Details of the dataset are shown in Table 1, and Figure 1. Shows samples of normal and pneumonia infected X-ray images within the dataset.

Table 1. The details of the data set

X-Ray images	# of Training Samples	# of Testing Samples
Pneumonia	3875	390
Normal	1341	231
Total	5216	624



Normal Pneumonia

Figure 1. Sample X-ray images selected from the dataset

Convolutional Neural Network

The convolutional neural network (CNN) is a popular method of deep learning. It is inspired by the biological neuron processes. The CNN structure consists of several layers. It starts with the input layer and ends with the output layer, and also there are many hidden layers between the first and last layers. CNN algorithm automatically learns spatial hierarchies of features. In order to get successive results, the algorithm needs large amounts of data. CNN is used in many different and diverse fields such as handwriting recognition, face detection, image classification and image segmentation [11].

In this study, CNN was used for the classification of X-ray images. The images in the dataset have different sizes. Therefore, all the images in the dataset were resized to 150x150. The proposed CNN model consists of 1 input, 1 output, 5 convolutional layers, and 1 fully connected layer.

Figure 2 shows the structure of the proposed CNN model, and Table 2 shows the parameters of the proposed CNN model used in this study.

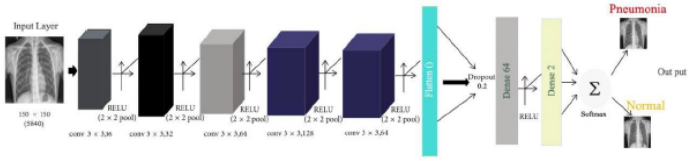


Figure 2. The structure of the proposed CNN model

Table 2. The parameters of the proposed CNN model

Layer	Parameters	Value
Input Layer	Input size	150x 150 x3
Convolutional Layer	Number of nodes Kernel size Activation functions	16 3 x 3 Relu
Pooling Layer (Max pooling)	Kernel size	2 x 2
Convolutional Layer	Number of nodes Kernel size Activation functions	32 3 x 3 Relu
Pooling Layer (Max pooling)	Kernel size	2 x 2
Convolutional Layer	Number of nodes Kernel size Activation functions	64 3 x 3 Relu
Pooling Layer (Max pooling)	Kernel size	2 x 2
Convolutional Layer	Number of nodes Kernel size Activation functions	128 3 x 3 Relu
Pooling Layer (Max pooling)	Kernel size	2 x 2
Convolutional Layer	Number of nodes Kernel size Activation functions	64 3 x 3 Relu
Pooling Layer (Max pooling)	Kernel size	2 x 2
Fully Connected Layer	Number of nodes Activation functions	64 Relu
Output Layer	Softmax	2 Class

Experiments and Results

In all experiments, a standard PC with an Nvidia GeForce GTX 1060 GPU card of 6 GB was used, and the algorithm was implemented in Python version 3.7. During the training process, the epoch number was 30 and dropout was 0.2. The accuracy and loss graphics generated during the training process are shown in Figure 3.

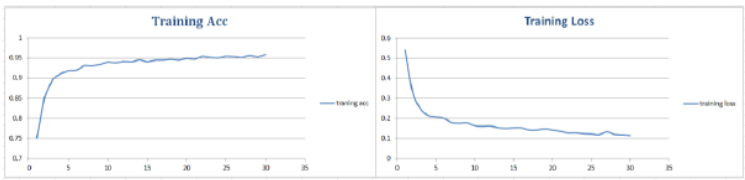


Figure 3. The accuracy and loss graphics during the training process

According to experiments, the test accuracy of the system was achieved as 87.65%. Table 3 shows the results of previous studies [10] together with the purposed CNN model.

Table 2. Performance evaluation CNN against state-of-the-art supervised algorithms on Pneumonia dataset

Algorithm	Accuracy
SMO*	76.76%
C4.5*	74.83%
3NN*	74.51%
Voting*	76.12%
WvEnSL3*	83.49%
Purposed CNN model	87.65%

* The results are taken from [10]

CNN is a modern algorithm used in image processing and classification. It is also a powerful method for classification problems and used in many other machine learning problems. In this study, a CNN model was used for the classification of Pneumonia by using X-ray images. According to the experimental results in this study, the proposed CNN model achieves more accurate results than the compared supervised algorithms.

References

[1] Suzuki, K. (2012). A review of computer-aided diagnosis in thoracic and colonic imaging. *Quant. Imaging in Med. Surg.,* **2**(3), 163-176.

[2] Anwar, S.M., et al., (2018). Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems,***42**(11), 226.

[3] Kermany, D.S., et al., (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell,***172**(5): 1122-1131.

[4]Anonymous, W*orld Health Organization*. https://www.who.int/maternal_child_adolescent/news_events/news/2011/pneumonia/en/. [Acces Date: 16/06/2019]

[5] Antin, Benjamin, Joshua Kravitz, and Emil Martayan (2017). Detecting Pneumonia in Chest X-Rays with Supervised Learning. *Semanticscholar.org.*

[6] Rajpurkar, Pranav, et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv*,1711.05225 .

[7] Park, M., Jin, J. S. ve Wilson, L. S. (2004). Detection of abnormal texture in chest X-rays with reduction of ribs. *Proceedings of the Pan-Sydney area workshop on Visual information processing*, 71-74.

[8] Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, **7**, e6201.

[9] Choudhari, S., & Biday, S. (2014). Artificial Neural Network for SkinCancer Detection. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* **3**(5), 147-153.

[10] Livieris, I. E., Kanavos, A., Tampakas, V., & Pintelas, P. (2019). A Weighted Voting Ensemble Self-Labeled Algorithm for the Detection of Lung Abnormalities from X-Rays. *Algorithms*, **12**(3), 64.

[11] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks. *An Overview and Application in Radiology. Insights into imaging*, **9**(4), 611-629.

A Novel Initial Centroid Selectıon Algorithm for Clustering

A . Namburu[1], M. Q. Hatem[2]

[1]VIT University Amaravathi, India, namburianupama@gmail.com
[2]Middle Technical University Dyala, Iraq, M_mu7@yahoo.com

Abstract

Image segmentation is a crucial and primary step in image processing and it has numerous applications in recognition and detection. Image segmentation is performed mainly using classification and clustering. Classification requires prior information and needs operator intervention in performing segmentation. Clustering is preferred as it is unsupervised and does not require prior information. However, the clustering algorithms require initial centriods in order to obtain the clusters. The wrongly chosen clusters results in local minima producing invalid segmentation regions. In this paper a novel initial centroid selection algorithm is presented which assists the clustering algorithm to result in the region close to ground truth in limited iterations.

Introduction

Image segmentation is the processes of extracting relevant regions from the image. The segmentation can be performed using clustering and classification techniques. The clustering is a unsupervised technique and preferred over classification for segmenting the regions. Clustering is the grouping of similar data items into same data set. K-means is the most popular among the clustering techniques. However, the clustering techniques require initial centroids to segment the image into clusters. Different initial centroids result in different clusters.Hence, it is essential to identify the right initial centroids that are consistent with the data distribution. Every clustering algorithm requires the initial centroids to obtain the clusters. Initially, the centroids were initialized by assigning the data elements to random clusters and finding the mean value of the clusters for calculating the initial centroids[1]. However random selection of centroids resulted in different results at every run of the clustering algorithm.

A modified random selection of centroids is presented in [2] that embeds euclidean distance to assign the data elements to the clusters. In [3] proposed an optimized centroid selection for k-means algorithm. In this algorithm the author has spread the initial centroids in feature space so that the distance between them are as afar as possible. This has the advantage of making the initial centroids as cluster centroids and are efficient that random initialization. In [4] a minimum attribute selection by converting the data sets into positive and negative space. The algorithm produced best results but computationally it is very expensive. Improved k-means algorithm is proposed in [5] with better initial centroids obtained using weighted average technique.The k-means algorithm based on the improved centroids resulted in less iteration in clustering. However, this technique requires desired cluster number as input.

Numerous algorithms were proposed in literature [6, 7, 8, 9, 10] to find the optimized centroids for clustering algorithm that improves the clustering efficiency, reduced iterations by faster convergence and reduced computational cost. However, no specific algorithm is suitable for all type of data.Hence, the initial centroid selection is still challenging as it has applications in data clustering, speech and image clustering as well. Anupama et al.,[10] has proposed optimized centroids based on histogram peaks. Researchers used means value of a histogram as centroid, maximum peak associated intensity value as a centroid. However, to compute the histogram peaks the range of histogram values need to be specified. In order to over come the problems associated with these algorithms a novel initial centroid selection is proposed in this method that computes the the centroids close to the distribution of the data. The organization of the later paper is as follows: The background of the proposed algorithm for image segmentation is discussed in Section 2. The proposed initali centroid selection is presented in Section 3. Implementation and experimental results of the proposed algorithm is presented in section 4. Conclusions and the future scope of the proposed algorithms are presented in Section 5.

Background

K-means clustering

[11] is an algorithm used to automatically partition individual of the given universe into n groups. The general steps in k-means clustering algorithm applied to MR brain image are shown in Algorithm 1.

Algorithm 1 Algorithm to segment MR brain image using k-means.

INPUT: Data set X, number of clusters k .
OUTPUT: k clusters extracted from brain.

1: Initialize the cluster centroids C_j for $j = 1, 2, ..., n$ clusters .
2: Determine the distance d_{ij} between every pixel x_i in the image X to the cluster centroids.

$$d_{ij} = argmin(|| x_i - C_j ||) \quad (1)$$

3: Cluster/group the individuals based on minimum distance.
4: Calculate the new centroid for new clusters.

$$C_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

5: Iterate steps 2 to 4 until the clusters are stable.
6: Extract the tissues from the stable clusters.

The advantages of k-means include:

1. Simple and fast clustering algorithm.
2. Works effectively to segment the round objects.
3. Simply uses distance measure to group the clusters.

In spite of these advantages, the k-means algorithm suffers from local minima problem, initialization of centroids and pixel can belong to single cluster at a time.

Proposed Method

The proposed method is explained in the following steps.

1. Obtain the unique data values that are repeated in the data set X.Let they be Unq(k).
2. Find euclidean distance between each xi in X to Unq(k).
3. For every xi we get k distances for every unique Unq(k). Let the distances be dik.
4. Find the minimum and maximum distances of every k distances of dik. We obtain x minimum distances and x maximum distances.
5. Find the average of minimum distances to find a threshold t1.
6. Find the average of maximum distance to find a threshold t2.
7. if the data xi is less than t1 group all into one cluster, else if data t1 <= xi <t2 group that data into one cluster else group them into other.
8. compute the average of all the clusters to find the initial centroids. This algorithm generates three centroids which is suitable to segment the data into three groups. If the algorithms needs to generate more number of clusters, the step 4 is changed to obtain minimum, average and maximum distance. With this we obtain t1, t2 and t3. This will generate four centroids.

Experimental Results

The proposed technique is applied to magnetic resonance brain images. The K-means algorithm is implemented and compared with random initialization[1], optimized histogram centroids[10] and the proposed method. Table 1 shows

Table 1: Comparison of proposed method for phantom image segmentation [12]

Initialization	No of Iterations	Execution time in sec
Histogram	10	1.87
Random	43	16.2
proposed	8	1.4

the execution of proposed method on phantom image for segmenting the image into white matter, grey matter and cerebro spinal fluid. Initial three centroids are calculated as with the proposed method and the k-means is performed iteratively to obtain the stable clusters. The proposed method of selecting the initial centroids is efficient when compared to the existing methods.

Conclusion

In this paper, a novel initial centroid selection algorithm is proposed to identify the centroids close to the distribution of data. The proposed method assist in avoiding local mimima problem and takes less time and iterations to make the clusters stable. The methods works very well for applications that need three centroids as input. The more the clusters needed the algorithm need to be modified for intermediate thresholds.

References

[1] R. O. Duda, P. E. Hart et al., Pattern classification and scene analysis. Wiley New York, 1973, vol. 3.
[2] F. Yuan, Z.-H. Meng, H.-X. Zhang, and C.-R. Dong, "A new algorithm to get the initial centroids," in Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826), vol. 2. IEEE, 2004, pp. 1191–1193.
[3] A. R. Barakbah and A. Helen, "Optimized k-means: an algorithm of initial centroids optimization for k-means," in Proc. Seminar on Soft Computing, Intelligent System, and Information Technology (SIIT), Surabaya, 2005.
[4] M. Yedla, S. R. Pathakota, and T. Srinivasa, "Enhancing k-means clustering algorithm with improved initial center," International Journal of computer science and information technologies, vol. 1, no. 2, pp. 121–125, 2010.
[5] M. S. Mahmud, M. M. Rahman, and M. N. Akhtar, "Improvement of k-means clustering algorithm with better initial centroids based on weighted average," in 2012 7th International Conference on Electrical and Computer Engineering. IEEE, 2012, pp. 647–650.
[6] M. Erisoglu, N. Calis, and S. Sakallioglu, "A new algorithm for initial cluster centers in k-means algorithm," Pattern Recognition Letters, vol. 32, no. 14, pp. 1701–1705, 2011.
[7] Y. Ye, J. Z. Huang, X. Chen, S. Zhou, G. Williams, and X. Xu, "Neighborhood density method for selecting initial cluster centers in k-means clustering," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2006, pp. 189–198
[8] W. Kwedlo and P. Iwanowicz, "Using genetic algorithm for selection of initial cluster centers for the k-means method," in International Conference on Artificial Intelligence and Soft Computing. Springer, 2010, pp. 165–172.
[9] L. Bai, J. Liang, and C. Dang, "An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data," Knowledge-Based Systems, vol. 24, no. 6, pp. 785–795, 2011.
[10] N. Anupama, S. Kumar, and E. Reddy, "Rough set based mri medical image segmentation using optimized initial centroids." International Journal of Emerging Technologies in Computational and Applied Sciences., vol. 6, no. 1, pp. 90–98, 2013.
[11] J. T. Tou and R. C. Gonzalez, "Pattern recognition principles," 1974.
[12] [Online]. Available: http://www.bic.mni.mcgill.ca/brainweb/

Control of Human-Robot Interactive Haptic-Teleoperation System by Fuzzy Logic and PID Control Methods

T. Abut[1], S. Soyguder[2]

[1]Mus Alparslan University, Mus, Turkey, tayfunabut@gmail.com

[2] Firat University, Elazig, Turkey, ssoyguder@firat.edu.tr

Abstract

Human-robot interaction (HRI) has become a very popular study topic in recent years. It is very important to control and overcome the problems of these systems. In this study, it is aimed to realize the bilateral control of the haptic-teleoperation system consisting of real and virtual Phantom Omni haptic robots. The dynamic equations of the robots that make up the system were obtained by using the Lagrange-Euler method. Based on the structural similarity between the real Phantom Omni master robot and the virtual slave Phantom Omni robot, the system is modeled. A visual interface is designed to visualize the movements of the slave robot. Fuzzy Logic and PID control methods were used to control the system in an experimental environment. These methods were carried out in an experimental environment and the results were compared and examined.

Introduction

Teleoperation systems are a robot technology that enables human-robot interaction and enables the user to prepare and develop tasks for unknown environments. Hapticteleoperation systems are widely used[1-4]. One of the most important sources of uncertainty in haptic-teleoperation systems is the dynamic models of robots due to their complex nonlinear structures. Therefore, the kinematic and dynamic robot model reveals parametric uncertainties that lead researchers to use adaptive algorithms in control design. Based on this observation in the literature, the fuzzy logic control method was preferred and tried to overcome the control problems arising due to the uncertainties of dynamic and kinematic connections. In this study, it is aimed to perform two-way control of teleoperation system consisting of real and virtual Phantom Omni haptic robots. The dynamic equations of the robots forming the system were obtained by using Lagrange-Euler method. The system is modeled based on the structural similarity between the actual Phantom Omni master and the virtual slave Phantom Omni robot. The MATLAB virtual reality toolbox is used to visualize the slave virtual robot. Two-way control of the system has been realized by using Fuzzy Logic and classical PID control methods. The parameter values used when creating a virtual slave robot are the factory production dimensions of the Phantom Omni haptic robot in our laboratory. Performance results are given graphically and examined.

Mathematical Model of System

Lagrange-Euler method was used to obtain the dynamic equations of the system. The equations of motion of the master and slave robot are given below.

$$M_m(q_m)\ddot{q}_m + C_m(q_m,\dot{q}_m)\dot{q}_m + G_m(q_m) = \tau_h + \tau_m \qquad (1)$$

$$M_s(q_s)\ddot{q}_s + C_s(q_s,\dot{q}_s)\dot{q}_s + G_s(q_s) = \tau_s - \tau_e \qquad (2)$$

The control of the robots was realized by using the equations of the first three axes, which are the basic axes of the robots. The equations of motion of the master and slave robot are given below. $q_i, \dot{q}_i, \ddot{q}_i$, and τ_i respectively represent position, speed, acceleration, and control torque. The $i \in \{m, s\}$ indices represent the master and slave robots respectively. $M(q_i) \in R^{3*3}$ is a positively defined symmetric matrix, the inertia matrix, $C(q_i) \in R^{3*3}$ Coriolis and centrifugal forces matrix, $G(q_i) \in R^{3*3}$ shows the weight forces. τ_m, τ_s refers to torques acting on master and slave robots respectively. τ_h and τ_e torques corresponding to the external forces exerted by the user and acting on the environment from the system. Figure 2 shows the Phantom Omni haptic robot and its rotational axes. Table I shows the physical parameter values of the Phantom Omni haptic robot. As shown in Figure 1, the Phantom Omni haptic robot has 6 rotating joints, but the first 3 joints are active and the 3 wrist joints are passive, ie not motor driven, but the robot has 6 encoders.

Table 1. Parameters of Haptic Robot

Description& Symbol	Units &Value
Gravity	~1,47 g
Dimension	~168W*203D mm
Inertia (I)	~45 g
Position resolution (θ)	~0.055 mm
Maximum improvement force (F_{max})	3.3 N
Force feedback (F)	x, y, z
Position measurement	x,y,z Pitch,roll,yaw
Interface	Parallel port
Supported Platforms	Intel or AMD based -PCs

The physical parameters of the slave virtual robot are obtained from the factory production dimensions of the 6-degree Phantom Omni haptic robot, which we use in real time in our laboratory. Figure 1 shows the Phantom Omni haptic robot, CAD model and visual interface.

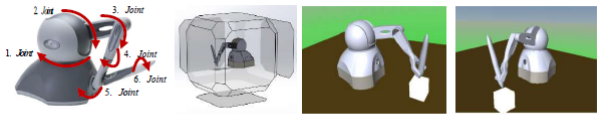


Figure 1 The Phantom Omni haptic robot, CAD model and visual interface.

Controller Design of the System

Fuzzy Logic (FLC) and classical PID (proportional integral derivative) control methods were used in the position control of the haptic-teleoperation system. With these controllers applied to the system, this error is tried to be minimized. In this study, traditional PID (proportional integral derivative) control method and Fuzzy Logic control method are used for the control of the system. u controller output is called Kp proportional gain Ki integral gain, Kd derivative gain, and e error signal. Fuzzy Logic and fuzzy set theory were introduced by the Azerbaijani professor in 1960. Lotfi A. Zadeh (California University, Berkeley). In this work, he attributed the reason that people can better control some systems than machines because people have the ability to make decisions by using certain (uncertain) information that cannot be expressed with certainty. The block diagram of a fuzzy logic controller is shown in Figure 2.

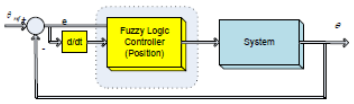
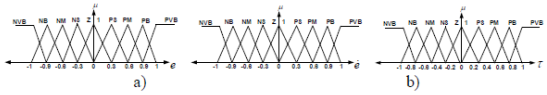


Figure 2. Block diagram of a fuzzy logic controller

Table 2. Rule Base Created for FLC

$\dot{e}$ / e	NB	NS	Z	PS	PB
NB	PB	PVB	PS	PM	PB
NS	PM	PS	PB	PS	PM
Z	PB	PM	Z	PM	PB
PS	PM	PS	PB	PVB	PS
PB	PM	PB	PS	PVB	PB



a) b)

Figure 3. Membership functions defined for input value a) e and b) $\dot{e}$ c) output value τ

Experimental Results

In this section, experimental studies are carried out using the equations of motion of the master and slave robots. The designed interface provides visual feedback to the user. The control variables of the system are joint angles and force values. The basic axes of the robots $\theta\square$, $\theta\square$, $\theta\square$ that is, the first three axes were checked. Simulation results obtained from the control of the system are given in the graphs below. $\tau\square = J\square\ \square\ F\square$ and $\tau\square = J\square\ \square F\square$ in the form. The relationship between the human operator and the environment is modeled as a spring-damper model of contact with a cube-shaped object formed in a virtual environment. Figure 4 gives an overview of the physical model modeled as a mass-spring-damper system, and the Haptic-Teleoperation system, the interaction model with the human operator and the environment.

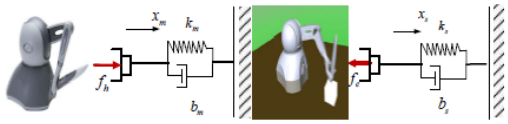


Figure 4. General view of the haptic-teleoperation system

$$f_h = k_m(x_m - x_{md}) + b_m(\dot{x}_m - \dot{x}_{md}) \quad (3)$$

$$f_e = k_s(x_s - x_m) + b_s(\dot{x}_s - \dot{x}_m) \quad (4)$$

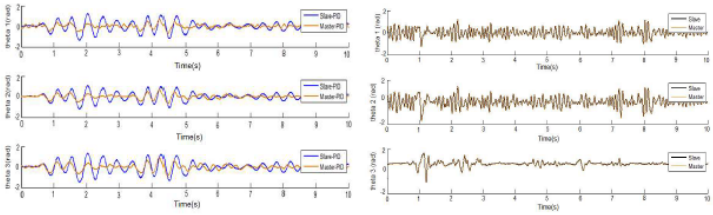


Figure 5. Position graphs of the system's joint angles obtained using a) PID and b)fuzzy logic control method

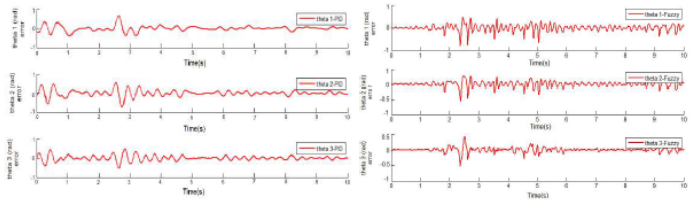


Figure 6. Position error graphs obtained using a) PID and b)fuzzy logic control method of joint angles of the system

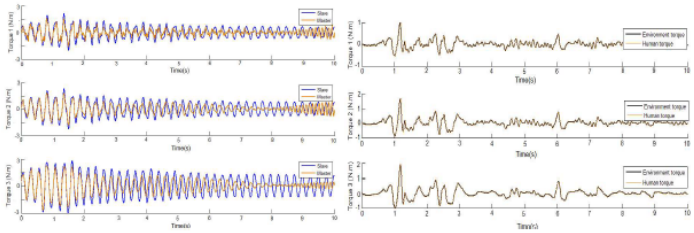


Figure 7. Torque values obtained by human-operator and environment when system interacts using a) PID and b)fuzzy logic control method

Results

In this study, a real-time bilateral haptic-teleoperation study was performed between the master 6-DOF Phantom Omni haptic robot and the slave virtual 6-DOF robot Phantom Omni haptic robot. As a result, the Fuzzy Logic control method performed better for the haptic-teleoperation system than PID control method. In addition, the PID control method showed poor performance as shown in the graphs. As a result of the experimental studies, it is shown in the graphs that the position and force values are largely followed by teleoperations between the real master robot-slave virtual robot.

References

[1] Ateş, G., Majani, R., & Dede, M. İ. C. (2019). Design of a teleoperation scheme with a wearable master for minimally invasive surgery. In New Trends in Medical and Service Robotics(pp. 45-53). Springer, Cham.

[2] Zadeh, L.A. (1988), "Fuzzy logic", Computer, Vol. 21 No. 4, pp. 83-93.

[3] Shahdi, A., & Sirouspour, S. (2012). Adaptive control of bilateral teleoperation with time delay. International Journal of Intelligent Mechatronics and Robotics (IJIMR), 2(1), 1-27.

[4] Soyguder, S., & Abut, T. (2016). Haptic industrial robot control with variable time delayed bilateral teleoperation. Industrial Robot: An International Journal, 43(4), 390-402.

Self-Tuning PID type Fuzzy Impedance Control and Performance Evaluation of a Teleoperation System

T. Abut[1], S. Soyguder[2]

[1]Mus Alparslan University, Mus, Turkey, tayfunabut@gmail.com

[2] Firat University, Elazig, Turkey, ssoyguder@firat.edu.tr

Abstract

Teleoperation systems are defined as systems that provide human-robot interaction (HRI). The control of these systems in the simulation environment is important in terms of preventing errors detected during the algorithm development stages. Parameter uncertainties and dead zone problems of robots, which are one of the main problems of these systems, constitute huge problems in the performance of the systems. In this study, self-tuning PID type fuzzy logic controller and PID control methods are used to overcome the mentioned problems. A virtual interface is designed to visualize the movements of the slave robot, one of the robots that make up the system. These methods were carried out in a simulation environment and the results were compared and examined.

Introduction

Teleoperation systems are robot technology that enables human-robot interaction and enables the user to prepare and develop tasks for unknown environments. A teleoperation system consists of a human operator, a master robot, a communication channel, a slave robot and the environment in which the dependent robot interacts. For this purpose, the master and slave robots exchange control signals over the communication channel and the force/power interaction is reflected back to the operator. The main purpose of such systems is to extend the user's manipulation capabilities to a remote environment. Hapticteleoperations studies have been studied by various researchers in the literature and are still continuing today[1-3]. One of the most important sources of uncertainty in hapticteleoperation systems is the dynamic models of robots due to their complex nonlinear structures. Specifically, it is difficult to identify and predict the friction and high-frequency dynamics of robotic systems and often results in errors. On the other hand, it is impossible to mathematically derive the joint angle of a robot manipulator from its final position and orientation in the work area. Therefore, the kinematic and dynamic robot model reveals parametric uncertainties that lead researchers to use adaptive algorithms in control design. Based on this observation in the literature, the fuzzy logic control method was preferred and tried to overcome the control problems arising due to the uncertainties of dynamic and kinematic connections. In this study, it is aimed to control the teleoperation system consisting of singles degree of freedom master and slave robot and two-way control of the
system was realized by using self-tuning PID type fuzzy and classical PID control methods.

Modeling of System

Lagrange-Euler method was used to obtain the dynamic equations of the system. Robots have a single degree of freedom. The control of the robot was realized by using the equations of the single axis. The equations of motion of the master and slave robots with single degree of freedom are given in equations 1 and 2 below.

$$I_m \ddot{q}_m + b_m \dot{q}_m = f_h + f_m \quad (1)$$

$$I_s \ddot{q}_s + b_s \dot{q}_s = f_s - f_s \quad (2)$$

The equations of motion of the master and slave robot are given below. $q_i, \dot{q}_i$, $\ddot{q}_i$, and τ_i respectively represent position, speed, acceleration, and control torque. The $i \in \{m, s\}$ indices represent the master and slave robots respectively. I_m and I_s represent the moment of inertia and b_m and b_s show the damping coefficients of the robots. f_m, f_s refer to torques acting on the master and slave robots respectively. f_h and f_e represent the force corresponding to the disturbing forces exerted by the user and acting on the environment from the system. The relationship between the human operator and the environment the contacts between a virtual wall and the end of the robot is modeled as a spring-damper system. Table I shows physical parameters of single degree of freedom robots.

Table I. Physical parameters of single degree of freedom robots

Description& Symbol	Units &Value
Mass (m_m)	0.257 kg
Length of arm (lm)	0.156 m
Inertia (J_m)	0.012 kgm^2
Spring coefficient (k_m)	0.01 N /m
Viscous damping coefficient(b_m)	0.0024 N s/m
Mass (m_s)	0.257 kg
Length of arm (l_s)	0.156 m
Inertia (J_s)	0.012 kgm^2
Spring coefficient (k_s)	0.01 N /m

Controller Design of the System

Self-tuning PID type fuzzy and classical PID (proportional integral derivative) control methods were used in the position control of the haptic-teleoperation system. With these controllers applied to the system, the error is tried to be minimized. u controller output is called Kp proportional gain Ki integral gain, Kd derivative gain, and e error signal. When designing the controller, it is aimed to follow the reference value of the angular moving single degree of freedom robot. Figure 1 shows self-tuning PID type fuzzy logic controller block diagram.

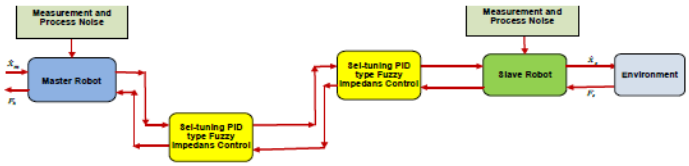


Figure 1. Self-tuning PID type fuzzy logic controller block diagram

The self-adjusting PID-type fuzzy controller is an automatic adaptive controller designed using a fuzzy logic controller to set the parameters of the PID controller with fuzzy control rules online. It constantly examines e and in the study, then provides the controller by finding the optimal values on the line adaptive on-line with the fuzzy control rules of three parameters (kp, kd, and ki) to ensure the better dynamic performance of the control parameters.

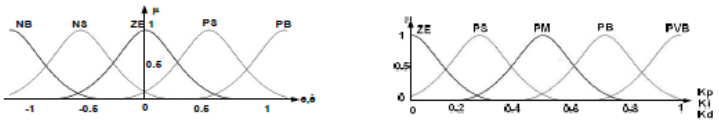


Figure 2. Membership functions defined for input and output value a) $e, \dot{e}$ and b) k_p, k_d, k_i

Table 2. Rule Base created for kp, ki, kd

kp / e \ ė	NB	ZE	PB
NB	NB	NB	ZE
ZE	NB	ZE	PB
PB	ZE	PB	PB

ki / e \ ė	NM	ZE	PM
NM	NM	NM	ZE
ZE	NM	ZE	PM
PM	ZE	PM	PM

kd / e \ ė	NS	ZE	PM
NS	NS	NS	ZE
ZE	NS	ZE	PM
PM	ZE	PM	PM

Simulation Results

In this section, experimental studies are carried out using the equations of motion of the master and slave robots. The designed interface provides visual feedback to the user. The variable parameters of the slave robot were transferred to the Quarc package program and the robot's 3D image was created. Table 3 shows comparison of control methods.

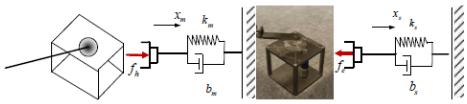
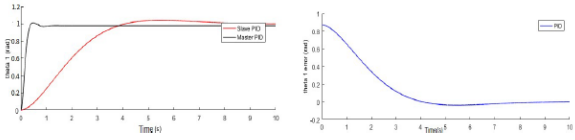
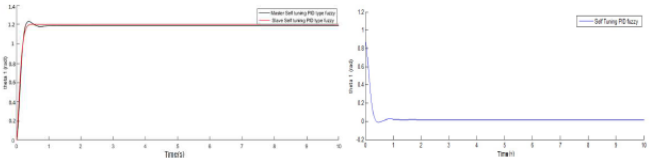


Figure 3. Simulation model of interaction of master and slave robots with environment



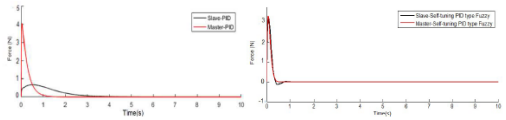
a) b)

Figure 4. a) Position and b) Position error response of joint angle obtained by the PID control method



a) b)

Figure 5. a) Position and b) Position error response of joint angle obtained by the Self-Tuning PID type fuzzy control method



a) b)

Figure 6. Force response of joint angle obtained by a) the PID control method and b) the Self-Tuning PID type fuzzy control method

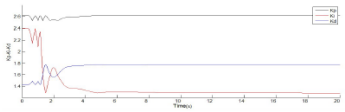
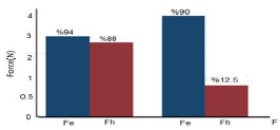


Figure 7. Control coefficients obtained by Self-Tuning PID type fuzzy control method

Table 3. Comparison of control methods

Control Method	Settling time (s)	Maximum exceed (%)	Steady-state error (rad)
Self-tuning PID type Fuzzy Control	1.3	3	0.08
PID	9	7	1



Results

A bilateral teleoperation study was conducted in a single degree of freedom simulation. As a result of the simulation studies, the reference position between the master robot and slave robot with teleoperations is shown in the graphs and tables that it mostly follows. Self-tuning PID type Fuzzy control method among the designed and applied control algorithms showed the best performance on the system. As a result of the simulations, useful information about the movement of the system is obtained.

References

[1] Basdogan, C., De, S., Kim, J., Muniyandi, M., Kim, H., & Srinivasan, M. A. (2004). Haptics in minimally invasive surgical simulation and training. IEEE computer graphics and applications, 24(2), 56-64.
[2] Yager, R. R., & Zadeh, L. A. (Eds.). (2012). An introduction to fuzzy logic applications in intelligent systems (Vol. 165). Springer Science & Business Media.
[3] Abut, T., & Soyguder, S. (2017). Real-time control of bilateral teleoperation system with adaptive computed torque method. Industrial Robot: An International Journal, 44(3), 299-311.

Effect of Type III Error on Decision Making Process

M. Mendeş[1], H. Mirtagioglu[2]

[1]Canakkale Onsekiz Mart University, Canakkale, Turkey, mmendes@comu.edu.tr

[2]Bitlis Eren University, Bitlis, Turkey

Abstract

Researchers generally focus on Type I (α) and Type II (β) error rates in their simulation studies which will be conducted to compare performances of different tests or approaches. Although majority of the simulation studies consider Type I (α) and Type II error rates (β), however, a third type of error (Type III error-) has been suggested in the literature. The Type III error rate is related to direction in the decision-making process. Therefore, making a rejection in the wrong direction is called a Type III error. A comprehensive simulation study was carried out to estimate Type III error rates for two independent group means under different sample size and effect size combinations under normal distribution. Results of the simulation study indicated that as sample sizes increased, the Type III error rates decreased. Decreases in Type III error rates were more prominent especially when effect size was greater than 0.60 regardless of sample size combinations. The effect of Type III error on interpretation of results was more obvious especially when sample size was small (n≤10) or studied with unbalance designs. γ

Introduction

Although Type I and II errors are the primary points of concern for researchers when conducting hypothesis testing, another type of error called Type III error (γ) may occur. The effect of this error will be more evident especially when sample size is small (Kaiser, 1960; Fowler, 1978; Leventhal and Huyhn, 1996; Leventhal, 1999; Mendeş, 2007; Heinz and Waldhoer, 2012; Rasch, 2012; Kaur and Stoltzfus, 2017). The Type III error occurs when a false null hypothesis is correctly rejected correctly but the claimed "direction" of truth is opposite of what it really is. Therefore, making Type III error may cause serious problems especially at the stage of interpretation of the results. For instance, suppose one is interested in investigating effect of two different treatments (A and B) on weight gain of broiler chickens.

Suppose the null (H0) and alternative (H1) hypotheses are set up as H0:μA=μB and H1:μA≠μB. For such cases, at the end of hypothesis testing, the researcher may be faced with making three types of error rates namely Type I, Type II, and Type III error rates. If a correct null hypothesis is rejected, then a Type I error is made. In other words, a Type I error would occur if the researchers conclude that treatment A is better than treatment B (or vice versa) when two treatments are equally effective in terms of weight gain in reality. In contrast, if a false null hypothesis is accepted, then a Type II error is made. In other words, a Type II error would occur if our sample data prompt us to claim that the treatment A is equal or similar to the treatment B when two treatments are not equally effective in terms of weight gain. Now the question is: how might a Type III error occur for this study? The Type III error would occur if a false null hypothesis is rejected but making a rejection wrong direction. In other word, making a rejection in the wrong direction is called Type III error (γ)(MacDonald, 1999). Suppose the treatment A is actually better than treatment B in terms of weight gain(μA>μB), but the sample data showed that mean of the treatment B is bigger than mean of the treatment A and this difference is statistically significant. Therefore, it is concluded that the treatment B is more effective than the treatment A (μB>μA) based on the sample evidence. In this case, Type III error would be occurred. That is, although it is known that the treatment A is actually better than the treatment B in terms of weight gain, based on our sample data set we conclude that the treatment B is better than the treatment A, and then a Type III error would occur. In this case, at the end of hypothesis testing the following results will be achieved for this study.

a) The null hypothesis is rejected. That is, there is a statistically significant difference is observed between means,
b) The sample data set indicates that μB>μA even though μA>μB in reality and
c) Based on the sample evidence, it is concluded that the treatment B is better than the treatment A in terms of weight gain.

In this case, the false null hypothesis is correctly rejected. However, the direction of our inference is not true. Direction of the inference is opposite from the real truth of the situation (Kimball, 1957; Kaiser, 1960; Shaffer, 1972; Hopkins, 1973; Leventhal, 1999; MacDonald, 1999; Sharon and Carpenter, 1999; Huynh, 2004). It is because the treatment A is truly better than the treatment B but it is concluded that the treatment B is better than the treatment A. That is, a Type III error is made. Therefore, the Type III error term is used for designation of this kind of inferential

error and it will be important especially at the stage of interpretation of the results. As it has been indicated by previous studies (Çamdeviren and Mendeş, 2005; Mendeş, 2007; Heinz and Waldhoer, 2012) making the Type III error may lead to get unreliable results, especially when sample size is small. Therefore, although it is generally ignored by the researchers in practice, previous studies showed that the Type III error rate might have an important impact on the reliability of the results especially when sample size is small. For example, Mendeş (2007) reported that the Type III error rate was affected by small sample size and effect size, whereas it was not affected by distribution shape. Likewise, Heinz and Waldhoer (2012) investigated the effect of Type III error rate in epidemiological maps and they reported that Type III error rate should be taken into consideration when interpreting results presented in epidemiological maps. Considering the studies on biological and agricultural sciences have generally been conducted with small sample size, investigating the effect of the Type III error rate might provide some extra information to the researchers. At this stage, it is very important to be aware of all error types (i.e. Type I, Type II, Type III) in hypothesis testing and decision making process. In this study, it has been aimed at investigating the effect of Type III error rate on the reliability of the results by considering different experimental conditions. For this purpose, a comprehensive Monte Carlo simulation study was carried out.

Materials and Methods

A Monte Carlo Simulation has been carried out to estimate Type III error rates and determine the factors that might affect the Type III error rate in comparing two independent group means across a variety of experimental conditions. Type III error rates have been evaluated under five different effect sizes (δ=0.20, 0.40, 0.60, 0.80, 1.0), various sample sizes patterns (2:2, 4:4, 6:6, 8:8, 10:10, 20:20, 30:30, 2:4, 10:15, 15:30), and variance ratios (homogeneity and heterogeneity of variances). For this purpose, random numbers generated from IMSL library of Microsoft FORTRAN Developer Studio. In order to create differences among the groups being compared, constant numbers in standard deviation form (δ=0.2, 0.4, 0.6, 0.8, 1.0) were added to the random numbers of the first population. The effect sizes (standardized mean difference) of 0.20 standard deviation approximate those suggested by Cohen (1988) to represent small effect sizes and the effect sizes of 0.8 and more standard deviation approximate to represent large effect sizes. In this study, we used 1.0 standard deviation to represent large effect size.

Huynh (2004) illustrated how to calculate Type III error rate by means of Proc Power in SAS/STAT. In the simplest case, two groups with equal variance, the Type III error rate can be analytically derived from the non-central t distribution. Therefore, Type III error probability:

$$\Pr(\text{Type III error}\,|\,\delta>0)=\Pr\left(\sqrt{\frac{n_1.n_2.(n_1+n_2-2)}{(n_1+n_2)((n_1-1)S_1^2+(n_2-1)S_2^2)}}.(\bar{y}_1-\bar{y}_2)<t_{n_1+n_2-2;\frac{\alpha}{2}}\right)=$$

$$F(t_{n_1+n_2-2;\frac{\alpha}{2}}\,|n_1+n_2-2(\frac{\delta}{\sigma}\sqrt{\frac{n_1.n_2}{n_1+n_2}}))$$

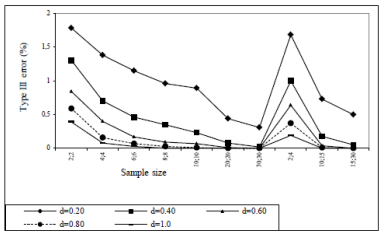
where $\bar{y}_1$ and $\bar{y}_2$ are the sample means, $F(t_{n_1+n_2-2;\frac{\alpha}{2}}\,|n_1+n_2-2(\frac{\delta}{\sigma}\sqrt{\frac{n_1.n_2}{n_1+n_2}}))$ is the distribution function of the non-central t-distribution with n_1+n_2-2 degrees of freedom.

For each experimental condition, random samples were drawn from normal populations and the t-test was performed. The t-value was then compared with the corresponding critical value, and the decision to reject, or fail to reject the null hypothesis was recorded. When population mean differences existed, rejections for the t-test were further classified as being either correct or incorrect. These steps were repeated 100,000 times for each combination of sample size, variance pattern, and effect size, and across all levels of k. Type III (rejection of a false null hypothesis in the wrong direction) error rates were computed for conditions where the null hypothesis was false. The predetermined alpha level was 0.05.

Table 1. Type III error rates estimates based on sample size and effect size

n	δ=0.20	δ=0.40	δ=0.60	δ=0.80	δ=1.0
2:2	1.78	1.30	0.84	0.59	0.39
4:4	1.38	0.70	0.40	0.16	0.08
6:6	1.15	0.46	0.17	0.07	0.03
8:8	0.96	0.35	0.09	0.03	0.00
10:10	0.89	0.23	0.07	0.01	0.00
20:20	0.44	0.08	0.01	0.00	0.00
30:30	0.31	0.02	0.00	0.00	0.00
2:4	1.68	1.00	0.64	0.37	0.19
10:15	0.73	0.18	0.04	0.01	0.00
15:30	0.50	0.05	0.00	0.00	0.00

Figure 1. Changes in the Type III error rates (%) based on sample size and effect size



Results and Discussion

Results of the simulation study are given in Table 1 and Figure 1. Table 1 contains the empirical Type III error rate of the two sample t-test at the α=0.05 significance level when samples are drawn from two normal distributions with various mean differences or effect size (δ) and sample sizes. When the sample size and δ were increased, the Type III error rates decreased. For example, when both effect size (δ=0.20) and sample size (2:2) were very small, the t-test rejected the null hypothesis in 5160 of the 100000 random samples. Of these sample rejections, 3380 would have led researchers to correctly conclude that the mean of population A was higher than the mean of population B. That resulted in a power estimate of 3.38 % ((1-β-γ) =3380/100000). The remaining samples, totaling 1770, would have led researchers to incorrectly conclude that the mean of population A was lower than the mean of population B, resulting in a Type III error rate of 1.78 % (γ=5160-3380/100000) (Table 1).
Under the same conditions, the t-test exceeded the critical value associated with the 0.05 level in 6.180 of the 100000 random samples of size (6:6). Of these samples, 5030 would have led researchers to correctly conclude that the mean of population A was higher than the mean of population B ($\mu_A > \mu_B$), which resulted in a power estimate of 5.03 %. The remaining samples, totaling 1150, would have led researchers to incorrectly conclude that the mean of population A was lower than the mean of population B, resulting in a Type III error rate 1.15 % ($\mu_A < \mu_B$). The t-test exceeded the critical value in 11530 of the 100000 random samples of size (30:30). Of these samples, 11220 would have led researchers to correctly conclude that the mean of population A was higher than the mean of population B ($\mu_A > \mu_B$) that resulted in a power estimate of 11.22 %. The remaining samples, totaling 310, would have led researchers to incorrectly conclude that the mean of population A was lower than the mean of population B, resulting in a Type III error rate 0.31 %. The Type III error rate was affected by total sample sizes rather than inequality in sample sizes. This effect was more pronounced for δ=0.40 and smaller. For example, for δ=0.20; n1=10 and n2=15 (total of 25 observations) resulted in Type III rate of 0.73 while n1=15 and n2=30 (total of 45 observations) resulted in Type III rate of 0.50. Though the ratio of sample sizes increased (n2:n1=15:10=1.5 to n2:n1=30:15=2), Type III error rate decreased as much as 0.23 percent. Those results are consistent with the findings reported by McDonald (1999), Leventhal and Huynh (1996) and Mendes (2007). On the other hand, Kaiser (1960) and Leventhal and Huynh (1996) reported that the Type III error rate was always less than 0.5α. Therefore, the difference between the tests was always less than 0.5α when two group means were compared. However, nothing has been reported for the comparison of more than two group means. In this simulation study, all Type III error estimates were found less than 0.5α as well.

Results of this study suggest that it is an important issue to be familiar about effect of all kind of errors on reliability of the results of our studies. Although the Type III error is rarely discussed in the literature, considering this kind of error along with Type I and Type II error rates in the hypothesis testing will be beneficial especially at the stage of interpretation of the analysis results. Otherwise, it may cause some problems related to the interpretation of the results. As it is stated by Cohen (2008), the reason for why the Type III error is so rarely mentioned is that the Type III error rate is thought to be always very small. However, when power is very low, nearly half of all significant results will actually be Type III errors.

Conclusion

Although the probability of a Type III error in most circumstances is so small and negligible level nevertheless, knowing the Type III error rate for our study can be beneficial for some cases. Heinz and Waldhoer (2012) reported that Type III error rate should be taken into consideration when interpreting results presented in epidemiological maps. The effect of Type III error on research conclusions and interpretations is more obvious especially when sample size is small and there is an imbalance in sample size. Therefore, the Type III error rate ought to be taken into consideration especially in small sample sizes. As a result, considering all kinds of error rates in decision making process will be beneficial in terms of getting more detailed and reliable results about our study.

References

[1]Cohen, B.H. (2008). Explaining Psychological Statistics, 3rd Edit.,John Wiley&Sons, ınc., Hoboken, New Jersey, USA. 807 p.
[2]Camdeviren, H., Mendes, M. (2005). A Simulation Study for Type III Error Rates of Some Variance Homogeneity Tests. Pakistan Journal of Statistics, **21(2):** 223-234.
[3]Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Second Ed. New Jersey:Lawrence Erlbaum Associates, Hillsdale.
[4]Heinz, H., Waldhoer, T. (2012). Relevance of the type III error in epidemiological maps. International Journal of Health Geographics, 11:34. doi:10.1186/1476-072X-11-34.
[5]Hopkins, B. (1973). Educational research and Type III errors. The Journal of Experimental Education, **41:** 31-32.
[6]Huynh, C.L. (2004). Estimation of Type III error and power for directional two-tailed tests under using PROC POWER. Paper 208-30. *www2.sas.com/proceedings/sugi30/208-30.pdf* . Access date: 7 Jan 2014.
[7]Kaiser, H.F. (1960). Directional statistical decisions. Psychological Review, **67:** 160-167.
[8]Kaur, P., Stoltzfus, J. (2017). Type I, II, and III Statistical Errors: A brief overview. International Journal of Academic Medicine, 3 (2):268-270.
[9]Kimball, A.W. (1957). Errors of the third kind in statistical consulting. J.Am.Stat. Assoc., **52:** 133-142.
[10]Leventhal, L., Huynh, C.L. (1996). Directional decisions for two-tailed tests: Power, error rates and sample size. Psychological Methods, **1(3):** 278-292.
[11]Leventhal, L. (1999). Answering two criticisms of hypothesis testing. Psychological Reports, **85:** 3-18.
[12]MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. The Journal of Experimental Education, **67:** 367-379.
[13]Mendes, M. (2007). The Effect of Non-normality on Type III error for comparing independent means. Journal of Applied Quantitative Methods, **2 (4):** 444-454.
[14]Rash, D. (2012). Hypothesis testing and the error of the third kind. Psychological Test and Assessment Modelling, **54(1):**90-99.
[15]Shaffer, J.P. (1972). Directional statistical hypothesis and comparisons among means. Psychological Bulletin, **77:**195-197.
[16]Sharon, S., Carpenter, K.M. (1999). The right answer for the wrong question: Consequences of Type III error for public health research. American Journal of Public Health, **89(8):** 1175-1181.

Open Source E-mail Forensic Tools: An Inside View

M. Harbawi[1], A. Varol[1]

[1]*Firat University, Elazığ, Turkey, malek@firat.edu.tr ,varol.asaf@gmail.com*

E-mail is a very effective method for exchanging digital messages using computer-based devices. Since the beginning of 2015, statistics estimate that, on average, there have been around 205 billion e-mails sent per day. Unfortunately, e-mail has been used as a carrier for malicious and spam contents as well as a medium of various criminal activities. The mechanisms used in e-mail make it a major contributor to digital forensics investigation, where the data and metadata carried by e-mails can be retrieved and analyzed for legislative processes. In an attempt of understanding e-mail forensic tools, this paper discusses and investigates a number of open-source software used for e-mail forensics investigation; followed by a practical application case showing the feasibility of the results reported by some of the tools. The results indicate that for the best outcome, e-mail forensics tools should be used in a combined way.

Background

E-mail is a very well-known communication technique that takes advantage of modern Information Communication Technology (ICT) to accomplish its required duty as a digital carrier between two parties. E-mail communication implies the use of a sender client-server system and a receiver client-server system as well. The communication process is done via a specified set of protocols, which directs the server on how and what to deliver and display to the involved parties. When a user creates and sends an e-mail, the workstation connects to the used e-mail server. Generally, Simple Mail Transfer Protocol (SMTP), a pick-and-drop server, is responsible for sending messages from the sender's e-mail to the destination e-mail address. The handed e-mail is delivered based on a unique Internet Protocol (IP) where the receiver's e-mail address belongs to, most likely, Post Office Protocol 3 (POP3) or Internet Message Access Protocol (IMAP); the e-mail will be stored in the deployed server until the receiver logs in the mailbox using username and password in order to check the incoming message.

The mechanism of e-mail communication is a key concept in digital forensics process. Generally, the uncommonly known information generated by e-mail exchanges is used for digital investigation purpose. Thus, digital investigators should be aware of these processes so that they can make the right decisions. Such decisions may cover which tools to use, what can be retrieved, and whether or not it is necessary to issue a warrant for Internet Service Provider (IPS) to hold a copy of the e-mails if there is any legal ground for such purpose.
Digital forensics deals with retrieving, extracting, scientifically analyzing and reporting in a readable way the indications and implications of digital evidence. In this regard, the application of digital forensics to the e-mail, as evidence source, is called e-mail forensics. This procedure has been practiced since the late 1990s and became popular in early 2000. The general process of digital forensics is literally applicable to e-mail forensics and the used steps are summarized in the following points [1, 2]:

- **Identifying and Retrieving**: As a standard procedure in digital forensics, digital forensics investigators initiate a warrant via the court for seizing any digital items related to the crime or the digital crime scene. Then a formal identification and retrieval procedure is taken to retrieve any vital information on the memory of the seized devices including e-mails copies.

- **Examining**: Next to the identification and retrieval, digital forensics investigators start examining the procedure of sent and received e-mails. The examining process includes the sender's e-mail address (X-Originating-E-mail, X-Sender, Return-Path, etc.), the initiation protocol (STMP or HTTP), message ID, e-mail hops, and IP address. In addition, determining and identifying the ISP for any possibility for server tracing of similar initiated e-mails [3].

- **Analyzing**: After inspecting and examining the e-mail header for vital information, forensics investigators usually use specialized tools to further analyze and report the forensic results. There are many tools and applications applicable for e-mail forensics, and the choice should be made based on the needs and circumstances. Nevertheless, the results of more than one tool can be combined for the optimal outcomes.

E-Mail Forensics Tools

In this study, we have considered only open-source tools used for e-mail forensics as a target of our testing and analyzing procedure. This is done for two main reasons: the availability of the tools and for the possibility of expanding the research by comparing the current results with the results obtained from other commercial or proprietary tools in the next part of the study. Even though there is a number of open-source tools suitable for e-mail forensics, it is impossible to cover all these tools in this paper. Thus, the choice made for this study is to investigate the following tools based on the provided criteria: *AccessData, eMailTrackPro, Paraben E-mail Examiner, Mail Xaminer and Aid4Mail*.

Case Study: Combined Tools Analysis Performance

One of the best-used methods while dealing with e-mail forensics tools is to combine the reports of more than one tool to complement the issued reports, verify the obtained results and clarify the outcomes. For this reason, we applied a practical analysis on a spam e-mail from an unknown source in two steps using two different tools as follows:

1. Analysis Using eMailTrackerPro

The typical version of the eMailTracnkerPro v10.0b standard edition was used on a machine equipped with Windows 10. The e-mail header was fed into the tool and the analysis was made. The analysis report is graphically presented in Figure 1 as a geographical map with a link between the sender and receiver, whereas Table I lists the detailed hops, IPs, and locations throughout the way from the sender to the receiver. As illustrated in Figure 1, eMailTracnkerPro is able to identify the initial source of the e-mail, which is a city in Australia in this case. However, the results may not be enough up to here. Therefore, to find more details that can assist in forensic analysis, we consider the information provided in Table II. The e-mail was initiated from Australia with the provided ISP and IP address, which are extremely important for the legal procedure. Validating the results using eMailTrackerPro cannot be achieved. In addition, it is impossible to further analyze each sole hop using the same tool. Thus, feeding the results into another e-mail forensics tool to further elaborate the outcomes can effectively enhance these results.

2. IP2LOCATION Analysis

To enhance the information provided by eMailTrackerPro, the hop IP when the e-mail had entered Turkey (5.23.8.21) is analyzed further with an online tool called IP2LOCATION [4]. This tool is used for extracting more details and verifying the obtained results. The provided results from IP2LOCATION is presented in Table II. These results are extremely important as they provide a verifying method for the results obtained from eMailTrackerPro as well as a more detailed hop analysis of the needed IP at any stage. A similar test can be applied using MxTOOLBOX [5] where very interesting results can be produced, such as identifying blacklisted IP addresses, their DNS, origins, as well as checking if the domain or the e-mail address is safe “healthy” or not.

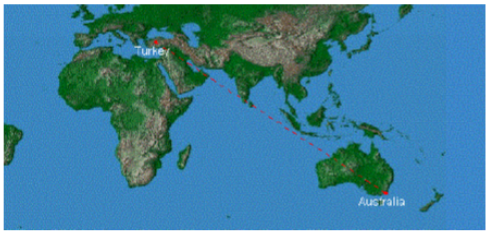


Figure 1. The map view of e-mail tracing using eMailTrackerPro

TABLE I. HOP DETAILS OF THE TRACKED E-MAIL

Hop IP	*Hop Name*	*Location*
10.6.1.1		(Private)
10.98.1.1		(Private)
10.6.1.254		(Private)
193.140.10.22		Turkey
193.140.10.221		Turkey
85.29.25.9	Host -85-29-25-9.resever.super online.net	Turkey
5.23.8.21	ix-6-0.trce1.it6-Ankara.as6453.net	Australia
195.219.50.165	If-37-3.t.cor1.fi0-frankfurt.as6453.net	Europe
195.219.50.138		Europe
89.149.181.1	et-7-1-0.atl11.ip4.gtt.net	Europe
173.241.130.54	total-server-solutions-gw.ip4.gtt.net	Australia
184.170.251.242	xe9-3.dist-aa.atl01.coloat.com	Australia
192.252.211.73	Colo.venatra.com	Australia

TABLE II. HOP DETAILS USING IP2LOCATION

IP Address	5.23.8.21
Country, Region, City	Turkey, Ankara, Ankara
Coordinates of City	39.91987, 32.85427 (39°55'12"E 32°51'15"N)
ISP	TATA Communications (Canada) Ltd.
Local Time	13 Jun, 2019 01:00 PM (UTC +03:00)
Domain	tatacommunications.com
Net Speed	(COMP) Company/T1
IDD & Area Code	(90) 0312
ZIP Code	12800
Weather Station	Ankara (TUXX0002)
Mobile Carrier	-
Mobile Country Code - MCC	-
Mobile Network Code - MNC	-
Elevation	876m
Usage Type	(DCH) Data Center/Web Hosting/Transit
Anonymous Proxy	No
Proxy Type	(DCH) Hosting Provider, Data Center or CDN Range
Proxy ASN	6453 Tata Communications (America) Inc
Proxy Last Seen	26 Days

Conclusion

In this paper, an insight on digital forensics from e-mail prospective was presented and discussed. In general, the paper addressed this issue from two aspects, that is e-mail metadata and suitable forensics tools applied to this metadata. There are various open-source software tools can be used for e-mail forensics, yet we have restricted the investigation to only five common tools in the field. The discussed tools vary in their abilities and features, which makes favoring one over the other is a challenging task. It is observed from the analysis that for optimal results, digital forensics investigators should apply more than one tool to the same e-mail analysis as they complement each other in the required tasks. To benefit from the presented work, further research should be done in the same field but on more e-mail forensics tools, especially proprietary tools. In addition, the research can be expanded in the direction of complementary online tools such as MxTOOLBOX that provides vital information on blacklisted IP addresses.

References

[1] Al-Zarouni, M. (2004, November). Tracing E-mail Headers. *In Australian Computer, Network & Information Forensics Conference* (pp. 16-30).

[2] Devendran, V. K., Shahriar, H. and Clincy, V. (2015). A comparative study of email forensic tools. *Journal of Information Security*, 6(2), 111-117.

[3] Sansurooah, K. (2009, March). A Forensics Overview and Analysis of USB Flash Memory Devices. In *Australian Digital Forensics Conference* (pp. 99-108).

[4] IP2LOCATION E-mail Analysis Tool. [Online]. Available at: http://www.ip2location.com/ (Accessed on June 13th, 2019)

[5] MxTOOLBOX E-mail Analysis Tool. [Online]. Available at: http://mxtoolbox.com/SuperTool.aspx (Accessed on June 13, 2019)

Using Production Time Data For Estimating The Setup Time Of The Cnc Machines

S. Kuzgun Akın[1], M. Sevüktekin[1]

[1]Bursa Uludağ University, Bursa, Turkey, sibel@yakamoz.net, sevuktekin@uludag.edu.tr

Abstract

In industry keeping track of production duration and numbers for each production order is important. It is preferable to collect all relevant data and calculate results accurately. However, collecting data is not an easy process always. In this article we tried to use multivariate linear regression analysis for estimating machine setup time and machining time for a single part. These values can be used for forecasting purposes for production planning.

Production data collection is a tricky process because most of the time blue-collar employees are involved in the process. Data collection should be simple and straightforward. If possible it should be automated using industry 4.0 solutions. In machining process CNC time can be collected directly from the CNC machine. However, the machine setup time is usually collected manually. Asking the employee to enter the setup time is open to faults. So, in this article we suggested to collect total production time (setup + CNC) and estimate the setup time.

Using real production data, we suggested 3 regression models to estimate setup time of the CNC/honing machines and forecast future total production time. We suggested to use machine learning to improve the parameters in time.

Introduction

By industry 4.0, collecting data from the CNC machines is considered as an automatical process. However, it is possible to read only machining time from the CNC machines. The setup time of the machine should be tracked manually. Most of the time the setup time is entered to the information system by the blue-collar workers. Usually it is convenient to collect total production time per work order, operation id and CNC machine. It is hard to make blue-collar workers to enter the setup time of the machine and machining time seperately. If it is applied, it will double the data collection costs.

On the other hand, it is important for a production planning engineer to know the machine setup time and unit machining time for producing a part for forecasting future production times per work order.

Total production time can be formulated as equals to machine setup time plus unit machining time multiplied by the number of parts.

In this paper, we tried to use available the total production time and the number of parts (quantity) data to estimate the machine setup time and the production time per part (unit production time) via linear regression analysis.

We started with surveying the literature about usage of regression analysis for production processes. We have seen that regression analysis is listed as one of the methods used for production cost estimation [4]. We have read that second degree multiple regression model and neural networks were applied for predicting surface rougness for CNC milling process of aluminium parts [2]. In another research [3], again for CNC surface milling, artificial neural networks and regression analysis were used to minimize machining time by configuring machine's working parameters. In the survey we couldn't have found any article about using regression analysis to estimate machine setup time.

Preliminary Work

The production data were collected for similar parts which were produced from the same metal alloy. All of the parts had a common form but their sizes differed. After filtering the data from the database using SQL, they were analyzed using Minitab. The irrelevant operation data like assembly, laser marking were deleted.

Before analying the data, in order to eliminate the possible data errors, extreme data were omited. For example it isn't possible a honing job to continue 94 hours. Most probably the operator forgot to end the job in the information system. If we consider the machine setup time, a job can't last less than 10 minutes. Extreme values like these weren't included into the analysis.

All data were divided into three groups based on the type of operation: Back surface machining, front surface machining and perforation, honing.

Units are as below:

Time: minute

Surface: dm^2

Machining speed: dm^2 / minute (Limit speed of the machine, 1 mm width is assumed)

Analysis of the back surface machining data

Firstly total time of a work order equation can be written as:

$$Time = \beta_0 + \beta\, quantity + u \quad (1)$$

Here quantity is the number of parts that were be produced in this work order.

Then the first regression coefficient β was defined as:

$$\beta = \beta_1 + \beta_2\, surface\ area / CNC\ speed \quad (2)$$

The regression equation for back surface machining becomes:

$$Time = \beta_0 + \beta_1\, quantity + \beta_2\, quantity\,.\,area / speed + u \quad (3)$$

If we calculate β_0 and β values, β_0 corresponds to the machine setup time and β corresponds to the unit production time.

Time is the response variable. Quantity, surface area, CNC speed are predictor variables. The number of data rows is 159.

At the end of the linear regression analysis [1], the equation was estimated as follows:

$$\widehat{Time} = 3.2 + 4.678\, quantity + 36.46\, quantity\,.\,area / speed \quad (4)$$

According to this formula, the machine setup time is 3.2 minutes. The unit machining time for a part with 1 dm^2 surface area is approximately 12 minutes using a CNC machine with 5 dm^2/minute maximum speed. These estimated coefficients are consistent with the real world observations.

Estimated	β_0	β_1	β_2
se	10.2	0.711	3.87
t	0.32	6.57	9.41
P	0.752	0.000	0.000

Table 1: The back surface machining coefficients

$R^2 = 70.57\%$

Adjusted $R^2 = 70.20\%$

Analysis of the front surface machining and perforation data

The total production time formula for the work order is the same.

$$Time = \beta_0 + \beta\, quantity + u \quad (5)$$

However, the definition of β coefficient was changed. Number of holes was added to the formula as another predictor variable.

$$\beta = \beta_1 + \beta_2\, surface\ area / CNC\ speed + \beta_3\, number\ of\ holes \quad (6)$$

Equation of the front surface machining and perforation operation:

$$Time = \beta_0 + \beta_1\, quantity + \beta_2\, quantity\,.\,area / speed + \beta_3\, quantity\,.\,holes + u \quad (7)$$

If the β_0 and β coefficients are estimated, β_0 will correspond to the machine setup time and β will correspond to the unit machining time.

Time is the response variable. Quantity, surface area, CNC speed, number of holes are predictor variables. The number of data rows is 149.

$$\widehat{Time} = 50.7 + 1.19\, quantity + 54.07\, quantity.area / speed + 0.08\, quantity.holes \quad (8)$$

The estimated machine setup time for the front surface machining and perforation operation is 50.7 minutes. It is longer than we expected. In the real world, the setup time is shorter. For a part with 20 holes and 1 dm^2 surface area (maximum speed of the CNC machine is 5 dm^2/minute) unit production time is 13.6 minutes and it is a reasonable time.

Estimated	β_0	β_1	β_2	β_3
se	12.8	1.30	6.29	0.0978
t	3.95	0.92	8.60	0.82
P	0.000	0.361	0.000	0.414

Table 2: The front surface machining and perforation coefficients

$R^2 = 82.91\%$

Adjusted $R^2 = 82.64\%$

Analysis of the honing data

The total production time formula for the work order is the same.

$$Time = \beta_0 + \beta\, quantity + u \tag{9}$$

In this equation β_0 stands for the machine setup time and β stands for the unit production time.

While formulating β, as there was only one honing machine in the factory, speed of the machine was excluded from the equation.

$$\beta = \beta_1 + \beta_2\, surface\ area \tag{10}$$

The equation of the honing operation:

$$Time = \beta_0 + \beta_1\, quantity + \beta_2\, quantity\,.\,area + u \tag{11}$$

Time is the response variable. Quantity and surface area are predictor variables. The number of data rows is 126.

In the first regression analysis of the honing data, β_0 (setup time) was calculated negative. It is unacceptable. So, we analyzed again excluding the β_0 intercept value from the equation. For the honing data, machine setup time couldn't have been estimated and assumed as 0. By the way, honing machines don't have program loading, plugging the appropriate tools, etc. In the real world, setup time of the honing machines is shorter than the CNCs'.

At the end of the regression analysis without the intercept coefficient:

$$\widehat{Time} = 7.46\, quantity + 18.89\, quantity\,.\,area \tag{12}$$

For a part with 1 dm^2 surface area, estimated honing time is nearly 26 minutes and it is an acceptable result.

Estimated	β_1	β_2
se	1.92	2.80
t	3.89	6.76
P	0.000	0.000

Table 3: The honing coefficients

$R^2 = 82.52\%$ Adjusted $R^2 = 82.25\%$

Analysis Results

At the end of these analyses, it has seen that all of the tree formulae are statistically meaningful. The best result was obtained from the front surface machining and perforation operation data analysis which has the veriest number of variables. We think that data can be reanalyzed by including additional variables to the other equations for better results.

Although we don't have seperately collected setup and unit machining times, we have seen that these times can be estimated using total production time and quantity. In such an analysis it is important to group the data for similar operations because depending on the operation type and machine used, setup and machining times may change.

These data have been taken from a manufacturing execution system's database and analyzed manually. If this regression analysis becomes an integral part of the information system, it may be applied easily to other operations. As available data increase by time, using a feedback loop, β values may be recalculated and it becomes a learning system. Using machine learning, machine setup time (β_0) and unit production time (β) estimations would become better.

While automating the analysis, the error in the data should be omitted automatically also. It is important to eliminate data errors originated by the users. It is suggested to define minimum and maximum acceptable values for

the production time while considering the setup time and production quantity. The data out of the range should be omitted.

Lastly, the analysis may be widened by including additional parameters to the equation, like the alloy type (steel, bronze, etc.), roughness, whether the surface has tempered or not, complexity of the operation. Like in the previous works for predicting surface roughness [2, 3], artificial neural networks may be used to enhance the estimation process.

References

[1] Sevüktekin M. (2013). Ekonometriye Giriş Teori ve Uygulamalar. (pp. 459 – 489).
[2]Sidda Reddy B., Padmanabhan G. and Vijay Kumar Reddy K. (2008). Surface Roughness Prediction Techniques for CNC Turning. *Asian Journal of Scientific Research*, 1: 256-264.
[3] Bhagora V. A., Shah S. P. (2015) Modelling and Optimization of Process Parameters for Turning Operation on CNC Lathe for ASTM A242 Type-2 Alloy Steel by Artificial Neural Network and Regression Analysis. *International Journal for Innovative Research in Science & Technology (*Vol. 1, Issue 10)
[4] Niazi A., Dail J. S., Balabani S., Seneviratne L. (2006) Product Cost Estimation: Technique Classification and Methodology Review. *Journal of Manufacturing Science and Engineering* (Vol. 128, pp. 563-575).

The Effects of Macroeconomic Variables on Housing Price Index: A Panel Data Approach

M. A. Arvas[1], K. Özen[2]

[1]Van Yüzüncü Yıl University, Van, Turkey, aarvas@yyu.edu.tr
[2]Atatürk University, Erzurum, Turkey

Abstract

In this study, for the Level 2 Regions based on Nomenclature of Territorial Units for Statistics (NUTS) it is aimed to analyze the impact of macroeconomic factors (such as income, unemployment, CPI, population) on housing price indicator. Data for the period of 2011-2017 belonging to Level 2 classification 25 regions were analyzed using panel data method. According to the results obtained from the fixed effect model, the increase in per capita income increases the housing price index while the increase in the unemployment rate decreases this index value. On the other hand, the effects of CPI and population variables were found statistically insignificant.

Introduction

Housing is defined as being more resistant to natural disasters, providing quality living conditions, rather than a four-walled environment that can provide an individual's livelihood. The housing, which is shown as an indicator of the existence of the city, is not only a shelter; in addition, it is an area where one has a private living space and can lead a life as he wishes within the standard of daily life. Houses are the main means of human existence. The need for housing can be explained in direct proportion to the financial adequacy of the occupants and one's desires. Therefore, these needs are heterogeneous. On the other hand, the built houses are quite different to meet the need for housing as well as to give people a sense of self-confidence and to achieve what people want. The factors that have any impact on the prices of different goods are explored through hedonic models. Hedonic is the satisfaction of the goods or services that emerge after a consumed product or service. Hedonic price is the amount that the individual is willing to disregard for his / her satisfaction (Bulut, Öner: 2015).

Hedonic price

Hedonic price generally explains the factors that affect changes in housing prices. In the hedonic price model, the primary premise is that consumers make pricing in line with the characteristics of the good or service rather than a good or service. Along with this premise, it is believed that housing properties have a permanent benefit to housing price and hedonic price strives to find this lasting contribution (Bulut, İslamoğlu & Öner: 2015).

Literature Review

Karahan (2009), in her study, aimed to create an integrative econometric model in the market of households and houses and find answers to the question of the direction of the housing demand market. In this direction, a large field study was conducted in Istanbul. In order to develop the model, in-depth interview technique approach was applied by using qualitative research methods. According to the findings, the individual's individual life ties are intertwined with the environment and social ties. While some of the analysis results overlap with the information in the sources, some of them do not. However, the findings obtained from the field study were not found in the literature.

Method and Data Set

Panel data analysis occurs when the number of cross-sectional units N is more than the number of periods T (N> T).

$$Y_{it}=a+\sum_{i=1}^{k} B_k X_{kit}+u_{it} \qquad i=1,.....,N;\ \ t=1,...T;\ k=1,2,....q \tag{i}$$

where Y is dependent variable, independent variables, α a constant parameter, B_k slope parameters and is the error term. The subscript i represents the cross-sectional units (city, individual, country, etc.) and the subscript t represents the time (year, day, month,).

The error term u_{it} is assumed to have a zero mean and constant variance. The constant and slope parameters in the model take values according to units as well as time. According to the above panel data model; it is assumed that all independent variables affect all of their horizontal sections to the same extent. An important point here is how

to define (B_1). In some cases, the starting point may be kept constant for all units, or it may be allowed to specify different starting points for other units.

As a result, in order to determine the effects of selected macroeconomic variables on the housing price index, the model to be estimated is as follows:

$$\dot{I}ndex_{it} = B_0 + B_1\dot{I}ncome_{it} + B_2\text{Unemployment}_{it} + \text{B}_3\text{CPI}_{it} + B_4\text{Population}_{it} + \text{u}_{it} \qquad \text{(v)}$$

In equation (v), Betas represent the coefficient to be estimated. The cross-sectional dimension of the subscripts represents the time t.

The Index represents the dependent variable housing price index, which includes indicators that are created in order to follow the price changes in the housing market index, Turkey. This variable covers the Level 2 (25) regions for the years 2011-2017 . In the model, the variables of Population, CPI, Income, Unemployment were used for the level 2 regions. Information about these variables is presented in the table below.

Conclusıon

Turkey in the housing sector in recent years has gained an important boost with the effect of urban renewal policies. Under the name of urbanization policy, almost all the countries showed the effect of urbanization. According to the results of the analysis conducted in 2011-2017 data of 25 regions in Level 2 classification, an increase in per capita income has a positive effect on the price index, while the increase in unemployment rate has a negative effect on the price index.

When the fixed effects method are applied, it is seen that the probability values of the variables other than income and unemployment variable are insignificant. Hence, increases in income and unemployment have a positive effect on the price index. When the random effects are applied, we can see that Income, CPI and population variables are statistically significant. Based on these results, it can be seen that an increase in CPI and population variables has a positive effect on price index.

References

[1] Ebru Karahan, A. Ş. (2009). Konut Kariyerini Etkileyen Faktörler Üzerine Nitel Bir Araştırma Yöntemi . İtü Fen Bilimleri Enstitüsü, Yapı Bilgisi Programı, 34469, Ayazağa.

[2] İslamoğlu, H. B. (2015). The Investgation Of The Factors Affecting On The Prices Of Real Estates In Samsun Via Hedonic Price Model. Alphanumeric Journal.

[3] Hasan Bulut, Y. Ö. (2015). The Investigation Of The Factors Affectıng On The Prices Of Real Estates In Samsun Via Hedonic Price Model . Alphanumeric Journal.

Effects of Uncertainty in Unified Design of Assembly and Disassembly Lines: A chance-constrained, piecewise-linear program

S.Mete[1], Z.A. Cil[2], F. Serin[3], M. Gül[4], E.Çelik[5]

[1]*Munzur University, Tunceli, Turkey, suleymanmete@munzur.edu.tr,*
[2]*Izmir Democracy University,, Izmir, Turkey, cilzeynelabidin@gmail.com*
[3]*Munzur University, Tunceli, Turkey, fserin@munzur.edu.tr*
[4]*Munzur University, Tunceli, Turkey, muhammetgul@munzur.edu.tr*
[5]*Munzur University, Tunceli, Turkey, erkancelik@munzur.edu.tr*

Abstract

This paper is proposed a stochastic mathematical model for hybrid line design based on assembly and disassembly lines. Assembly and disassembly tasks performed on the same product are unified in a common hybrid production system. Disassembly environment and operations cannot be standardized easily due to product conditions. Factors such as the conditions of use of the product, the amount of abrasion of the parts, environmental effects can be considered for this. Disassembly process and task times has more uncertainty relatively assembly. Therefore, disassembly task times are normally distributed with known mean and variance, and assembly task times are considered as deterministic. Hence, effect of stochastic task times to balancing quality of proposed line design is analyzed using chance constrained programming approach.

Introduction

Disassembly environment and operations cannot be standardized easily due to product conditions. There are a lot of factors can be considered for the disassembly environment like the amount of abrasion of the parts, the conditions of use of the product and environmental effects. Hence, disassembly process and task times has more uncertainty relatively assembly [1]. Therefore, large time variations for the same task on the disassembly line can be caused by uncertainty in the quality of the used products [2,3]. In the disassembly line balancing problem literature, there are some researches under stochastic task time variation with generally assumed that disassembly task times are normally distributed with known mean and variance [4, 5].

Although there are some mathematical models [6,7 and 8] to solve disassembly line balancing (DLB) problems, after the proof of DLB's NP-hard nature by McGovern and Gupta [9] different heuristic/meta-heuristic approaches are developed for getting a solution such as particle swarm optimization [10], beam search algorithm [11]. Besides, different objective functions are considered in the DLB problem literature like number of workstations [12]; maximize profit[13] and so on. The literature on the DLB problem and its variants is rich, and the reader is referred to the comprehensive survey by Ozceylan et al. [14]. In this paper, a novel line design, which is proposed by Mete et al. [2], is considered. The novel design contains assembly and disassembly task on the same line as parallel. For this, a mathematical model was proposed to develop an optimization support for unified design of assembly and disassembly lines. Tasks times for assembly and disassembly were considered as deterministic. Hence, in this paper, due disassembly environment and uncertainty of used product, disassembly tasks are considered as stochastic and assembly tasks times are taken as deterministic for hybrid production line. A toy car example is solved using chance constraint programming. For more detail related toy car example can be found in Mete et al. [2]. The rest of paper is given as follows: next part, proposed solution approach is examined. Then, computational results for mentioned example are analyzed and discussed. Finally, directions for future research and conclusions are argued in the last section.

The Proposed Approach

In this study, effect of stochastic task times to balancing quality of proposed line design is analyzed using chance constrained programming (CCP) approach. Only disassembly tasks are assumed that normally distributed with known mean and variance. However, deterministic tasks times are considered for assembly processing. The model for deterministic case is proposed by Mete et al. [2]. For the stochastic case, while objective function and other constrains are the same, only cycle time constraint is changed as follows:

$$\sum_{i=1}^{I} ta_i * a_{ik} + \sum_{j=1}^{J} \mu d_j * d_{jk} + z_{1-\alpha}\sqrt{\sum_{j=1}^{J} \sigma d_j^2 * d_{jk}} \leq c * Z_k \quad \forall_k = 1,2..K$$

The objective function is described minimizing the cost related to the non-assignment of similar tasks to the same workstation and total opened workstation. ta_i ; processing time of assembly task i ; td_j processing time of disassembly task j.

where $z_{1-\alpha}$ is the standard normal distribution value for $(1-\alpha)$ probability level; μd_j is the mean time of disassembly task$j$ and σd_j is the standard deviation of disassembly task j. Cycle time constraint ensures that completion probability of tasks in workstation within the cycle time is greater or equal than $(1-\alpha)$ probability level. Effect of stochastic times is analyzed on toy car instance with CCP model verification. Model is linearized using most recently proposed approach of Altekin [15] which is piecewise linear approximation.

Illustrative Example

In this part, a toy car example from Mete et al. [2] is solved. Deterministic times for disassembly tasks are taken as mean time. Moreover, assembly tasks time are taken as deterministic. More detail related precedence relations, tasks time and similar task for toy car example can be found in Mete et al. [2].

Test case example is formed in a similar way of Ağpak and Gökçen [16] and Altekin [15]. Variances are generated randomly for two maximum coefficients of variation (CoV) level, 0.25 and 0.50. Example problem is solved for three α, two CoV and four cycle time level. Therefore, we have a total of 24 instances. Number of break points for linearization is 40. Maximum variance value for each CoV value is found with solving knapsack problem disregarding precedence relations at cycle time 120. Linearized CCP model results separately for low and high CoV levels are given in Table 1. The number of workstations opened and similarity index (SI) values with CPU times are presented in pairs for each cycle time and Z_1-α combination.

Table 1. Stochastic case results for toy car example

		CoV=0.25			*CoV=0.50*		
Cycle Time	$Z_{1-\alpha}$	*No. of workstations*	**SI distance**	*CPU (s)*	*No. of workstations*	*SI distance*	*CPU (s)*
	1.280	7	3	1495	7*	6*	+3600
75	1.645	7*	10*	+3600	7*	7*	+3600
	1.960	7*	14*	+3600	7	5	3042
	1.280	7	2	483	7	3	740
80	1.645	7	2	650	7	3	2721
	1.960	7	2	334	7	4	1813
	1.280	6	2	64	6	1	88
96	1.645	6	2	116	6	2	213
	1.960	6	2	64	6	2	58
	1.280	5	0	8	5	0	79
120	1.645	5	0	16	5	0	6
	1.960	5	0	6	5	0	76

Out of 24 instances, optimal results are obtained in 20 instances. Under low CoV level setting, out of the 12 instances integer solutions are found for two instances for cycle time 75 seconds and Z1-α being 1.645 and 1.96 combination. In a similar manner, two instances for cycle time 75 seconds and Z_1-α being 1.280 and 1.645 combination provide integer solutions under high CoV level. The average CPU time to solve an instance to optimality has been 869 and 1336 seconds for low and high CoV values, respectively. So, increasing the CoV value also increases the required time to get an optimal solution. As expected, increasing cycle time decreases the CPU time for both low and high CoV level solutions.

Under low and high CoV level setting, all cycle time and Z_1-α combination have yielded the same number of workstations in all of the solutions. Conversely, instances under low CoV level improve the SI results obtained by high value CoV level by the overall average improvement 25% for the optimal results. When the results of Table 1 are compared with the deterministic solutions in Mete et al. [2], objective are worsened function values in stochastic case as expected. Deterministic and stochastic cases yield the same results for only cycle time of 120 seconds. Another difference between deterministic and stochastic cases is the requirement more CPU time for stochastic solution.

Conclusion

In this paper a new line design, which is proposed by Mete et al. [2], is analyzed and disassembly tasks time are taken as stochastic. Effect of uncertainty in unified design or hybrid design is examined and compared with deterministic results. Test case example is formed in a similar way of Ağpak and Gökçen [16] and Altekin [15]. For future research directions, large size test problem can be generated, and heuristic approaches can be developed.

References

[1] Kongar, E. and Gupta, S. M. (2006). Disassembly to order system under uncertainty, *Omega*, 34(6), 550–561.
[2] Mete, S., Çil, Z. A., Özceylan, E., Ağpak, K., & Battaïa, O. (2018). An optimisation support for the design of hybrid production lines including assembly and disassembly tasks. *International Journal of Production Research*, 56(24), 7375-7389.
[3] Altekin, F.T., Bayındır, Z.P., Gümüşkaya, V. (2016). Remedial actions for disassembly lines with stochastic task times, *Computers and Industrial Engineering*, 99, 78–96.
[4] Aydemir-Karadag, A., Turkbey, O. (2013). Multi-objective optimization of stochastic disassembly line balancing with station paralleling, *Computers & Industrial Engineering*, 65, 413–425.
[5] Bentaha, M.L., Battaïa, O., Dolgui, A. (2015). An exact solution approach for disassembly line balancing problem under uncertainty of the task processing times, *International Journal of Production Research*, 53, 1807–1818.
[6] Altekin, F.T., Kandiller, L., Özdemirel, N.E. (2008). Profit-oriented disassembly line balancing, International *Journal of Production Research*, 46 (10) 2675–2693.
[7] Koç, A. Sabuncuoğlu, I. Erel, E. (2009). Two exact formulations for disassembly line balancing problems with task precedence diagram construction using an AND/OR graph, *IIE Transactions*, 41(10), 866–881.
[8] Mete, S., Çil, Z. A., Özceylan, E., & Ağpak, K. (2016). Resource constrained disassembly line balancing problem. *IFAC-PapersOnLine*, 49(12), 921-925.
[9] McGovern, S.M., Gupta, S.M. (2007). A balancing method and genetic algorithm for disassembly line balancing, *European Journal Operational Research*, 179(3), 692–708.
[10] Kalayci, C.B., Gupta, S.M. (2013). A particle swarm optimization algorithm with neighborhood-based mutation for sequence-dependent disassembly line balancing problem, *The International Journal of Advanced Manufacturing Technology*, 69(1-4), 197-209
[11] Mete, S., Çil, Z. A., Ağpak, K., Özceylan, E., & Dolgui, A. (2016). A solution approach based on beam search algorithm for disassembly line balancing problem. *Journal of Manufacturing Systems*, 41,188-200.
[12] Koç, A. Sabuncuoğlu, I. Erel, E. (2009). Two exact formulations for disassembly line balancing problems with task precedence diagram construction using an AND/OR graph, *IIE Transactions*, 41(10), 866–881.
[13] Altekin, F.T., and Akkan, C. (2012). Task-failure-driven rebalancing of disassembly lines, *International Journal of Production Research*, 50 (18), 4955–4976.
[14] Özceylan, E., Kalayci, C. B., Güngör, A., & Gupta, S. M. (2018). Disassembly line balancing problem: a review of the state of the art and future directions. *International Journal of Production Research*, 1-23.
[15] F. T. Altekin (2016). A piecewise linear model for stochastic disassembly line balancing, *IFAC-PapersOnLine*, 49 (12) 932–937.
[16] K. Ağpak and H. Gökçen. (2007). A chance-constrained approach to stochastic line balancing problem, *European Journal of Operational Research*, 180 (3), 1098–1115.

A novel occupational risk assessment approach based on fuzzy VIKOR and k-means clustering algorithm

M. Gul[1], F. Serin[2], S. Mete[3], E. Celik[4]

[1]*Munzur University, Tunceli, Turkey, muhammetgul@munzur.edu.tr*
[2] *Munzur University, Tunceli, Turkey, fserin@munzur.edu.tr*
[3] *Munzur University, Tunceli, Turkey, suleymanmete@munzur.edu.tr*
[4] *Munzur University, Tunceli, Turkey, erkancelik@munzur.edu.tr*

Abstract

In this paper, we proposed a novel occupational risk assessment approach based on fuzzy multi-criteria decision-making (MCDM) and clustering. The Fine-Kinney method which is considered as a basic and commonly applied risk assessment method is used as the basis of the approach. The three risk parameters of consequence, exposure and probability are taken into consideration in assessing risks. Risk prioritizations are performed using fuzzy VIKOR. As a final step for the proposed approach, the prioritized risks are classified by the aid of *k*-means clustering algorithm. By doing this, corrective and preventive actions can be arranged with respect to these classes. A numerical case study is provided to show the applicability of the proposed approach for risk assessment in a gun and rifle manufacturing facility. Besides the numerical study, a sensitivity analysis is carried out to test the validity of the approach.

Introduction

Occupational risk assessment, which is an important concept for occupational health and safety (OHS), is defined as the whole of the efforts regarding determination of hazards related to the workplace environment detection of damages to workers, workplace and environment and taking measures against hazards. Depending on the structure of the sector, the selection of the appropriate risk assessment method is important in reducing losses and altering the risks to acceptable levels. In both academia and industry, stakeholders apply various risk assessment methods to their process. Decision makers of the industry mostly prefer easier to use methods such as decision matrix (in other words 5X5 matrix), Fine-Kinney, failure mode and effect analysis (FMEA), Ridley's method, hazard and operability analysis (HAZOP), fault tree analysis (FTA) and event tree analysis (ETA). Through academia, more complex and improved methods are proposed to eliminate some shortcomings of the abovementioned classical methods. At this point, fuzzy set theory is merged with multi-criteria decision making (MCDM) concept. MCDM is one of the most important concepts of operations research. Basically, it is divided into two main sections as multi-attribute decision making (MADM) and multi-objective decision making (MODM) based on the structure of the problem. In this paper, classical Fine-Kinney method is improved using fuzzy set theory and VIKOR method (it is an MCDM method and the abbreviation of *Vise Kriterijumska Optimizacija I Kompromisno Resenje* in Serbian). The fuzzy VIKOR denotes all assessment regarding hazards and risk parameters in fuzzy numbers by the aid of fuzzy linguistic terms. While prioritization of hazards is determined by fuzzy VIKOR, on the other side, the prioritized risks are classified by the aid of *k*-means clustering algorithm. This incorporation is performed for the first time in the literature. This is also the main contribution of the current paper to the literature. A numerical case study is also demonstrated to show applicability of the novel approach. The case study concerns occupational risk assessment in a gun and rifle manufacturing facility. A sensitivity analysis is also carried out to test the validity of the approach.

The remaining organization of the paper is structured as follows: Section 2 represents the proposed method. Section three provides the case study and the obtained numerical results. The final section includes conclusion and future research agenda.

The proposed approach

The Fine-Kinney method, which constitutes the conceptual basis of the proposed occupational risk assessment approach, is a simple technique and can be easily adopt by safety stakeholders to any industry from manufacturing to service. In this method, risk score is measured by multiplying the parameters of consequence (*C*), exposure (*E*) and probability (*P*). The formula of risk score (*RS*) is as follows: *RS=C*E*P*. Parameter *P* is the risk probability of the hazard-event. Parameter *E* is defined as the frequency of occurrence of the hazard-event. Parameter *C* is defined as the most likely results of a potential accident, including injuries and property damage [1-2].

At the core of the current study, Fuzzy VIKOR is used. We follow a methodology that includes six steps in total. The detailed steps are explained in the studies of [3-5]. Due to the space limitation in this paper, we no need to explain here all these steps with mathematical background in details. However, it is useful to know that the main headings of these six steps are as follows: (1) Defuzzification of the elements of the fuzzy decision matrix for the risk parameter weights and the hazards into crisp values, (2) determination of the best and worst values of all risk parameter and hazard ratings, (3) calculation of VIKOR-specific indexes (S and R), (4) calculation of Q values, (5) prioritization of hazards by sorting the values S, R and Q in ascending order and (6) compromised solution checking using conditions (acceptable advantage and acceptable stability in decision-making).

The *k*-means algorithm, which is used in addition to fuzzy VIKOR, is one of the simplest and popular unsupervised machine learning algorithms. This algorithm starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. In this approach, we used this algorithm to classify the risks (e.g. very high risk, high risk, substantial risk, etc.)

In general, the flow diagram of the novel proposed approach that covers fuzzy VIKOR and *k*-means clustering is shown in Figure 1. In this approach, the risk parameters for risk assessment are derived from Fine-Kinney method. Three risk parameter values are then weighted. In the following step, all the hazards are prioritized using fuzzy VIKOR method. Finally, the prioritized risks are classified using *k*-means clustering algorithm.

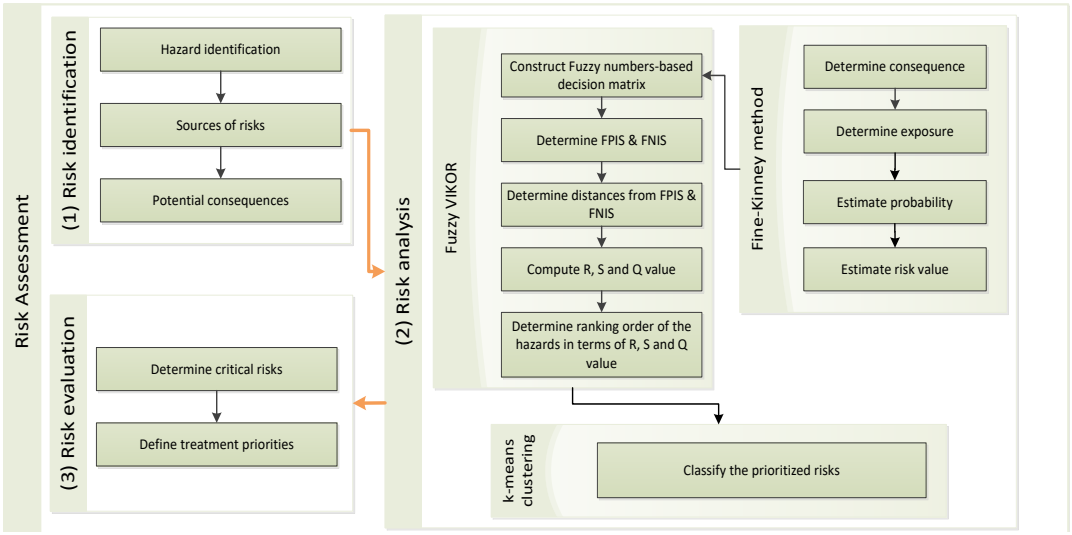


Figure 1. The proposed novel approach

Case study

A case study in the gun and rifle manufacturing factory is provided for the effectiveness of the model. The observed factory under study is stationed in Istanbul/Turkey. It is the first company in manufacturing semi-auto shotguns through the country. It manufactures civilian and defense-law enforcements. Inside the manufacturing environment, several hazards have emerged in the manufacturing facility because of product variety and rapid growing factors. The analyst group who is responsible for occupational health and safety management identified forty-three hazard descriptions. The list of hazards can be reachable in [4]. Expert group assign ratings for the occupational risk assessment in fuzzy environment and linguistic terms. The weights of risk parameters are considered as (0.418, 0.293, 0.289) for *C*, *E* and *P*, respectively. After assigning these values for risk parameter weighting, risk ranking of the identified hazards with fuzzy VIKOR is performed. The results related to the fuzzy VIKOR-specific indexes are given in Figure 2.

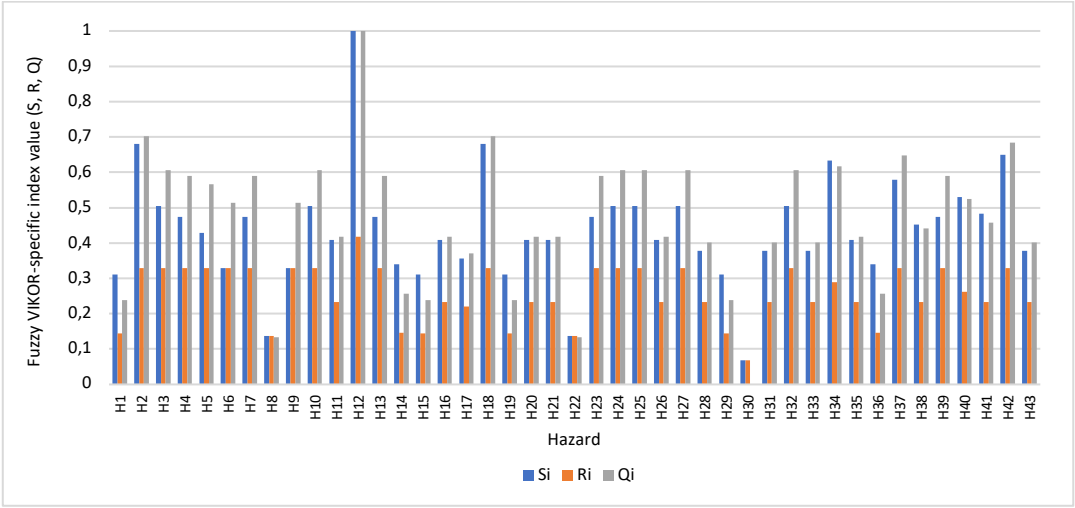


Figure 2. The results of fuzzy VIKOR

According to the fuzzy VIKOR results, H30, H22 and H8 are the most preemptive risks with lower Q values than the others. On the other side, H12 is the least serios risk among 43 hazards (Its Q value is 1). After applying fuzzy VIKOR algorithm, k-means clustering algorithm is run based on seven different combinations. We run the algorithm based on single, double and triple combinations of R, S and Q values. Also, we run these seven combinations with respect to the number of clusters generated which variated from k=2 to k=41. Considering that the classical Fine-Kinney occupational risk assessment method suggests five risk classes, we provide the results of k=5 in this paper. Table 1 shows the results of k-means clustering algorithm in classifying the prioritized hazards.

Table 1. Classifying hazards in k=5 clusters in terms of fuzzy-VIKOR Q value

Hazard	Risk cluster
H6, H9, H11, H16, H17, H20, H21, H26, H28, H31, H33, H35, H38, H41, H43	Cluster-0
H2, H3, H4, H5, H7, H10, H13, H18, H23, H24, H25, H27, H32, H34, H37, H39, H40, H42	Cluster-1
H8, H22, H30	Cluster-2
H12	Cluster-3
H1, H14, H15, H19, H29, H36	Cluster-4

Conclusion

This paper presents a novel occupational risk assessment approach based on fuzzy VIKOR and k-means clustering algorithm. Benefiting from the risk parameters of a classical occupational risk assessment method named Fine-Kinney, a novel model is adapted using fuzzy set theory and one of the most important MCDM methods "VIKOR". This model yields priorities of hazards. At the final stage of the proposed approach, by using k-means clustering, the prioritized hazards are classified in terms of riskiness.

References

[1] Fine, W. T. (1971). Mathematical evaluations for controlling hazards (No. NOLTR-71-31). NAVAL ORDNANCE LAB WHITE OAK MD.
[2] Kinney, G. F., & Wiruth, A. D. (1976). Practical risk analysis for safety management [Final Report].
[3] Gul, M. (2018). Application of Pythagorean fuzzy AHP and VIKOR methods in occupational health and safety risk assessment: the case of a gun and rifle barrel external surface oxidation and colouring unit. *International Journal of Occupational Safety and Ergonomics*, 1-14.
[4] Gul, M., Guven, B., & Guneri, A. F. (2018). A new Fine-Kinney-based risk assessment framework using FAHP-FVIKOR incorporation. *Journal of Loss Prevention in the Process Industries,* 53 3-16.
[5] Gul, M., Guneri, A. F., & Baskan, M. (2018). An occupational risk assessment approach for construction and operation period of wind turbines. *Global Journal of Environmental Science and Management,* 4(3) 281-298.

Results from Combination of ATA Method and ARIMA on M4 Competition Data Sets

T. Ekiz Yilmaz[1], G. Yapar[1]

[1]*Dokuz Eylul University, Department of Statistics, İzmir, Turkey, ekiztugce@gmail.com , guckan.yapar@deu.edu.tr*

Abstract

Recently; ATA method has been proposed which is a new forecasting method that has a similar form to exponential smoothing (ES) but innovatively the weights depend on the sample size. The optimization of ATA is quite simpler and faster compared to ES since the smoothing parameters are restricted to discrete values even though ATA still can capture all patterns in time series at least as well as ES does. Its forecasting performance is better than all the benchmarks for both the M3 and the M4 data sets. In this study, detailed comparisons of the results from combination of ATA method and ARIMA (i.e. ATA/ARIMA) and the other methods' will be discussed in M4 competition. It will be shown that despite the fact that ATA is a purely statistical method that neither utilize any advanced data pre-processing techniques nor any transformations, it is very accurate, simple and fast compared to the other methods and also some important benchmarks.

Introduction

Forecasting is an important issue that expands almost every fields such as business and industry, government, economics, environmental science, medicine, social science, politics, finance and so on. In order to obtain better predictions remain the foundation of all science, we need some basic properties: *accurate* (i.e. better supply chain management), *simple* (i.e. forecasting procedures should always be simple enough for forecast users to understand it), *fully automatic* (i.e. since many business and industries need thousands of forecasts every week or month) and lastly *fast* (namely; it can be need that determine just a few minutes).

In forecasting area, there are two main forecasting methods in the literature: exponential smoothing and ARIMA models [1]. In order to obtain the best results, if we compare these two methods, ARIMA models are more general than exponential smoothing. In addition; linear exponential smoothing models are all special cases of ARIMA models such as simple exponential smoothing (SES) is counterpart of ARIMA(0,1,1) and likewise Holt's trended method is as similar as ARIMA(0,2,2) and finally many ARIMA models have no exponential smoothing counterparts.

M4-Competition

The M-Competitions are empirical studies where the performance of a large number of major time series methods are compared based on their predictive performances on time series data. The first of M-competitions was initiated in 1979 by Makridakis and Hibon [2]. Seven participants took part in the first official competition "M1-competition" where they had to provide forecasts for 1001 time series of different time intervals [3]. For the second competition, M2-competition, only 29 data sets were provided to the 16 participants [4]. The third competition called the M3-competition was intended to both replicate and extend the features of the M-competition and M2-Competition, by including of more methods and researchers (especially researchers in the area of neural networks) and more time series [5]. A total of 3003 time series were used.

Finally; the fourth competition, M4, was announced in November 2017 and the attendance of competition started in Jan 1, 2018, ended in May 31, 2018. Initial results were published in the International Journal of Forecasting website [6]. The content of M4-competition is not as same as previous competitions but it almost replicated and extended in terms of the results of the previous three competitions, using an extended and diverse set of time series to identify the most accurate forecasting methods for different types of predictions. The most important purpose of the M4-competition is to get answers on how to improve forecasting accuracy and identify the most appropriate methods for each case. In order to get precise and compelling or convincing answers, the M4 Competition utilized 100,000 real-life series and incorporated all major forecasting methods in the literature including artificial intelligence methods (i.e. Machine Learning), and also traditional statistical ones (i.e. exponential smoothing, ARIMA etc.).

ATA Method

As large number of time series data are involved in the analyzing and forecasting process, the importance of robust, fast and accurate forecasting techniques is increasing. In this case, forecasting competitions play a critical role in moving towards the forecasting of large numbers of in real-life time series data. When we regard almost all methods for forecasting time series, there are still some shortcomings such as initial value problem. Recently, proposed new technique ATA method will appeal a wide range of attention by its simplicity, easy optimization and surprisingly good performance [7]–[9] .The novel method ATA can be performed non-seasonal time series or seasonalized time series which is applied by classical decomposition technique. The ATA method has similar properties to exponential smoothing method but there is considerable distinctness in ATA method which eliminates the initialization problem and is more feasible to optimize compared to its counterpart ES or ARIMA models [7].

ATA method's additive formula are following:

$$S_t = \frac{p}{t} X_t + \left(\frac{t-p}{t}\right)(S_{t-1} + T_{t-1}) \quad (1)$$

$$T_t = \frac{q}{t}(S_t + S_{t-1}) + \left(\frac{t-q}{t}\right) T_{t-1} \quad (2)$$

$$\hat{X}_t(h) = S_t + hT_t \quad (3)$$

where $t > p \geq q$. $S_t = X_t$ for $t \leq p$. $T_t = X_t - X_{t-1}$ for $t \leq q$. $T_1 = 0$ where $p \in \{1, \dots, n\}, q \in \{0,1, \dots, n\}$ and $p \geq q$. S_t is the equation of level; similarly T_t is the trend equation and lastly $\hat{X}_t(h)$ is the equation of h-step ahead forecast and where p and q are the smoothing parameters of level and trend respectively.

Discussion

In M4-Competition, our proposed model ATA-2 has competed very well to its counterpart models according to Table 1. Algorithm of this method is very simple compared to other major techniques. Firstly, we have introduced to data set purifying the seasonality effect i.e. deseasonalized with using classical multiple decomposition method if it is necessary and then estimated p and q parameters in order to find initial value of level and trend respectively and lastly we have applied again seasonality effect i.e. reseasonalized to obtain accurate forecast values.

According to Table 1, results from M4 competition informally are shown in detail on ranking with respect to sMAPE metrics' results. In addition to this metric; there are some other metrics that are used such as mean absolute scaled error (MASE) proposed by [10] and overall weighed average (OWA) which refers to simple average of sMAPE and MASE. Because the ATA/ARIMA model, which is obtained by using simple average of forecasts' results between ATA-2 and ARIMA, have not attended to M4, these results are informal. If we exclude to ATA/ARIMA results from table 1, then the remainders are the official results according to GitHub results [11]. As we can see at this table, ATA/ARIMA model have shown considerable success compared to all benchmarks, even though ATA method is purely statistical method that not use any pre-processing analyzing or any outlier detection and excluding methods. Since the first one is hybrid model, second and third ones are the combination of statistical and machine learning algorithms as well as the fourth and fifth ones are again statistical combination; these results are relatively better than ATA/ARIMA models' that refers to combination of only two statistical model: ATA and ARIMA. On the other hand; if we see at the only machine learning algorithms' results, these ranks are 57 and 59 that account for RNN and MLP respectively. Namely, we can easily say that machine-learning algorithms are not sufficient without using any statistical methods. However; if we look at the Table 1's results, we can say that some combinations of machine learning and statistical methods can be enhanced the model accuracy and reliability.

Conclusion

In this study, we have dealt with the achievement of ATA method and its simple combination: ATA/ARIMA. These two methods are successful on their own, but their combination is quite better than almost all benchmarks such as THETA, COMB, ARIMA and DAMPED etc.

Thus; ATA method is rather successful method obtaining forecast but it can be improved by optimizing with respect to different error criteria, modelling seasonality separately, using data pre-processing techniques and holdout techniques, using transformations if it is necessary and lastly applying combination and machine learning techniques

in terms of model accuracy and robustness. These and more will be our future working that include applying to some machine learning algorithms i.e. artificial neural network (ANN) and any other combination techniques.

Table 1. Results from M4-competition informally

Team Members	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total	Rank (sMAPE)
Smyl, S.	13.176	9.679	12.123	7.817	3.170	9.328	11.374	1
Jaganathan, S.	13.712	9.809	12.487	6.814	3.037	9.934	11.695	2
Montero-Manso, P.	13.528	9.733	12.639	7.625	3.097	11.506	11.720	3
Fiorucci, J., A.	13.673	9.816	12.737	8.627	2.985	15.563	11.836	4
Pawlikowski, M.	13.943	9.796	12.747	6.919	2.452	9.611	11.845	5
ATA/ARIMA	13.847	9.987	12.653	7.607	2.998	11.942	11.859	6*
ATA-2	13.930	10.292	12.936	8.540	3.095	12.851	12.098	12
THETA	14.953	10.311	13.002	9.093	3.053	18.138	12.309	21
COMB	14.848	10.175	13.434	8.944	2.980	22.053	12.555	22
ARIMA	15.168	10.431	13.443	8.653	3.193	12.045	12.661	24
DAMPED	15.198	10.237	13.473	8.866	3.064	19.265	12.661	26
ETS	15.356	10.291	13.525	8.727	3.046	17.307	12.725	27
SES	16.396	10.600	13.618	9.012	3.045	18.094	13.087	37
HOLT	16.354	10.907	14.812	9.708	3.066	29.249	13.775	39
RNN	22.398	17.027	24.056	15.220	5.964	14.698	21.152	57
MLP	21.764	18.500	24.333	21.349	9.321	13.842	21.653	59

References

[1] De Gooijer, J. G. and Hyndman, R. J. (2005). 25 Years of Iif Time Series Forecasting: A Selective Review.

[2] Makridakis, S., Hibon, M. and Moser, C. (1979). Accuracy of Forecasting: An Empirical Investigation, *J. R. Stat. Soc. Ser. A*.

[3] S. Makridakis *et al.*, "The accuracy of extrapolation (time series) methods: Results of a forecasting competition," *J. Forecast.*, 1982.

[4] S. Makridakis and M. Hibon, "Exponential smoothing: The effect of initial values and loss functions on post-sample forecasting accuracy," *Int. J. Forecast.*, 1991.

[5] S. Makridakis and M. Hibon, "The M3-competition: Results, conclusions and implications," *Int. J. Forecast.*, 2000.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: Results, findings, conclusion and way forward," *Int. J. Forecast.*, 2018.

[7] G. Yapar, "Modified simple exponential smoothing," *Hacettepe J. Math. Stat.*, 2018.

[8] G. Yapar, S. Capar, H. T. Selamlar, and I. Yavuz, "Modified holt's linear trend method," *Hacettepe J. Math. Stat.*, 2018.

[9] G. Yapar, I. Yavuz, and H. T. Selamlar, "Why and How Does Exponential Smoothing Fail ? An In Depth Comparison of ATA- Simple and Simple Exponential Smoothing .," vol. 01, no. 1, pp. 30–39, 2017.

[10] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006.

[11] M4, "GitHub," *2019*. [Online]. Available: https://github.com/M4Competition/M4-methods.

Investigation Of Skills and Trainings of Big Data Specialists in Turkey: Linked-In Data Mining Application

S.Ö. Rençber[1], A. Özdemir[2],

[1]Van Yüzüncü Yıl University, Van, Turkey, rencber.serhatomer@gmail.com
[2]Atatürk University, Erzurum, Turkey, abdulkadir@atauni.edu.tr

Abstract

The use of big data technologies is becoming widespread; however, the need for specialists to work is increasing. Specialist candidates who want to work in this field, theoretical and application training is carried out on various platforms. These trainings are usually postgraduate courses, online trainings, training seminars. These trainings bring various skills to specialist candidates. These skills must be in line with the private sector needs. Thus, specialist candidates will benefit from the trainings. The aim of this study, after the definition of big data concept, using data science and data mining techniques, the abilities to get information from Linked-in profiles of the specialists working in the field of big data in Turkey. It is to find the skill information required to work in the field of big data and to make recommendations to big data specialist candidates.

Introduction

In the future, big data is expected to be accepted as an indicator of economic growth and competition between countries is expected to be shaped around big data rather than traditional issues [1,2,3]. As big data technologies are developing rapidly, the need of experts in this field is increasing in parallel. Specialists working in the field of big data receive various trainings to improve their theoretical and application in this field. Postgraduate trainings, online trainings, training seminars help big data experts develop their skills in their fields. In the study of Rençber and Özdemir [3], there is no difference in perceived ease of use and perceived benefits of big data analytic graduate candidates who have graduate education in big data analytics, they also reported that postgraduate training contributes more to the specialization process than online training and other training.

Big data specialists has been determined with using keywords big data and turkey in Linkedin search engine. Search results has been download and convert to Microsoft Excel program format with program prepared in python. The datas of 600 profiles in Microsoft Excel format is prepared for analysis. The skills, universities and trainings of the specialists are determined and shown in tables by using Pandas and Numpy libraries in Python.

Big Data

It is a collection of data from traditional and digital sources within and around businesses, representing a source for ongoing exploration and analysis. When defining big data, it is also important to understand the mix of unstructured and multiple structured data that make up the volume of information. Since the concept of big data is used by companies in many disciplines and different sectors, it also changes depending on the technological tools and data production methods that develop according to time. In literature reviews, it is difficult at present to make a general definition of big data [3].

From sources that define big data in terms of volume, velocity and variety Gürsakal [4], with large volume, great velocity and great variety; It is defined as information requiring new forms of information processing that will enhance decision-making capabilities and improve insight and process optimization. Beyer ve Laney [4], These are information assets that are cost-effective and require information formats for decision making and accurate output, and include high volume, velocity and variety concepts. Schroeck arkadaşları [5] add verification to these definitions, in today's electronic market, it is defined as a combination of volume, variety, velocity and verification that provides companies competitive advantages and opportunities. In the definition made by adding the value property, Poulovassilis [6] describes a phenomenon involving complex and dynamic growth of data. Scientists conceptualize big data on structural and functional dimensions. The structural dimension of the big data reported that it includes volume, velocity, variety, validation and value elements [3].

Method

Data Detection

The frequency values of talent, education and universities were taken separately for each individual. The process is processed according to the model established in Figure-1.

Using the search engine on LinkedIn's social network, individuals with big data specialists and their education and talent information available in their profiles were identified for each individual, list of the detected person's saved for use in later steps.

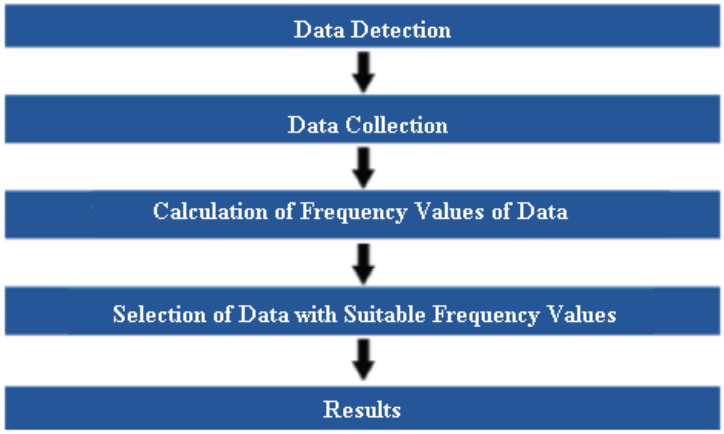


Figure 1. Applied model

Data Collection

The talent information of the talented individuals was drawn with LinkedinScraperProject program and made into a single file in MS Excel. In this way, for each individual on a single table in the form of 4 columns, experts, undergraduate education, graduate education and skills were listed.

Calculation of Frequency Values of Data:

On the data collected in the previous step, repetition numbers (frequency) were determined for each MS Excel column using Python numpy and pandas libraries.

Selection of Data with Suitable Frequency Values:

For each expert, a lower limit is determined by considering the frequencies of the undergraduate, graduate, doctorate and talents and the repetition rate of the data having the highest frequency.

Tools

LinkedinScraperProject: This program provides automatic access to the phone number, training information, location information, skills information, e-mail information in the profile information of the occupation or people we want to search with the keyword written to the LinkedIn search engine through the LinkedIn account that is logged in or Json format.

MS Excel: It is a calculation program that keeps the data in tables or lists and performs all the calculations and analyzes we need for this data. Graphs can be drawn to the results obtained from data with Excel, report can be created.

Jupyter Notebook : Analyzing the data in MS Excel, an application that writes and runs Python code. The results of the analysis were taken in this program.

Results

Training and skills of the big data experts working in Turkey will be examined under separate headings.

Since this study was carried out on a single computer with limited hardware resources, the LinkedInScaperProject program used at certain intervals was used to extract data from the LinkedIn profile of 600 randomly selected people. For this reason, the results are only based on the abilities of these 600 people.

Skills and Bachelor's Degrees of The Big Data Specialists

According to the results; All of specialists have bachelor's degree. most of the specialists graduated from Yıldız Technical, Istanbul, Istanbul Technical, Bilkent, Middle East Technical universities. The most of the specialists graduated from Computer Engineering, Statistics, Computer Science, Mathematics, Industrial Engineering departments. The most of skills of big data specialists are Java, Python, big data analytics and SQL.

Skills and Master's Degrees of The Big Data Specialists

246 of the specialists have master degree, most of the specialists graduated from Istanbul Technical, Bosphorus, Yıldız Technical, Bahçeşehir and Sabancı universities. The most of the specialists graduated from Computer Engineering, Industrial Engineering, Data Analytics, Computer Science, Big Data Analytics departments. The most of skills of big data specialists are Java, SQL, Python, big data analytics,data mining, data analysis, machine learning, R, Busines Intelligence.

Skills and Doctoral Degrees of The Big data Specialists

40 of the specialists have doctoral degree, most of the specialist graduated from Kadir Has, Istanbul Technical, Istanbul and Gazi universities. The most of the specialists graduated from Computer Engineering, Management Information Systems, Computer Science, Industrial Engineering departments. The most of skills of big data specialists are big data analytics, Python, data mining, Java, machine learning, Apache Spark.

Conclusions

In this research, it was found that big data specialists graduated from many different undergraduate and postgraduate programs. Big data specialists have a high rate of repetition of their programming language skills, such as Java and Python but it is noteworthy that the skills of big data analytics platforms such as Hadoop and Spark are low.

This study was carried out on a single computer with limited hardware resources, the analysis of LinkedscraperProject program used according to the profile of 600 people was performed at a certain interval. The results obtained were based solely on the abilities of these 600 people.

References

[1] Jin, X., Wah, B. W., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research,* 2 (2) 59-64.

[2] Çiğdem, Ş., & Seyrek, İ. H. (2015). İşletmelerde Büyük Veri Uygulamaları: Bir Literatür Taraması *2. Ulusal Yönetim Bilişim Sistemleri Kongresi* (s. 45-46). Erzurum: Atatürk Üniversitesi,

[3] Rençber, S.Ö. (2019). Büyük Veri Kullanımının Teknoloji Kabul Modeli ile İncelenmesi: Lisansüstü Eğitim Örneği. Doktora Tezi, Atatürk Üniversitesi, Erzurum, Türkiye.

[4] Gürsakal, N. (2014). *Büyük Veri.* Dora Basım Yayınevi.Bursa, Türkiye.

[5] Beyer, M., & Laney, D. (2016). *The Importance of 'Big Data': A Definition.* Gartner: https://www.gartner.com/doc/2057415/importance-big-data-definition.

[6] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. *IBM Global Business Services ,* 12 1-20.

[7] Poulovassilis, A. (2016). *Big Data and Education*. BirBeck Üniversitesi of London: http://www.dcs.bbk.ac.uk/oldsite/research/techreps/2016/bbkcs-16-01.pdf.

Community Pharmacies' Views and Attitudes Towards Surplus

Metin E. Kartal[1], Gülbin Ozcelikay[2]

[1,2] *Ankara University, Ankara, Turkey*

Abstract

Surplus concept is very important for Turkish pharmaceutical sector hence it it the way of increasing operational profits of all components. However, as the economic conditions became worse in the last years, surplus concept depreciated. The most important point that the surplus partially was started to be given by consumers, not by industry as a result of price odds. Also, it is known for a long time that surplus can result in unethical relationships between pharmacist-doctor-pharmaceutical industry. So, this paper aims to investigate pharmacists' opinions about surplus concept deeply.

Introduction

The margin of profit of community pharmacies basically based on the surplus concept. The legally given margin of profit isn't enough for pharmacies to continue their operations. On the other hand, the surplus given to community pharmacies by the industry was marginally cut off due to financial stricts. This research aims to investigate community pharmacists' views and attitudes towards surplus concept.

Method

We prepared 5 likert type 21 questions questionnaire. The questionnaire was issued at the Sadeceeczacı group in the Facebook through SurveyMonkey. The analysis of the questionnaire was handled by SPSS. Basically non-parametric tests were used due to heterogenous distribution of the data. Mann-Whitney U, Kruskall Wallis and chi-square of independence tests were used to analyze the data.

Results

Three hundred eighty-seven community pharmacists participated in the research. %49,1 of the participants is women, and %50,9 is the men. %23,30 of the participants have a master and doctorate level. %49,5 of the participants have been continuing the profession for 1-10 years, %28,5 is for 11-20 years, %14 is for 21-30 years, and %8 is for 30 years and above. %41,6 of the participants serves towards family health care services, %15,1 is at the main streets, %20 is towards hospitals and %23,1 is in the central districts. Participants marked that %18,26 of the expired medicines at the box level consists of surplus at the end of the year. Participants pointed out that the margin of profit has to be increased by approximately %17,68. Participants thought that surplus contributes to the margin of profit nearly %16,32.

Non-Parametric Analysis Results

The Mann Whitney U test was carried out to evaluate statistical significance between surplus and ethical problems related to surplus and participant gender. According to the analysis males (N196 mean rank:218,04) prefers surplus rather than females (N:189 mean rank:167,04), U:13615 z:-4,725 p:0,000. Females (N:187 mean rank:201,52) agree more than males (N:196 mean rank:182,92) that more surplus which can cause more price odd results in the ethical problem, U:16546 z:35852 p:0,082. Females (N:187 mean rank:207,65) believe more than males (N:196 mean rank:177,07) that surplus can cause unethical relationship between a pharmacist-doctor-drug company, U:15400 z:34706 p:0,005. Males (N:196 mean rank:212,84) believe more than females (N:187 mean rank:170,16) that pharmacies can't continue operating with the existing margin of profits if the surplus is abolished, U:14241,5 z:-4,100 p:0,000.

The Kruskal Wallis Test was carried out to evaluate statistical significance between discounts made by pharmacies and pharmacists' opinions about the surplus. The statistical significance was determined between those two factors, x2(3):7,172 p:0,067. Subsequently, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted p-values are

presented. The difference was found at the pharmacies which make %0 discount (mean rank:211,02) and the pharmacies which make %2,75 discount (mean rank:168,06), p:0,08.

The Kruskal Wallis Test was carried out to evaluate statistical significance between profession time and pharmacists' opinions about the surplus. According to analysis, statistical significance was determined between the profession time and the more surplus which can cause more price odd, x2(3):24,672 p:0,000. Subsequently, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted p-values are presented. Statistically, pharmacists who practice for 11-20 years (N:110 mean rank:195,43) are more agree than pharmacists who practice for 21-30 years (N:54 mean rank:142,95) p:0,014; pharmacists who practice 1-10 years (N:191 mean rank:212,73) are more agree than pharmacists who practice 21-30 years (N:54 mean rank:142,95) p:0,000 and finally pharmacists who practice 1-10 years (N:191 mean rank:212,73) are more agree than pharmacists who practice (N:31 mean rank:150,61) p:0,012 with that drugs have more price odd which have more surplus than generics.

A chi-square test of independence was conducted between profession time and discount made by pharmacy to social security institution. %93,8 of the expected cell frequencies were greater than five. There was a statistically significant association between profession time and discount made by pharmacy to social security institution, χ2(9):23,170 p:0,006. The association was small (Cohen, 1988), Cramer's V = .142. As the profession time increases, discount made by pharmacy increases. A chi-square test of independence was conducted between profession time and the condition that drug represant can cause relationship between pharmacist-doctor-drug firm. All expected cell frequencies were greater than five. There was a statistically significant association between profession time and the condition that drug represant can cause relationship between pharmacist-doctor-drug firm, χ2(3):11,142 p:0,011. The association was small (Cohen, 1988), Cramer's V = .170. Pharmacists who practice for 1-10 years are more agree with that condition than who practice for more years. A chi-square test of independence was conducted between gender and perception as an ethical problem of not taking price odd due to more surplus. All expected cell frequencies were higher than five. There was a statistically significant association between gender and perception as an ethical problem of not taking price odd due to more surplus, χ2(1):7,092 p:0,008. The association was small (Cohen, 1988), Cramer's V = .136. Women are more acceptable to this condition than men. A chi-square test of independence was conducted between pharmacy location and pharmacist- doctor collaboration to the prescription of the drugs which have more surplus. All expected cell frequencies were higher than five. There was a statistically significant association between pharmacy location and pharmacist- doctor collaboration to the prescription of the drugs which have more surplus, χ2(2):5,517 p:0,063. The association was small (Cohen, 1988), Cramer's V = .120. A chi-square test of independence was conducted between pharmacy type and discount made by the pharmacy to the social security institution. %80 of the cell frequencies were higher than five. There was a statistically significant association between pharmacy type and discount made by the pharmacy to the social security institution, χ2(12):40,476 p:0,000. The association was small (Cohen, 1988), Cramer's V = .187. Pharmacies towards hospitals make more discount than pharmacies at the main streets.

Discussion & Conclusion

The surplus concept is very important for community pharmacies to continue their operations. On the other hand, most of the pharmacists tend to increasing of margin of profits instead of surplus concept. Also, it is determined that the surplus concept can cause unethical results or relationships (1). In addition to this, surplus can result in unnecessary stock in pharmacy organizations or can indireclty increase consumption of drugs unnecesarily in pharmacy organizations which decreases the efficiently usage of sources of the country economy. At this point, there are lots of things for legislators to improve this condition.

References

[1] Kartal M. E., Ozcelikay G. (2019). The Ethical Problem in Community Pharmacies Due to Drug Price-Based Policy**.** *J Lit Pharm Sci.*, 8(1) 43-50, doi: 10.5336/pharmsci.2018-61939

Optimization of Holder Table Test Function using Dynamic Mutated Genetic Algorithm

E.N. Bak[1], R. Saraçoğlu[2], S. Şehribanoğlu[3]

[1] Van YüzüncüYıl University, Van, Turkey, nudabak@gmail.com
[2] Van YüzüncüYıl University, Van, Turkey, ridvansaracoglu@yyu.edu.tr
[3] Van YüzüncüYıl University, Van, Turkey, sanem@yyu.edu.tr

Abstract

The holder table function is one of the most popular test functions used in evaluating of optimization algorithms. Heuristic methods can be used for the optimization of this multimodal test function. In this study, it was tried to find the minimum value of the function by using a genetic algorithm. In addition to the classical genetic algorithm, a solution has been proposed in which the mutation parameter changes dynamically. According to the experimental results for various population sizes, the proposed approach improved the performance of the genetic algorithm.

Introduction

Today's difficult and complex problems have led us to seek quick and easy solutions to solve these problems [1]. Examples are easily accessible with GA. GAs are applied in many areas. Many of the problems require screening of a large solution basin. GA is used to obtain an acceptable result in a short time [2].

In order to understand the functioning of this algorithm, it is necessary to understand the process of natural evolution which is based on it. Evolution is a process that makes living things adapt to the environment in the long term and makes them "best" in time. One of the basic laws of evolution is the basis of this algorithm; only the "best" always lives and multiplies. Failed individuals cannot breed and are eliminated. This is why the deterioration becomes difficult and the patient gets better over time. This algorithm does not have specific standards that can be used in all problems except some adaptive superficial standards (principles of evolution). The algorithm is shaped according to the problem [3]. In computer language, arrays and subarrays are used to implement this natural process. Sequences contain the population, individuals and genes of individuals.

The term "individuals" refers to all elements in the solution space, that is, the cluster within the solution. Each element is represented by an array of bits of length that can include all elements. Individuals make up the population and each individual has a fitness value.

In this study, we aimed to increase GA success. In order to achieve this, the mutation parameter is especially focused. A method for dynamically changing this parameter is proposed. In this way, the value of the mutation parameter may change according to the current situation as the generations of GA progress. Thus, the algorithm may have a greater mutation rate when needed. If necessary, the genetic process will continue with a smaller parameter value. The optimization test function selected to test this change in GA is the Holder Table. The success of the proposed approach with experimental results was tried to be demonstrated.

Literature Review

Türkay and Artaç developed a new GA that makes optimum distribution network design. Line design, line and distribution transformer station size selection is made in the realized design. In the model, line and distribution transformer stations facility and loss costs are taken into account. GA developed in MATLAB was applied to a real network. The population size was 50, the cross rate was 0.7, and the mutation rate was 0.03 [4]. Bolat et al. discussed the definition and working principle of GA and examined the crossover and mutation operators used. The literature for the applications of GA in the field of engineering has been made and information about these studies has been presented. In the last part of the study, an explanatory example of GA is given. The population size was 4 [5]. Özdemir's [6] study is to obtain highly predictive regression models by using the least number of independent variables in a data set containing multiple variables. The population size was 100, the maximum generation number was 1000, the probability of crossing was 0.90, and the probability of mutation was 0.02. .In Keskintürk's study, DGA was introduced and its stages were explained. In the sample problem in this study, the mutation rate was 0.8 [7].

In the study of Ok, the definition of the frequency assignment problem in cognitive radio networks was made and heuristic methods in the literature related to the solution of the problem were examined. The population size was between 14 and 20. The cross rate was 0.4 and the mutation rate was 0.6 [8]. In the study of Doğan et al. in the study; They have aimed to obtain more realistic results while taking into account the center and depth of the earth in search security studies due to the 3D sphere surface of GA method. Since the population size is 250, the cross rate is 0.80, the mutation rate is 0.01[9].In their study, Kaya and Güler designed a fuzzy and fuzzy-genetic algorithm based biaxial solar tracking system to maximize the voltage at the output of the photovoltaic (PV) solar panel. The population size was 30. Mutation ratio was taken as 0.1[10].In the study of Çalışkan et al. the aim is to achieve creative and innovative results in designs. The crossing rate was taken as 0.5 and the mutation rate as 0.5 [2].

Keleş and Keleş examined the usage areas of artificial intelligence systems in construction management which forms the basis of construction projects. Population size was 20, the generation number was 20, crossover rate was 0.6 and the mutation rate was 0.033 [11].

Experimental Results

In this section, it is implemented a GA to found the global minimum of Holder Table test function using the MATLAB programming. In the implementation mutation rate can be change between 1/population size and 0.1. Population size is 50, Crossover rate is 0.8, Number of generations are 100, used Encoding method is reel-value coding, Crossover method is the arithmetic crossover and Selection method is Tournament.

Run the GA algorithm on test functions 30 times, and calculate the mean values and standard division of each criterion, as shown in Table 1.

Table 1. Experimental results of classical GA

Population size	Mean values	Standard deviation
10	-14,5762878155409	4,96134323309051
20	-15,6269298783485	4,75626014021893
30	-16,5342411928070	4,32791270093712
40	-17,4306579005379	3,62983987840932
50	-17,5552971669934	3,66735101448196
60	-17,8419657499838	3,32996476352924
70	-18,4582822644438	2,45514856509566
80	-17,8920447363726	3,34797321867986
90	-18,1172462263352	2,95730895825531
100	-18,4950653091144	2,45303409482343
110	-18,4328636537354	2,46611957793065
120	-18,5434973412561	2,45730667262185
130	-18,8409035666708	1,76555094721672
140	-18,8524004522245	1,77078823796826
150	-18,9588037143475	0,869935989340489
160	**-19,1724042678677**	**0,0759887184853001**
170	-18,8417183760559	1,54137997898496
180	-18,8446005927085	1,76765658814904
190	-18,7833853152065	1,77995062875416
200	-19,1708370993604	0,0561576347580374

Run the dynamic mutated GA algorithm on test functions 30 times, and calculate the mean values and standard division of each criterion, as shown in Table 2.

Table 2. Experimental results of dynamic GA

Population size	Mean of Fitness values	Standard deviation
10	-14,9314809964471	4,46493272191508
20	-15,3913628805987	4,70687219947882
30	-17,5533650262360	3,66456371986615
40	-16,8709352456594	4,14467459927270
50	-16,8328985444561	4,13774587057959
60	-16,8579901070908	4,15252643141574
70	-18,8404417547125	1,76533630742165
80	-18,5054471452532	2,44974566743147
90	-18,1535723228268	2,94107428934369
100	-19,1705671255063	0,05943719875706
110	-18,4297983218545	2,44511825022417
120	-19,1159684569459	0,24426249803585
130	-18,7484846012345	1,78541340501270
140	-19,1512942189698	0,14373770418994
150	-19,1580647639702	0,09967116256716
160	**-19,1955202094816**	**0,01987204732334**
170	-19,1304424871261	0,11185083627831
180	-18,3959913814814	2,44780451659151
190	-18,4961845591313	2,44676068122793
200	-19,1420342286092	0,15699774468318

Conclusion

In this study, it was investigated the effect of mutation rate on the performance of the genetic algorithm. We used the Holder Table function as a test function. It is studied the population size between 10 and 200. the average and Standard Deviation of a different number of populations was calculated, and the system performance were evaluated. We have presented previous studies on the effect of algorithm parameters on their performance in finding the optimal solution to the problem. In this study, we have noticed that the best number of populations is 160. This study may serve as a basis for future studies of the mutation rate to improve the performance of the genetic algorithm.

References

[1] Emel and Taşkın. (2002). Genetik Algoritmalar Ve Uygulama Alanlari, *Uludağ University Journal of Economy and Society*, 21 (1), 127-152.

[2] Çalışkan, F., Yüksel, H., Dayık, M. (2016). Genetik Algoritmaların Tasarım Sürecinde Kullanılması. *SDU Journal of Technical Sciences,* 6(2), 21-27.

[3] Güracar, B. (2017) Genetik Algoritmalar, https://docplayer.biz.tr/108210219-Genetik-algoritmalar-busra-guracar.html. Last Access date: 03.05.2019

[4] Türkay, B., Artaç T., (2003). Dağıtım Şebekesinin Genetik Algoritma İle Optimum Tasarımı. *Bilgisayar Mühendisliği 10. Ulusal Kongresi*, 95-98.

[5] Bolat, B., Erol, K.O. and İmrak, C.E, (2004). Genetic Algorithms in Engineering Applications and The Function of Operators, *Journal of Engineering and Science,* 4, 264-271.

[6] Özdemir, M, (2017). Predictive Model Selection In Linear Regression By Genetic Algorithms.*Pamukkale University Journal of Social Sciences Institute*, 28, 214-233.

[7] Keskintürk, T. (2006). Differential Evolution Algorithm, *İstanbul Ticaret University Journal of Science* 5 (9), 85-99.

[8] Ok, A. (2017). Solving the Frequency Assignment Problem Using Heuristic Methods for Cognitive Radio Networks, Ankara University, Master Thesis.

[9] Doğan, Ş., Koca, G.Ö, Yılmaz, H., (2017). Application of Genetic Algorithm Method for 3D Surface to Ensure Optimal Distribution of Tools at Search and Rescue Operations. *Karaelmas Science and Engineering Journal,*7 (2),577-585.

[10] Kaya T., Güler H., (2016). The Design of Fuzzy-Genetic Algorithm Based Sun Tracking System for Maximizing Output Voltage, *Science and Engineering Journal of Fırat Univ.*28 (2), 99-108.

[11] Keleş, A.E, Keleş, M.K. (2018). Prediction of Employee Leadership Perception in Construction Management Using Feature Selection with Genetic Algorithm, *Çukurova University Journal of the Faculty of Engineering and Architecture,* 33 (4), 97-110.

Comparison of Particle Swarm Optimization and Teaching-Learning Based Optimization Algorithms from Swarm Based Metaheuristics for Dynamic Berth Allocation Problem with Port Structure Constraints

S. Mutlu[1], B. Bilgen[2], M. Ghallali[3]

[1]*Eskişehir Technical University, Eskişehir, Turkey, serkanmutlu@eskisehir.edu.tr*
[2]*Dokuz Eylül University, İzmir, Turkey, bilge.bilgen@deu.edu.tr*
[3]*Dokuz Eylül University, İzmir, Turkey, mouadgh@gmail.com*

Abstract

Due to its cheapness and safety, sea transport is usually the preferred mode of transportation of goods to far away locations. Therefore, improving the container terminals where vessels load and unload their goods is of a tremendous importance. Berth allocation is one of the main operations in terminal containers. Improving such an operation is by allocation the optimal berth location to upcoming vessels, in other words, finding the best solution to the berth allocation problem, abbreviated as BAP. We propose in this study a new mathematical model in order to solve BAP. We consider that vessels arrive dynamically and the study is limited within the specific structural constraints of the harbor. Since the solution is not possible in polynomial time, we compare two swarm based metaheuristic algorithms such as Particle Swarm Optimization (PSO) and Teaching-Learning based Optimization (TLBO) by adapting them to the problem. When analyzing the computational results, we notice that similarity in effectiveness between TLBO and PSO in small and medium sized cases, however, once the case study's size begins to increase, PSO becomes more effective than TLBO.

Introduction

BAP is defined as the allocation of suitable vessels to appropriate berths and was first introduced in literature by 1997, Imai et al. [1]. After this study, interest in BAP increased and sea transport became more and more effective. In 2015, Bierwirth and Meisel [2] published a survey for BAP showing several studies published until 2015. Their study has focused on the various algorithms used to solve BAP and the results reached. In the following years, the newly developed and hybridized metaheuristic algorithms were emphazied, as in [3], [4], [5] and [6]. In this study, we propose a new mathematical formulation for BAP, where ship arrivals are dynamic and contain harbor-specific structural constraints. Furthermore, we develop two metaheuristic algorithms, the first being Particle Swarm Optimization (PSO) while the second is Teaching-Learning based Optimization (TLBO).

Dynamic Berth Allocation Problem Definition

When ship arrivals become dynamic, early or late assignment of ships to the port arises. In this case, the BAP becomes the Dynamic Berth Allocation Problem (DBAP). Many studies on DBAP have been done for ports with discrete berths, yet very few were published considering hybrid berths. BAP with a discrete berth structure means that a single vessel is assigned to a single berth, while a hybrid dock structure is defined by assigning a vessel to multiple berths simultaneously. In this study, we consider a DBAP with hybrid berth structure for ports with rectangular structure. DBAP was modeled using sets, parameters and variables in Table 1.

Table 2. Sets, Parameters and Variables

Parameters	Definition	Variables	Definition
V	Number of vessel	$x_{i,j}$	If vessel *i* is assigned to berth *j* 1, o.w. 0
B	Number of berth	s_i	Starting time of vessel *i*
CB	Set of corner berths	e_i	Ending time of vessel *i*
$i, i' \in \{1, \dots, V\}$	Indices of vessel	$\delta_{i,j}$	If vessel *i* is not assigned to berth *(j+1)* while vessel *i* assign berth *j* or v.v. 1, o.w. 0
$j \in \{1, \dots, B\}$	Indice of berth	$\mu_{i,i'}$	if at dock j, ship i is being processed before ship i' 1, o.w. 0
$cost_i^1$	Early berthing unit cost of vessel *i*	ΔEAR_i	Early arrival time of vessel *i*
$cost_i^2$	Late berthing unit cost of vessel *i*	ΔLAT_i	Late arrival time of vessel *i*
EST_i	Earliest starting time of vessel *i*		
ETA_i	Expected time arrival of vessel *i*		
l_i	Length of vessel *i* (**unit of:** berth)		
p_i	Processing time of vessel *i*		

$$\boldsymbol{Min}\ Z = \sum_{i=1}^{V}(cost_i^1 \Delta EAR_i + cost_i^2 \Delta LAT_i) \quad (1)$$

Subject to;

$\sum_{j=1}^{B} x_{ij} = l_i$	$\forall i \in \{1, \dots, V\}$	(2)
$\sum_{j=1}^{B-1} \lvert x_{ij} - x_{ij+1} \rvert \leq 2$	$\forall i \in \{1, \dots, V\}$	(3)
$\sum_{j=1}^{B-1} \delta_{ij} \leq 2$	$\forall i \in \{1, \dots, V\}$	(3.1)
$x_{ij} - x_{ij+1} \geq -\delta_{ij}$	$\forall i \in \{1, \dots, V\}, \forall j \in \{1, \dots, B\}$	(3.2)
$x_{ij} - x_{ij+1} \leq \delta_{ij}$	$\forall i \in \{1, \dots, V\}, \forall j \in \{1, \dots, B\}$	(3.3)
$x_{i1} + x_{iB} \leq 1$	$\forall i \in \{1, \dots, V\}$	(3.4)
$e_i = s_i + p_i$	$\forall i \in \{1, \dots, V\}$	(4)
$s_i \geq EST_i$	$\forall i \in \{1, \dots, V\}$	(5)
$M\left(3 - x_{ij} - x_{i'j} - \mu_{ii'}\right) + s_i \geq e_{i'}$	$\forall i, i' \in \{i \neq i' \lvert 1, \dots, V\}, \forall j \in \{1, \dots, B\}$	(6)
$M\left(2 - x_{ij} - x_{i'j} + \mu_{ii'}\right) + s_{i'} \geq e_i$	$\forall i, i' \in \{i \neq i' \lvert 1, \dots, V\}, \forall j \in \{1, \dots, B\}$	(7)
$x_{ij} + x_{ij+1} \leq 1$	$\forall i \in \{1, \dots, V\}, \forall j \in CB$	(8)
$\Delta EAR_i \geq ETA_i - s_i$	$\forall i \in \{1, \dots, V\}$	(9)
$\Delta LAT_i \geq e_i - (ETA_i + p_i)$	$\forall i \in \{1, \dots, V\}$	(10)
$s_i, e_i, \Delta EAR_i, \Delta LAT_i \geq 0$	$\forall i \in \{1, \dots, V\}$	(11)
$x_{ij}, \delta_{ij}, \mu_{ii'} \in \{0,1\}$	$\forall i, i' \in \{i \neq i' \lvert 1, \dots, V\}, \forall j \in \{1, \dots, B\}$	(12)

The objective function (1) minimizes early arrival and late departure costs. Constraint (2) explains that ships will occupy as much space as their length. Hybrid berth structure is defined in constraint (3). Constraints (3.1), (3.2), (3.3) and (3.4) were created to convert constraint (3) into linear programming structure. Constraint (4) shows the departure time and departure time of the ship. Constraint (5) ensures that the arrival time of a vessel at the port is greater than the earliest arrival time. Constraints (6) and (7) explain that the vessel assigned to the same berth do not overlap. Constraint (8) is a structural constraint and ensures that vessels cannot be assigned to corners. Constraints (9) and (10) calculate the early arrival time and the late departure time in the objective function. Constraints (11) and (12) are signal constraints. This problem cannot achieve an optimal result in polynomial time, so it is included in the NP-hard class. To create a solution, the PSO and TLBO solution algorithms shown in the next section have been developed.

Metaheuristic Algorithms

We propose two swarm intelligence algorithms called Particle Swarm Optimization (PSO) and Teaching-Learning Based Optimization (TLBO). In the algorithms, mutations were applied to 20% of the population once every 5 replications. Stop criterion; the number of iterations and a quarter of the number of iterations were used to improve the solution.

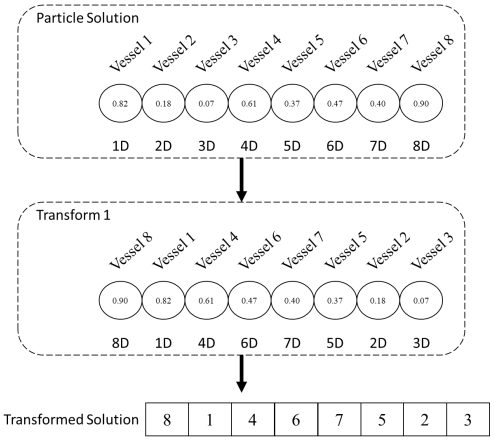


Figure 6. Encoding Phase

Algorithm 1. Decoding Algorithm
Input: Individual Solutions (Transformed Solutions), Berth Locations, EST_i, ETA_i, $cost_i^1$, $cost_i^2$, l_i, p_i
For all individual u
  For all vessels i in Transformed Individual Solutions (in order)
    If there is until l_i any empty berth location suitable for corner constraints between ETA_i and $ETA_i + p_i$ times:
    Then assign the vessel to the first berth locations that are empty
    Else:
      If $(\Delta EAR_i + 1)cost_i^1 \leq (\Delta LAT_i + 1)cost_i^2$ **and** $ETA_i - (\Delta EAR_i + 1) \geq EST_i$:
        ΔEAR_i+= 1
        If there is until l_i any empty berth location suitable for corner constraints between $ETA_i - \Delta EAR_i$ and $ETA_i + p_i - \Delta EAR_i$ times:
        Then $\Delta LAT_i = 0$ and assign the vessel to the first berth locations that are empty
      Else:
        ΔLAT_i+= 1
        If there is until l_i any empty berth location suitable for corner constraints between $ETA_i + \Delta LAT_i$ and $ETA_i + p_i + \Delta LAT_i$ times:
        Then $\Delta EAR_i = 0$ and assign the vessel to the first berth locations that are empty
    Memorize:
      $\Delta EAR_i, \Delta LAT_i$
      $x_{ij} = Settlements\ 1, other\ 0$
      $s_i = ETA_i - \Delta EAR_i + \Delta LAT_i$
      $e_i = ETA_i + p_i - \Delta EAR_i + \Delta LAT_i$
  EndFor
  Reset port and vessel
EndFor
Output: DBAP Individual Solutions

Figure 7. Decoding Phase

Figure 1 and 2 show the coding and encoding phase. PSO and TLBO algorithms generated by mutation in Genetic algorithms are respectively shown in Figure 3 and Figure 4.

Algorithm 2. PSO algorithm for DBAP
Algorithm Inputs: Number of iterations, Number of particles, Maximum space limit (MSL), w, c_1, c_2
DBAP Inputs: Number of berths, Number of vessels, p_i, EST_i, ETA_i, $cost_i^1$, $cost_i^2$, l_i
Create initial solutions for each particle;
$x_i(0) = rand(-MSL, +MSL) \times (x_i^1, x_i^2, \dots, x_i^D)$
Encoding to solutions for each particle
Decoding to solutions for each particle
Evaluate $f(x_i)$ with Equation (1) for each particle
Repeat
  For all particles i
    Update velocities: $v_i(t) = w \times v_i(t-1) + \rho_1 \times c_1 \times (p_i - x_i) + \rho_2 \times c_2 \times (g - x_i)$
    Update to new position: $x_i(t) = x_i(t-1) + v_i(t)$
    if $Mod(iteration, 5) = 0$, ***then*** *select random %4 particle and* $x_i(t) = rand(-MSL, +MSL) \times (x_i^1, x_i^2, \dots, x_i^D)$
    Encoding to solutions for each particle.
    Decoding to solutions for each particle
    Evaluate $f(x_i)$ with Equation (1) for each particle
    if $f(x_i(t)) \leq pbest_i$, ***then*** $pbest_u = f(x_i(t)), p_u = x_u(t)$
    if $f(x_i(t)) \leq gbest$, ***then*** $gbest = f(x_i(t)), g = x_i(t)$
  EndFor
Until Stopping Criteria

Figure 8. PSO for DBAP

Algorithm 3. TLBO algorithm for DBAP
Algorithm Inputs: Number of iterations, Number of students, Maximum space limit (MSL)
DBAP Inputs: Number of berths, Number of vessels, p_i, EST_i, ETA_i, $cost_i^1$, $cost_i^2$, l_i
Create initial solutions for each student;
$Z_i = rand(-MSL, +MSL) \times (Z_i^1, Z_i^2 \dots, Z_i^D)$
Encoding to solutions for each student
Decoding to solutions for each student
Evaluate $F(Z_i)$ with Equation (1) for each student
Calculate Z_{mean}
Repeat
  For all students i
    Teacher phase: $Z_{new\ i} = Z_{new\ i} + r \times \left(Z_{teacher} - \left(T_f \times Z_{mean}\right)\right)$
    Student phase:
      if $F(Z_a) \leq F(Z_b)$; $Z_{new\ a} = Z_{old\ a} + r_i \times (Z_a - Z_b)$
      if $F(Z_b) < F(Z_a)$; $Z_{new\ a} = Z_{old\ a} + r_i \times (Z_b - Z_a)$
    if $Mod(iteration, 5) = 0$, ***then*** *select random %4 student and* $Z_i = rand(-MSL, +MSL) \times (x_i^1, x_i^2, \dots, x_i^D)$
    Encoding to solutions for each student.
    Decoding to solutions for each student
    Evaluate $F(Z_i)$ with Equation (1) for each student
    Calculate Z_{mean}
    if $F(Z_i) \leq F(Z_{teacher})$, ***then*** $Z_{teacher} = Z_i$
  EndFor
Until Stopping Criteria

Figure 9. TLBO for DBAP

Computational Experiments and Conclusion

The PSO and TLBO algorithms were coded in Python and run 50 times for small, medium and large scale casings on a 32 GB ram computer with Intel Xeon E5 3.50 GHz processor and the results (minimum - average - maximum) are given in Table 2. At the same time, the results of First in First Out (FIFO) used by many port operators are added to the table. All data and codes are shared with https://github.com/serkanmutlu87/Paper33_DMS2019.

Table 3. Computational Experiments

Case Type	nV-nB	PSO		TLBO		FIFO
		Obj. F.	CPU (sec.)	Obj. F.	CPU (sec.)	
Small	10-10	0 – 0 – 0	0.0 – 0.04 – 0.1	0 – 0 – 0	0.0 – 0.02 – 0.1	0
	15-10	113 – 115 – 133	3.1 – 3.1 – 6.0	113 – 116 – 128	2.4 – 3.8 – 7.8	194
	20-10	282 – 298 – 341	4.6 – 8.1 – 18.4	292 – 310 – 364	4.7 – 7.9 – 17.1	1487
Medium	20-20	0 – 0 – 0	0.0 – 0.05 – 0.1	0 – 0 – 0	0.0 – 0.05 – 0.1	0
	30-20	25 – 35 – 61	4.1 – 7.7 – 15.6	29 – 40 – 57	4.7 – 8.1 – 13.7	38
	40-20	32 – 38 – 47	3.9 – 6.4 – 10.5	32 – 40 – 48	4.1 – 6.9 – 13.2	96
Large	30-30	0 – 0 – 0	0.1 – 0.1 – 0.1	0 – 0 – 0	0.1 – 0.1 – 0.1	0
	45-30	12 – 14 – 21	5.2 – 8.8 – 18.5	12 – 17 – 35	5.7 – 9.0 – 21.7	38
	60-30	89 – 117 – 150	11 – 23.9 – 41	100 – 149 – 193	10 – 17.1 – 32	222

Both algorithms (PSO and TLBO) provide better results than FIFO algorithm used by existing enterprises. In addition, PSO gives better results than TLBO when Case size increases. In future studies, the effect of these algorithms on the results can be investigated by adding local search method.

References

[1] Imai, A., Nagaiwa, K., and Tat, C.W. (1997). Efficient planning of berth allocation for container terminals in Asia. *Journal of Advanced Transportation*, 31 (1), 75-94.
[2] Bierwirth, C.., and Meisel, F. (2015). A follow-up survey of berth allocation and quay crane scheduling problems in container terminals. *European Journal of Operational Research*, 244 (3), 675-689.
[3] Şahin, C., and Kuvvetli, Y. (2016). Differential evolution based meta-heuristic algorithm for dynamic continuous berth allocation problem. *Applied Mathematical Modelling*, 40 (23-24), 10679-10688.
[4] Correcher, J.F., and Alvarez-Valdes, R. (2017). A biased random-key genetic algorithm for the time-invariant berth allocation and quay crane assignment problem. *Expert Systems with Applications*, 89 (15), 112-128.
[5] Wang, R., Nguyen, T.T., Li, C., Jenkinson, I., Yang, Z., and Kavakeb, S. (2019). Optimising discrete dynamic berth allocations in seaports using a Levy Flight based meta-heuristic. *Swarm and Evolutionary Computation*, 44 (1), 1003-1017.
[6] Barbosa, F., Priscila, C., Rampazzo, B., Yamakami, A., and Camanho, A.S. (2019). The use of frontier techniques to identify efficient solutions for the Berth Allocation Problem solved with a hybrid evolutionary algorithm. *Computers & Operations Research*, 107 (1), 43-60.

Analyzing The Effect Of Various Factors On Having Coronary Heart Disease

S.Ö. Rençber[1], E. Biçek[2], E. Kına[3]

[1]Van Yüzüncü Yıl University, Van, Turkey, rencber.serhatomer@gmail.com
[2]Van Yüzüncü Yıl University, Van, Turkey, emre.bicek@yyu.edu.tr
[3]Van Yüzüncü Yıl University, Van, Turkey, erol.kina @yyu.edu.tr

Abstract

In the medical examination, blood parameters and blood pressure are the indicators when defining the risk of heart diseases. The aim of this study is to compare the effect of gender, age, heart rate by correlating some blood tests to predict the risk factor for having heart diseases. Data is extracted from healthy human population and this data is analyzed by using data science and data mining techniques.

Introduction

Cardiovascular Diseases (CVD) mortality rate increased gradually in developed countries until 1960s, then slowed down. However, it is still one of the main causes of death.[1] Since the 1960s, deaths of cardiovascular diseases have fallen in industrialized countries. The main reason for this is thought to decrease the risk factors for Coronary Heart Disease (CHD) [2,3].

Individuals who are at risk for cardiovascular diseases can now be identified more easily with technology. The incidence of Coronary Heart Disease (CHD) is higher in men than in women. In both sexes, the risk of CHD increases with age, but the increase in women is more pronounced. There is a significant difference between genders in the risk of Coronary Heart Disease (CHD) [4].

Age, sex, blood pressure and total cholesterol are effective at the risk of death in coronary heart disease [5,6]. Blood test results and blood pressure are indicators that determine the risk of coronary heart disease.

The aim of this study is to compare some blood tests with the data of cardiovascular disease and healthy individuals by considering gender, age, heart rate and other values to estimate the risk factor for cardiovascular disease.

Method

The dataset used in the study is the data set presented by the Cleveland Clinic in Ohio State of the United States, where personal information is extracted and shared on the internet. The data set for 303 people is in MS Excel format as shown in Table 1and . Table 2. These data include the data of 165 cardiovascular patients and 138 healthy individuals.

Table 1. Cleveland Clinic Patients Datas

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Table 2. Descriptions of columns in the dataset

Coulmn No	The definition of columns
1	Age = How old is he/she
2	Sex= Gender
3	Chest Pain Type (4 values) = The meaning of Values (4 Values, 1=The Usual Pain, 2=Unusual Pain 3= Not a sign of Chest Pain 4= Not pain related to disease
4	Resting Blood Pressure - For your resting heart rate, the target is between 60 and 100 beats per minute (BPM)
5	Serum Cholestoral in mg/dl - serum cholesterol levels in milligrams per deciliter (mg/dL).
6	Fasting Blood Sugar > 120 mg/dl - If fasting blood glucose> 120 mg / dl, the value is shown as = 1. Otherwise, the value is 0.
7	Resting Electrocardiographic Results (values 0,1,2) - electrocardiographic (ECG) results at rest (values 0,1,2)
8	maximum heart rate achieved - elde edilen maksimum kalp atış hızı
9	Exercise Induced Angina - angina due to exercise (angina = chest pain due to spasm in the heart muscles) If value = 1 there is chest pain at the time of exercise.
10	Oldpeak = ST depression induced by exercise relative to rest - oldpeak = exercise-induced ST depression according to rest (a test that shows a high risk of cardiovascular disease)
11	the slope of the peak exercise ST segment
12	number of major vessels (0-3) colored by flourosopy
13	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14	Target = Status (not Patient if Status = 0, patient if Status = 1)

Tools

Python programming language was used for data analysis. Python programming language has been used frequently in data science and data mining studies in recent years. In the Python programming language, libraries such as Pandas, Numby, Seaborn can handle data processing and visualization. In this study, data were analyzed using the advanced features of these libraries. Analysis results are shown graphically.

Results

In the data set, the number of male individuals was 207 and the number of female individuals was 96. Figure 1 shows that individuals have a higher rate of disease between the ages of 41 and 54 years.

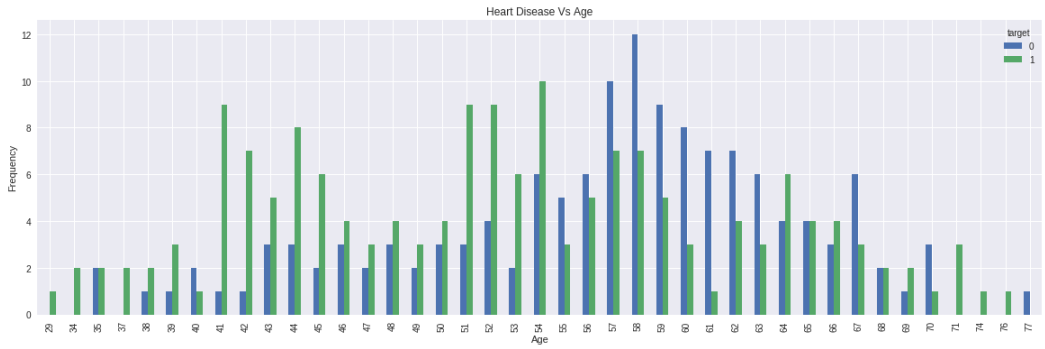


Figure 1. The number of individuals with cardiovascular disease by age.

According to the values in Figure 2; Blood pressure between 120 and 140 values of men with the disease was remarkable. It was found that the number of individuals carrying the disease was much higher in individuals with cholesterol values between 200 and 300 in men and women. Interestingly, blood sugar was found to be low in individuals with the disease.

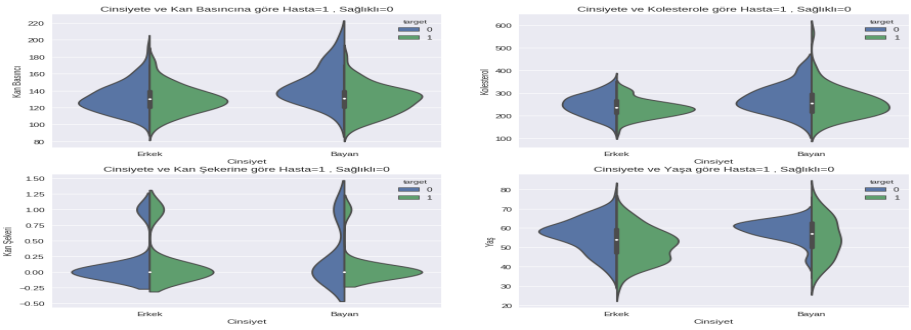


Figure 2. Disase rates by gender and other values

In Figure 3, it was determined that the heart rate of patients with and without high blood sugar levels was high.

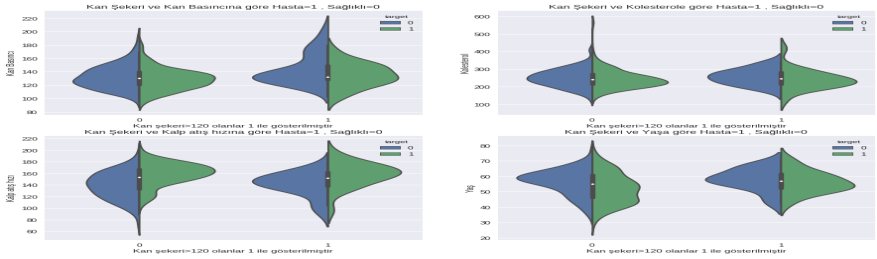


Figure 3. Disease risks according to the blood sugar and other values

Conclusions

According to the results, it was determined that many factors directly or indirectly affect the risk of cardiovascular diseases. Badıllıoğlu et al. [7] in their research on 314 individuals in Güzelbahçe district of İzmir, the incidence of CHD increases significantly with age. In the study group, the incidence of CHD was found to be higher in patients with diabetes, family history of CHD, smokers, and total cholesterol and triglyceride levels.

In our study, in the data set between the ages of 41-54 years, despite the intensive cardiovascular disease appears less after 54 years. In addition, there was no difference in CHD risk with a glucose level above or below 120. The limitation of this study was evaluated according to the data of 303 individuals with CHD and healthy individuals. Different results can be obtained with more data. It is thought that this study will contribute to similar studies.

References

[1] McGovern PG, Pankow JS, Shahar E (1996). Recent trends in acute coronary heart disease–mortality, morbidity, medical care, and risk factors. *N Engl J Med*, 334(14), 884–90.
[2] Capewell S, Ford ES, Croft JB, Critchley JA, Greenlund KJ, Labarthe DR (2010). Cardiovascular risk factor trends and potential for reducing coronary heart disease mortality in the United States of America. *Bull World Health Organ,* 88(2), 120–30.
[3] Capewell S, Hayes DK, Ford ES (2009). Life-years gained among US adults from modern treatments and changes in the prevalence of coronary heart disease risk factors between 1980 and 2000. *Am J Epidemiol*, 170(2), 229–36.
[4] Jousilahti, P., Vartiainen, E., Tuomilehto, J., & Puska, P. (1999). Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in Finland. *Circulation*, 99(9), 1165-1172.
[5] Castelli W.P. (1984). Epidemiology of coronary heart disease: the Framingham Study. *Am J Med.*,76, 4–12.
[6] Jackson R, Chambless L, Higgins M, Kuulasmaa K, Wijnberg L, Williams D (1997). Sex difference in ischaemic heart disease mortality and risk factors in 46 communities: an ecologic analysis. *Cardiovasc Risk Factors*, 7, 43–54.
[7] O. Badıllıoğlu, B. Ü. Toğrul, Ş. Reyhan Uçku (2011). İzmir, Güzelbahçe'de koroner kalp hastalığı beş yıllık insidansı ve risk faktörleri ile ilişkisi. *Türkiye Halk Sağlığı Dergisi,* 9(3), 129-132.

Using Decomposition-based Approaches to Time Series Forecasting in R Environment

Z. Pala[1]

[1] *Muş Alparslan University, Computer Eng Dept. Muş, Turkey, z.pala@alparslan.edu.tr*

Abstract

A time series measured in any area is the basis for characterizing an observed system and predicting its future behavior. Specifically, the purpose of a predictive model used for time series is based on estimating the value of an unknown variable. It does this by looking at historical data and analyzing the relationships between historical data. In this study, we estimate the number of men who died of lung cancer in England between 1974-1979, using a monthly time series consisting of 72 records. In the estimation process, decomposition-based approach and non-decomposition-based approaches were compared. In the comparison process, method estimations for Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were used and the results were discussed.

Introduction

Time series are widely used in many areas because they form an important link between the past and future. In literature, it is possible to find publications in a wide range of fields such as business economics, statistics, engineering, environmental sciences or physical sciences [1]. Time series are different from machine learning (ML) approaches in terms of structures and forms of analysis. The input data for machine learning may be independent, but this is not true for the time series. Each entry of the time series is associated with the next entry. Each entry is also the output of the previous observation. For example, in most ML models, a model is trained, tested, retrained if necessary until satisfactory results are obtained, and then evaluated with the new data set. Once satisfactory results are obtained, the model at hand is available for new predictions. But for time series models, the situation is different. Whenever a new prediction is desired, the model must be retrained using the latest dataset of the time series.

It is accepted that errors are independent of each other when creating a linear regression model. However, in the time series, the situation is different and it is assumed that the errors affect each other. In other words, the terms of the error are associated with the previous, depending on the time. Therefore, Machine Learning methods can be disadvantageous with respect to time series techniques [2].

In this study, a decomposition-based approach is used which is a very simple but very robust method for modeling and estimating time series. The basic logic of this method is to model time series data as trend, seasonal and remainder. A monthly dataset in R datasets was used. Using the monthly dataset consisting of 72 records, future estimates based on the knowledge of male patients who died of lung cancer in England between 1974-1979 were made and the results obtained were evaluated.

Time series and Modeling

If we display the data of a time series with x, it is possible to display each data with x_t. Here, t represents the time shown as the index. Where t = 1 indicates the first observation value, t = T indicates the last observation value. The whole set of time can be expressed as the observation period t = 1,2,3, ..., T. Observations are measured at the same time interval. This time interval can be hourly, hourly, daily, weekly, monthly, seasonal or annual [3]. The future forecast horizon for a time series can be displayed as T + 1, T + 2, ..., T + h. Here, h = 1, 2, ..., H can be expressed as a forecast horizon. It is possible to separate a time series into three or four components with the help of the decomposition model:

$$\hat{X}_t = T_t + S_t + R_t + \in_t \quad (1)$$

Here $\hat{X}_t$, represents the modeled or predicted value at time t, T_t, the trend component at time t, S_t, the seasonal component at time t, R_t, the remainder component at time t and the term $\in_t$, error at time t.

In the time series, the decomposition method is considered as a step of the analysis before proceeding with the estimation process. It can also be considered as an analysis method. In order to predict the future of a time series

using a decomposition model, the future values of the trend, seasonal and remainder components are calculated separately. These components are then reassembled. The challenge in this method is to find the best model for each of the components. Here we use the Local regression (LOESS) method to make decomposition based estimation in the R environment. STL is a versatile and robust method for separating time series [4]. The STL algorithm was used in 1990 by Cleveland et al. [5]. The STL algorithm uses polynomial regression to model trend and seasonal components with the help of the LOESS method.

Then, using a standard algorithm such as Arima, ETS, Naive and Rwdrift, the residual R_t component is estimated. However, a h-step forward estimation is made for the remainder component (R_{t+h}). Finally, the results obtained by h-step forward estimation are collected for T_{t+h} and S_{t+h}components. Thus, the final estimate is made.

A variety of statistical measures have been calculated, including the Average Absolute Percent Error (MAPE), the root mean square error (RMSE), and the mean absolute error (MAE) to examine the performance of the models using our time series numerical values [6]. MAPE represents the percentage of the mean absolute error.

Material and Method

The dataset, which forms the basis of this study, was taken directly from the datasets library of the R programming language (version 3.5.1) used for analysis. The dataset included the number of male patients who died of lung cancer in England between 1974-1979 and consisted of 72 records.

For the time series to be trained and to evaluate the results realistically, part of the series is reserved for the training and the remaining part is for testing. For this purpose, the first 54 records (75%) were reserved for training, while the last 18 records (25%) were reserved for the test procedure.

Results and Discussion

Here, Arima, ETS, Naïve and Rwdrift models of the R programming language and environment were used with the STL method defined in the forecast library. The models were first trained with the training data and 18-month future estimates were made. As shown in Figure 1, 18-month estimates are given at the end of the training data.

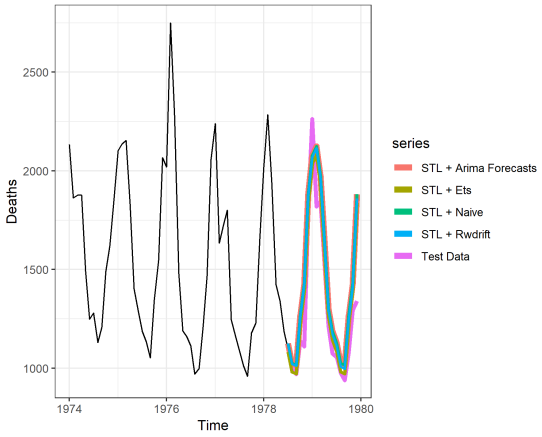


Figure 1. Comparative estimates for monthly time series

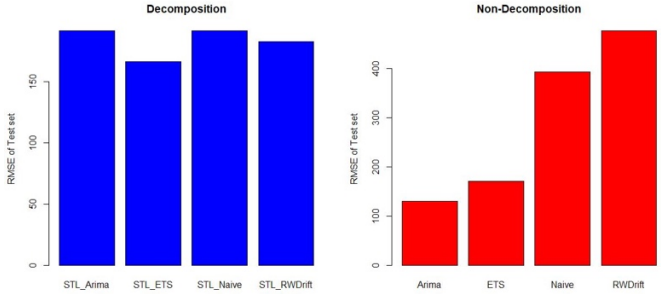


Figure 2. RMSE test set performance graphs for Decomposition and Non-decomposition approaches

Figure 2 is generated to illustrate the performance of the algorithms visually and more clearly with the support of measurement metrics. In addition, RMSE, MAE and MAPE metric values of two different methods are given in Table I. With the exception of the Arima model, when the data in both the table and the performance graphs are analyzed, it is seen that the decomposition approach makes more successful predictions in the non-decomposition approach. However, we can say that the decomposition-based approach for the widely used dataset in the R datasets library, Airpassengers, used the MAPE metric to make up to 3.25% less error than the normal approach.

Table 1. Comparison metrics of forecasting methods

	Decomposition			Non-Decomposition		
Method/Metric	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Arima	191.37	144.18	10.97	131.08	87.30	6.73
ETS	166.32	109.26	7.99	171.69	115.05	8.48
Naive	191.37	144.18	10.96	393.07	280.38	18.74
Rwdrift	182.41	133.60	10.06	476.13	334.66	21.30

Within the four algorithms used for estimation, the best estimates were made by ETS, Rwdrift, Naive and Arima respectively. The ETS model made a 7.99% error with the decomposition approach with the MAPE metric while making an error of 8.48% with the non-decomposition approach. The performance of the same algorithm as MAPE metric was 92.01% and 91.52% for two models, respectively. Therefore, the decomposition approach has made less than 0.49% errors for the ETS model, so it has been more successful than the non-decomposition approach.

Conclusions

In this study, we estimate the number of men who died of lung cancer in England between 1974-1979, using a monthly time series consisting of 72 records. For analysis, the analysis of the decomposition-based approach using the open-source R programming language was performed with the help of evaluation metrics such as RMSE, MAE and MAPE. The decomposition-based approach was found to be more successful than the normal approach, especially on seasonal data.

References

[1] Pala, Z., Atici. R. (2019). Forecasting Sunspot Time Series Using Deep Learning Methods. *Solar Phys,* 294 (50), 1-14.

[2] Pala, Z., Ünlük, İ.H., Şahin, Ç. (2018). Forecasting Low Frequency Electromagnetic Fields Values Time Series Using Python, *International Conference on Innovative Engineering Applications* (CIEA' 2018), 20-22 Sep 2017, Sivas, Turkey.

[3] Mills, T.C. (2019). *Applied Time Series Analysis A Practical Guide to Modeling and Forecasting*, Academic Press, 125 London Wall, London EC2Y5AS, United Kingdom.

[4] Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.), Monash University, Australia.

[5] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–33.

[6] Namin, S. S., Namin, A. S.(2018). *Forecasting economic and financial time series: Arima vs. LSTM*, Texas Tech University, Lubbock, TX, USA, 2018.

Time Series Analysis of Radiological Data of Outpatients and Inpatients in Emergency Department of Mus State Hospital

E. Yaldız[1], Z. Pala[2]

[1] Muş Municipality Information Processing Office Muş, Turkey, erkan1yaldiz@gmail.com
[2] Muş Alparslan University, Muş, Turkey, z.pala@alparslan.edu.tr

Abstract

Time series is a series of discrete or continuous observations that are time dependent. Time series analysis and modeling is a dynamic research area that has long attracted the attention of the research community. The main aim of time series modeling is to develop a suitable model by observing the past of the time series and to predict the future accordingly. The aim of this study is to estimate the future for the number of patients referred to the radiology unit and to contribute to the management of the hospital in terms of better service quality. We have obtained the dataset from Mus State Hospital. The dataset with 60 months of data includes the number of patients referred to the radiology unit from the emergency department between 2014-2018. With the help of the classical and artificial neural networks algorithms used in the R programming language, the estimated future results are evaluated. RMSE, MAE and MAPE metrics were used to evaluate the estimation results obtained by using different algorithms.

Introduction

Time series are sequences of measurements obtained at discrete or continuous and consistent time intervals connected with each other over a specified period of time [1]. A time series cannot be created from data that is not collected specifically or within a specific order. Even if generated, estimates cannot be made as expected.

In the literature, it is possible to come across publications in various fields related to time series. We can include this in various fields such as statistics, engineering, finance, business economics and physical sciences [2]. Especially in the health sector, time series are of particular importance.

The dataset used in this study was obtained from Muş State Hospital. The dataset used belonged to a five-year period and consisted of 60 records. The aim is to estimate the number of radiological images taken by the patients presenting from the Emergency Department of the Muş State Hospital between 2014-2018 using the available dataset and to make future plans.

Time series and Modeling

If we display the data of a time series with x, it is possible to display each data with x_t. Here, t represents the time shown as the index. Where t = 1 indicates the first observation value, t = T indicates the last observation value. The whole set of time can be expressed as the observation period t = 1,2,3, ..., T. Observations are measured at the same time interval. This time interval can be hourly, daily, weekly, monthly, seasonal or annual [3].

Various statistical metrics have been used to examine the performance of models using our time series numerical values, including Mean Absolute Percent Error (MAPE), root mean square error (RMSE), and mean absolute error (MAE) [4]. MAPE represents the percentage of the mean absolute error.

Material and Method

The dataset used was composed of 60-month records including the number of outpatients and inpatients whose radiological images were requested during 2014-2018. Part of the time series is reserved for training and the remaining part is for the testing process in order to evaluate the results. In order to better understand the link between training and test data, two training and two test data were studied. For this purpose, the first 42 records (70%) were reserved for training and the last 18 records (30%) were reserved for the test procedure. After the first evaluation result, the second 48 records (80%) are reserved for training and the last 12 records (20%) are reserved for testing.

Results and Discussion

Arima, ETS, Snaive and Tbats models were used together with STL method defined in the forecast library of R programming language and environment [5].

The models were first trained with training data. Then, 18 and 12 months future forecasts were made, and then the results were compared with the test results and performance evaluation was performed. Figures 1 and 2 show 18-month and 12-month forecasts, respectively.

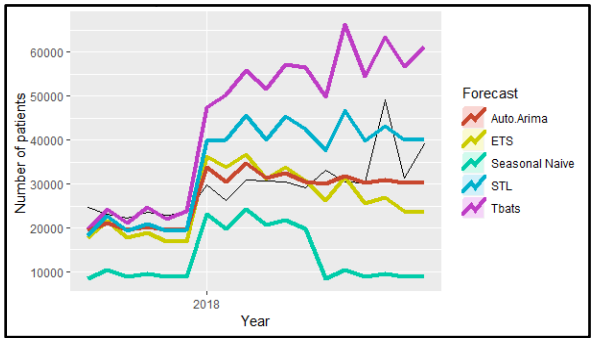


Figure 1.12-month forecasts

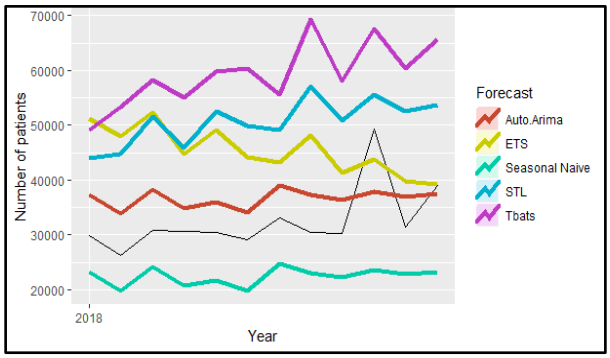


Figure 2. 18-month forecasts

It is produced in Figure 3 and Figure 4 to visually and more clearly show the performance of the algorithms with the support of measurement metrics. In addition, RMSE, MAE and MAPE metric values showing the performance of two different methods are given in Table I and Table II.

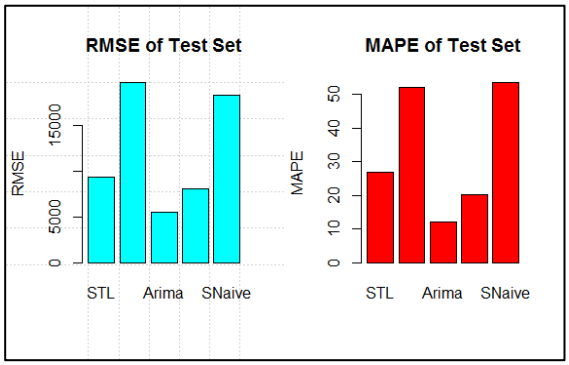


Figure 3. Dataset training/testing(70% and 30%)

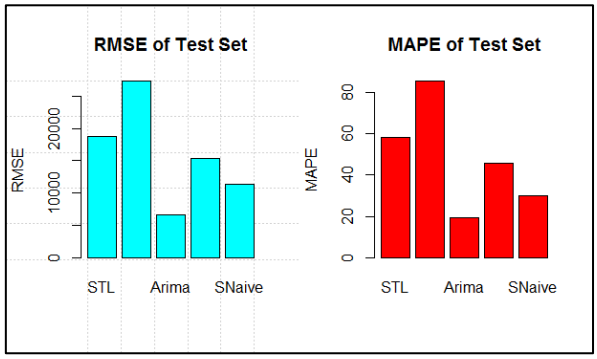


Figure 4. Dataset training/testing (80% and 20%)

Table 1. Comparison metrics of forecasting methods (70% and 30%)

	Training Set			Test Set		
Method/Metric	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ARIMA	1706,57	588,07	4,44	5557,47	3852,58	12,21
ETS	739,25	411,93	5,09	8108,77	6291,81	20,24
SNAİVE	6440,21	5229,96	51,34	18306,13	16137,83	53,55
STL	861,13	289,21	2,06	9348,91	7890,93	26,88
TBATS	838,06	325,90	3,04	19643,99	16112,57	52,09

Table 2. Comparison metrics of forecasting methods (80% and 20%)

	Training Set			Test Set		
Method/Metric	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ARIMA	1796,91	702,95	4,66	6611,02	6217,77	19,37
ETS	1165,22	633,39	6,00	15296,86	13787,34	45,55
SNAİVE	8236,93	6705,77	52,85	11396,42	10121,91	29,77
STL	1251,06	533,55	3,63	18730,16	18054,35	58,31
TBATS	1231,49	574,20	4,18	27330,28	26818,05	85,43

When the two analyzes performed on the dataset are evaluated, it is possible to see both the accuracy of the algorithms and how the performance sequences on the dataset change.

For the Training / Test (80%, 20%) approach, the best estimates were made by ARIMA, SNAIVE, ETS, STL, TBATS, respectively. For example, ARIMA model made 19,37% error according to MAPE metric.

When we compare the metric values in Table I (first approach: 70%, 30%) and Table II (second approach: 80%, 20%), we can say that the first approximation is more successful than the second approach.

Conclusions

In this study, a monthly time series consisting of 60 records including outpatient, inpatient and radiological images of the patients who came to Muş State Hospital Emergency Department between 2014-2018 was estimated. RMSE, MAE and MAPE were used to evaluate the analysis results.

References

[1] Lewis, N.D., (2017). N*eural Networks for Tıme Serıes Forecastıng with R: An Intuitive Step by Step Blueprint for Beginners*, CreateSpace Independent Publishing Platform.
[2] Pala, Z., Atici. R. (2019). Forecasting Sunspot Time Series Using Deep Learning Methods. *Solar Phys*, 294 (50), 1-14.
[3] Terence C. Mills, (2019). *Applied Time Series Analysis A Practical Guide to Modeling and Forecasting*, Academic Press, 125 London Wall, London EC2Y 5AS, UK.
[4] Namin , S. S., Namin ,A. S. (2018). *Forecasting economic and financial time series: Arima vs. LSTM*, Texas Tech University, Lubbock, TX, USA.
[5] Hyndman R. J., Athanasopoulos G.(2018). *Forecasting: Principles and Practice* (2nd ed.), Monash University, Australia.

Prediction of monthly electricity consumption used in Muş Alparslan University Complex by means of Classical and Deep Learning methods

İ.H. Ünlük[1], Z. Pala[2]

[1] Muş Alparslan University, Muş, Turkey, ih.unluk@alparslan.edu.tr
[2] Muş Alparslan University, Muş, Turkey, z.pala@alparslan.edu.tr

Abstract

Time series modeling and prediction are of great importance for a variety of practical areas. Recently, interest in this area has increased. Predicting the future based on the past is one of the fundamental tasks of the time series. Many important models have been proposed in the literature for time series modeling and prediction.

In this study, a monthly dataset for the electricity consumed in Muş Alparslan University campus between 2014-2019 was obtained from VEDAŞ. With the help of the dataset which has been transformed into time series, the active energy consumed in MŞÜ campus is estimated. The prediction process uses built-in prediction algorithms in the R language. R language and software environment play a big role in creating, analyzing and estimating time series. In addition to classical algorithms, the predicted results were compared by using deep learning algorithms and the results obtained within the framework of the estimated performance measures were evaluated.

Introduction

With the discovery of electricity, we have met inventions that change, facilitate and direct our lives. Today, electric energy is among the indispensable sources due to its environment-friendly nature. It is a very important problem that large institutions such as universities estimate their own electricity costs and take measures accordingly.If the future is planned from today, if it is predicted in a sense, the importance to be taken will be meaningful.

Various algorithms are used in the literature for electrical energy estimation. The lower the estimated error rate, the better the performance of planning [1].

In this study, it is aimed to estimate the electricity consumption spent between 2014-2019 in MUSU campus. Classical and deep learning algorithms will be used for the training and testing of the dataset, which is obtained from VEDAŞ and consists of 62 records. The results will be evaluated with RMSE, MAE and MAPE evaluation metrics.

Time series and Modeling

Time series are sequences of measurements obtained at discrete or continuous and consistent time intervals connected with each other over a specified period of time [2].

If we display the data of a time series with x, it is possible to display each data with x_t. Here, t represents the time shown as the index. Where t = 1 indicates the first observation value, t = T indicates the last observation value. The whole set of time can be expressed as the observation period t = 1,2,3, ..., T. Observations are measured at the same time interval. This time interval can be hourly, daily, weekly, monthly, seasonal or annual [3].

Various statistical metrics have been used to examine the performance of models using our time series numerical values, including Mean Absolute Percent Error (MAPE), root mean square error (RMSE), and mean absolute error (MAE) [4]. MAPE represents the percentage of the mean absolute error.

Material and Method

In this study, classical and deep learning algorithms in the forecast library of the R programming language were utilized. The dataset used for estimation was obtained from VEDAŞ. The time series graph of the dataset consisting of 62 records is given in Figure 1.

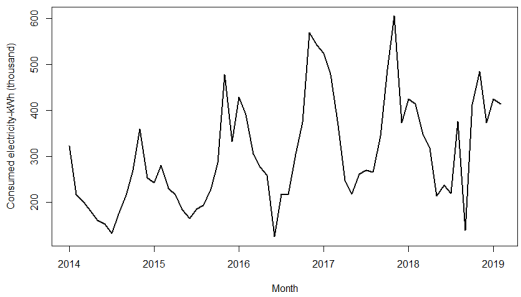


Figure 1. Monthly time series for electricity consumption

Results and Discussion

Arima, ETS, Snaive and Tbats models were used together with STL method defined in the forecast library of R programming language and environment [5]. Using two different approaches, the existing dataset was divided into two for training and testing. In the first approach, the dataset was divided into 67% and 33% for training and testing, respectively. Predictions of this approach are shown in Figure 2. In the second approach, the dataset was divided by 87% and 13% for training and testing, respectively. Predictions of this approach are shown in Figure 3.

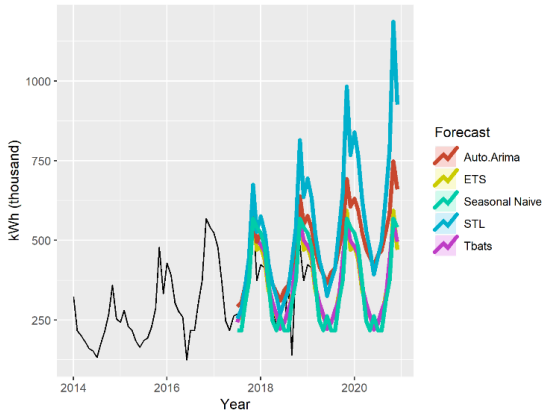


Figure 2. Estimates using classical algorithms (for training 67%, for testing 33%).

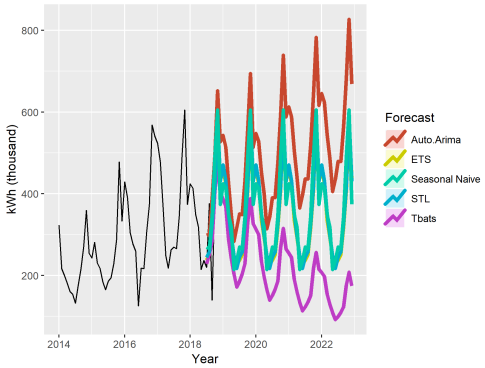


Figure 3. Estimates using classical algorithms (for training 87%, for testing 13%).

Furthermore, the metric graphs and values of these two approaches are given in and Figure 4, Table I respectively.

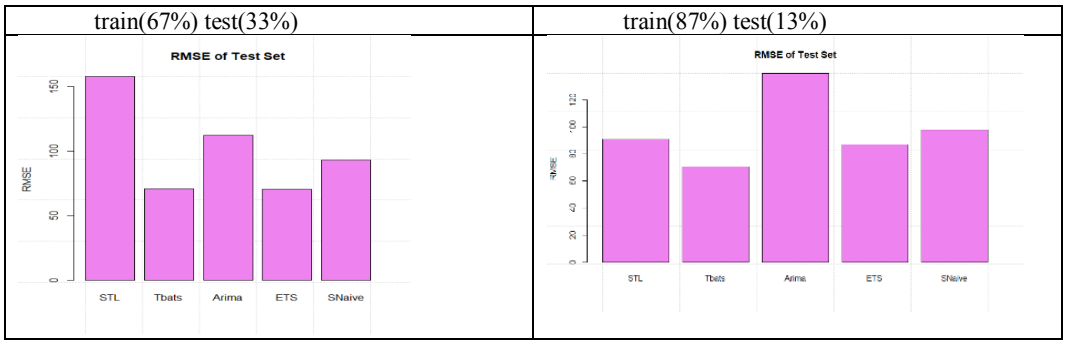


Figure 4. Comparison of evaluations metrics

Table 1. Comparison metrics of forecasting methods

	train(67%) test(33%)			train(87%) test(13%)		
Method/Metric	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ARIMA	112.36	92.49	32.84	139.48	127.81	47.58
ETS	70.52	51.44	18.02	87.11	65.07	26.90
NAIVE	92.82	75.95	23.68	97.96	70.12	30.20
STL	157.74	123.32	39.33	90.98	69.77	28.58
TBATS	70.92	55.77	19.10	70.74	49.91	21.12

Evaluation metrics for deep learning algorithms are given in Table II.

Table 2. Comparison metrics of forecasting methods

	train(67%) test(33%)			train(87%) test(13%)		
Method/Metric	RMSE	MAE	MAPE	RMSE	MAE	MAPE
NNETAR	91.08	79.04	23.72	103.96	71.29	35.06
MLP	130.57	101.03	34.97	109.79	81.48	37.28
ELM	153.02	125.37	32.26	82.21	64.82	27.90

When we compare the metric values in Table I (first approach: 67%, 33% second approach: 87%, 13%), we can say that the first approximation is more successful than the second approach. The performance of the algorithms used may vary according to the division rate of the dataset.

Conclusions

In addition to classical algorithms, the predicted results were compared by using deep learning algorithms and the results obtained within the framework of the estimated performance measures were evaluated. In this study, a monthly dataset for the electricity consumed in Muş Alparslan University campus between 2014-2019 was obtained from VEDAŞ. The performance of the algorithms used may vary according to the division rate of the dataset.

References

[1] Hamzaçebi, C , Kutay, F . (2013). Yapay Sinir Ağlari İle Türkiye Elektrik Enerjisi Tüketiminin 2010 Yilina Kadar Tahmini. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 19 (3).
[2] Lewis, N.D., (2017). Neural Networks for Tıme Serıes Forecastıng with R: An Intuitive Step by Step Blueprint for Beginners, CreateSpace Independent Publishing Platform.
[3]Terence C. Mills, (2019). *Applied Time Series Analysis A Practical Guide to Modeling and Forecasting*, Academic Press, 125 London Wall, London EC2Y 5AS, UK.
[4] Namin, S. S., Namin, A. S. (2018). *Forecasting economic and financial time series: Arima vs. LSTM*, Texas Tech University, Lubbock, TX, USA.
[5] Hyndman, R. J., Athanasopoulos G. (2018). *Forecasting: Principles and Practice* (2nd ed.), Monash University, Australia.

An application of Genetic Algorithm with Changeable Population Size for Optimization of The Schwefel Function

E. Soygüder[1], A. O. Kızılçay[2], R. Saraçoğlu[3]

[1]Van Yüzüncü Yıl University, Van, Turkey, erensygdr@hotmail.com
[2] Van Yüzüncü Yıl University, Van, Turkey, oguzkizilcay@yyu.edu.tr
[3] Van Yüzüncü Yıl University, Van, Turkey, ridvansaracoglu@yyu.edu.tr

Abstract

Population size is one of the important factors affecting the success of the genetic algorithm. The best value of this parameter varies according to the problem. In this study, an approach in which the population size can change in the genetic process during the genetic algorithm study is proposed. The population size is changed depending on whether the algorithm achieves a better solution than the previous generation. The optimization test function selected for the experimental study is the Schwefel function. The results showed that the proposed approach has achieved successful results.

Introduction

The Genetic Algorithm (GA), which was inspired by the genetic processes of living things and still being developed by many researchers, was used for the first time in Bagley's [1] doctoral study, then Holland [2] brought the theoretical findings of GA into the literature. De Jong [3] is then shown on several experiments that GA can be used for optimization purposes. To date, many researchers have contributed to the improvement of the solution by doing various studies on GA.

When the Traditional GA study structure is examined, it is seen that each individual in a fixed population is exposed to basic operations such as selection, passage, mutation. There are many coding, selection, crossing and mutation methods in the literature today. It is not important what methods are used in this study because we focus on the impact of the population.

Related Works

To date, many researchers have carried out studies on the change in population size. In [4], the authors proposed a variable population Genetic Algorithm by introducing the concepts of "age" and "lifetime" for chromosomes. Since the selection procedures were performed independently of the suitability values in the proposed algorithm, chromosomes were subjected to death process in accordance with age and lifetime parameters. At the same time, population size has been reduced for generations as well-adapted individuals live longer. In [5], the authors self-adapted the intermediate population using three separate populations. In the following generations, the population size was changed according to the most appropriate value. In [6], again it is aimed to change the population size by procedures such as lifetime and death. In addition, the crossover and mutation rate were adjusted to self-adaptive. In [7], parents are kept alive after reproduction. Population size increases until it reaches a certain level between generations. When the upper limit is reached, deaths occur according to the suitability values of the individuals and return to the initial population size. In [8], the population size is altered according to the suitability value, similar to our suggestion. Here, however, new individuals are obtained by cloning good individuals selected by tournament selection from the actual population.

Proposed Method

Traditional GAs try to achieve the best solution by stabilizing the population. The number of individuals to be crossed and the number of elitist individuals are constant for each generation, and the intergenerational population size does not change. In this study, we aimed to increase the number of parents and the number of elite individuals to be selected if the solution did not develop by comparing the best

result of the previous generation with the current generation, and decrease if the solution developed. Individuals added to or removed from the population are provided by increasing or decreasing the number of individuals to be selected for reproduction from existing individuals. Likewise, the number of elitist individuals is increasing or decreasing by being affected by this situation. The variable population genetic algorithm that works in this way has produced a better solution for each trial than the traditional GA in studies on the Schwefel function.

In this works, a real coded genetic algorithm is used [9]. The initial population consists of random numbers. From the first generation, the eligibility of all individuals is calculated for the selection operations of the parents. The individuals to be selected for reproductive are determined using the roulette wheel method. New individuals are created by linear crossover method. The operation of transferring the best individuals to the next generation takes place in line with the crossover ratio. The chromosomes of the individuals are then mutated based on a certain ratio. The mutation occurs by random chromosome generation at the initial determined search range.

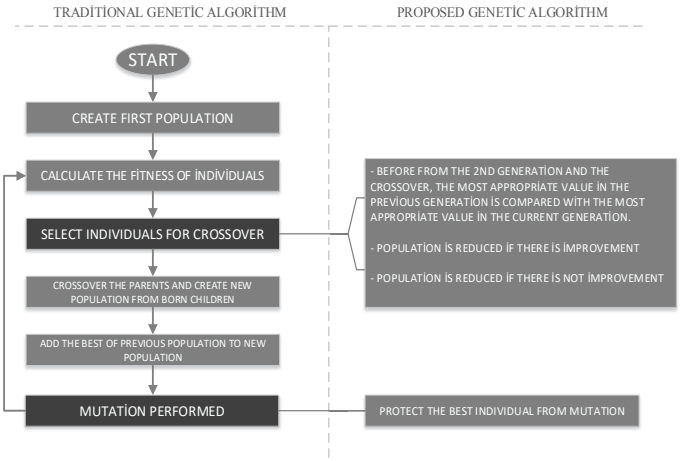


Figure 1. Traditional GA and proposed GA flow chart

Test Function and Parameters

Schwefel function was used as test function. Since the distance between local minimum and global minimum points is high, the probability of early convergence of algorithms is quite high.

The algorithm searches the [-500, 500] range for the Schwefel function. The spherical minimum of the function is 0. Here, when the function reaches zero, all roots take the value of [420.9687,…, 420.9687].

Schwefel: $f(\mathbf{x}) = 418.9829d - \sum_{i=1}^{d} x_i \sin(\sqrt{|x_i|})$

	Traditional GA	Proposed GA
Population Size	80	Veriable
Crossover Rate	0.8	0.8
Mutation Rate	0.2	0.35
Max. Generation	200	200
Dimensions	3	3

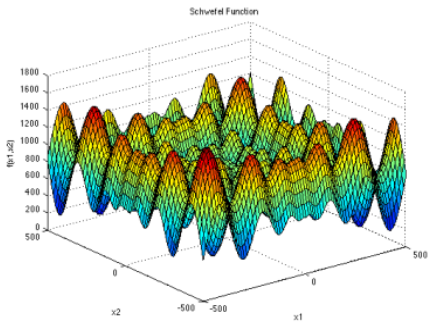


Figure 2. Schwefel function and parameters used & 3D simulation of Schwefel

The lower limit of the population is 50 and the upper limit is 100 for the proposed study. In case of non-healing, 2 individuals were added and in case of recovery, 20 individuals were excluded. As a result

of these process, the average number of individuals per generation was calculated as 80. It is also important to note that mutation rates are selected differently for the two studies. Here we wanted both algorithms to provide the best possible solution. We searched the mutation rate with a sensitivity of 0.01 in the range [0.01, 0.5] and used the values that produced the best solution. The mutation rates used are shown in the table in Figure 2.

Experimental Results

While calculating the results for both algorithms, 100 independent studies were performed. The means, standard deviations and best solutions of these studies are given in Figure 3. As can be seen from the results, the proposed method has achieved better results in every respect than the traditional GA. It is particularly important that the standard deviation is smaller. It means that the solutions found are closer to the average.

	Mean	Best	STD
Traditional GA	0.01764	1.01071e-05	0.02942
Proposed GA	**0.00161**	**4.54747e-13**	**0.0047**

Figure 3. Experimental results

Conclusion

In this article, we have presented a more understandable and useful suggestion, unlike the recommendations made on the size of the population until today is quite complex and difficult to implement. We used different methods to make fair comparisons for both algorithms. As a result, GA with variable population produced better solutions in every respect than traditional GA. In the future, different approaches can be applied for the further development of this study. Particularly, self-adaptive adjustment of real parameters will contribute to the solution. Other selection, crossover and mutation techniques can also be applied for the proposed study. Particularly, self-adaptive adjustment of real parameters will contribute to the solution. Other selection, crossover and mutation techniques can also be applied for the proposed study.

References

[1] Bagley, J. D. (1967). *The behavior of adaptive systems which employ genetic and correlation algorithms*, PhD thesis, University of Michigan.

[2] Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan press.

[3] De Jong, K. A. (1975). *An analysis of the behavior of a class of genetic adaptive systems Dissertation Abstracts International*, PhD thesis, University of Michigan.

[4] Arabas, J., Michalewicz, Z., and Mulawka J. (1994). GAVaPS – a genetic algorithm with varying population size, *In Proc. of the First IEEE Conf. on Evolutionary Computation*, Piscataway, NJ, 73–78.

[5] Hinterding, R., Michalewicz, Z., and Peachey, T C. (1996). Selfadaptive genetic algorithm for numeric functions, *in Parallel Problem Solving from Nature*, PPSN IV, 420–429.

[6] Rajakumar B.R., and Aloysius G., (2013) APOGA: An Adaptive Population Pool Size based Genetic Algorithm, *AASRI Procedia*, (4), 288-296.

[7] Shi, X. H., Wan, L. M., Lee, H. P., Yang, X. W., Wang, L. M., and Liang, Y. C. (2003) An Improved Genetic Algorıthm With variable Population size and A PSO-GA Based Hybrıd Evolutıonary Algorıthm, *Proceedings of the Second International Conference on Machine Learning and Cybernetics, IEEE*.

[8] Eiben, A. E., Marchiori, E., and Valko. V. A. (2004). Evolutionary algorithms with onthe-fly population size adjustment, *Parallel Problem Solving from Nature PPSN VIII, LNCS 3242*, 41–50.

[9] Wright A. (1991). Genetic algorithms for real parameter optimization, *Foundations of genetic algorithms*, (1) 205- 218.

LDA-Based Aspect Extraction from Turkish Hotel Review Data

K. Bayraktar[1]

[1]Gazi University, Ankara, Turkey, bayraktarkivanc@gmail.com

Abstract

Thoughts are the most important element affecting human life and enabling institutions and businesses to shape their future plans. As technology improves, we find opportunity to employ user data acquired through web resources to determine our daily basis, habits and decisions. In this study, LDA-based (Latent Dirichlet Allocation) aspect extraction methods have been proposed to identify single and multi-word aspects (MWA) for Turkish datasets, automatically. Introduced methods have been tested on a fragment of hotel dataset obtained via TripAdvisor. Experimental results show that proposed methods more successful than classical LDA. Classical LDA has 48% f-score value while LDA-WSBFE and LDA-C-value approaches obtained 74% and 72% f-score values, respectively.

Introduction

Internet has an important position in daily life and people can easily disseminate their thoughts through the internet. Therefore, data on the internet has been an effective resource which can be employed in decision making process. Along with the rapidly increasing data sizes, data processing has become notably challenging. Therefore, the concept of sentiment analysis has emerged. Sentiment analysis is defined as the area of research that analyzes people's views, feelings, thoughts and attitudes about products, services, organizations, individuals, subjects and activities [1].

Sentiment analysis has divided into three as (i) document level, (ii) sentence level, and (iii) aspect-based sentiment analysis [1]. Document level sentiment analysis focuses on determine whether the whole document expresses positive or negative emotions. Sentence level sentiment analysis aims to determine whether each sentence in the document expresses positive or negative emotions. However, both of these two analyzes lack the ability to find what people like or dislike. Aspect-based sentiment analysis is applied for a more detailed analysis. Aspect-based sentiment analysis consists of two stages: target extraction and target classification [1]. In target extraction step, while evaluating the subjects in the text, the opinions of the different objectives are classified as positive, negative or neutral in target classification step. In this study, two different hybrid target extraction approaches utilizing LDA-based Web Search Based Feature Extraction and C-value techniques have been proposed.

Related Works

Frantzi et al. suggested domain-independent method for automatic extraction of multi-word terms by improving sensitivity on multi-word and nested terms [2]. Turney proposed a simple unsupervised learning algorithm to classify comments as advisable or not. Firstly, opinion phrases were extracted using rules. Then semantic orientation value of each phrase was calculated by subtracting the PMI value between the given word phrase and the words *"excellent"* and *"poor"*. Semantic orientation of whole document was calculated by collecting of each phrase's semantic orientation [3]. Siqueira and Barros presented the *WhatsMatter* system for domain independent target extraction. This system consists of frequent nouns identification, relevant nouns identification, feature indicators mapping and unrelated nouns removal. PMI technique was used to remove irrelevant nouns in the last stage [4]. Kama et al. proposed *WSBFE (Web Search Based Feature Extraction)*, domain-independent and unsupervised feature extraction method for feature-based sentiment analysis in Turkish texts. With this method, it is aimed to increase the performance of frequency-based feature extraction using search engine [5]. Ekinci et al. used n-gram model, finite state machine and the PMI to perform multi-word targets extraction from Turkish user comments. After the candidate targets were found with n-gram model, those whose frequency values were below a certain value were excluded from the candidate targets. Elimination process was carried out by using finite state machine and PMI criteria using Turkish grammar rules and multi-word targets were obtained [6].

Background

Latent Dirichlet Allocation (LDA) is a three-level hierarchical Bayesian model for the collection of discrete data. The LDA is based on the fact that the documents are represented as random mixtures over hidden topics and each topic is characterized by a distribution over words [7].

Pointwise Mutual Information (PMI) is used to measure the degree of statistical dependence between two terms [8]. It is calculated with the following equation (1):

$$PMI(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}. \quad (1)$$

P(word₁, word₂) represents the probability of coexistence of *word₁* and *word₂*, while *P(word₁)P(word₂)* refers to the probability that the two terms coexist when they are statistically independent.

C-value is a domain-independent method for aims to improve nested term subtraction by combining linguistic and statistical information with multi-word aspect recognition [2]. This method extracts candidate multi-word terms from the dataset by using various filtering rules, as follows (2):

$$C - value = \begin{cases} \log_2 |a| . f(a), & a \text{ is not nested term} \\ \log_2 |a| . \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)\right), & a \text{ is nested term} \end{cases} \quad (2)$$

where a refers candidate string, $|a|$ refers count of string, *f(a)* refers frequency of occurrence in dataset, T_a refers set of extracted longer candidate terms that contain *a*, $P(T_a)$ refers number of candidate T_a, $\sum_{b \in T_a} f(b)$ refers total frequency by which *a* appears longer strings.

Problem Definition and Proposed Method

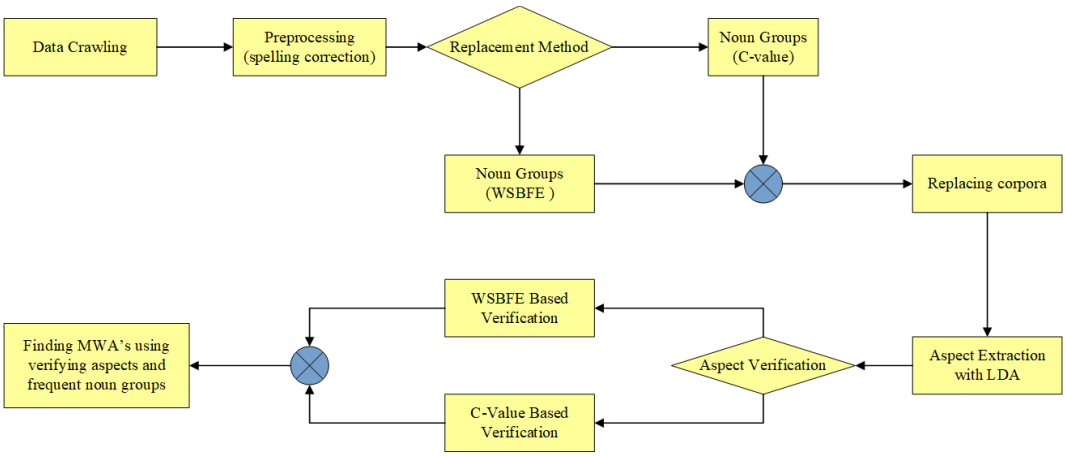


Figure 10. Flowchart of the Proposed Method

LDA is a topic modeling method utilizing bag of words (BoW) to uncover hidden topics within the dataset [7]. BoW increases the efficiency, however, causes loss in inter-word semantic relationship information. In this study, two different LDA-based aspect extraction methods are proposed which utilize WSBFE and C-value techniques for Turkish data to consider these relationships without the need of human annotation. In both methods, firstly, words inflectional suffixes are removed, then noun groups found with with WSBFE or C-value techniques are replaced in the dataset. For example, *"otel personel"* noun group is found, then *"otel"* and *"personel"* consecutive pairs in the dataset are converted into *"otel_personel"*. After dataset changed, aspects are founded with LDA. Aspects found with LDA are verified adopting WSBFE or C-value. Aspect list is expanded by using frequently approved noun groups and approved aspects in the dataset. The overall methodology is shown in Figure 10.

Experimental Results

In order to evaluate performance of proposed methods *TripAdvisor*[1] user reviews were automatically collected using the *jsoup*[2] library and stored in XML format. Dataset includes 1691 comments and 8233 sentences. The spelling errors in the data were removed with the help of *Zemberek-NLP*[3] and *Yandex.XML*[4].

The experimental results showed that the proposed methods were more successful than the classical LDA method, in terms of precision, recall and f-score (see Table 4).

LDA α, β, iteration and topic count values were determined as 0.1, 0.01, 50 and 20, respectively. WSBFE word frequency threshold was decided as 10, search engine query result count threshold was selected as 800. C-value single word threshold was selected as 10, multi-word threshold was chosen as 2.5. All of the input values were determined through manual observations.

Table 4. Experiment Results

Method	Precision (%)	Recall (%)	F-score (%)
LDA	62	36	45
C-Value	36	70	47
WSBFE	60	55	57
LDA enhanced with C-value	66	78	72
LDA enhanced with WSBFE	70	79	74

Conclusion

Aspect extraction is the first step in aspect-based sentiment analysis. In this study, two novel LDA-based methods are proposed. Experimental studies have shown that introduced methods provide relatively higher success in terms of reported performance metrics than standalone application of LDA does.

References

[1] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

[2] Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, *3*(2), 115-130.

[3] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

[4] Siqueira, H., & Barros, F. (2010, October). A feature extraction process for sentiment analysis of opinions on services. In *Proceedings of International Workshop on Web and Text Intelligence* (pp. 404-413).

[5] Kama, B., Ozturk, M., Karagoz, P., Toroslu, I. H., & Ozay, O. (2016, September). A web search enhanced feature extraction method for aspect-based sentiment analysis for Turkish informal texts. In *International Conference on Big Data Analytics and Knowledge Discovery* (pp. 225-238). Springer, Cham.

[6] Ekinci, E., Türkmen, H., & Omurca, S. İ. (2017). Multi-word Aspect Term Extraction Using Turkish User Reviews. *International Journal of Computer Engineering and Information Technology*, *9*(1), 15.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[8] Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

[1] https://www.tripadvisor.com.tr/

[2] https://jsoup.org/

[3] https://github.com/ahmetaa/zemberek-nlp

[4] https://tech.yandex.com.tr/xml/

Stock Exchange Prediction via Long Short-Term Memory Networks

B. Unal[1], C. H. Aladag[2]

[1]*İskenderun Technical University, Hatay, Turkey, bakiunal@gmail.com*
[2] *Hacettepe University, Ankara, Turkey, chaladag@gmail.com*

Abstract

Stock market prediction is a challenging task. Random walk hypothesis claims that stock market prices follow a random walk and history of a stock price cannot be used to predict its future movement. Efficient market hypothesis states that asset prices fully reflect all available information and no profit can be made from information-based trading. There are many studies in the literature which demonstrates that stock market prices follow random walk and unpredictable. However financial models in these studies cannot able to capture complex non-linear dependencies. Recently it is revealed that machine learning techniques are capable of identifying non-linear structures in stock market data. As a machine learning technique, artificial neural networks are successfully used in prediction of stock market prices. With deep learning artificial neural networks' prediction accuracy is increased. Long short-term memory networks are advanced deep learning architecture for sequence learning. Therefore, İstanbul stock market is forecasted by using long short-term memory networks in this study. For evaluating the performance of these networks, its forecasting results are compared to those obtained from Box-Jenkins' models.

Introduction

Artificial intelligence was born in the 1950s, when pioneers from the field of computer science started asking whether computers could be made to think [1]. Some of these pioneers are John McCarthy, Marvin Minsky, Allen Newell, and Herbert A. Simon. For a long time, many experts believed that human-level artificial intelligence could be achieved by writing enough amount of computer code. This approach is named as symbolic artificial intelligence (AI). It is proved that symbolic AI is suitable to solve well-defined, logical problems, such as playing chess. However, this approach failed to solve complex, fuzzy problems, such as image classification, speech recognition, and language translation. A new approach, machine learning (ML), arose to take symbolic AI's place.

ML was born with the questions such as: "Could a computer go beyond what we know how to order it to perform and learn on its own how to perform a specified task? Could a computer surprise us?" These questions led to a new programming paradigm. In classical programming paradigm, such as symbolic AI, human programmers write the rules as computer programs and input data is processed according to these rules, and outputs are obtained as answers. However, in machine learning humans input data as well as the answers expected from the input data to ML model and obtain the rules. After that the obtained rules can then be applied to new data and original answers can be generated.

Deep learning (DL) [2] is a subfield of machine learning. DL models are based on artificial neural networks (ANNs) and contain many layers of ANNs. In DL by each successive layer increasingly meaningful representations are generated. Number of layers in a DL model is called the depth of the model. The main idea behind the DL models is to calculate a loss score which measures the difference between true targets and the predictions of DL model and to use this loss score as a feedback signal to adjust the value of the parameters (weights) in the layers in a direction that will lower the loss score. This adjustment is done by an optimization algorithm.

Stock market prediction is an attempt to forecast the future value of a company stock or other financial instrument traded on an exchange. Stock market prediction methodologies fall into three categories: Fundamental analysis, technical analysis and technological methods. Sometimes these categories overlap. ML models fall in technological methods. The most prominent techniques involve the use of ANNs [3]. The most common form of ANN which is used for stock market prediction is the feed forward neural network (FFNN).

Long Short-Term Memory Networks (LSTMs) are a special kind of Recurrent Neural Networks (RNNs). Unlike standard FFNNs, LSTMs have feedback loops. LSTMs were introduced by Hochreiter and Schmidhuber [4]. LSTMs can remember information for long periods of time and this makes them suitable for sequence learning. LSTMs are successfully used in time series prediction [5,6].

Neural Networks

In this section, FFNN and LSTM are presented.

Feed Forward Neural Networks

FFNN is an ANN where connections between the nodes do not have any cycle. In FFNNs the information moves in only one direction, forward, from the input nodes, through the hidden nodes and to the output nodes. FFNNs were the first type of artificial neural network invented and are simpler than RNNs. Multi-Layer Perceptron (MLP) networks are a kind of FFNNs which contains multiple layers of computational units interconnected in a feed-forward way. In multilayer networks various learning techniques are used. Most popular learning technique is the back-propagation. In this technique a loss function which measures the difference between the network's answer and correct answer is calculated. Then this calculated difference fed back to the network. According to this information an optimization algorithm adjusts the weights of each connection.

Long Short-Term Memory Networks

LSTM is an RNN architecture. RNNs suffer from short-term memory. If a sequence is long enough RNNs cannot carry information from earlier time steps to later ones. So, if you are trying to predict a time series RNNs may exclude important information from the beginning. In longer sequences RNNs can forget thereby RNNs have short-term memory. LSTMs were created as the solution to short-term memory. LSTMs have gates which regulate the flow of information. These gates can learn which data in the sequence must be kept or thrown away. In this way LSTMs can pass important information down the long chain of sequences to make predictions. LSTMs are successfully used in robot control, time series prediction, speech recognition, rhythm learning, music composition, grammar learning, handwriting recognition, human action recognition, sign language translation, protein homology detection, predicting subcellular localization of proteins, time series anomaly detection, semantic parsing, and object co-segmentation.

Stock Market Prediction with LSTM

In this work we predict Istanbul Stock Exchange BIST 100 Index by using ARIMA, FFNN, and LSTM. Then we compared prediction accuracies by using root mean square error (RMSE) and mean absolute percentage error (MAPE). Our data cover BIST 100 Index daily closing prices. Our training set spans between 01.01.2002 and 31.08.2012. Our test set spans between 03.09.2012 and 02.09.2018. The training set contains 2756 observations and the test set contains 1420 observations.

First, we fit an ARIMA (0,1,1) model to training set for comparison purposes and we obtained 13844.2 RMSE value. And then we fit several FFNN models to training data. These FFNNs include one hidden layer with 8, 16, 32, 64, and 128 units and two hidden layers with 8-8, 16-16, 32-32, 64-64, and 128-128 units. In our models we utilized root mean square propagation (RMSPROP) and adaptive moment optimization (ADAM) algorithms. RMSE results for one hidden layer FFNN models are given in Table 1 and for two hidden layer FFNN models are given in Table 2.

Table 1. FFNN model with one hidden layer

FFNN	**8**	**16**	**32**	**64**	**128**
RMSE (RMSPROP)	83,6261	108,5901	89,7922	104,6501	94,7635
RMSE (ADAM)	101,5778	101,8051	102,6049	102,1194	101,6244

Table 2. FFNN model with two hidden layers

FFNN	**8×8**	**16×16**	**32×32**	**64×64**	**128×128**
RMSE (RMSPROP)	71,401	80,3982	76,7834	77,9592	74,1096
RMSE (ADAM)	101,6163	101,6163	101,6163	101,5853	101,6092

Table 3. LSTM model with one hidden layer

LSTM	**8**	**16**	**32**	**64**	**128**
RMSE (RMSPROP)	72,0395	66,5377	59,379	64,2658	77,4254
RMSE (ADAM)	101,6015	101,6854	101,8966	102,0132	101,744

Table 4. LSTM model with two hidden layers

LSTM	8×8	16×16	32×32	64×64	128×128
RMSE (RMSPROP)	56,395	56,7346	55,8313	75,474	105,3692
RMSE (ADAM)	101,6163	101,6162	101,6163	101,6163	101,6163

We also fit several LSTM networks to our training data which involve one hidden layer with 8, 16, 32, 64, and 128 units and two hidden layers with 8-8, 16-16, 32-32, 64-64, and 128-128 units. In these models we also used RMSPROP and ADAM optimization algorithms. RMSE results for one hidden layer LSTM models are given in Table 3 and for two hidden layer LSTM models are given in Table 4.

Conclusion

In this work we used LSTM networks to predict Borsa İstanbul stock market BIST 100 Index. For the prediction we also utilized FFNNs and an ARIMA model for comparison purposes. According to our findings RMSPROP learning algorithm produces better results for both FFNN and LSTM. Using two hidden layers gives better forecasting performance for both FFNN and LSTM. Using two hidden layers gives better forecasting performance for both FFNN and LSTM. As a result of the application, the best forecasting results are obtained from LSTM. Our results are summarized in the Table 5 below. It is clearly seen that LSTM gives the most accurate forecast in terms of both criteria RMSE and MAPE.

Table 5. Best RMSE and MAPE values for the models

Models	RMSE	MAPE
ARIMA(0,1,1)	13844.2	0.1146238
FFNN 8x8	71,401	0.0007852
LSTM 32x32	55,8313	0.0007176

References

[1] McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27 (4), 12-12.

[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[3] Aladag, C.H., Egrioglu, E., Editors, (2012) *Advances in time series forecasting*. Bentham Science Publishers Ltd.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9 (8), 1735-1780.

[5] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270 (2), 654-669.

[6] Gers, F. A., Eck, D., & Schmidhuber, J. (2002). Applying LSTM to time series predictable through time-window approaches. *In Neural Nets WIRN Vietri-01* (pp. 193-200). Springer, London.

Service Development with Service Oriented Architecture

E. Doğaç[1], R. Saraçoğlu[2]

[1]Van Yüzüncü Yıl University, Van, Turkey, emine.dogac@gmail.com
[2] Van Yüzüncü Yıl University, Van, Turkey, ridvansaracoglu@yyu.edu.tr

Abstract

Today, it is one of the most invested areas of organizations to ensure the interoperability of existing corporate applications and newly developed or purchased applications. Developed systems are designed to be shaped according to the needs of the businesses that use it. Classical systems are automation systems that operate from point to point and from other platforms in the enterprise without interaction. Unlike these, the service-oriented architecture (SOM) contains a structure that is ready to interact. In this study, a system has been designed in order to reduce the integration costs between the applications in the long term by using SOM approach, to shorten the business processes in the institution and to reduce the economic costs. In the project, workpieces were published as a method of service over the service layer. Thus, the platform independent main application is accessible to each workpiece. At the same time, these workpieces can come together and be reconfigured according to the request of the enterprise. Some of the advantages achieved by using SOM are: Providing improved work processes for services and actors, ensuring collaboration, utilizing from existing and reuse with savings, providing business flexibility by responding to changing business needs.

Introduction

Service Oriented Architecture (SOA) enables users to reduce integration costs between implementations in the long run, while shortening business processes in the organization and reducing economic costs. Most importantly, it prevents you from writing more codes and re-use of coding. Thanks to these features, which come from the working principle of virtualization, the information is delivered to the right users at the right time and speed. Thus, it also contributes to information reliability very well.

Analysis Of Field Literature

Channabasavaiah et al. highlighted that the managers of the organization suppress for the more efficient use of corporate resources and supported the idea that SOA is one of the biggest solutions to this problem [1].

Beklen stated that SOA is an approach with unique rules that are implemented in different fields such as finance, insurance and public, and requires detailed analysis and modelling in order to define services well. SOA has been successfully implemented for different sectors all over the world and has increased the development level of architecture. It is stated that the solution to application problems by layers abstraction and using the methods of platform-independent integration are the most important advantages of architecture. Large-scale software projects are developed by applying software engineering disciplines to solve different business problems in various sectors. However, each institution has used different technologies and approaches in its infrastructure. In some projects, applications were developed without using any architectural approach. Such situations have transformed applications, which have grown and become more complex over time, to the applications that cannot be integrated with each other, with difficult maintenance, repetitive work, or platform-dependent applications. SOA, which aims to solve such problems by isolating the application layers from each other, has been examined under subheadings such as architecture, service, and definition [2].

Dongsu et al. reported that SOA facilitates the collection, organization, and maintenance of institutional solutions in order to respond rapidly to changing needs. In an application using SOA, reusability is high. Every service created can be reused even in different projects. In case of a change in business requirements, this change only has an effect on the application side. In other words, the relevant change is made only in this service and the applications that use this service have revealed that applications are not affected by the change to a great extent [3].

In their study, Çopur argued that SOA has an architectural style for creating software applications that use the services offered in a network like a web. Applications in SOA are built based on services. It promotes a loose connection between services so that they can be reused. A service is an application of well-defined business functionality, and such services can then be consumed by customers in different applications or business processes.

Using SOA, businesses can attain significant effectiveness in development costs and quickly adapt to changing business conditions by reusing and restructuring existing services in the development of new applications. SOA has demonstrated better integration of enterprise IT resources, including pre-isolated application silos and legacy systems [4].

Kreger found out that it reduces application complexity by defining Web services and SOA service interfaces, as well as providing Web services, full-time integration, and interoperability of legacy applications [5].

Materials And Methods

The basic idea of SOA is service. Service is defined as a discrete unit of business functionality that can be done through a service agreement. The service agreement specifies the interaction between the service provider and the service user. These are service interface, interface document, service rules, quality of service and performance.

SOA services are described in standard definition language, allowing multiple platforms and protocols. They have a common interface. By supporting similar processes, they conduct the activities by communicating with each other [6].

In SOA, each function is defined as a service and All services operate autonomously [7].

The lifecycle of SOA is as follows; Model (identify, design, analyse) requirements), Merge (create, merge, test), Engage (people, processes, information integration) and Manage (manage applications and processes, monitor process, business unit, and information technology coordination).

Developed Services

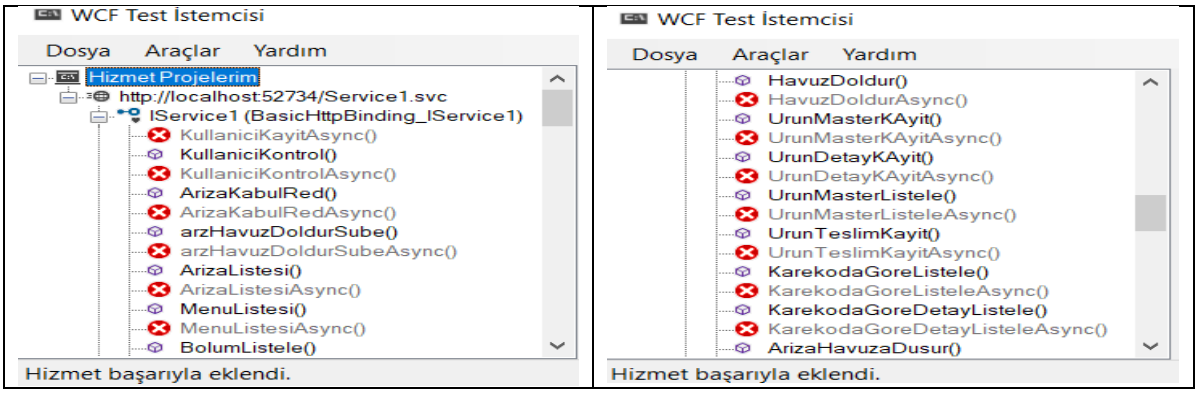


Figure 1. Service Service Methods

Figure 4.6 shows the service layer image of SOM. Each workpiece written in the main application was published as a method of service over the service layer. Therefore, the platform independent main application is accessible to every thread. At the same time, these threads came together to create new screens without re-creating rules and database layer.

Conclusions

It has been revealed that, through SOA, the same thread can be used in different projects without the need to redevelop the thread since the threads are written as the methods of service. It is also seen that the cost of code development due to reusable service methods may decrease. Moreover, it is evident that thanks to the structure of application developed by SOA, integration can be made with each project since it responds to every request of the end user with notation such as Simple Object Access Protocol (SOAP), json, etc. Hence, it provides platform independent development opportunity by being integrated to each platform. It was seen that a meaningful and consistent data source could be created by using a database and that the incremental code has been developed to give more control over the application and that independent components, objects, and services work together. It has been also seen that the services originating from the SOA structure have been brought together by information technology employees like lego, new screens and new applications have been developed.

References

[1] Channabasavaiah K., Holley K., Edward M. Tuggle, Jr. (2004). Migrating to a service-oriented architecture, IBM *Journal of Research Development*,3-22.
[2] Beklen A., 2009. *Kurumsal Servis Odaklı Mimari Kavramı, Teknolojisi Ve Tasarımı*. Maltepe University, Institute of Science and Technology, Istanbul.
[3] Dongsu, K. and Doo-Kweon, B., 2010, Bridging Software Product Lines and Service-Oriented Architectures for Service Identification using BPM and FM, *9th IEEE/ACIS International Conference on Computer and Information Science.* Yamagata, Japan.
[4] Çopur, D., 2011, *An Applıcatıon Of Servıce Orıented Archıtecture (Soa) Approach*. Baskent University, Institute of Science and Technology, Ankara.
[5] H. Kreger. (2003). Fulfilling the Web services promise. *Communications of the ACM*, 46(6), 29-ff.
[6] T. Francis, E. Herness, R. High Jr, J. Knutson, K. Rochat and C. Vignola. (2004). *Professional IBM WebSphere 5.0 Application Server*.
[7] M. P. Papazoglou and W.-J. Heuvel. (2007). Service oriented architectures: approaches, technologies and research issues. *The VLDB Journal—The International Journal on Very Large Data Bases,* 16(3), 389-415.

Robust Portfolio Optimization in Stock Exchange Market

B. Unal[1], C. H. Aladag[2], B. Senoglu[3]

[1]*İskenderun Technical University, Hatay, Turkey, bakiunal@gmail.com*
[2] *Hacettepe University, Ankara, Turkey, chaladag@gmail.com*
[3] *Ankara University, Ankara, Turkey, senoglu@science.ankara.edu.tr*

Abstract

Harry Markowitz took a first step for quantitative management of portfolios. Markowitz suggested an optimization problem which minimizes the risk of the portfolio for a given expected return level. The inputs of Markowitz's optimization problem are expected returns of assets and dependence structure between assets. These inputs must be estimated from historical data. Markowitz suggested that expected returns of the assets can be estimated from historical mean values of assets and dependence structure between assets can be estimated from traditional sample variance-covariance matrix. However, these estimates are very sensitive to outliers and outliers can lead to suboptimal results. In financial time series, outliers and fat tailed distributions are commonly encountered. To overcome distorting effects of outliers, robust statistical methods have been proposed. In this study, we applied robust estimators to Markowitz's portfolio optimization problem. We used the rolling window approach to compare the performance of original and robust optimization methods. We also investigated the effects of different rolling window sizes on the results.

1. Introduction

A major step in the direction of the quantitative management of portfolios was made by Harry Markowitz in his paper "Portfolio Selection" published in 1952 in the Journal of Finance [1]. Mean-variance analysis which is developed by Markowitz supplies a framework to construct and select portfolios, based on the expected performance of the investments and the risk appetite of the investor. Mean-variance analysis is the process of weighing risk against expected return. By looking at the expected return and risk of an asset, investors seek the lowest risk for a given expected return or seek the highest expected return for a given risk level.

Markowitz claims that investors should make an efficient choice based on trade-off between risk and expected return. Expected return of an asset is defined as the expected price change plus any additional income over the time horizon considered. Markowitz suggested that risk should be measured by the variance of returns. Markowitz asserted that for any given level of expected return, a rational investor would choose the portfolio with minimum variance from amongst the set of all possible portfolios.

To choose the optimal portfolio's weights, an investor must solve an optimization problem which minimizes the variance (risk) portfolio for a target level of expected return (μ_0). This problem is a quadratic optimization problem with equality constraints and given below.

$$\min_{w} w'\Sigma w$$

Subject to the constraints:

$$\mu_0 = w'\mu$$

$$w'\iota = 1 \quad \iota' = [1,1,\dots,1]$$

In the formulation above w is vector of portfolio weights of assets, Σ is variance-covariance matrix and μ is expected return vector of assets. To solve the optimization problem above expected returns of assets (μ) and variance-covariance matrix (Σ) must be estimated. However, these estimates are subject to estimation error. Inferior estimates, such as sample means and sample covariance matrices may lead to poor performance. Mean-variance optimization problem is very sensitive to the accuracy of input estimators.

Normality assumption quite often does not hold for stock market return data. When the distributional assumption is not met, the estimators lose their desirable properties. In fact, the arithmetic mean is sensitive to extreme observations, such that the estimate does not reflect the bulk of the data well. On a similar note, the dependence between two random variables can be highly distorted by a single outlying data pair.

2. Estimators to Markowitz's portfolio optimization

There are different ways to address the estimation error issue in Markowitz's optimization problem. One solution is to use robust estimators instead of classical estimators. Robust estimators [2] are less sensitive to outliers, and other sampling errors. In this work we utilized two robust estimators: Minimum Volume Ellipsoid (MVE) [3] and Minimum Covariance Determinant (MCD) [4]. In MVE it is considered any ellipsoid containing half of the data. The basic idea is to search among all such ellipsoids for the one having the smallest volume. Once this subset is found, the mean and covariance matrix of the corresponding points are taken as the estimated measure of location and scatter, respectively. In MCD half of data that has the smallest generalized variance is searched. The MCD estimator searches for the half of the data that is most tightly clustered together. The generalized variance of a p-dimensional random vector variable X is defined as the determinant of its variance-covariance matrix.

3. The Implementation

In the implementation part of the study, robust estimators (MVE and MCD) and traditional (classical) estimators are used in Markowitz's portfolio optimization problem. We perform a back-testing simulation with different estimators and for various window sizes. Different window sizes from 25 to 260 are used. All portfolio optimization methods are applied to the data for 24 different window sizes. Hence, the data is analyzed for 72 times totally. We assume that investors choose the global minimum variance portfolio.

Our data spans daily closing prices between 1/1/2014 and 6/27/2018 for following ten stocks: DOHOL, EREGL, ISCTR, KOZAA, PETKM, SAHOL, SODA, TCELL, TOASO, and YKBNK. In Figure 1 means of returns with different estimators and for different window sizes are given. In Figure 2 standard deviations of returns with different estimators and for different window sizes are shown. We also utilized hypothesis tests to compare return means for different estimators. In this regard we used t-tests and trimmed means tests. Obtained p-values are shown in Table 1.

Table 1. Hypothesis tests for comparison MVE, MCD and Classical estimators

	T-test p values			**Robust p values**		
Window Size	Classic-MCD-p	Classic-MVE-p	MCD-MVE-p	Classic-MCD-p	Classic-MVE-p	MCD-MVE-p
25	0.8425	0.6994	0.4695	0.7253	0.6319	0.3020
30	0.5937	0.9576	0.5856	0.7037	0.2004	0.2667
40	0.7602	0.9991	0.7334	0.9182	0.6582	0.5440
50	0.6123	0.6792	0.2403	0.6701	0.7716	0.8605
55	0.5205	0.2783	0.5484	0.3395	0.0981	0.3864
60	0.2532	0.8194	0.2369	0.3728	0.6357	0.5752
65	0.1610	0.7334	0.1722	0.3525	0.8084	0.1223
70	0.3622	0.7890	0.4490	0.4967	0.9780	0.3566
75	0.3007	0.9941	0.1287	0.3460	0.9131	0.1484
80	0.7201	0.2525	0.2612	0.8054	0.5333	0.6001
85	0.0982	0.2308	0.4524	0.2796	0.2293	0.8705
90	0.2059	0.3450	0.5742	0.1871	0.7439	0.1329
100	0.3454	0.3349	0.9953	0.8235	0.7577	0.8989
120	0.1478	0.1880	0.8157	0.7929	0.4474	0.1105
130	0.1728	0.5834	0.1036	0.4502	0.6292	0.6069
140	0.3265	0.1713	0.4780	0.6871	0.3869	0.4200
160	0.1778	0.0721	0.5590	0.6076	0.2184	0.2760
180	0.1225	0.3968	0.1475	0.2806	0.1992	0.7377
200	0.2301	0.2264	0.8421	0.3517	0.4340	0.8527
220	0.3928	0.4627	0.7411	0.7132	0.5671	0.0870
230	0.8025	0.5481	0.4586	0.9944	0.7425	0.5302
240	0.6864	0.7520	0.8284	0.8616	0.6856	0.2339
250	0.6757	0.3721	0.4083	0.9588	0.7224	0.4458
260	0.4880	0.7597	0.4224	0.7712	0.7090	0.8832

Conclusion

According to Figure 1, it is seen that classical estimator has lower mean return values for almost all window sizes. From Figure 2 we see that when window sizes increase standard deviation of returns decrease. Standard deviations of classical method are lower so classical method generates more stable results. When the window size is greater than 65-70 standard deviation converges. Since small window sizes have lower return means and higher standard deviations window sizes between 50 and 100 can be used. We compared the means of returns for different estimators with t-tests and find no statistically significant difference among estimators. We also used robust hypothesis test (trimmed means) to compare means of returns for different estimators and again find no statistically significant difference among estimators.

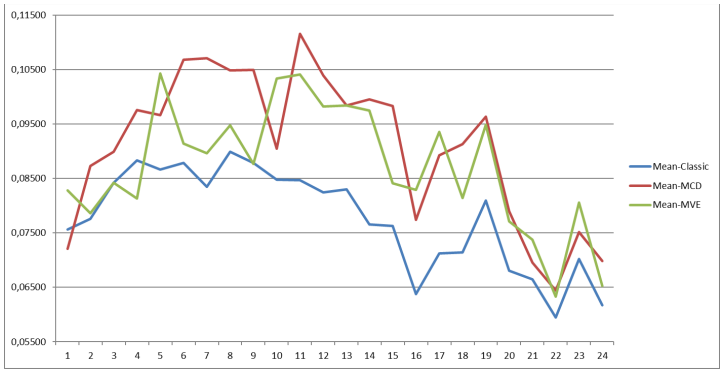


Figure 1. Means of returns for different window sizes

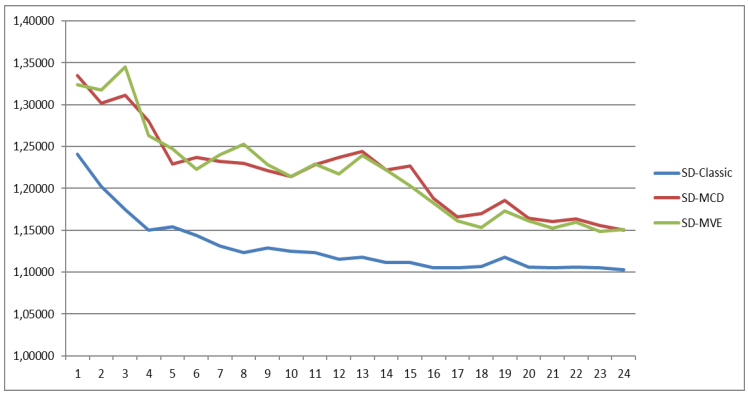


Figure 2. Standard deviations of returns for different window sizes

References

[1] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77-91.
[2] Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.
[3] Rousseeuw P. (1985) Multivariate estimation with high breakdown point In Mathematical Statistics and Applications (ed. Grossmann W., Pflug G., Vincze I., and Wertz W.) vol. B Reidel Publishing Dordrecht pp. 283–297.
[4] Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection* (Vol. 1). New York: Wiley.

Support Vector Machine Algorithm for Predicting the Bus Arrival Time in Public Transportation

F. Serin[1], S. Mete[1], M. Gul[1], E. Celik[1]

Munzur University, Tunceli, Turkey fserin@munzur.edu.tr, suleymanmete@munzur.edu.tr, muhammmetgul@munzur.edu.tr, erkancelik@munzur.edu.tr

Abstract

The long waiting times in bus stations directly affects the passenger satisfaction level in public transportation. The passengers want to know the real time information about transit buses such as the current location, arrival time of buses at the bus stops. Therefore, the prediction of bus arrival time determines the accuracy of travel time that presents high level passenger satisfaction. In this paper, we presented a supervised learning algorithm as support vector machine (SVM) for prediction the bus arrival time. It is based on kernel functions and it is considered as a nonparametric technique. An SVM Model contains points that separated by a clear gap in space. These points represent example data and the gap (margin width) between them should be as wide as possible. A case study is presented for a bus line in İstanbul public transportation systems. The mean absolute error (MAE), the mean absolute percentage error (MAPE), mean squared error (MSE), the root mean square error (RMSE), the residual sum of squares (RSS) are calculated for evaluating the accuracy of the proposed algorithm.

Introduction

The passengers generally prefer to use public transportation in place of their private car due excessive and unreliable travel times [1]. Therefore, the passengers of public transportation expect short waiting times at bus stops. The prediction of bus arrival time is a vital problem to know the public transportation system information technology, and it can affirmatively affect for development of public transportations system. The bus arrival time information can reduce the passenger waiting time, make passenger understandably arrange their trip plans [2]. For this reason, it is very important to give exact information about the time of bus arrival to passengers on public transportation.

SVM models are also applied to predict bus travel time that is a specific learning algorithm improved based on learning theory. SVM is one of the applied methods to the bus arrival time prediction [3]. Yang et al. [2] developed SVM with genetic algorithm to forecast time of bus arrival. Bin et al. [3] examined the applicability and feasibility of SVM to predict bus travel time. In this paper, we applied Support Vector Machine Regression (SVMR) for predicting bus arrival time. The model has been evaluated for bus arrival time prediction at bus stop by real data in Istanbul, Turkey. In this case, one of the most crowded and longest line (500T: Tuzla Şifa Mah.-Cevizlibağ) is selected for evaluating the developed prediction model and it consists so many bus routes and bus stop with high demand on bridge crossing every day. Five different performance measures, which are MAE, MAPE, MSE, RMSE, RSS are considered to evaluate the results of the prediction for bus route 500T of Istanbul.

Support Vector Machine Regression

Support Vector Machine is a supervised learning model which is developed by Vapnik [5]. Because the method is based on kernel functions, it is considered as a nonparametric technique. An SVMR contains points that separated by a clear gap in space. These points represent example data and the gap (margin width) between them should be as wide as possible. The mathematical formulation of the SVMR is given as

Goal: minimize $\frac{1}{2}||w||^2$ (1)

Constraints: $y_i - wx_i - b \leq \varepsilon \; and \; wx_i - y_i + b \leq \varepsilon$ (2)

SVMR provides a model that depends only on a subset of the training data. Because data is close to the model prediction is ignored by the cost function for building model. Then the solution is presented by means of this small subset of training data. It has some advantages as available to work in the high dimensional feature space instead of linear functions and having a high generalization capability with high prediction accuracy. On the other hand, the disadvantage is to high algorithmic complexity and extensive memory requirements in large scale tasks

Case study

The SVMR has been applied for bus arrival time prediction at bus stop by real data in Istanbul, Turkey. Istanbul is crowded city and it has more than 15 million population and about 78% of the Istanbulers are commuted via road transportation. In addition, %27.26 of Istanbulers directly used Bus, Metrobus or Private Buses. About 13 million passengers per daily uses public transportation in İstanbul [6]. In this case, one of the most crowded and longest line (500T: Tuzla Şifa Mah.-Cevizlibağ) is chosen for evaluating the developed prediction models. This line is selected because it consists so many bus routes and bus stop with high demand on bridge crossing every day. This line starts from Europe and ends in Asia and consists of a total of 76 bus stations (Figure 1). There is 20 different lines in this bus station. The length of the line is approximately 73.6 km. The main aim of the study is to predict the arrival time for 163 segments of T500 bus line applying SVMR.

Five different performance measures, which are MAE, MAPE, MSE, RMSE, RSS are considered to evaluate the results of the prediction for bus route 500T of Istanbul.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \quad (3)$$

$$MAPE = \left(\frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t}\right) * 100 \quad (4)$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2} \quad (6)$$

$$RSS = \sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|^2 \quad (7)$$

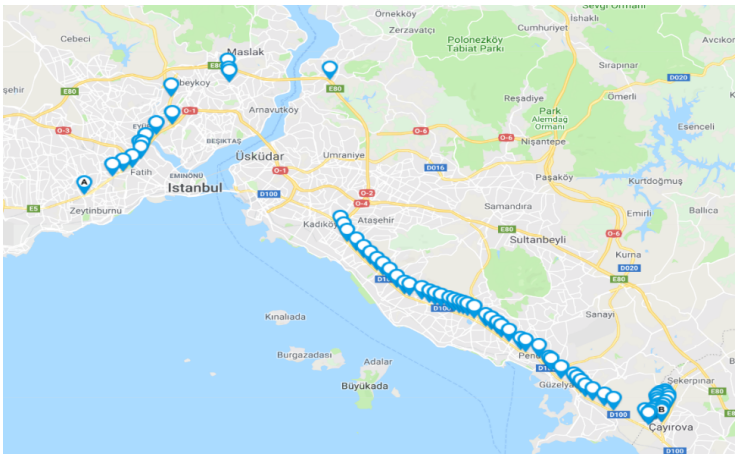


Figure 1. 500T: Cevizlibağ Tuzla-Şifa Mah. Route

The output data obtained from this study contains the following information: (1) route id and segment number, (2) departure and arrival station id (3) sample size and signal size, and (4) five performance measures as MSE, RMSE, MAE, MAPE, RSS.

Table 1. The results of the SVMR

Performance Measures	**Results**
MAE	0.2690
MSE	0.2005
RMSE	0.3201
RSS	0.9164
MAPE	21.6123

The result of the SVMR for five performance measures is presented in Table 1. MAE, MSE, RMSE and RSS values in model are 0.2690, 0.2005, 0.320 and 0.9164). The other performance measure is MAPE with a value of 21.6123. In the SVMR, we did not apply any normalization procedure. If any normalization procedure as min-max normalization, principal component analysis or 0-1 normalization is applied, the performance measure will be also analysed.

Conclusion

This paper examined the prediction of bus arrival time at bus stop on route 500T of Istanbul using SVMR. The five performance measures as MSE, RMSE, MAE, MAPE, and RSS are used for evaluating the results of the prediction for bus route 500T of Istanbul. The prediction model used in the study can be implemented to any public transportation. For future studies, we will apply machine learning algorithms for

References

[1]. Celik, E., Bilisik, O. N., Erdogan, M., Gumus, A. T., & Baracli, H. (2013). An integrated novel interval type-2 fuzzy MCDM method to improve customer satisfaction in public transportation for Istanbul. *Transportation Research Part E: Logistics and Transportation Review*, 58, 28-51.

[2]. Yang, M., Chen, C., Wang, L., Yan, X., & Zhou, L. (2016). Bus arrival time prediction using support vector machine with genetic algorithm. *Neural Network World*, 26(3), 205-217.

[3]. Yu, B., Lam, W. H., & Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, 19(6), 1157-1170.

[4]. Bin, Y., Zhongzhen, Y., & Baozhen, Y. (2006). Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems*, 10(4), 151-158.

[5]. Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media, New York.

[6]. IETT, (2015). Public Transportation in Istanbul, https://www.iett.istanbul/en/

Technology Integration Process of a Math Teacher

Ö. Özgüner[1], Ç. Güler[2]

[1]University of Van Yuzuncu Yil , Van, Turkey, ozgeozguner1992@gmail.com
[2]University of Van Yuzuncu Yil , Van, Turkey, cetin@yyu.edu.tr

Abstract

Changes in learning approaches are also reflected in the activities carried out during the education process. It is a question of whether the innovations provided by technology will be effective in education. Regardless of the individual differences in mathematics, it is seen that all students are educated as mathematics literate. In this study, one of the most challenging subjects of mathematics teachers learning in Van Turgut Reis Secondary School and students in the 8th grade was developed by using computer technology. The process of using this developed material in the classroom is examined. In this study; For the Geometry of Transformation Geometry in the 8th grade mathematics course, the development of course material with the GeoGebra program and the changes of the teacher's integration of this material on the students' attitude towards the course were examined. Case study pattern was used in the study and the research data were; Interviews with students consist of researcher and teacher diaries as well as course materials developed by the teacher. The data were analyzed by content analysis method. According to the findings of the analysis, it can be said that the materials used in the course increase students' attitude and participation in the course. In addition, it can be stated that there will be changes in teachers 'understanding of material development with the spread of online environments, producing materials that will attract students' attention and attention, increasing the participation and success of the course, as well as creating more effective learning environments by sharing them over the internet. Therefore, it can be argued that teachers' use of online environments such as GeoGebra will not only enable them to develop materials that will contribute to the course, but will also have a positive impact on the effective participation and academic success of their students.

Introduction

Technology; It is seen by many educators, teachers and researchers as indicators of quality in education and it is observed that the use of technology is increasing day by day in schools. Teachers should be able to use technological tools and materials effectively in order to reach individuals who use / access information. In the literature studies, there is no standardized definition of technology integration. Although there is no clear definition of technology integration in primary schools, teachers can be considered to use all kinds of technology to increase the success of students in the classroom [1]

When the literature was searched, the effect of GeoGebra on student achievement was examined and as a result of comparisons between the tests and groups, it was found that GeoGebra had a positive effect on students' learning and achievement [2]. The prospective mathematics teachers think that GeoGebra software will contribute positively to students' learning mathematics and they want to use dynamic software.

It was determined that teachers found GeoGebra to be preferable and usable in real classroom environments with its prominent features such as being free, using in Turkish, user-friendly interface, ease of use and the potential to dynamically reveal the relationships between geometry and algebra [3]. In line with the findings, the material development process was determined by taking into account the individual differences and learning speed of the mathematics teacher. In this way, the necessary training was provided through the GeoGebra software in order to develop the materials suitable for the acquisition of “Transformation Geometry 8. of the 8th grade mathematics course by the teacher and the teacher was designed to design the material. Does the use of materials affect the attitude and participation of the students? was searched for.

Purpose

The aim of this study is to develop material for the subject of “Transformation Geometry 8. in the 8th Grade Mathematics course and to determine the changes of the teacher's integration of this material on the student's attitude to the class. Studies show that pre-service teachers' opinions about using computers in mathematics education affect the use of computers in mathematics education. For this reason, it was aimed to determine the quality of the materials that the mathematics teacher created with GeoGebra software and to reveal the perspectives of mathematics teaching using dynamic mathematics software. For this purpose, the following questions were sought:

• What are the views of the students when the mathematics lesson is compared with the material prepared in GeoGebra with the traditional narrative technique?
• What are the views of the mathematics course on the material prepared with the GeoGebra program?
• What are the students' views on what subjects of this mathematics lesson they would like to see?
• What are the views of the teacher about the material developed using the GeoGebra program?
• What are the students' attitudes and reactions to the lesson?

Methods

Case study design was used in this study. Factors related to a situation (environment, individuals, events, processes, etc.) are investigated with a holistic approach and they focus on how they affect the situation and how they are affected [4].

Study group

The mathematics teacher who participated in the study to develop and integrate the material was selected with criterion sampling method. The research was carried out with 20 students studying in 8th grade in Turgut Reis Primary School in 2016-2017 spring term.

Data Collection Tool

The research data consists of interviews with students, researcher and teacher diaries as well as materials developed by the teacher. Each stage of the process is included in the journal of the researcher and the teacher.

Results

The teaching of the program, material production, teachers' and researchers' diaries and students' opinions were analyzed and the data obtained were classified according to the purpose of the research.

Teacher's Attitudes Towards Process

It was observed that mathematics teacher was very willing and positive behaviors towards teaching the program and himself during the material production process. The diary of November 21, 2016, in which the teacher is taught the geometric animation preparation environment (GeoGebra), contains the following notes;

Researcher notes:
"I told the math teacher about the GeoGebra program, he listened with interest, even though he was a little uneasy".
"He often asked questions to understand the general logic of the program."
"My students were weak in understanding transformation geometry. As a result of the materials I will develop with this program, they will be able to learn better." It was observed that the notes in the diaries of the researcher and the teacher on 21 November 2016 were parallel to each other.

Student Opinions about the Process

A: Researcher S: Student
A: What technique would you prefer when you compare the traditional lecture technique with the mathematics course with the material prepared in GeoGebra?
S1:" I prefer both. There's nothing for me. "
A: Have you ever seen animation on the subject in mathematics classes?
S2: "I've seen animations like this from my computer at home, but it's about science."
A: What are your views on the material prepared with the GeoGebra program? Do you think it would be useful to use such animations in mathematics?
S5: "It's almost the same as watching a video, but it's definitely useful to use this kind of stuff in class."
A: Which subjects of mathematics are you happy to see these animations?
S6: "We worked on the coordinate system last year and didn't understand much. It makes me happy to have these animations."

As it is understood from the students' opinions, they stated that animation is effective in their learning and should be used in more subjects. According to these findings, the teacher has a positive attitude towards the learning activity and follows this program with interest. It was concluded that the use of this material had a positive effect on the attitudes and participation of the students. Thus, the aim of the research was achieved and the process was successfully completed.

Interpretation, Discussion and Conclusion

In the study, the mathematics teacher was trained in material development with the GeoGebra program. In this training process, importance was given to the achievements of the subject and material production process was realized in accordance with the objectives determined by the teacher. GeoGebra teaching and material development process was conducted at the appropriate time and place for the teacher and the teacher was fully involved in the lessons. In this material development process, which was supported by the findings of the study, positive opinions were received from the students with the ability of the teacher to control their own learning and to produce the material and apply it in the classroom. It can be said that this online material leads to an increase in student attitude and achievement. In this study, mathematics teachers who want to stand out from traditional teaching methods can apply the GeoGebra software, whose effects are examined, in their courses. They can learn the GeoGebra software easily by using the advantage of having Turkish menu. They should know how to properly integrate such software into their courses. In order to achieve this, it is very important for teachers to be guided and to carry out in-service training activities for teachers. With such studies, perceptions that use computer-aided software in the course take time are prevented.

References

[1] Hew, K. F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. *Education Technology Research & Develeopment*, 55, 223–252.
[2] İçel, R. (2011). Bilgisayar Destekli Öğretimin Matematik Başarısına Etkisi: GeoGebra Örneği. Selçuk Üniversitesi Yüksek Lisans Tezi.
[3] Kabaca, T., Aktümen, M., Aksoy, Y., &Bulut, M. (2010). Matematik Öğretmenlerinin Avrasya GeoGebra Toplantısı Kapsamında Dinamik Matematik Yazılımı GeoGebra ile Tanıştırılması ve GeoGebra Hakkındaki Görüşleri. *Turkish Journal of Computer and Mathematics Education,* 1(2), 148-165
[4] Yıldırım, Ali, ve Şimşek, Hasan. *Sosyal Bilimlerde Nitel Araştırma Yöntemleri*. Seçkin Yayıncılık, 2013.
[5] Tatar, E., Akkaya, A. & Kağızmanlı, T.B. (2011). İlköğretim Matematik Öğretmeni Adaylarının GeoGebra ile Oluşturdukları Materyallerin ve Dinamik Matematik Yazılımı Hakkındaki Görüşlerinin Analizi. *Turkish Journal of Computer and Mathematics Education*, 2(3), 181-197.

Prospective Teachers' Opinions of Computer and Instructional Technologies for STEM Education

Ö. Özgüner[1], Ç. Güler[2], H. Çavuş[3]

[1]University of Van Yuzuncu Yil , Van, Turkey, ozgeozguner1992@gmail.com
[2]University of Van Yuzuncu Yil , Van, Turkey, cetin@yyu.edu.tr
[3]University of Van Yuzuncu Yil , Van, Turkey, hcavus@yyu.edu.tr

Abstract

Rapid development of technology; it is seen that traditional learning methods and environments are insufficient to meet the expectations of digital natives born and growing in a digital age. Therefore, individuals need to acquire 21st century skills. In our country, studies on STEM education, which is defined as giving the disciplines in an integrated way rather than separately, are focused on. In the literature review, it was observed that the studies related to STEM education generally focused on prospective science teachers. A study was conducted on the views and attitudes of the prospective teachers of Computer and Instructional Technologies towards STEM education as STEM education has a direct relationship with the discipline of "Technology". The aim of this study was to collect the opinions of 30 computer and instructional technology prospective teachers who are studying in 4th grade at Yuzuncu Yil University by using semi-structured interview method. The collected data were analyzed by content analysis, codes and themes were created and the findings were interpreted. As a result of the analysis, it has been determined that prospective teachers of Computer and Instructional Technologies do not have knowledge about STEM education and that they can associate with their departments due to "technology" discipline. They stated that science, technology, engineering and mathematics cannot have any relationship except that they are only numerical courses. It was found that they do not have enough information about STEM education and most of them heard this concept for the first time. In line with the data obtained, it is important to disseminate the content and training of STEM education and to bring the prospective teachers to a level that can ensure their competence in this subject. Candidates emphasize that they will work more efficiently with their students after starting to teach in line with the increase in their competence in this subject.

Introduction

In recent years; A new technological phase has been introduced thanks to robots that have been used in both education and production technology platforms. It is foreseen that robotic systems will be inevitable in the programming and execution of the education and training processes of this generation (digital natives). It is stated that robots contribute to science and engineering education and become applicable in different education levels with technology transfers [1].

This technology, called 'Robotics otik, which is integrated with many disciplines, is required to be a part of interdisciplinary STEM or STEAM education obtained by the participation of art education in some countries, especially science, technology, engineering and mathematics education [2].

Purpose

The aim of this research is to determine the attitudes and views of pre-service teachers of Computer Education and Instructional Technology Department in terms of STEM education and applications, engineering, technology and the relationship between these disciplines in the inde Technology Applications in Educational Environments 'course of Computer Education and Instructional Technologies Department. Accordingly, answers to the following questions will be sought:

1. What are the opinions of prospective teachers about STEM before and after the experimental procedure?
2. What are the contributions of STEM training and practices to prospective teachers?
3. What are the attitudes of teacher candidates towards the relationship between STEM disciplines as a result of STEM training and practices?

Methods

Study group

The study group of the study consisted of 30 students studying in the 4th grade of the Department of Computer Education and Instructional Technology, Faculty of Education, Van Yüzüncü Yıl University in 2018-2019 spring term. In the research, the names of the preservice teachers were not explained in accordance with the ethical principle. Pre-service teachers were coded as ÖA1, ÖA2, ÖA3… ..ÖA30.

Data analysis process

The semi-structured interview form, which will be prepared to examine in depth the effects of STEM training and practices on the prospective teachers' thoughts, will be applied to the research group before and after the STEM education and applications. At the end of the application, qualitative data will be analyzed using content analysis method and the answers will be examined and codes will be determined to simplify the raw data.

Results

This section contains the findings of the research questions. In order to support the findings, the answers of the preservice teachers were conveyed exactly.

Table 1. *'What do you think of STEM?'* Answers to the question before the application

Themes	Codes	*f*	Participants
Conceptually STEM	First time hearing	22	ÖA1, ÖA2, ÖA3, ÖA4, ÖA5, ÖA6, ÖA7, ÖA8, ÖA9, ÖA10, ÖA11, ÖA12, ÖA15, ÖA16, ÖA20, ÖA21, ÖA22, ÖA23, ÖA24, ÖA25, ÖA27, ÖA28
STEM in terms of definition	Interdisciplinary relationship	8	ÖA13, ÖA14, ÖA17, ÖA18, ÖA19, ÖA26, ÖA29, ÖA30

Table 2. *'What is the relationship between science, technology, engineering and mathematics?'* Answers to the question given before the application course:

Thenes	Codes	*f*	Participants
Relationships	Numerical courses	15	ÖA1, ÖA2, ÖA5, ÖA10, ÖA12, ÖA16, ÖA20, ÖA21, ÖA23, ÖA24, ÖA26, ÖA27, ÖA28, ÖA29, ÖA30
	Integration of other courses with technology	11	ÖA2, ÖA4, ÖA5, ÖA6, ÖA8, ÖA9, ÖA11, ÖA19, ÖA20, ÖA22, ÖA29
	Blending of disciplines	5	ÖA13, ÖA14, ÖA15, ÖA16, ÖA20
	Production oriented	2	ÖA18, ÖA25
	Developing collaborative and creative thinking	1	ÖA17
	Unable to connect	1	ÖA7

Table 3. *'Could you please tell us what you understand from STEM?'* The answers to the question after the application:

Themes	Codes	*f*	Participants
Definitions	Interdisciplinary approach	18	ÖA1, ÖA2, ÖA3, ÖA5, ÖA6, ÖA7, ÖA8, ÖA9, ÖA13, ÖA14, ÖA16, ÖA18, ÖA19, ÖA20, ÖA21, ÖA22, ÖA26, ÖA29, ÖA30
	Project based learning	11	ÖA1, ÖA2, ÖA4, ÖA5, ÖA6, ÖA7, ÖA9, ÖA14, ÖA18, ÖA23, ÖA25
	21st century skills development	8	ÖA2, ÖA8, ÖA11, ÖA13, ÖA22, ÖA26, ÖA27, ÖA30
	Collaboration	5	ÖA2, ÖA4, ÖA15, ÖA20, ÖA28
	Finding the solution to the problem	5	ÖA5, ÖA6, ÖA14, ÖA28, ÖA30
	Meaningful learning	5	ÖA4, ÖA5, ÖA12, ÖA13, ÖA17, ÖA30
	Increasing creativity and imagination	3	ÖA10, ÖA24,ÖA27
	Student-centered approach	1	ÖA12

Table 4. *'What are the benefits of STEM applications during a semester?'* answers to the question after the application

Themes	Codes	*f*	Participants
Benefits of STEM training	Internship and work as a teacher	15	ÖA1, ÖA2, ÖA3, ÖA4, ÖA7, ÖA8, ÖA9, ÖA12, ÖA14, ÖA16, ÖA17, ÖA18, ÖA19, ÖA21, ÖA27
	Integrating Arduino into STEM training	11	ÖA2, ÖA4, ÖA5, ÖA10, ÖA13, ÖA14, ÖA16, ÖA23, ÖA24, ÖA27, ÖA30
	Gaining an interdisciplinary perspective	8	ÖA2, ÖA3, ÖA5, ÖA8, ÖA13, ÖA14, ÖA15, ÖA29
	Add the student to the lesson more actively	7	ÖA1, ÖA6, ÖA7, ÖA8, ÖA14, ÖA28, ÖA30
	Self-development in technology and engineering	7	ÖA2, ÖA4, ÖA5, ÖA9, ÖA17, ÖA25, ÖA30
	Positive contribution to improving teaching environment	6	ÖA1, ÖA2, ÖA8, ÖA11, ÖA17, ÖA20
	21st century mastery of skills	5	ÖA10, ÖA14, ÖA16, ÖA26, ÖA29
	Guiding students and keeping themselves up to date	4	ÖA3, ÖA4, ÖA20, ÖA28
	Lack of innovation	3	ÖA12, ÖA19, ÖA21
	Production oriented	3	ÖA9, ÖA18, ÖA22
	Reduced prejudice to engineering	1	ÖA6

Interpretation, Discussion and Conclusion - V

In line with the first sub-problem of the study, pre-service teachers' opinions were evaluated before and after STEM education application process. As a result of this study, teacher candidates did not have any information about STEM training; problem solving skills, interest and curiosity, 21st century. life skills, high-level thinking skills, creativity and curiosity will develop the sense of guidance and guidance to the students after they are useful for their work. In addition, participants attended STEM training; They stated that science, technology, engineering and mathematics will be blended and that an interdisciplinary education will be given to students. The prospective teachers stated that it would be beneficial to include a course on STEM education and practices in the curriculum and it would be in the best interest of the student to give this education from a younger age. STEM education students 21st century. it is thought to be effective in gaining skills [3]. Since the gains in science, engineering and mathematics in the education process contribute to the technology learning, they think that providing disciplines under STEM education will have positive contributions to IT and Software courses. They emphasized that STEM education was limited in the interview with the prospective teachers and they lack the materials used in the education process, the insufficiency of the time, the crowded class size and the teaching environment is not suitable for STEM education and practices. Looking at the literature, Siev et al. (2015) stated that STEM activities and applications are costly, time-consuming and materially equipped. Therefore, it can be said that eliminating these limitations will have positive results in STEM education and approach.

References

[1] Mataric, M. J. (2004). Robotics Education for All Ages. Paper presented at the American Association for Artificial Intelligence Spring Symposium on Accessible, Hands-on AI and Robotics Education.
[2] Cameron, R. G. (2005). Mindstorms robolab: Developing science concepts during a problem based learning club(Unpublished master's thesis). Canada: The University of Toronto.
[3] Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. Technology and Engineering Teacher, 70(1), 30-35.

Drought Characterization of Van Lake Basin Using Standardized Precipitation Index (SPI)

G. M. Perihanoğlu[1], H. Karaman[2]

[1] *Van Yüzüncü Yıl University, Van,Turkey, gm.perihanoglu@yyu.edu.tr*
[2] *Istanbul Technical University, Istanbul, Turkey karamanhi@itu.edu.tr*

Abstract

Drought is the most dangerous of disasters that have not yet been fully understood and whose effects have not been adequately assessed, despite the fact that the drought is increasing in the world. Drought analysis is significant for prevention of drought disaster. There is a need for algorithms that can be used in conjunction with GIS and integrated into the geographic information system in order to monitor the danger of drought and foresee future prospects. SPI provides an important role for monitoring the drought hazard and using it with GIS to estimate future droughts.

In this study, a meteorological drought analysis of Van Basin was performed by using the Standardized Precipitation Index (SPI) with the data. Precipitation data were taken from Turkish State Meteorological Service. From 1975 to 2017, the same periodic data are obtained available from 8 meteorological stations for the Van Lake basin drought assessment. To study meteorological drought and classification drought mapping in this paper Standardized Precipitation Index (SPI) and spatial interpolation in GIS is used. To study the meteorological drought, records from 8 stations within the same period of 42 years (1975-2017) have been get. According to the SPI method, index values of drought severity were obtained at 12 and 24 months periods. In the second stage of the study, the temporal trends in these drought indices at different periods were determined by the Mann-Kendall (M-K) method. The statistical significance levels of the trend results obtained from the M-K trend test applied to the 12 and 24 month SPI drought indices were calculated and their geographical distribution was analyzed.

Introduction

Drought is one of the most serious, complicated and slow-growing natural disaster [1]. There are various indices and methodologies for monitoring drought by using in different parameters such as soil moisture, rainfall, vegetation index, and temperature [2]. The scientific research on drought in Turkey began with the 1940s. Drought events in Turkey in 1984, 1989-1991, 1999-2000, occurred in 2001-2004 and 2007-2008 as severe [3]. Due to its geographical location and structure, our country has very different climatic zones and microclimate areas.

This paper is about the characterization of drought and determination of trend trends of Lake Van Basin. The commonly used standard precipitation index is used to characterize drought, while commonly used non-parametric trend tests are used to test the significance of the drought severity and hence the significance of the trend in the magnitude of the effects of drought. SPI was applied to the data which get from each station. Maps of drought analysis were produced using interpolation techniques with Arcgis software for the different time scales. Time series drought maps of the Van province have been collected through a geographic information system (GIS) method using SPI drought indices.

Study Area and Data Used

Van Lake Basin is located between 37 ° 55'- 39 ° 24 'north latitude, 42 ° 05 - 44 ° 22' east longitude. In Figure 1, location of Lake Van basin is given. Moreover the map showing the settlements and meteorological stations in the Van Lake Basin is shown.

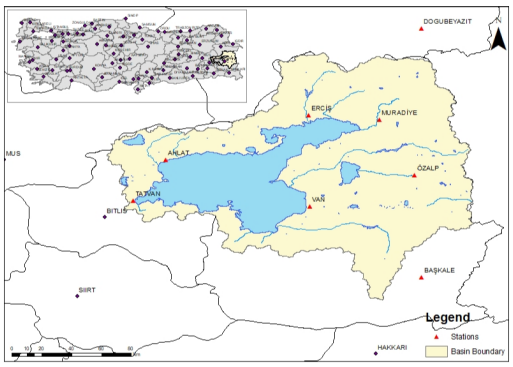


Figure 1. Location of the Van Basin and Stations.

The common time period for the analysis was 1975 to 2017 (42 years). Monthly precipitation datasets were checked for homogeneity.

Methods

Drought indices are very important tools for drought analyzing. Calculation of the SPI indice for any region or location is depend on the record long-term precipitation [4,5]. Mathematically, to calculate SPI value, is taken the difference of the precipitation from the mean for a particular time scale, this difference is divided by the standard deviation. The SPI values classifies according to the basic combination of the two classes: one $-1.0 \leq SPI \leq 0.0$ (light drought) and the other $0.0 < SPI \leq 1.0$ (light wet). The purpose of the trend analysis is to determine whether a series of observations of a random variable may decrease or increase over time [6]. Positive values indicate an increase in trend over time and negative values indicate a decrease in trend [7]. For spatial distribution of SPI values calculated for 8 stations across the entire study area, reverse distance weighting (IDW) interpolation was used.

Results and Discussions

The SPI values for the study area were produced between 1975 and 2017 on an annual and 24 months time scales basis. The total values obtained after examining the 12 and 24 month SPI values for each station and the most severe year and month of drought among these values are shown in Table 1.

Table 1. Peak intensity of drought values in SPI values for all stations.

	Observed Peak Intensity (1975-2017)					
	12 Month SPI			**24 Month SPI**		
Stations	**Value**	**Year**	**Month**	**Value**	**Year**	**Month**
Ahlat	-2.89	2014	5	-1.83	2000	9
Başkale	-2.77	2012	5	-1.58	2012	6
Doğubeyazıt	-2.88	2012	4	-2.06	2012	12
Erciş	-2.86	2001	2	-2.43	2013	12
Muradiye	-2.74	2001	5	-2.46	2000	12
Özalp	-2.98	2012	6	-1.97	2012	12
Tatvan	-2.64	2014	3	-2.25	2000	12
Van	-2.56	2001	2	-2.04	2013	12

The SPI values were calculated 12 and 24 months time scales to all stations in order to observe the severity long term effect of the drought. The Kendall's tau and p value were calculated. The state stage of the trend analyzes, the dry-humid periods within the station itself were determined and the common periods of these periods were determined. In Mann-Kendall analysis, there is no trend for H0 the hypothesis of no trend in the relevant series. In the 95% confidence interval, this hypothesis cannot be rejected if the significance level is higher than 5%. Mann-Kendall analyzes were examined in this context. As a result, while there was no trend in terms of increase or decrease in the stations throughout the years, there was a tendency to decrease in Ahlat and Muradiye stations. At Tatvan station, there was no trend in the SPI time series of 12 months, while there was a decrease in the SPI time series of 24 months. The drought characterization map of 1977, 2000 and 2017 using GIS is shown in Figure 2.

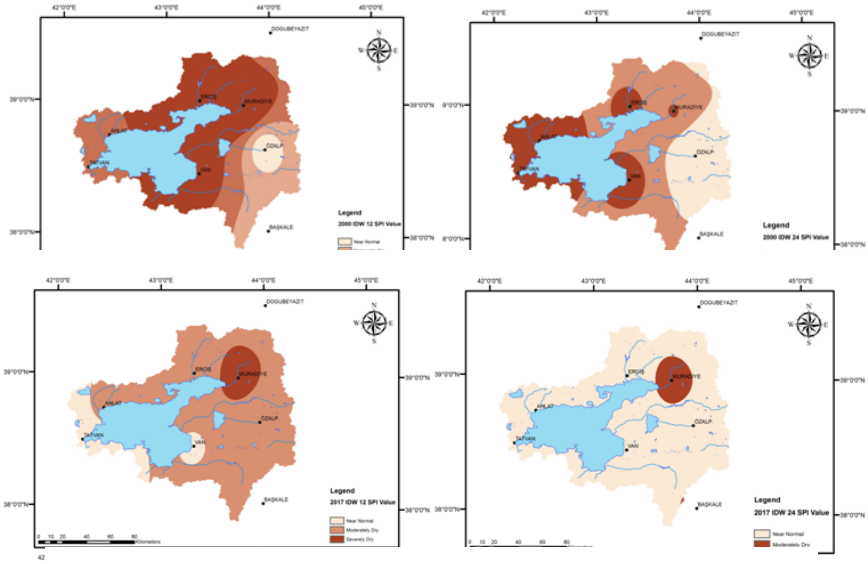


Figure 2. Spatial-temporal drought assessment

Extreme drought was experienced in Özalp station with the highest SPI value (-2.98) in June 2012. From fig. 2 ,SPI for 8 stations in Van Basin with the period of 1975-2017 was analyzed which indicate that in 2000 SPI dropped as low as -3.05 is related to Muradiye station in the north of the Basin. The results are mostly concentrated in the drought class, which is close to normal in drought. When the results are examined on visuals, it is seen that normal droughts are concentrated throughout the basin. When the 24-month period is compared to the 12-month period, it is seen that the drought approaches to normal in the 24-month period.

Conclusion

This study aimed to determine the possible drought trends in Van Basin by using Mann-Kendall tests. In the study, SPI values were calculated 12 and 24 month time scales for a total of 42 years of data from 1975 to 2017 for classifying severity of drought. Through inverse distance weighting (IDW) algorithm was applied to the whole study area for indicate the spatial distribution. As seen in the SPI trend analysis, it is clear that the stations generally show rainfall decreases in similar periods except for some periods. These reduction periods can be generalized as 1995-2001 and 2014 considering all the stations in and around the basin. After finding the severity classes of each station, severity maps were generated by inverse distance weighting method. When creating drought severity and risk maps, classes were separated four classes. These classes; normal and above, mild, moderate and severe arid classes. Considering all the drought severity classes in the Lake Van Basin, it is seen that there is a high probability of normal and above drought.

References

[1] Mathbout, S., Lopez, B., Martin- Vide, J., Bech, J., Rodrigo, F. (2018). Spatial and temporal analysis of drought variability at several time scales in Syria during 1961–2012, *Atmospheric Research,* 153-168

[2] McKee,T.B.,Doesken,N.J.and Kleist,J.(1993),The relationship of drought frequency and duration to time scales, *Eighth Conference on Applied Climatology*,179-184.

[3] Oğuztürk, G., Yıldız, O. (2014). Drought Analysis for Different Time Periods In The City of Kırıkkale, *International Journal of Engineering Research and Development*, 6 (2),1-7

[4] Wang, H., Pan, Y., Chen, Y., (2017). Comparison of three drought indices and their evolutionary characteristics in the arid region of northwestern China. *Atmos. Sci. Lett.* 18, 132–139.

[5] Bachmair, S., Stahl, K., Collins, K., Hannaford, J., Acreman, M., Svoboda, M., Knutson, C., Smith, K.H., Wall, N., Fuchs, B., Crossman, N.D., Overton, I.C., (2016). Drought indicators revisited: the need for a wider consideration of environment and society. *WIREs Water,* 3, 516–536.

[6] Rahmat, S.N., Jayasuriya, N., Bhuiyan, M. (2012). Trend analysis of drought using Standardised Precipitation Index (SPI) in Victoria, Australia, *34th Hydrology and Water Resources Symposium* (pp:441-448)

[7] Suryabhagavan, K.V. (2017). GIS Based climate variability and drought characterization in Ethiopia over three decades, *Weather and Climate Extremes,* 15, 11-23.

Mathematical Model of Flow Shop Scheduling Problems and Solution with Metaheuristic Methods

S. Kaya[1], A. Çelik[2], İ. H. Karaçizmeli[3], İ. B. Aydilek[4], A. Gümüşçü[5], M.E.Tenekeci[6]

[1]Harran University, Şanlıurfa, Turkey, serkankaya@harran.edu.tr
[2]Harran University, Şanlıurfa, Turkey, aysecelik@harran.edu.tr
[3]Harran University, Şanlıurfa, Turkey, hkaracizmeli@harran.edu.tr
[4]Harran University, Şanlıurfa, Turkey, berkanaydilek@harran.edu.tr
[5]Harran University, Şanlıurfa, Turkey, agumuscu@harran.edu.tr
[6]Harran University, Şanlıurfa, Turkey, etenekeci@harran.edu.tr

Abstract

Flow shop scheduling is the type of scheduling that occurs when n jobs are processed in the same order on m machine. In the beginning, the problem was to schedule n jobs on two machines. With the increase in the number of jobs and machinery, the problem has entered in the NP-Hard scope. In this study, mathematical model is presented for the solution of flow shop scheduling problems. Small sized problems are solved by using mathematical model and a model based on particle swarm optimization algorithm is proposed for the solution of medium and large sized problems. The results showed that particle swarm optimization algorithm yields effective results in solving such problems.

Introduction

One of the most commonly used types of scheduling is flow shop scheduling, which generally aims to minimize makespan. The problems that have different *m* machines and *n* jobs and each job is done in the same order are defined as flow shop scheduling problems (FSSP). Mathematical model is enough to reach optimal solution for small size problems. However, with the increase in the number of machines and jobs, the problem falls within the scope of NP-hard and it becomes very difficult to reach an optimal solution [1].

Ruiz et al. [2], Chen et al. [3] and Chen et al. [4] proposed genetic algorithms for FSSP, Liu and Liu [5] presented hybrid artificial bee colony algorithm, Fernandez-Viagas and Framinan [6] proposed NEH based heuristic algorithm, Liu et al. [7] improved NEH heuristic algorithm, Tayeb et al. [8] proposed a hybrid algorithm from genetic and artificial immune algorithms for FSSP. Wang et al. [9] proposed cuckoo search algorithm, Liu et al. [10] presented a hybrid algorithm of memetic and particle swarm optimization algorithms, Yang and Deb [11] presented a particle swarm optimization algorithm for FSSP. Abedinnia et al. [12] presented a NEH based local search algorithm for the solution of the same problem for total flow time. Taşgetiren et al. [13] presented the discrete bee colonies algorithm and Li and Yin [14] bee colonies algorithm for FSSP.

In this study, a proposal was made to minimize the objective of makespan in FSSP. While it is shown that the solution can be reached by using the mathematical model for small size problems, it has been shown that with the growth of the problem size, the mathematical model is inadequate and a value close to the target can be reached by meta heuristic methods. Particle Swarm Optimization (PSO) algorithm was used as the meta-heuristic method in the study. In the second part of the study, the mathematical model of FSSP, in the third part, PSO, in the fourth part experimental results, in the fifth part conclusion are given.

Mathematical Model

The mathematical model which aims to minimize the makepan value of the FSSP is as follows.

Indices
i: jobs *(i= 1, 2, ……., I)*
j: machines *(j= 1, 2, ……., J)*
*k:*sequence position *(k= 1, 2, ….., K)*
Parameters
t_{ij} *:* processing time of job i on machine j *(∀i ∈I, ∀j ∈J)*

Decision Variables
x_{ik}*:* if job i is performed k. position 1, otherwise 0 *(∀i ∈I, ∀k ∈K)*
C_{kj}*:* Completion time of job k on machine j *(□k ∈K, □j ∈J)*
P_{kj}*:* Processing time of job k on machine j *(∀k ∈K, ∀j ∈J)*

C_{max}: Time for the last job to leave the last machine

Mathematical Model

$$Min\ C_{\max} \quad (1)$$

$$\sum_{k\in K} x_{ik} = 1 \quad ,\forall i \in I \quad (2)$$

$$\sum_{i\in I} x_{ik} = 1 \quad ,\forall \mathrm{k} \in \mathrm{K} \quad (3)$$

$$P_{kj} = \sum_{i\in I} x_{ik} t_{ij} \quad ,\forall k \in K, \forall j \in J \quad (4)$$

$$C_{k1} = C_{(k-1)1} + P_{k1} \quad ,\forall \mathrm{k} \in K / \{1\} \quad (5)$$

$$C_{kj} \geq C_{k(j-1)} + P_{kj} \quad ,\forall j \in J \ / \{1\}, \forall \mathrm{k} \in K \quad (6)$$

$$C_{kj} \geq C_{(k-1)j} + P_{kj} \quad ,\forall j \in J \ / \{1\}, \forall \mathrm{k} \in K / \{1\} \quad (7)$$

$$C_{11} = P_{11} \quad (8)$$

$$x_{ik} \in \{0,1\} \quad ,\forall \mathrm{i} \in \mathrm{I}, \forall \mathrm{k} \in \mathrm{K} \quad (9)$$

$$C_{kj}, P_{kj} \geq 0 \quad ,\forall k \in K, \forall j \in J \quad (10)$$

In the model, equation (1) aims at minimizing the makespan. Equation (2) shows that each job can only be assigned to one position, and equation (3) indicates that only one job can be assigned to each position. Equation (4) k. means that the processing time of the job performed in machine j is equal to the processing time of that job on that machine. Equation (5) k. the time to complete the next job on the first machine, equation (6) and equation (7) k. shows the time when the next job is completed on other machines. Equation (8) states that the completion time of the first job on the first machine is equal to the processing time of that job on that machine. Equation (9) and equation (10) are the sign constraints of decision variables.

Particle Swarm Optimization

One of the meta-heuristic methods used to solve FSSP is the PSO algorithm. PSO is an optimization technique developed by Kennedy and Eberhart [15] in 1995 based on the movements of flocks of birds and fish during food search [16]. Compared to other evolutionary and mathematically based algorithms, PSO is an algorithm that does not require much memory, has effective computational capabilities, is easy to implement and has fast convergence features [17].

Experimental Results

Mathematical model and PSO algorithm for the solution of FSSP were solved with 6 job 6 machine problem from Benavides and Ritt [18] studies and the first 3 problems of Taillard's [19] 20 job 5 machine data set. The problem [18] is solved primarily with the mathematical model and PSO algorithm for the minimization of makespan and the results obtained are compared with the results (Benavides and Ritt, 2018). According to the result obtained with the mathematical model, the job sequence of the problem was realized as *{J_5-J_6-J_1-J_3-J_4-J_2}* and C_{max} was *41*. The job sequence obtained by Benavides and Ritt (2018) was obtained as *{J_5-J_4-J_6-J_2-J_1-J_3}* and C_{max} was *43*.

The data of the first 3 problems were solved with the mathematical model and developed PSO algorithm from Taillard's [19] 20 job 5 machine problem sets. In the PSO algorithm, the best solutions obtained after 3 trials were considered. The comparison was made according to the makespan values obtained in Table 1.

Table 1. Comparative makespan results.

Problem Nr.	Taillard Results [19]	Mathematical Model Results	PSO Best Results		
1	1278	1278	1312	1297	1278
2	1359	1359	1359	1359	1360
3	1081	1081	1089	1098	1098

When Table 1 is examined, it is seen that the mathematical model can find Taillard's optimal result for all three problems. The developed PSO algorithm was able to find the optimal result for 2 of 3 problems.

Conclusion

In this study, a mathematical model and PSO algorithm are presented for the makespan solution of FSSP. The solution of the small size problems obtained from the literature was realized with mathematical model and PSO algorithm. It has been seen that the mathematical model yields optimal results for small size problems. Since the mathematical model could not provide the optimal solution as the problem size increased, PSO algorithm was developed for the solution of these problems. The developed PSO algorithm was tested on 3 problems solved by mathematical model. When the results obtained in the first 3 trials for these problems were examined, it was seen that the optimal solution for the 2 problems could be achieved.

Acknowledgements

This study is a part of the project which is supported by The Scientific and Technological Research Council of Turkey (TUBITAK). The authors thank the TUBITAK for financial support of this work under the grant number 118E355.

References

[1] Baker, K. R., and Trietsch D. (2009). *Principles of sequencing and scheduling*. John Wiley&Sons.
[2] Ruiz, R., Maroto, C., Alcaraz, J. (2006). Two new robust genetic algorithms for the flowshop scheduling problem. *Omega,* 34 (5), 461–476.
[3] Chen, S. H., Chang, P. C., Cheng, T. C. E. and Zhang, Q. (July, 2012). A Self-guided Genetic Algorithm for permutation flowshop scheduling problems. *Computers & Operations Research*, 39 (7), 1450-1457.
[4] Chen, Y.M., Chen, M.C., Chang, P.C. and Chen, S.H. (2012). Extended artificial chromosomes genetic algorithm for permutation flowshop scheduling problems. *Computers & Industrial Engineering*, 62, 536-545.
[5] Liu, Y.F. and Liu, S.Y. (2013). A hybrid discrete artificial bee colony algorithm for permutation flowshop scheduling problem. *Applied Soft Computing*, 13, 1459-1463.
[6] Fernandez-Viagas, V. and Framinan, J.M. (2014). On insertion tie-breaking rules in heuristics for the permutation flowshop scheduling problem. *Comput. Oper. Res.*, 45, 60–67.
[7] Liu, W., Jin Y. and Price, M. (2017). A new improved NEH heuristic for permutation flowshop scheduling problems. *Int. Journal of Production Economics*, 193, 21-30.
[8] Tayeb, F. B., Bessedik, M., Benbouzid, M., Cheurfi, H. and Blizak, A. (2017). Research on Permutation Flow-shop Scheduling Problem based on Improved Genetic Immune Algorithm with vaccinated offspring. Procedia Computer Science, 112, 427-436.
[9] Wang H., Wang W., Sun H., Cui Z., Rahnamayan S. and Zeng S. (2017). A new cuckoo search algorithm with hybrid strategies for flow shop scheduling problems. *Soft Computing*, 21, 4297–4307.
[10] Liu B., Wang L. and Jin Y. H. (2007). An effective pso-based memetic algorithm for flow shop scheduling. *IEEE Trans Sys Man Cybern Part B Cybern*, 37(1), 18–27.
[11] Yang X. S. and Deb S. (2009). Cuckoo search via Lévy flights. In: World congress on nature and biologically inspired computing IEEE, pp 210–214.
[12] Abedinnia, H., Glock, C. H. and Brill, A. (2016). New simple constructive heuristic algorithms for minimizing total flow-time in the permutation flowshop scheduling problem. *Computers & Operations Research*, 74, 165-174.
[13] Taşgetiren M.F., Liang Y. C., Şevkli M., Gençyılmaz G. (2007). A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem. *Eur J Oper Res*, 177(3), 1930–1947.
[14] Li X. and Yin M. (2012). A discrete artificial bee colony algorithm with composite mutation strategies for permutation flow shop scheduling problem. *Scientia Iranica*, 19(6),1921–1935.
[15] Kennedy J. and Eberhart R. C. (1995). Particle swarm optimization. *International Neural Networks*, 95, 1942-1948.
[16] Dahal K.P., Tan K. C. and Cowling, P.I. (2007). *Evolutionary Scheduling*. Springer-Verlag, New York.
[17] Kaya S. and Fığlalı N. (2018). Çok Amaçlı Esnek Atölye Tipi Çizelgeleme Problemlerinin Çözümünde Meta Sezgisel Yöntemlerin Kullanımı. *Harran Üniversitesi Mühendislik Dergisi*, 3(3), 222-233.
[18] Benavides, A. J. and Ritt, M. (2018). Fast heuristics for minimizing the makespan in non- permutation flow shops. *Computers and Operations Research*, 100, 230–243.
[19] Taillard, E. (2004). Best known lower and upper bounds of the PFSSP for Taillard's instances. http://mistic.heig-vd.ch/taillard/problemes.dir/ordonnancement.

A Practical Approach to Calculation of Shielding Effectiveness of Frequency Selective Surfaces

E. Soygüder[1], A. O. Kızılçay[2]

[1] Van Yüzüncü Yıl University, Van, Turkey, erensygdr@gmail.com
[2] Van Yüzüncü Yıl University, Van, Turkey, oguzkizilcay@yyu.edu.tr

Abstract

Frequency selective surfaces (FSS) made of composite materials are widely used in electromagnetic shielding problems. Debye model is one of the methods used to specify FSS characteristics. In order to achieve required Shielding Effectiveness (SE), this model's parameters are usually determined by trial and error method which takes time. In this study, an analytical solution method was applied to Debye model in order to eliminate this problem. The results were analyzed graphically using Matlab software. Hence, it was clearly seen that the shielding performance of FSS can be obtained in a more effective way.

Introduction

Today, the importance of non-homogeneous artificial composite materials is rapidly increasing. Composite material design with advancing technology is crucial to solve various problems related to electromagnetic compatibility, electromagnetic immunity and signal integrity. Such materials are characterized by their high protection against electromagnetic fields ranging from DC to optical frequencies, as well as a low reflection coefficient in a defined frequency range. Examples of applications are (FSS), filters, integrated optical microwave guides, thin films, memory devices, new antennas, radar-absorbing materials. One of the most commonly used homogenizing theories for material design is the Maxwell Garnett (MG) mixing rule [1]. This paper presents a new and easy method for solving complex formulas used in defining SE parameters of the Debye model. This paper presents a new and easy method for the solution of complex formulas used in defining the SE parameters of the Debye model, covering the common equation solutions of MATLAB.

Related Works

Assuming that a two-phase mixture is uniformly dispersed in a material, the Maxwell Garnett mixing rule, which represents the effective permeability of such a compound, can be written as follows [2]:

$$\varepsilon_{Re} = \frac{\omega^2\tau^2\varepsilon_\infty + \varepsilon_s}{1+\omega^2\tau^2} \tag{1}$$

$$\varepsilon_{Im} = \frac{\omega\tau(\varepsilon_s - \varepsilon_\infty)}{1+\omega^2\tau^2} \tag{2}$$

$$\varepsilon_D = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1+j\omega\tau} \tag{3}$$

Here, the $\varepsilon_{D=}\ \varepsilon_{Re\ -\ j}\varepsilon_{Im}$ represents the complex, frequency-dependent single-degree relative permeability of the equivalent homogeneous material described in (3). ε_s is the static relative permittivity, ε_∞ is the relative high-frequency limit permittivity, $\omega=2\pi f$ is angular frequency and τ is the relaxation time. For a composite media layer laced in the air, the SE in dB is the same as in (4) [3]. The total permeability given in (5) can be written from the ε_D type [4].

$$SE = -20\log\left(|T|\right) \tag{4}$$

$$T_{(\varepsilon_D)} = \frac{2\sqrt{\varepsilon_D}}{\begin{pmatrix} 2\sqrt{\varepsilon_D}\cos\left(\omega W\sqrt{\varepsilon_D}\sqrt{\varepsilon_0\mu_0}\right) \\ +j(1+\varepsilon_D)\sin\left(\omega W\sqrt{\varepsilon_D}\sqrt{\varepsilon_0\mu_0}\right)\end{pmatrix}} \tag{5}$$

In (5), ε_0 and μ_0 are the dielectric permittivity and magnetic permeability in free space. It is necessary to define the lower (6) and upper (7) frequency limit of SE [5]. Equation (6) is the frequency at which the difference of the imaginary part of Debye to the real part is maximum. It represents the lower frequency limit of the known frequency range and depends only on the relaxation time (τ).

$$f_{ri} = \frac{1+\sqrt{2}}{2\pi\tau} \tag{6}$$

$$f_n = \frac{\sqrt{2}c_0}{4W\sqrt{|\varepsilon_D| + \varepsilon_{Re}}}(2n+1) \quad n = 0,1,2 \ldots \tag{7}$$

Equation (7) is the upper frequency limit and as it depends on the frequency $f_{n=0}$ must be solved numerically. Where c0 is the velocity of the electromagnetic waves in the free space.

$$|T_{SE\_max\_\infty}| = \frac{4sinh(b_\infty)\varepsilon_\infty + 2\,cosh(b_\infty)\sqrt{\varepsilon_\infty}\,(\varepsilon_\infty+1)}{\left[\begin{array}{c}4\,(cosh(b_\infty)^2-1)\varepsilon_\infty + cosh(b_\infty)\,. \\ [4sinh(b_\infty)\sqrt{\varepsilon_\infty}(\varepsilon_\infty+1) + (cosh(b_\infty)(\varepsilon_\infty+1)^2]\end{array}\right]} \tag{8}$$

$$|T_{SE\_min\_\infty}| = \frac{4cosh(b_\infty)\varepsilon_\infty + 2\,sinh(b_\infty)\sqrt{\varepsilon_\infty}\,(\varepsilon_\infty+1)}{\left[\begin{array}{c}4\,cosh(b_\infty)^2\,\varepsilon_\infty - (\varepsilon_\infty+1)^2 + cosh(b_\infty)\,. \\ [4sinh(b_\infty)\sqrt{\varepsilon_\infty}(\varepsilon_\infty+1) + (cosh(b_\infty)(\varepsilon_\infty+1)^2]\end{array}\right]} \tag{9}$$

$$b_\infty = \frac{(\varepsilon_s - \varepsilon_\infty)W}{2c_0\tau\sqrt{\varepsilon_\infty}} \tag{10}$$

Where, equation (10) represents the difference between the magnitude of $|\varepsilon_D|$ and its real part ε_{Re} for $\omega\to\infty$. Equation (8) and (9) determine the maximum and minimum of SE in infinity. For more information a ; equations we have given so far, see also (De Paulis et. al., 2014).

The main purpose is to determine parameters of the material (ε_∞, ε_s, W, τ), using the formulas given above. In [5], the authors gave numerical values to the lower (6) and upper (7) frequency limits, b_∞ parameter (10) and maximum level at infinity (8) of SE (Example: f_{ri}= 50 MHz, $f_{n=0}$= 5 GHz, b_∞= 0.7, $|T_{SE\_max\_\infty}|$= 10 dB.). (τ) was calculated by equation (6). Equation (8) was used to calculate the (ε_∞). Equations (7) and (10) are solved together to obtain the (ε_s) and (W). In [6], similar methods were used, but the Newton Method (Isaacson and Keller, 1994) was used to obtain the (ε_∞) from equation (8).

Proposed Method

The method proposed in this section is explained by a numerical example. The gains to be obtained as a result of the example is to determine the real parameters of the homogenized composite material. In the example, the pseudocode of the proposed method is shown in Fig. 1.

```
Begin
% create symbolic variables.
    syms εω εs bω W
% define fixed parameters.
    ε0 = 8.85x10^-12; μ0 = 4πx10^-7; c0 = 3x10^8;
% Obtain (τ) from equation (6).
    fri = 50 MHz;
% Solve equations (8) and (9) together.
    |T_SE_max_ω| = 15;   |T_SE_min_ω| = 14,3;
    vpasolve([|TSE-max-ω|,|TSE-min-ω|],[ εω,bω]);
% Solve equations (7) and (10) together.
    vpasolve([bω, fn=0],[ εs,W]);
% Set the frequency range. Draw the SE curve.
    f = 10^-1:0.1:10^6; SE = -20log10(|T|);
    semilogx(f,SE);
    end
```

Figure 1. Pseudocode of the example applying the proposed method

The frequency-dB graph is plotted to observe the behaviour of the SE in the desired frequency range (Fig. 2).

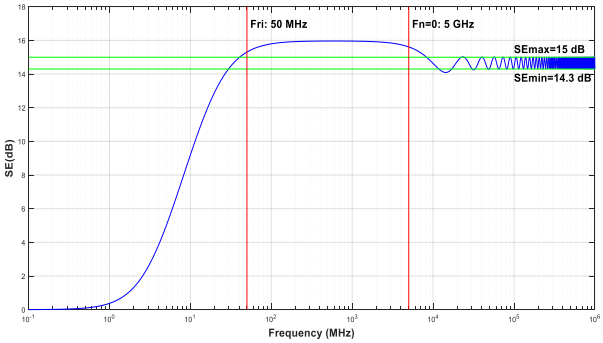


Figure 2. SE of the example applying the proposed method

First, the relaxation time from equation (6) is calculated as (τ) = 7.68 ns. Then equations (8) and (9) are solved together. The relative high frequency limit permittivity and b_∞ parameter was calculated as (ε_∞) = 20.7816 and b_∞= 1.1601. Then equations (7) and (10) are solved together. The static relative permittivity and panel thickness was calculated as (ε_s)= 1.2179e+04 and (W)= 2 mm. The SE plot of the material with these characteristics is as in Figure 2. It is seen that the material displays approximately 16 dB shielding in the range of 50 MHz - 5 GHz.

Results

In this article, we propose an easier and controllable solution for obtaining one-dimensional Debye model parameters, although there are similar parts to the suggestions we mentioned in the related studies section. The designed model is used in the production of composite material including cylindrical inclusions. It is believed that the techniques used herein will facilitate the design of different and complex materials in the future.

References

[1] Garnett, J. C. M. (1904). *Colors in metal glasses and metal films*. *Trans. R. Soc.*, vol. CCIII.
[2] Nisanci, M. H., De Paulis, F., Koledintseva, M. Y. and Orlandi, A. (2011 August). full-wave EMC simulations using Maxwell Garnett model for composites with cylindrical inclusions. *Proc. Int. IEEE Symp. Electromag. Compat.*, USA.
[3] Koledintseva, M.Y., Ravva, P.C., DuBroff, R.E., Drewniak, J.L., Rozanov, K.N. and Archambeault, B. (2005 August). Engineering of composite media for shields at microwave frequencie. *in Proc. IEEE Symp. Electromag. Compat.,* Chicago, IL, vol. 1, pp. 169-174.
[4] De Paulis, F., Nisanci, M. H., and Orlandi, A. (2012). Evaluation of dielectric permittivity for homogeneous materials from transmittance requirements. *in Proc. IEEE Int. Symp. Electromagn. Compat.,* Pittsburgh, PA, USA, Aug.
[5] De Paulis F., Nisanci M. H., Orlandi A., Koledintseva M. Y. and Drewniak J. L. (2014). Design of homogeneous and composite materials from shielding effectiveness specifications. *IEEE Transactions on Electromagnetic Compatibility*, 56(2), 343-351. doi: 10.1109/TEMC.2013.2280463
[6] Dogusgen, C., Kent, S. (2017). An Analytıcal Approach For Materıal Synthesıs Based On Shıeldıng Effectıveness Characterıstıcs. *Uludag University Journal of The Faculty of Engineering,* 22(1.1)
[7] Isaacson, E. and Keller, H. B. (1994). Analysis of Numerical Methods. *Courier Corporation*, New York.

Statistical Modeling of Lead (Pb) Adsorption on Clay Minerals of Çelebibağ

C. Demir[1], A.R. Kul[2], Y. Demir[3], H. Yıldız[4], S. Keskin[5]

[1] Van Yüzüncü Yıl, University, Van, Turkey, canandemir@yyu.edu.tr
[2] Van Yüzüncü Yıl, University Yıl, Van, Turkey, alirizakul@yyu.edu.tr
[3] Van Yüzüncü Yıl, University, Van, Turkey, ydemir@yyu.edu.tr
[4] Van Yüzüncü Yıl, University, Van, Turkey, erzenyildiz5665@gmail.com
[5] Van Yüzüncü Yıl, University, Van, Turkey, skeskin@yyu.edu.tr

Abstract

The aim of this study is to statistical modeling of lead (Pb) adsorption on clay minerals Çelebibağ. Modeling, depending on time, was performed to determine for lead (Pb) adsorption level at fixed pH 5.5 for various concentration and temperatures in clay minerals Çelebibağ. One-way analysis of variance was used for comparison to various temperature and concentration levels. Logarithmic, quadratic, cubic and logistic models as well as linear were used to determine adsorbed lead (Pb) amount at different temperature levels and heavy metal concentration. Differences between various time and concentration levels were found statistically significant; however, there were no significant differences between temperature and concentration levels. In addition, cubic model had higher R^2 values for each concentration and temperature levels.

Introduction

Lead (Pb) is a heavy metal commonly found in nature [1]. Heavy metals get involved in the food chain through agricultural products and affect the whole ecological system as well as affecting human health negatively [2]. The lead taken to the human body in various ways such as inhalation, nutrients and water is distributed primarily to the soft tissues and parenchymal organs and then stored in the bones by replacing calcium. It is reported that it affects many systems and organs such as hematologic system, central nervous system, kidneys and liver [3]. Due to its high toxicity, the adsorption of lead is of great importance for both environmental and human health.

One of the substances used as adsorbent material is clay. Clay is very common in Van and more economical than other adsorptive substances. In case of using clay as an adsorbent agent, determination of the amount for adsorbed lead at different concentrations (25, 50, 75 and 100ppm) and temperatures (25, 35, 45°C) is important. Although many studies have been conducted about change of adsorbed substances amount to temperature and time, it can be said that studies on modeling of these substance amount with regard to time and temperature are not sufficient. Therefore, in addition to the linear regression model, four nonlinear regression models were performed to obtain estimation equations and to determine the usability of these equations.

Statistical Analysis

Regression is expressed as a function of independent variables thought to be associated with the dependent variable. The functional form of the relationship between variables is examined by regression models. The regression model that should be used differs according to the structure of the data. Using the wrong model can lead to erroneous results. The general expression of the regression equation is written as follows [4].

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

Y: Dependent (result) variable is assumed to have a certain error
X: Independent (cause) variable is assumed to be measured without error
β_0: Y is the value that is when X= 0. It is the point where the line breaks the y-axis
β_1: It is the slope of the line or the regression coefficient. X refers to the amount of change in Y when a unit changes
ε: It is the random error value [5].

Least Squares Method: The most commonly used method for estimating the regression coefficients is the least squares method. Least squares method is based on finding a curve equation so that the sum of squares of errors will be minimum.

$$Min \sum e_i^2 = \sum \left(Y - (\beta_0 + \beta_1 X)\right)^2 \quad (2)$$

For minimum of a second order function derivative are equalized to zero. Accordingly, regression constant and regression coefficient are obtained as follows [6].

$\beta_0 = \bar{Y} - \beta_1 \bar{X}$ $\beta_1 = \frac{\sum (Y-\bar{Y})(X-\bar{X})}{\sum (X-\bar{X})^2}$

Descriptive statistics for the studied variables were expressed as mean, standard deviation, minimum and maximum values. Variance analysis was used to determine whether there was a difference between both time and temperature levels for three concentrations. In order to identify different groups, Tukey's multiple comparison test was used following ANOVA. In addition, logarithmic, quadratic, cubic and logistic models as well as linear model were used to estimate the amount of adsorbed heavy metals at different temperatures and heavy metal concentrations.

Linear; $Y = a_0 + a_1 X$

Logarithmic; $Y = a_0 X^{a_1} => lnY = a_0 + a_1 lnX$

Quadratic; $Y = a_0 + a_1 X + a_2 X^2$

Cubic; $Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3$

Logistic; $Y = a_0 + a_1 (logX) + a_2 (logX)^2$

Accordingly, the most properly model was determined and the approximate curves representing the model on the spread diagrams were shown [7]. R-square (R^2, Determination coefficient) was considered to determine goodness of fit the models. Level of significance was taken 5% for all statistical tests and comparisons and SPSS statistical software was used for the all statistical computations.

Findings

Difference between the times for each concentration was found to be statistically significant ($p<0.01$). Accordingly, when lead ion shaken for 4 minutes, average adsorption amount was 4.15 in 25ppm concentration however, average concentration was 22.02 when it shaken for 185 minutes in the same concentration. At 50ppm, when shacked for 4 min, the average was 11.82, while it was found 41.73 by increasing of shaking duration to 185 minutes in the same concentration. When lead ions concentration was 75ppm, adsorbed amount of the substance was 18.35 for 4 min shaking duration, however, this value reached to 60.04 for 185 min shaking duration. When lead ions concentration was 100ppm, adsorbed amount of the substance was 29.17 for 4 min shaking duration, however, this value reached to 77.71 for 185 min shaking duration. There are no differences among the amounts of temperature levels (25°C, 35°C, 45°C) for adsorption on Çelebibağ clay of lead ions in different concentrations.

As seen Table 1, cubic and quadratic models with 99% R^2 value were the best models for 25ppm and 25°C temperature. This model was followed by logarithmic model with 95% and linear model with 81% R-square values. The logistic model was the last with 68%. Cubic and quadratic models were also the best models with 99% R^2 value at 35°C. This model was followed by logarithmic model with 96% and linear model with 82% R^2 value. The logistic model was included in last with 70%. When temperature was 45°C, the best estimations were made by cubic and quadratic models with 99% R^2 value. These models were followed by logarithmic model with 94% and linear model with 82% R-square value. The lowest estimation was made by logistic models with 74% R^2 value. When we look at the table values to 50ppm and 25°C temperature, cubic model was the best models with 97% R-square value. This was followed by quadratic model with 92%, logarithmic model with 91%, linear model with 64% and logistic model with 52%. For the same concentration and 35°C temperature, cubic model had the highest predictive value with 97% R-square value. R-square values for quadratic, logarithmic, linear and logistic models were found 93%, 92%, 66% and 54%, respectively. Similarly, for the same concentration and for the 45°C temperature, cubic model had the highest (98%) value. This model was followed by the quadratic model with 95%, logarithmic model with 93%, linear model with 69% and logistic model with 61% R-square value.

Estimation of the cubic model was the best (99%) for 75ppm concentration and 25°C temperature. This was followed by quadratic (96%), logarithmic (92%) and linear (69%) models. The lowest (58%) estimation had the logistic model. On the other hand, for the same concentration, cubic model provided the highest R-square value (99%) at 35°C. R-square values of quadratic, logarithmic, linear and logistic models were found 93%, 89%, 62% and 54%, respectively. For the same concentration and 45°C, cubic model was the best model with 97% R-square value. This was followed by the quadratic (91%), logarithmic (86%), linear (59%) and logistic (54%) models. When we

look at the table values to 100ppm and 25°C temperature, cubic model was the best model with 99% R-square value. This was followed by quadratic model with 97%, logarithmic model with 92%, linear model with 69% and logistic model with 63%. For the same concentration and 35°C temperature, cubic model had the highest predictive value with 98% R-square value. R-square values for quadratic, logarithmic, linear and logistic models were found 93%, 91%, 64% and 58%, respectively. Similarly, for the same concentration and for the 45°C temperature, cubic model had the highest (95%) value. This model was followed by the quadratic model with 88%, logarithmic model with 86%, linear model with 59% and logistic model with 53% R-square value.

Table 1. Model Summary and Parameter Estimates

ppm	Heat	Model	R^2	Constant	b1	b2	b3	ppm	Heat	Model	R^2	Constant	b1	b2	b3
25	25 ^{0}C	Linear	0.81	6.208	1.028			75	25 ^{0}C	Linear	0.69	27.195	2.161		
		Logarithmic	0.95	0.444	7.749					Logarithmic	0.92	12.764	17.409		
		Quadratic	0.99	-0.718	3.007	-0.099				Quadratic	0.96	8.062	7.628	-0.273	
		Cubic	0.99	-1.399	3.370	-0.143	0.001			Cubic	0.99	-0.852	12.378	-0.852	0.019
		Logistic	0.68	0.166	0.916					Logistic	0.58	0.040	0.942		
	35 ^{0}C	Linear	0.82	7.070	1.000				35 ^{0}C	Linear	0.62	33.431	1.873		
		Logarithmic	0.96	1.475	7.530					Logarithmic	0.89	19.696	15.682		
		Quadratic	0.99	0.445	2.892	-0.095				Quadratic	0.93	14.359	7.323	-0.272	
		Cubic	0.99	-0.235	3.255	-0.139	0.001			Cubic	0.99	2.892	13.433	-1.017	0.025
		Logistic	0.70	0.143	0.924					Logistic	0.54	0.032	0.954		
	45 ^{0}C	Linear	0.82	8.236	0.959				45 ^{0}C	Linear	0.59	37.388	1.708		
		Logarithmic	0.94	3.023	7.147					Logarithmic	0.86	24.721	14.368		
		Quadratic	0.99	2.047	2.727	-0.088				Quadratic	0.91	19.371	6.856	-0.257	
		Cubic	0.99	1.953	2.777	-0.095	0.000			Cubic	0.97	8.316	12.747	-0.975	0.024
		Logistic	0.74	0.119	0.934					Logistic	0.54	0.028	0.961		
50	25 ^{0}C	Linear	0.64	20.781	1.348			100	25 ^{0}C	Linear	0.69	42.532	2.330		
		Logarithmic	0.91	11.156	11.161					Logarithmic	0.92	27.015	18.745		
		Quadratic	0.92	8.002	5.000	-0.183				Quadratic	0.97	21.454	8.352	-0.301	
		Cubic	0.97	0.464	9.017	-0.672	0.016			Cubic	0.99	13.101	12.803	-0.844	0.018
		Logistic	0.52	0.053	0.947					Logistic	0.63	0.024	0.958		
	35 ^{0}C	Linear	0.66	22.893	1.321				35 ^{0}C	Linear	0.64	47.981	2.088		
		Logarithmic	0.92	13.588	10.877					Logarithmic	0.91	33.026	17.308		
		Quadratic	0.93	10.844	4.764	-0.172				Quadratic	0.93	28.008	7.794	-0.285	
		Cubic	0.97	3.858	8.487	-0.626	0.015			Cubic	0.98	15.829	14.285	-1.076	0.026
		Logistic	0.54	0.047	0.953					Logistic	0.58	0.022	0.964		
	45 ^{0}C	Linear	0.69	23.264	1.366				45 ^{0}C	Linear	0.59	49.056	2.092		
		Logarithmic	0.93	14.104	11.021					Logarithmic	0.86	33.566	17.585		
		Quadratic	0.95	11.156	4.825	-0.173				Quadratic	0.88	28.176	8.058	-0.298	
		Cubic	0.98	5.530	7.823	-0.538	0.012			Cubic	0.95	13.920	15.655	-1.224	0.031
		Logistic	0.61	0.045	0.955					Logistic	0.53	0.021	0.963		

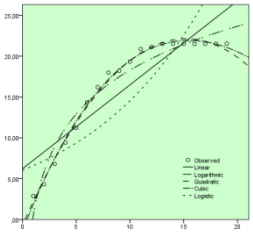


Figure 1. 25ppm 25^{0}C

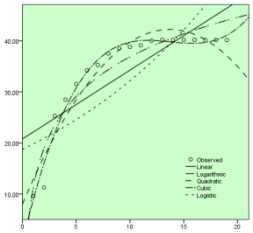


Figure 4. 50ppm 25^{0}C

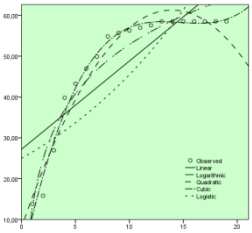


Figure 7. 75ppm 25^{0}C

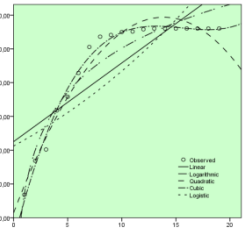


Figure 10. 100ppm 25^{0}C

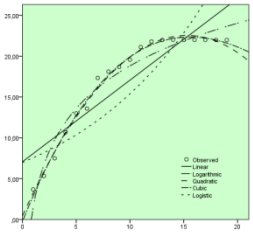


Figure 2. 25ppm 35^{0}C

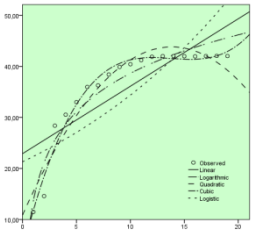


Figure 5. 50ppm 35^{0}C

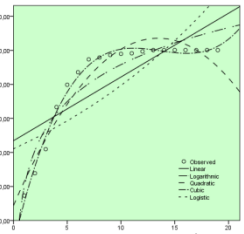


Figure 8. 75ppm 35^{0}C

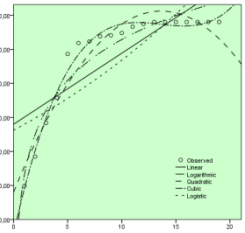


Figure 11. 100ppm 35^{0}C

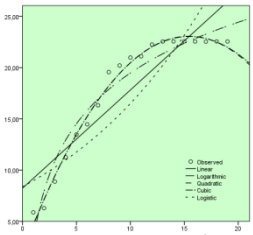


Figure 3. 25ppm 45^{0}C

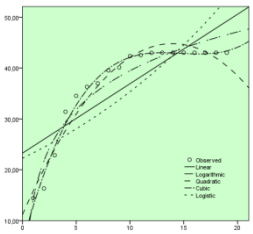


Figure 6. 50ppm 45^{0}C

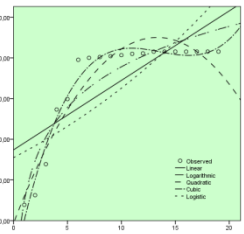


Figure 9. 75ppm 45^{0}C

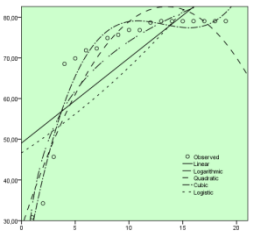


Figure 12. 100ppm 45^{0}C

Discussion and Conclusions

In this study, the adsorption of lead ions on Çelebibağ clay were examined at various temperatures (25oC, 35oC, 45oC), time (4 , 8, 12,…, 185 min) and (25ppm, 50ppm, 75ppm, 100ppm) concentrations. Difference between the durations for each concentration was found statistically significant ($p<0.001$). Similar results were reported by [2; 8]. It was observed that the adsorption efficiency did not change with temperature. Based on the result in table 1 to determine the adsorption of lead ions on the Çelebibağ clay; the models having the highest estimation coefficient of determination (R^2) and the lowest standard error values were indicated as the most appropriate models. Both for all temperature values (25oC, 35oC, 45oC) and for all concentrations (25ppm, 50ppm, 75ppm, 100ppm), cubic model provided the best estimation of the R^2 value which ranges from 95% to 99%. This model was followed by quadratic model which varies from 88% to 99% R^2 value, logarithmic model which varies from 86% to 96% R^2 value, linear model which varies from 59% to 82% R^2 value and logistic model which varies from 52% to 74% R^2 value.

References

[1] Seven, T., Can, B., Darende, B. N., and Ocak, S. (2018). Heavy Metals Pollution in Air and Soil. *National Environmental Science Research Journal*, 1(2), 91-103.

[2] Özcan, A. S. (2010). The Characterization of Natural Bentonite and Its Adsorption Ability of Lead(II) Ions. *Journal of Balıkesir University Institute of Science and Technology*, 12 (2), 85-97.

[3] Çaylak, E. (2010). Lead Toxication and Oxidative Stress in Children and Antioxidant Effects of Thiol Compounds. *J Child,* 10 (1), 13-23.

[4] Arı, A., and Önder, H. (2013). Regression Models Used For Different Data Structures. *Anadolu J Agr Sci,* 28 (3), 168-174.

[5] Hamzaoğlu, S. (2013). *Power analysis in multiple regression methods* (Master's Thesis). University of Ondokuz Mayıs, Samsun.

[6] Demirelli, H. (2018). *Investigations of reasons for district disasters* (Master's Thesis). University of Fırat, Elazığ.

[7] Spiegel, M. R, and Stephens L. J. (2018). *Statistics* (8. Edition). McGraw-Hill Education.

[8] Tekir, O. (2006). *Preparation of activated carbon from the hazelnut husk and adsorption of some heavy metal ions* (Master's Thesis). University of Sakarya, Sakarya.

Households Electricity Consumption Estimation:A Dynamic Linear Model Application

E. Doruk[1], T. Tasci[2]

[1]*Sakarya Üniversitesi, Sakarya, Türkiye, emredoruk@sakarya.edu.tr*
[2]*Sakarya Üniversitesi, Sakarya,Türkiye, ttasci@sakarya.edu.tr*

Abstract

Electrical energy as a fundamental requirement in today's world, has always been facilitating human life at the base of rapid population growth, industrialization, technological developments and social welfare. Due to the fact that electrical energy needs to be supplied to consumers on-demand, there appears a certain need for managing transmission process. Any approach addressing this process should also have proper functions ensuring service continuity and quality as well as cost efficiency. This study aims presenting a method for cost prediction by estimation of energy demands. In the study, 3-year electricity consumption data of households in Sakarya, Turkey is used. Data is modelled as one-variable time series within a Dynamic Linear Model framework considering the effects of trend, seasonality, cyclicality and randomness together. The error rate of the results is measured as 5.08% (MAPE) in comparison with the realized values which provides a remarkable improvement against the methods employed so far in Sakarya province by the operator corporation.

Giriş

İnsanoğlu tarihin ilk çağlarından itibaren yaşamını kolaylaştırmak ve yaşam kalitesini arttırmak için çevresindeki nesnelerden en iyi şekilde faydalanmayı öğrenmiştir. Ancak yoğun nüfus artışı, kentleşme ve insanlığın yaşamını kolaylaştırmak ve yaşam kalitesini arttırmak için geliştirdiği her ürün ve yöntem beraberinde yeni ihtiyaçları da ortaya çıkarmıştır. Bu ihtiyaçlar için gerekli arzın karşılanmasında kaynakların etkin kullanılması ve planlanması gerekmektedir. Planlama işlemi ise ancak elde yeterli bilgi varsa mümkün ve gerçekçi olmaktadır. Bu noktada planlama öncesi veri elde etmek için tahmin işlemleri gerçekleştirilir.

Tahmin işlemleri bir topluluğun ya da kişinin kararlarına göre ya da matematiksel ve istatistiksel yöntemlerle gerçekleştirilebilir. Bu yöntemlerden birisi olan zaman serisi analiz yöntemleri, gözlenen herhangi bir değişkenin, gözlem süresince belirli zaman noktalarında oluşturduğu serinin tanımlanması, modellenmesi ve tahmin edilmesi sürecini kapsar. Zaman serileri tahmininin ise bakıldığında istatistik, ekonomi, üretim, tıp, iklim analizleri, işletme ve verimlilik gibi farklı alanlarda uygulama alanı vardır. Uygulama alanlarından bir diğeri ve bu çalışmanın konusu olan elektrik talep tahminidir.

Elektrik talep tahmini ile öngörülen enerji talebini karşılamak için mevcut sistemlerin yük kapasitesi arttırılarak ya da yeni elektrik santralleri ile gereken enerji sağlanabilir ve bu doğrultuda iletim-dağıtım sistemlerinde de geliştirmeler yapılabilir. Bu sayede kapasite artırımına ilişkin gerekli maliyet hesaplamaları önceden planlanarak daha etkin bir çalışma sergilenebilir.

Elektrik enerjisinin üretimi, iletimi ve dağıtımı ile sorumlu bulunan kurumların temel amacı, üretilen enerjinin tüketicilere ucuz ve kaliteli olarak sunulmasıdır. Talep tahmininin düşük hatalarla yapılması, bu talebe karşılık gelecek arzın oluşturulması ve enerji darboğazlarının yaşanmaması için üretim, iletim ve dağıtım sistemlerinin düzenli olarak planlamalarının yapılması, gerçek değerlerle aykırılığı ortadan kaldırarak elektrik piyasası rekabetinde avantajı da beraberinde getirecektir.

Bu çalışmada elektrik sektöründe elektrik talep tahmini işlemlerinde sıklıkla kullanılan tahmin yöntemlerine ek olarak Dinamik Lineer Modeller ile tahmin uygulaması gerçekleştirilmiş, test verileri ile karşılaştırılarak başarım sonuçları elde edilmiştir.

Materyal ve Yöntem

Materyal

Yapılan uygulamalarda SEDAŞ aracılığı ve izni ile elde edilen Sakarya iline ait elektrik tüketim verileri kullanılmıştır. Veri setinde Sakarya ili için gerçekleşen 3 yıllık (2015-2017) hanehalkı tüketim verileri bulunmaktadır. Ocak 2015/Kasım 2017 arası veriler test verisi, Aralık 2017 verisi test verisi olarak kullanılmıştır.

Elde edilen veri günümüzdeki elektrik tüketim dilimleri olan ve daha önemli olduğu düşünülen T1, T2 ve T3 zaman dilimlerine uygun bir biçimde düzenlenmiştir. T1 zaman dilimi 06:00 ile 17:00 saatleri arası tüketimi, T2 zaman dilimi 17:00 ile 22:00 saatleri arası tüketimi ve son olarak T3 zaman dilimi 22:00 ile sonraki gün 06:00 saatleri arası tüketimi ölçmektedir.

Yöntem

Bu çalışmada tahmin yöntemi olarak kullanılan Dinamik lineer modeller, 1960'lı yılların başında, dinamik sistemleri izlemek ve kontrol etmek için geliştirmiştir. Zaman serilerinin modellenmesinde kullanılan Dinamik lineer modeller ise Harrison ve Stevens'ın dinamik lineer modeller üzerinde geliştirdikleri yaklaşımdır. Bu yaklaşım, bir değişkenin yerel düzeyinin, değişme oranının, değişkenin seviye atlayıp atlamadığını ya da süreksiz olduğunu belirlemek için Kalman filtreleme yöntemini kullanır. West and Harrison (1997), Harvey (1989), Durbin ve Koopman (2001), Rob J. Hyndman (2008), Petris G. (2009) ve'ın yapmış olduğu çalışmalar zaman serilerinin dinamik lineer modeller olarak ele alınması ve tahmin uygulamalarının gerçekleştirilmesine öncülük etmiştir [1].

Genel olarak bakıldığında DLM 'yi oluşturan denklem elemanları: F_t, (mxp) boyutlu t zamanında bilinen bağımlı değişkenler matrisi, v_t , (mx1) boyutlu gözlem hatası matrisi, θ_t , (px1) boyutlu süreç parametre vektörü, Y_t , (mx1) boyutlu süreç gözlem vektörü, w_t, (pxp) boyutlu bilinen sistem hatası matrisi, G_t, (pxp) boyutlu bilinen sabit geçiş(evrim) matrisi şeklinde olmak üzere bir DLM durum ve sistem denklemi olmak üzere iki eşitlikle ifade edilir. Sistem denklemi;

$$Y_t = F_t\theta_t + v_t \quad , \quad v_t \sim N(0, V_t), \quad t = 1,2 \ldots\ldots, \tag{1}$$

şeklindedir ve durum denklemi ise;

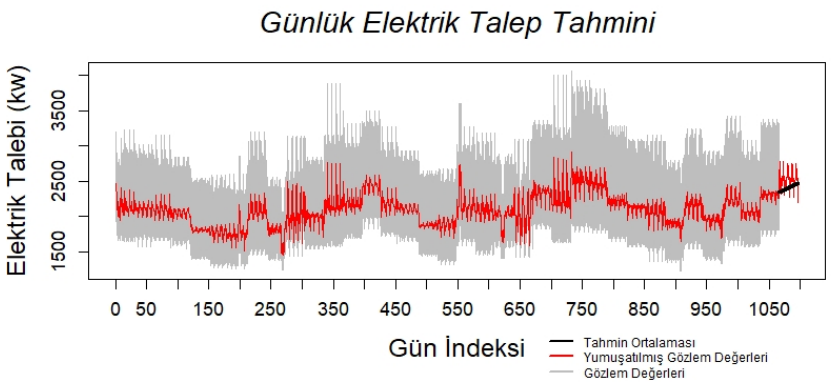
$$\theta_t = G_t\theta_{t-1} + w_t \quad , \quad v_t \sim N(0, W_t), \quad t = 1,2 \ldots\ldots, \tag{2}$$

olarak yazılır. Burada v_t , w_t birbirinden bağımsız normal dağılımlı, sıfır ortalama ve bilinen bir varyans matrisidir [3,4].

Denklem (1) gözlem denklemi olarak isimlendirilir ve θ_t durumları için Y_t' nin örneklem dağılımını tanımlamaktadır. Y_t terimi önceki gözlem ve parametre değerlerinden bağımsızdır. θ_t ve Y_t arasında dinamik ilişki mevcuttur. Gözlem denklemi hatası v_t başlangıç durumu için bilinmekle beraber zamanla değişme olasılığına sahiptir. Denklem (2) durum denklemi olarak isimlendirilmektedir. Durum vektörü θ_t' nin zaman içindeki gelişimini tanımlamaktadır. θ_{t-1}, G_t ve W_t sistemin bilinen girdi verileri olurken G_t, sistemin deterministtik geçiş matrisidir.

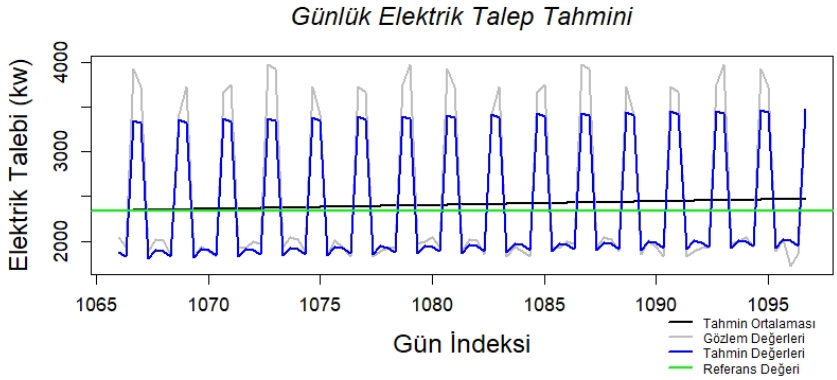
Bulgular ve Değerlendirme

Yapılan çalışmada Sakarya ili hanehalkı verileri kullanılarak elde edilen tahmin sonuçları aşağıda paylaşılmıştır. T1,T2 ve T3 zaman aralıkları ayrı ayrı tahmin edilmek yerine tahmin yönteminin başarısını belirtmek amacı ile tahmin modelinde birlikte ele alınmıştır. Grafiklerde tahmin edilen değerleri gösterilmiştir.



Şekil 2. Aralık-2017 için Elektrik Talep Tahmini ve Yumuşatılmış Gözlem Değerleri

Şekil 1.'de tahmin edilen Aralık ayının tahmin ortalama değerleri ve kurulan DLM modeline göre yumuşatılmış Kalman Smoothing gözlem verileri vardır. Grafikte görüldüğü üzere gerçek değerler ve yumuşatılmış veriler uyum içindedir.



Şekil 2. Tahmin Bölgesi için Tahmin Ort., Gözlem D. ve Tahmin D.

Şekil 2.'ye bakıldığında ise tahmin bölgesi olan 2017 Aralık ayı için elde edilen tahmin değerleri, tahmin ortalaması, gözlem değerleri ve tahmin ortalaması için referans çizgisi gösterilmiştir. Bakıldığında tahmin değerlerinin trende sahip olduğu görülmektedir. Kurulan DLM ile zaman serisinin sahip olduğu trend ve sezonsallık bileşeninin yanında rassal etkilerinde modellenmesi tahmin değerlerinin, gözlem değerlerine yakın bir dağılıma sahip olmasını sağlamıştır.

Tablo 1.'de elde edilen tahmin değerlerinin, Aralık-2017 ayı için gerçekleşmiş olan elektrik tüketim değerleri ile kıyaslanması sonucu ortaya çıkan tahmin hatası değerleri verilmiştir.

Tablo 1. Tahmin Değerleri & Test Verisi Tutarlılık Sonuçları

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Test Seti	124,3914	230,7329	154,637	3,4665	5,0836	0,3144	0,2105

Sonuç

Sakarya ili hanehalkına ait elektrik tüketim verileri kullanılarak gerçekleştirilen tahmin uygulamasından elde edilen sonuçlara ve test verilerine bakıldığında, zaman serilerinin dinamik lineer modeller aracılığı ile tahmininin etkin bir tahmin yöntemi olarak kullanılabileceği görülmektedir. Uygulamada günlük elektrik verisinin 3 farklı zaman dilimine ayrılarak bir aylık zaman dilimini kapsayan 93 adet (Aralık 2017) verinin tahmin işlemi gerçekleştirilmiştir. DLM'de tek değişkenli zaman serisi kullanılmıştır ve sonuçlara bakıldığında MAPE hata ölçüm yöntemine göre %5,08'lik bir hata oranı bulunmuştur.

References

[1] Petris G., Petrone S., Campagnoli P. , Dynamic linear models. In Dynamic Linear Models with R. , Springer , 2009, 978-0-387-77237-0.

[2] Karagöz K. (2004). Öngörü ve Zaman Serileri Analizinde Bayesyen Yaklaşım, İstanbul Üniversitesi, Sosyal Bil. Ens., Ekonometri Anabilim Dalı.

[3] Gürkan, H. Y. (2013). Markov Zinciri Monte Carlo Yönteminin Dinamik Doğrusal Modellere Uygulanması, Hacettepe Ü., Fen Bil. Ens., İstatistik Böl., Yüksek Lisans Tezi.

[4] Petris G., Petrone S., Campagnoli P. (2009). *Dynamic linear models. In Dynamic Linear Models with R.* Springer, 978-0-387-77237-0.

[5] Honjo, K., Shiraki, H., & Ashina, S. (2018). Dynamic linear modeling of monthly electricity demand in Japan: Time variation of electricity conservation effect. PloS one, 13(4), e0196331.

Multiclass Classification with Decision Trees, Naive Bayes and Logistic Regression: An Application with R

E. Bilgic[1], F. Esen[2]

[1]*Kayseri University, Kayseri, Turkey, emrahbilgic@kayseri.edu.tr*
[2] *Istanbul Medeniyet University, Istanbul, Turkey, fevzi.esen@medeniyet.edu.tr*

Abstract

The developments in Data Science have also enabled the emergence of data which can support the decision making processes of the companies. Data Mining (DM), which includes the processes of data acquisition, storage and analysis, has been successfully implemented in different business problems. In this study, the size of the firms that the shares belong to will be estimated by classification analysis by means of variables such as the type of the transaction, quantity and amount of the shares. This analysis is thought to be useful in exploring investor behaviors. In this context, three different classification techniques, Decision Trees (DT), Logistic Regression (LR) and Naïve Bayes (NB) were applied to a dataset of financial transactions using R programming language. As a result of the study, when the performance of the classification techniques was compared, it was found that the DT technique made 72 % correct classification.

Introduction and Purpose

Risk and cost are one of the most important factors that explain investor behavior in capital markets. Investors' fully and rapid access to information minimizes the negative effects of risk and cost in the markets. This also draws attention to the importance of investors to develop various strategies based on their own knowledge and experience, depending on the structure of the business, sector and the factors related to functions of the stock exchanges and the market conditions [1,2].

The data regarding cost of funding, profitability, weight of tangible assets, operational risk and the size of company which may affect investors' decisions are discussed in theoretical and empirical frameworks [3,4,5]. Considering the effects of company size on equity and asset structure, the cost of funding or debts and liabilities of a company are important factors that affect investor behavior. For example, while investors frequently execute their transactions for small companies in smaller periods, they perform transactions in wider periods in large companies. Besides, it is known that investors buy more stock in small companies and they tend to make more sales in large companies [6]. In addition, it is stated that the frequency of the purchase transactions decreases as the size of company increases, and investors sell twice as much as they purchase. Investors also perform larger volumes of shares in larger companies and the profitability of the shares increases with the transaction volume [6,7]. Therefore, it can be stated that the type, amount and size of the transactions are related to the size of the issuer of the stock and this relationship is one of the important components explaining the return from purchases and sales.

In this study, it is aimed to classify company sizes by using stock purchases and sales data. In this context, the performance of three different classification techniques is compared. Investors' transactions are obtained from the companies listed on USA exchanges. The data set includes 500.000 sales and purchases transactions.

Method: Classification Analysis

Classification analysis, which is one of the most commonly used techniques in the field of DM, aims to assign the correct class labels to all records assuming that the records in the data set belong to predetermined classes. For this purpose, the model is trained on a given training data set and in the following stage, trained model is used for classification. Therefore, classification appears as a supervised learning technique. In classification analysis the step of model selection is one of the most important step [8]. Examples of these models can be given as statistical, non-parametric statistical, artificial intelligence and mathematical programming [9]. In this study, DT, LR and NB techniques are applied to a financial data set and their classification performances are compared.

Application and Findings

In this study, investors' transactions at NYSE, AMEX and NASDAQ are obtained and analyzed. As independent variables: transaction type, amount of the transaction and the cost of the stock in dollars are selected. The dependent variable is market capitalization of the company (size of the company) on the date of the transaction. Based on the general classification of market capitalization in the literature, the companies are determined as "small" if the size of the company is below two billion dollars, "middle" if the size of the company is between two and ten billion dollars, "big" if the size of the company is above ten billion dollars. *party* package in R programming language is used for classification analysis with DT [10]. 70 % of the data set is divided into as training data and the remaining 30 % is as test data. The distribution of the companies in the training and test data sets according to their size is given in Table 1.

Table 1. Distribution of Company Size in Training and Test Data

	Big	**Small**	**Middle**
Training Data Set	139.733	105.094	104.990
Test Data Set	60.267	44.906	45.010

After training the data set, the estimation of the test data was performed by decision trees and regression technique. In Table 2, the confusion matrix obtained for company sizes is given. Accordingly, 48,207 of 60,267 big companies, 40,510 of 44,906 small companies and 19,193 of 45,010 medium sized companies are classified correctly in the test data. In this case, the classification accuracy was 72% which is a good ratio for further analysis. The precision values and recall values are also calculated and the precision values are respectively 68%, 84% and 60% for big, small and middle-sized companies while recall values are 80%, 90% and 42%. Precision values tell us that when the algorithm evaluates the transactions of the investors (variable values), 68% of the companies that it predicts as big companies are really big. Therefore, 32% of the transactions performed according to the algorithm are not similar to those performed for large companies. Same situation exists for middle-sized companies since the precision value is 60%. These values indicate that there seem some problems in investor behaviors which should be addressed since the transactions made by investors do not resemble the characteristics of investments made in large and middle-sized companies.

Table 2. Confusion Matrix

	Big	**Small**	**Middle**
Big	48.207	1.479	21.173
Small	2.656	40.510	4.644
Middle	9.404	2.917	19.193

Naïve Bayes classifier is also performed with *e1071* package in R programming language. The classification accuracy is 30% with this classifier. Since the accuracy is very low no need for further analysis. Multinominal logistic regression classifier is also performed using *nnet* package in R and the classification accuracy is very low which is calculated as 34%.

Results and Discussions

In this study, the performances of three commonly used techniques of classification analysis, one of the DM tools, were compared on a data set with dependent variable of three categories (small, medium, large). As a result, it has been found that DT technique provides higher classification accurate (72%) for the data set used in this study. Furthermore, considering that the type, amount and volume of transactions of the investors are important variables explaining the size of the company in which they conduct transactions, the correct classification of investor transactions is considered to be an important tool that can be used in the development of new investment strategies, which may also affect investor behavior. In this respect, it is recommended to apply different techniques to transform data into information for investment evaluation and strategy planning in financial markets. Furthermore 28% error rate indicate that there seem some problems in investor behaviors which should be addressed since the transactions made by investors do not resemble the characteristics of investments especially made in large and middle-sized companies.

References

[1] Duxbury, D., and Yao, S. (2017). *Are investors consistent in their trading strategies? An examination of individual investor-level data.* International Review of Financial Analysis, 52, 77–87.

[2] Ng, L., and Wu, F. (2006). *Revealed Stock Preferences of Individual Investors: Evidence from Chinese Equity Markets*, Pacific-Basin Finance Journal 14(2), 175- 192

[3] Bondt, W.F.M., and Thaler, R.H. (1995). *Financial decision-making in markets and firms: A behavioral perspective*, Handbooks in Operations Research and Management Science, 9, 385-410.

[4] Aoun, D., and Hwang, J. (2008). The effects of cash flow and size on the investment decisions of ICT firms: A dynamic approach, *Information Economics and Policy*, 20(2), 120-134.

[5] Gala, V., and Julio, B. (2016). Firm Size and Corporate Investment. Retrieved from http://repository.upenn.edu/fnce_papers/30

[6] Seyhun, H. (1998). *Investment intelligence from insider trading*, The MIT Press, England.

[7] Fauzi, R., and Wahyudi, I. (2016). The effect of firm and stock characteristics on stock returns: Stock market crash analysis, *The Journal of Finance and Data Science*, 2(2), 112-124

[8] Gerlein, E. A., Mcginnity, M., Belatreche, A., and Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach, *Expert Systems with Applications*, 54, 193-207.

[9] Peng, Y., Wang, G., Kou, G., and Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction, *Applied Soft Computing*, 11(2), 2906-2915.

[10] R Core Team (2013). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Opinions of Community Pharmacies About Probiotic Use

Y. Kılıçdağı[1], G. Özçelikay[1]

[1]*University of Ankara, Ankara, Turkey, yagmursfl89@hotmail.com*
[1]*University of Ankara , Ankara, Turkey, gozcelikay@ankara.edu.tr*

Abstract

Probiotics are usually microorganisms in the human intestine that are beneficial to health. They can be found everywhere along the gastrointestinal tract, where they play a significant and protective role. Probiotics are used for the prevention or treatment of some diseases. 80% of probiotics are marketed in Turkey from pharmacies. For this reason, this study was planned to be done in order to determine the opinions of pharmacists with community pharmacy in the city center of Ankara. The main purpose of this study is to determine the opinions of communıty pharmacists about probiotic use of patients. The results of this study verified the views of pharmacists on the use of probiotics and compared with studies conducted in various countries. As a result, there were conflicting views among healthcare team members, uncertainties about the guidelines and uncertainties about safety concerns, dosing or application of probiotic use.

Introduction

Probiotics are living microorganisms that positively affect the health of the host when taken orally in sufficient amounts.[1] There is now an increasing interest in the potential health benefits of probiotics by both health professionals and consumers. The market for food supplements containing probiotic strains is vast and is still growing. For example, probiotics are available in a variety of foods such as commercially available functional foods and beverages (dairy products, cereals, baked foods, fermented meat products, dry food probiotics), nutritional supplements (food supplements and nutrients).[2] It is very important to publicize the health benefits of food products through health professionals, educators, the media and the food industry.[3] The focus of most scientific research on this issue has been to understand the mechanism of action of probiotics and to determine the health benefits of probiotics. There are no studies on the factors affecting probiotic consumption by patients and consumers in the literature.

The knowledge of healthcare professionals directly affects the outcome of any treatment.[4] Pharmacists, the closest health advisor of the public, constitute an important part of healthcare providers. Few data are available on the perception of community pharmacists on probiotic and functional foods. The importance of health workers' knowledge of probiotics has been emphasized by many researchers before.[5] It is important that free pharmacists have sufficient knowledge about probiotics, which play an important role in the protection and promotion of health. In this study, the opinions of community pharmacists about probiotics were tried to be determined by a questionnaire. In addition, some sources of information were suggested to pharmacists by giving examples from the studies of probiotic use of pharmacists in the world.

Material and Metod

The sample of the study consists of commnity pharmacists registered to Ankara Chamber of Pharmacists in Ankara city center. Data were collected by conducting a face-to-face questionnaire application to pharmacists located in Ankara city center. The population of the study consists of 1992 community pharmacists in Ankara province registered in Ankara Chamber of Pharmacists. Since the generalizability of the research findings is important, simple random sampling method was used to determine the sample.

In this study, a questionnaire containing information about the probiotic content, usage and behavior of community pharmacists was compiled as a face to face interview method. The questionnaires containing informed consent were left to pharmacists and the forms were collected after filling. Fifty-four volunteer pharmacists participated in the study. All data collected were entered electronically. SPSS 25.0 package program was used to evaluate the data.

Results

According to the results of the study, 80% of the pharmacists who participated in the survey were self-employed for more than 15 years. Of the pharmacists, 61% owns a district pharmacy, 23% across the hospital, and 15% owns a pharmacy in other places. 46.2% of the pharmacists surveyed have probiotics in their pharmacy. 53.6% of the beneficiaries recommend probiotics.

88% of those who came to the pharmacy and demanded probiotics are women, and 35% prefer probiotics more than age. It was determined that 84.6% of probiotic users were not health workers. It is seen that the price range of the most consumed probiotics is between 51-100 TL and pharmacists obtain information about probiotics from the companies by 30.8%. Pharmacists stated that most beneficiaries applied to probiotics because of constipation.

Discussion and Conclusion

In this study, it was found that probiotic use is common among pharmacists, and most participants have more or less knowledge about the concept of probiotics. In a study conducted in India, it was reported that approximately 93% of healthcare workers knew the term probiotic, and about 80% of respondents correctly identified probiotics.[6] Other studies in the United States, Europe and Asia have shown that the term probiotic is widely known among health care providers. Soni et al. Stated that 85% of the doctors, 62.5% of the nutritionists and 69.4% of the pharmacists had good knowledge about the probiotic doses to be used, their mechanism of action and their effects on health.[7]

In conclusion, parallel to the increasing number of studies showing the positive effects of probiotics on health, it is necessary to inform about the rational use of probiotics. Considering that physicians and pharmacists are the main factors in transferring and applying this information, it is thought that existing information should be supported through various trainings.

References

[1] Arora, M., Sharma, S. and Bald, A. (2013). Comparative insight of regulatory guidelines for probiotics in Usa, India and Malaysia: a critical review. *International Journal of Biotechnology for Wellness Industries*, 2, 51-64.
[2] Kaur, P.S. and Pandit, R.K. (2016). Probiotic products in Punjab region. *International Journal of Applied Biology and Pharmaceutical Technology*. 7(1), 154-160.
[3] Diplock, A.T. (1999). Scientific concepts of functional foods in Europe consensus document. *British Journal of Nutrition*; 81, 1-27.
[4] Bjerrum, M., Tewes, M. and Pedersen, P. (2011). Nurses' self-reported knowledge about and attitude to nutrition before and after a training programme. *Scandinavian Journal of Caring Sciences*; 26(1), 81-89
[5] Oliver L, Rasmussen H, Gregoire MB, Chen Y. (2014). Health care provider's knowledge, perception and use of probiotics and prebiotics. *Topics in Clinical Nutrition*; 29(2), 139-149.
[6] Menrad K. (2003).Market and marketing of functional foods in Europe. *Journal of Food Engineering*; 56(2/3), 181-188.
[7]Richa Soni, Kamna Tank, Nayan Jain. (2018). Knowledge, attitude and practice of health professionals about probiotic use in Ahmedabad, India. *Nutrition & Food Science*; 48(1), 125-135.

A General Evaluation on Drug Distribution and Automation Systems in Pharmacy

B.Kıran[1] N.Şencan[2]
[1]Ege Üniversitesi, İzmir,Turkey, kiran.bulent@gmail.com
[2]Acıbadem Ünivesitesi, İstanbul,Turkey, nazlisencan@gmail.com

Abstract

In this study, the advantages and disadvantages of main automation systems have been examined, which have a tendency to become widespread in drug distribution in hospitals around the world. It was determined that the costs of automation systems are high; on the other hand, these are effective in reducing the workload of pharmacists and hospital accreditations, and also in preventing drug waste. It was also determined that the automation systems enable pharmacists to make drug counseling more effective by saving time for drug distribution, and increase patient safety by reducing the risks associated with medication errors [1]. It was evaluated that determining the size of pharmaceutical automation systems according to the number of beds in hospitals will be more affordable in terms of cost efficiency. In pharmaceutical services where technology usage is increasing, it is considered to be beneficial to include pharmaceutical automation systems in undergraduate, graduate and on the continuing education programs.

Introduction

The use of technology in pharmaceuticals and pharmacy services is becoming increasingly common around the world. The aim of the study is to evaluate drug automation systems used in hospital pharmacies in the world, their advantages and disadvantages and the relevant literature, and to draw attention to the importance of the subject. Problems arising from drug management processes, including many procedures such as purchase, storage, distribution of the drugs, and also the administration of the drugs safely to patients in clinics lead to frequently encountered medical malpractices in hospitals [1].

On the other hand, hospitals should be accredited in order to engage international health tourism. In addition, canceling out the medical malpractices for drug distribution and administration is one of the most important criteria for accreditation requested by the International Accreditation Institution- JCI (Joint Commission International) [1]. In this context, drug automation systems are today becoming increasingly common in internationally accredited hospitals [1-3].Automation systems reduce the workload of pharmacists, increase the patient safety and allow for more effective counseling for patients and health care professionals on their rational drug use [1-5].On the other hand, it is reported that the use of drug automation systems is positively affected the quality of service in community pharmacies and that this technology-based system of the future will become common in both community and hospital pharmacies [4-5].

Systems Used for Automated Drug Distribution and Unit Dose Drug Management in the World

Although the systems that deliver unit dose drugs are designed differently according to various brand characteristics, they basically work on almost the same principle. Briefly; it is an automatic cabinet including touch screen, monitor, keyboard and safe storage spaces. A user ID and password are required to access this storage. First, the user selects a patient profile from the list and the medical treatment that should be applied to that patient after entering its user ID and password. Then the distribution cabin locks the drawer containing the drug, and when it is time for the drug administration it allows the user to access the drawer. If the drawer does not close properly and in case of undesired situations, the system warns by signaling.

As a result, drugs prepared according to the treatment plan directives prescribed by the doctor become available for the administration of nurse or relevant health care personnel. Automatic drug distribution cabinets are generally found in nursing stations, emergency services and surgery rooms [1].

The most widely known unit dose drug automation systems in the world are Pyxis, Swisslog and Rowa Speedcase automation systems. Their advantages and disadvantages can be summarized as follows [6-8] ;

Advantages and Disadvantages of Pharmaceutical Automation Systems:

Advantages of Pyxis Automation System [6].

1. Fast access to first dose of the drug
2. Shifting the workload in the pharmacy to the floors of the clinics
3. Only authorized persons can access the drug by fingerprint verification
4. Able to make stock tracking
5. Able to receive all reports about the drug
6. Ensuring access to the right drug at the right time

Disadvantages of Pyxis Automation System [6].

1. Obligation to perform deblister
2. Drug wastes during deblister
3. The drug expires within 2-3 weeks due to deblister,
4. Medication errors as there is more than one drug in each drawer during drug intake
5. Bringing extra workload to nurses
6. Storage of parenteral forms, without barcodes, such as ampules, vials, syringes
7. Failure to make lot tracking as unit dose
8. Errors that may result from restoring the returned drug
9. Failure to check the expiry date of the same medicines in the drawer
10. Data transfer problems
11. Barcode verification of tablet drugs only at the bedside
12. Failure to completely prevent medication errors
13. Unable to save labor in the pharmacy or on the floors (on the contrary, extra labor is required)

Advantages of Swisslog system [7].

1.Ability to barcode each type of drug as unit dose

2.Able to write patient and drug information on the bags delivered by the device

3.Unit dose stock tracking of all kinds of drugs

4.Able to receive all reports about the drug

5.Ensuring access to the right drug for the right patient

6.Drug preparation process is automated

7.The expiry date of the drugs as unit dose can be checked

8.Ensures savings by having to return unused drugs to the pharmacy

9.Prevents unnecessary stock on the floors because unit dose drug is prepared for the patient

10.Significantly reduces the labor requirement (60%)

11.Minimizing the waste of time by reducing the workload in pharmacies and nurses

12.No need for deblistering the drug

13.Due to the fact that there is no need to deblister, the drug expiry date remains the same as the expiry date during production

14.No loss of data transfer due to its central structure

Disadvantages of the Swisslog system: Large volume drugs such as serum and formula cannot be prepared as unit dose for the patient. [7].

Advantages of Rowa Speedcase System [8].

1- Reduction in department expenses

2- Reduction in medication errors

3- Reducing drug distribution times to patients

4- Ensures time to staff for direct patient care

5- Increased service reliability

6- Easier stock control

Disadvantages of Rowa Speedcase System [8].

Rowa Vmax system cannot make unit dose packaging and send drugs to the clinic.

Today the most important medical malpractice and economic losses in hospitals are drug-induced errors. As well as the advantages of the proposed drug automation systems instead of the classic drug delivery systems, due to their contribution to accreditation processes that facilitate international patient admission of hospitals, there are also major disadvantages according to their brand and model. In addition, while the need for qualified personnel to use these systems is increasing, it is suggested that it will increase unemployment problem due to drastically reduce the need for existing labor force.

In addition, it is thought that training of pharmacists on drug automation systems, working in hospitals affiliated to the Ministry of Health with the largest drug consumption in the public sector and the establishment of these systems in city hospitals with large number of patient beds are important issues in terms of both preventing economic losses and reducing medical malpractices.

Conclusion and Recommendations

Although automated systems for unit dose drug and medical equipment distribution are costly, there are increasing claims that these systems are highly increased the patient safety and ensured savings with very good planning and dedicated solutions to the needs of each hospital. As hospitals may prefer a fully automated pyxis or swisslog system, which includes the high-cost complete automation, it is also possible for them to move into this system step by step according to the hospital characteristics and budget facilities These steps can be performed step by step with storage systems, and automation systems for unit tablet dosage only and so on [1-11]. The usage in private hospitals targeting to receive accreditation in the future as well as large-scale university and public hospitals can be expected to become widespread. Therefore, it is considered to be beneficial to include pharmaceutical automation systems in undergraduate, graduate and on the continuing education programs.

References

[1] Kıran, B. (2014). Evaluation of Pharmaceutical Automation Systems, KİTDER. *Journal Yuklenim*, 10(14), 11-6.

[2] Black, A. Brice, S. Hardy, L. Longshaw, R.(2006). Validation of cold storage shelves in an automated dispensing system. *Hospital Pharmacist,* (13), 372.

[3] Brice, S. , Hardy, L. , Longshaw, R.(2006). Evaluation of automatic loading devices with a ROWA speedcase system. *Hospital Pharmacist,* (13), 375-378.

[4] Lauren B. A., Dale B.C., Stefanie P. F (2005). Impact of Community Pharmacy Automation on Workflow, Workload, and Patient Interaction**.** *JAPhA***,** 45(2), 138–144.

[5] James O. B.(2010).Analysis of the Medication Management System in Seven Hospitals; medBPM Whitepaper, *CareFusion research*.

[6] The Unit Dose System, "PYXİS"; (Access:12.05.2019). http://www.simeks.com.tr/haberler/birim-doz-ilac-uygulayan-sistem-pyxis.htm

[7] Pharmacy Automation System Swisslog, (Access:1205.2019), http://www.swisslog.com/hcs-pharmacyautomation.pdf

[8] Goundrey-Smith S. (2008). Pharmacy robots in UK hospitals: the benefits and implementation issues. *Pharm* J; 280, 599-602.

[9] Analysis of the Medication Management System in Seven Hospitals (2008). medBPM Whitepaper, CareFusion research,

[10] Chan E.,A.Ramudhin (2007).An Evaluation of Various Business Process Modeling Techniques Applied to Healthcare.*ISEM 07*,Beijing.

[11] Şardan, Y.Ç. (2008). Pharmaceuticals and Materials Management Roadmap. *Journal of Medical Sağlık Ve Tıp Tekn.* (77), 1-4.

Analysis and Price Forecasting the Used Cars Property with Multiple Linear Regression in Turkey

Y.Yiğit[1], E.Avcı, M.Karabatak[3]

[1]*Bitlis Eren University , Bitlis, Turkey, yyigit@beu.edu.tr*
[2] *Firat University, Elazig, Turkey*, enginavci@firat.edu.tr
[3] *Firat University , Elazig, Turkey, muratkar@hotmail.com*

Abstract

Data generated by computer systems are worthless, because they do not make sense when viewed with the naked eye [1, 2]. When this data is processed for a particular purpose, it starts to make sense. Therefore, it is important to use techniques that can process large amounts of data. Transformation of this raw data into information or meaning can be done with data mining. Data mining includes techniques for identifying or estimating meaningful information that is not previously known [1, 3]In the study, 720 cars used in the internet and advertisements were used. The data set created with 720 adverts was analyzed by multiple linear regression and a model to make a price estimate for second hand ads was tried to be obtained.

Introduction

Nowadays, the amount of information on the earth doubles every 20 months with the beginning of the storage of data in digital media, and the number of databases is increasing at a similar or even higher rate. [4, 5]. As a result of the low cost of high-capacity processing, data storage has become easier and the data itself has become cheaper.

With the increase in electronic commerce and online shopping mechanisms, which are becoming widespread today, the studies of companies competing in this field highlight the importance of data mining. [6, 7]. With the increase of digital data in recent years and the storage of these data in large databases, there is a need to benefit from this data in the most efficient way. Data mining is mainly examined under two main headings. The first one is the predictive used for the estimation of the data whose results are unknown and the other is the descriptive which enables the identification of the data at hand. [8, 9, 10, 11]. Data mining models can be analyzed under three main headings according to the functions they see. These;

- Classification and Regression,

- Clustering,

- Association rules and Sequential time patterns

In statistics, linear regression is a linear approach to model the relationship between scalar response and one or more explanatory variables. The state of an explanatory variable is called simple linear regression. For multiple descriptive variables, the process is called multiple linear regression. In multiple linear regression, the aim is to explain the total change in the response variable using descriptive variables or regressors. [12, 13].

Multiple linear regression model for sample,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \;\; i = 1,2, \dots, n \qquad (1)$$

It is defined as. Where I = 1,2, .. n and j = 1,2, k. Value; xij, j. explanatory variable i. level value; β_j, j. the regression coefficient and ε_ (i), i. indicates the error term. ε_i ~ N (0, σ ^ 2) has normal distribution. The multiple linear regression model contains k explanatory variables or regressors. In this case, 2^k candidate models can be formed by using the available regressors for the selection of the best regression medele. In the case of multiple linear regression models, if the number of explanatory variables or regressors is small, classical or stepwise methods are used to select the best regression model. There are two goals in choosing the best model. First, the generated model is requested to include all possible regressors. Thus, the information content in these factors affects the generated response values. Second, it is desirable that the model generated contains a minimum number of regressors. Because the variance of the generated response values increases with the increase in the number of regressors. In addition, increasing the number of regressors in the model means more data collection.

MATERIAL AND METHOD

In this study, 720 used car advertisements and 36 properties belonging to these advertisements were used. Linear regression method was used to analyze the variables and meaningless variables were removed from the dataset. 15 significant variables were used to create a model and a model of these variables was created. The data used is shown in table1.

Table 1. Car Classifieds and Features

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	Y
1989	2	1	200.000	0	0	0	0	0	0	0	0	0	0	0	10.000
1989	2	1	243.000	0	0	0	0	0	0	0	0	0	1	0	11.000
1990	2	1	220.000	1	0	0	0	0	0	0	0	0	0	0	11.500
1990	1	1	237.000	1	0	0	0	0	0	1	0	0	1	1	12.400
1989	1	1	328.000	0	0	0	0	0	0	0	0	0	0	0	12.500
1991	2	1	300.000	0	0	0	0	0	0	0	0	0	0	0	12.750
1990	2	3	290.000	0	0	0	0	0	0	0	0	0	1	0	13.000
1990	2	1	300.000	1	0	0	0	0	0	0	0	0	0	0	13.500
1990	2	1	290.000	1	0	0	0	0	0	0	0	0	0	0	13.500
1990	2	1	278.000	1	0	0	0	0	0	0	0	0	0	0	13.500
1990	1	1	120.000	1	0	0	0	0	0	0	0	0	0	0	13.750
1990	2	3	265.000	1	0	0	0	0	0	0	0	0	0	0	14.000
1990	1	1	1.000	1	0	0	0	0	0	0	0	0	0	0	14.000
1989	1	1	294.000	0	0	0	0	0	0	0	0	0	0	0	15.000
1990	2	1	208.000	0	0	0	0	0	0	0	0	0	0	0	15.500
1988	2	1	290.000	0	0	0	0	0	0	0	0	0	0	0	15.500
1990	2	1	90.000	0	0	0	0	0	0	0	0	0	0	0	16.000
1990	2	1	225.000	1	0	0	0	0	0	0	0	0	1	0	16.000
1990	2	1	303.000	0	0	0	0	0	0	0	0	0	0	0	16.000

Y = Price X1 = Model Year X2 = Fuel Type X3 = Gear Type
X4 = Speedometer X5 = ABS X6 = EBA X7 = EBD
X8 = BAS X9 = YOKUSKALKISDESTEGI X10 = ARMORED CAR
X11 = AIR BAG (DRIVER) X12 = Fatigue Detection System
X13 = ISOFIX X14 = CENTRAL LOCK X15 = IMMOBILIZER

Result And Discussion

The dataset used was analyzed using linear regression and the model in Equation 2 was obtained.

$$Y = 2626.0973 * X1 + -4291.9349 * X2 + 2703.6757 * X3 + -51.5241 * X4 + 3584.5705 * X5 + -14778.2157 * X6 + 14301.5509 * X7 + -17144.8495 * X8 + 9051.5038 * X9 + 54720.7231 * X10 + -2935.3447 * X11 + 11535.607 * X12 + 2495.2702 * X13 + 4267.8106 * X14 + -8349.5943 * X15 + -5191156.7623 \quad (2)$$

In the model obtained, the coefficients show how much the given variables affect the price. Some of the coefficients are negative. Negative coefficients mean that the price has negative effects. Figure 1 shows that there is a direct proportional relationship between the model year and price of the car in the obtained model.

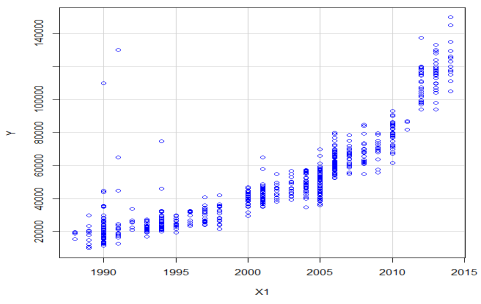


Figure 1. Relationship between X1 and Y

When the graph in Figure 1 is analyzed, it is seen that the prices of the vehicles increase as the model year approaches. The model year is an important factor affecting the price.

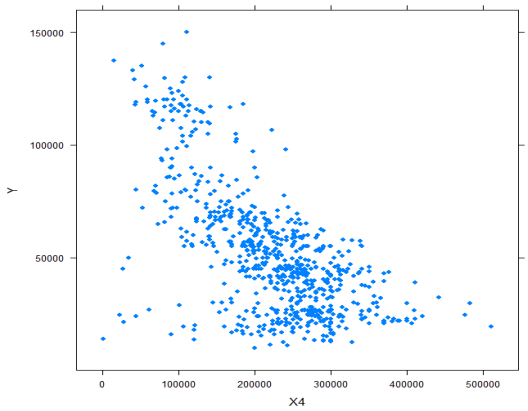


Figure 2. Relationship between X4 and Y

Figure 2 shows the graph showing the relationship between the current mileage and prices of vehicles. When the graph is examined, it is seen that there is an inverse ratio between price and vehicle mileage. In other words, the price of the vehicle decreases as the mileage increases.

All data and graphs of this study could not be placed due to limitations. With the help of the model created in the study, an estimated price is created according to the features mentioned in the future announcements. While analyzing used car advertisements on the internet, price analysis can be done with the help of the model. In addition, with the help of the estimated price generated with the help of the model, when looking for used car ads, it is possible to classify the prices as low and normal according to the prices.

References

[1] Karabatak, M. ve İnce, M.C. (2004). Apriori Algoritmasi ile Öğrenci Başarisi Analizi.
[2] Koç, M. ve Karabatak, M. (2012). Sosyal Ağlarin Öğrenciler Üzerindeki Etkisinin Veri Madenciliği Kullanilarak İncelenmesi, *e-Journal of New World Sciences Academy*, 155-164.
[3] İyi, P. ve Erol, H. (2008). Selection Of The Best Model In Multiple Linear Regression,Yüksek Lisans Tezi, 2008.
[4] Ç. Kurt, Ç.ve Erdem, O. A. (2012). Öğrenci Başarısını Etkileyen FaktörlerinVeri Madenciliği Yöntemleriyle İncelenmesi, *Politeknik Dergisi*, 111-116,.
[5] Batan, M. (2012). Diyarbakır ile Batman İlleri İklim Verilerinin Lineer Regresyon ile Karşılaştırılması ve Ilısu Barajı Sonrası Batman İlinin Gelecek İklim Verilerinin Elde Edilmesi, *Batman Üniversitesi Yaşam Bilimleri Dergisi*, 225-232.
[6] Tekin, A. ve Öztekin, Z. (2018). Eğitsel Veri Madenciliği ile İlgili 2006-2016 Yılları Arasında Yapılan Çalışmaların İncelenmesi, Educational Technology Theory and Practice, 67-89.
[7] Çodur, M. Y., Tortum, A. ve Çodur, M. (2013). Genelleştirilmiş Lineer Regresyon ile Erzurum Kuzey Çevre Yolu Kaza Tahmin Modeli, Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 79-84.
[8] Terzi, S. (2012). Hile ve Usulsüzlerin Tespitinde Veri Madenciliğinin Kullanımı, *Muhasebe ve Finansman Dergisi*, 51-64.
[9] Ozgan, E. (2008). Karayolu Araç Tipi ve Kaza Şekli ile Kaza Sonuçları Arasındaki İlişkilerin Analizi, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 97-104.
[10] Özkul, F. U. ve Pektekin, P. (2009). Muhasebe Yolsuksuzlarının Tespitinde Adli Muhasebecinin Rolü ve Veri Madenciliği Teknınlerinin Kullanılması, *World of Accounting Science*, 57-89.
[11] Yalçın Ateş, M. K. (2017). Nicel Birliktelik Kuralları İçin Çoklu Minimum Destek Değeri, *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 57-65.
[12] Albayrak, A. S. ve Yılmaz, S. K. (2009). Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 31-52.
[13] Vahaplar, A. ve İnceoğlu, M. M. (2001). Veri Madenciliği ve Elektronik Ticaret, Türkiye'de İnternet Konferansları, İstanbul.

Support Vector Machines and Logistic Regression Analysis on Predicting Financial Distress Model

S. Doğan[1], D. Koçak[2], M. Atan[3]

[1] *Karamanoğlu Mehmetbey University, Karaman, Turkey, dogans@kmu.edu.tr*
[2] *Ankara Hacı Bayram Veli University, Ankara, Turkey, deniz.kocak@hbv.edu.tr*
[3] *Ankara Hacı Bayram Veli University, Ankara, Turkey, murat.atan@hbv.edu.tr*

Abstract

Financial distress and bankruptcy that goes along with it are rather costly and destructive processes for all units of economy. Predicting financial distress in advance is critically important not only for the general functioning of economy, but also for firm shareholders, investors and creditor institutions. The aim of the study is to develop a useful and significant prediction model using Support Vector Machines (SVM) to predict the future success or distress situations of firms. To test the durability of the method, the Logistic Regression Analysis (LRA) has been preferred. This method has also been taken as a feature selection technique to increase the classifying performance of SVM and thus the rates that best reflect the uncertainty level have been determined. Another factor that contributes to the classifying performance of SVM is to optimize the parameters. To do so, the Grid-search technique, which is effective and easy to apply, has been preferred. According to empirical results, both SVM and LRA achieved a quite high level of prediction accuracy rate. However, the classifying performance of SVM is better than the logistic model, as is expected. In addition, SVM model which has been run by the highly informing new feature sub-set obtained from the LRA has turned out to be the model with the best performance.

Financial Distress and Background

Financial distress might occur due to a large spectrum of external and internal factors from the general economic conditions to poor decisions made by the production and management departments of enterprises. This issue, which is referred to as financial distress, bankruptcy, and business failure prediction, has been the focus of several studies and the process of financial distress of businesses has been defined in various ways. In the classical literature, financial distress has been defined as the case when an enterprise cannot pay currently due liabilities, when an enterprise go bankrupt, or when the business is put into run-off upon the request of creditors (Refs. [4, 2, 8, 13]). Insolvency is a situation in which the liabilities of an enterprise exceed its assets; bankruptcy, on the other hand, is the final stage of financial distress and it is the situation in which an enterprise cannot at all meet its liabilities and this is evidenced by the bankruptcy court and the enterprise is liquidated or restructured. Among the causes of a firm's financial distress that have been put forth in recent literature are loss-making for two consecutive years or the case that the net active value per share drops below the book value per share, and the performance of the firm's stock compared to the general index of the stock market in which it is traded (Refs. [3]). The concept of financial distress in enterprises requires a thorough multi-dimensional analysis in which all relevant factors should be considered together. Financial distress prediction is of critical importance to take timely precautions for business managers, investors, creditors, regulatory bodies, independent auditors, and employment agencies (Refs. [1]).

Financial distress has long been one of the important subjects for research on finance. Several solutions have been proposed for this frequently studied issue through statistical methods with single/multiple variables, mathematical modeling, and, recently, Artificial Intelligence (AI) techniques. The use of statistical models in studies on financial distress prediction dates back to the 1960s (Refs. [4, 2]). After 1960s, several multiple-variable models such as the multiple-regression analysis developed by Meyer and Pifer (Refs. [11])., LRA developed by Ohlson (Refs. [13]), and the probit model proposed by Zmijewski (Refs. [15]). were started to be implemented one after another. Due to the strict assumptions of traditional methods, AI techniques are used such as Inductive Learning (Refs. [10]), Rule-Based Learning (Refs. [6]) and Artificial Neural Networks (ANN) (Refs. [5]). Min and Lee (Refs. [12]). were first to apply SVM to bankruptcy prediction problems. According to the results of the study, SVM, which yields similar or better results compared to ANN with back propagation algorithm, can learn through relatively fewer training sets.

Support Vector Machines

For the data that cannot be separated linearly, the soft margin optimization can be applied. In this case, the problem can be linearly separated by accepting a certain error to be assigned for incorrectly classified samples. The problem is formulated as finding the hyperplane that minimizes the training errors through slack variables:

$$\underset{w,b,\xi}{Min}\ \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i$$
$$subject\ to: y_i(\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i \ and\ \xi_i \succ 0,\ \ i = 1,...,l \tag{1}$$

Here, C is the penalty parameter on training errors, and ξ_i is non-negative slack variable. This optimization problem can be solved using the Lagrange function. The dual model, in which the Lagrange variables are maximized, is given below:

$$\underset{\alpha}{Max}\ L_d(\alpha) = \sum_{i=1}^{m}\alpha_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$
$$subject\ to:\ \sum_{i=1}^{n} y_i \alpha_i = 0\ \ and\ \ 0 \leq \alpha_i \leq C,\ \ i = 1,...,m \tag{2}$$

For non-linear SVM, the mapping function ϕ is used for training samples. Designating an appropriate kernel function based on inner product so that data could be linearly separable is generating a non-linear transformation of data from the input space to a high dimensional (it could also be unrestricted) feature space. The Kernel function, $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, uses the space of the inner product. Then, the decision function is obtained by $f(x) = y = sig\left(\sum_{i=1}^{m} \alpha_i^* y_i \langle K(x_i, x_j) \rangle + b^*\right)$. There are a lot of kernel functions that support SVM to find the optimal result. The most commonly used of this function is radial basis kernel $K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$. (For detailed explanations about SVM, please refer to Refs. [9, 15, 7]).

Application

Different model propositions have been made both to test the success of the method and to strengthen the method. The proposed models, LRA, SVM and SVM with feature selection by LRA. The results of these models are discussed and all of the methods are compared.[5]

Data Set

The firms that we are going to predict the financial distress are operating in the manufacturing industry and its sub-sectors, (172 of those firms), and they are traded in the BIST stock market. 24 different financial ratios in 6 groups (**Growth Rates:** Asset growth, net sales growth, share equity growth; **Valuation Ratios:** Market value/book value, market value/net sales; **Operating Ratios:** Stock turnover, accounts receivable turnover, asset turnover; **Financial Structure Ratios:** Fixed assets/assets, short term loans/assets, short term loans/share equity, short term loans/total loans, share equity/assets, total loan growth, loan capital ratio; **Profitability Ratios:** Real operating profit margin, net profit margin, return on assets, profit capital, gross real operating profit margin; **Liquidity Ratios:** Current ratio, quick ratio, current assets/total assets) comprise the sampling set of the study. In the study, the success or distress situations of the firms have been used as classifying variables. 71 of the firms in our data set are classified as distressed firms and 101 as nondistressed ones. The whole data set was divided into two groups in order to test the validity of the training and the developed model. The data set has been randomly divided into two as 80% training and 20% test sets.

Study Design and Empirical Results

In the present study only the radial basis kernel function is going to be used since it would be an effective choice to use RBF for SVM to find the optimal result. Grid search technique has been preferred to optimize C and gamma parameter. In this technique, the value space for C and gamma has been divided into 110 parts. While the limits for C is $[2^{-5},2^{15}]$, for gamma the limit changes in the $[2^{-15}, 2^{3}]$ interval. The optimal parameter C and gamma are determined by *k*-fold cross validation technique. The classification accuracy rate of SVM is, also certain factors such as the quality of the data set. Leaving an important feature out of the model by chance might decrease the accuracy rate of the classifier. However, the existence of certain features in the data set might not at all affect the solution results or it might include a high amount of noise. LRA was chosen as the feature selection technique based on the approach proposed by Min and Lee (2005). To evaluate the performances of the models, accuracy rate, sensitivity, specificity, precision, and Matthews correlation coefficient (MCC) criteria have been used. In Model 1 (***GridSVM***), all variables are included in the analysis. In Model 2 (***Logit***), too, all of the determined financial ratios have been

[5] SVM model developed in the study has been designed using MATLAB 9.4 and LIBSVM (Chang and Lin, 2011) software system. For LRA, IBM SPSS Statistic-21 package program has been used.

used to do LRA. Finally in Model 3 (***Logit* + *GridSVM***), LRA has been seen as the feature selection technique and the analysis was done using SVM model with the feature sub-set that was found to be significant and would provide useful information.

The empirical results of the application will be examined under the following framework: LRA outputs, GridSVM and GridSVM feature selection model output, the performances of the proposed model. The empirical results of the logistic regression model are given in Table 1. As shown in Table 1, asset growth, real op. profit margin, net profit margin, gross real op. profit margin, current ratio, quick ratio are variables (financial ratios) of *Logit* model and B is the parameter of this model. Classification performance of *Logit* model was discussed in Table 3. These financial ratios, also use as a feature subset for *Logit* + *GridSVM* model.

Table 1. Logistic Model Outputs

	B	S.H	Wald	s.d.	*p*-value	Exp(B)	95% confidence interval	
							Low.	Upp.
Asset Growth	-1.761	.408	8.664	1	.000	.172	.077	.382
Real Op. Profit Margin	-1.947	.767	6.453	1	.011	.143	.032	.641
Net Profit Margin	-1.750	.772	5.136	1	.023	.174	.038	.789
Gross Real Op. Profit Margin	.746	.377	3.916	1	.048	2.108	1.007	4.412
Current Ratio	1.465	.705	4.310	1	.038	4.326	1.085	17.242
Quick Ratio	-2.728	1.096	6.197	1	.013	.065	.008	.560
Constant	-1.031	.355	8.455	1	.004	.357		

The theoretic model represents the data and the model is fully significant according to the statistical results of the prediction model -2LogLikelihood=86.949; chi-square=12.493; s.d=8; p=0.131). In the statistical results regarding the coefficients (chi-square=100.654; s.d.=6; p=0.000) the coefficients have been concluded to be significant.

The analysis outputs given in Figure 1 shows how SVM affects the results of the SVM parameters (C and gamma).

(a) (b)

Figure 1. (a) GridSVM and (b) Logit + GridSVM models cross validation rates

Figure 1(a) shows that the highest accuracy rate is GridSVM results with 87.21%. This rate has been obtained when for C 2^{11} and for gamma 2^{-13} values have been assigned. In Table 2, the impacts of performing feature selection on SVM results are summarized.

Table 2. Empirical Results Regarding GridSVM and GridSVM with Feature Selection Models

Models	Accuracy Rate			Cross-Validation Rate		
	mean	s. d.	max	mean	s. d.	Max
GridSVM	83.28%	(± 0.0597)	90.63%	70.39%	(± 11.25)	87.21%
Logit+GridSVM	85.44%	(± 0.5520)	93.75%	74.80%	(± 11.63)	90.06%

Table 3 presents the results of the performance criteria chosen for the study. These results are the performance criteria of the classifier that gives the best value in all iterations (100 iterations) for each model. It has been seen that the logit model takes the lowest value in other performance criteria as well. However, LRA provides important information regarding the selection of the new feature sub-set and it increases the performance of SVM run by this new feature sub-set.

Table 3. Performance Criteria of the Proposed Models

	GridSVM		Logit		Logit + GridSVM	
	Training	Test	Training	Test	Training	Test
Accuracy	0.9282	0.9063	0.9000	0.8000	**0.9424**	**0.9375**
Sensitivity	0.9310	0.9000	0.8545	0.6153	**0.9375**	**0.9444**
Specificity	0.9268	0.9091	0.9294	**0.9411**	**0.9452**	0.9286
MCC	0.8539	0.7896	0.7893	0.6018	**0.8831**	**0.8730**
Precision	0.9000	0.8182	0.8867	0.8888	**0.9615**	0.9444

Conclusion

In the present study, distress prediction has been made by SVM. The two important parameters of SVM, C and gamma, have been optimized by grid search technique and it has been shown how the accurate identification of this parameter pair affects the results. Furthermore, feature selection for SVM is another factor which considerably influences the results. LRA is selected as a feature selection method. The results of the study have shown that parameter optimization and the feature selection have contributed positively to the classification performance of SVM. Financial distress prediction has been made based on a real data set of firms operating in the manufacturing industry in Turkey. The proposed models have been compared based on this data set. The results obtained from LRA are satisfactory. However, SVM has achieved a higher level of success. The superiority of SVM has once again been proved by the present study. When the results of the proposed models are compared, it has been seen that the classifying success of SVM, which allows both for parameter optimization and for feature selection, is higher. In brief, an effective early-warning model has been developed via SVM, which is a newly developed machine learning technique for financial distress prediction problem.

References

[1] Alifiah, M. N. (2014). Prediction of financial distress companies in the trading and services sector in Malaysia using macroeconomic variables. *Procedia-Social and Behavioral Sciences*, 129, 90-98.

[2] Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance,* 23 (4), 589-609.

[3] Altman, E. I., Zhang, L. and Yen, J. (2007). *Corporate financial distress diagnosis in china.* New York University Salomon Center Working Paper.

[4] Beaver, W. (1966). Financial ratios as predictors of failure, journal of accounting research. *Empirical Research in Accounting: Selected Studies,* 4 (1), 71-111.

[5] Boritz, J., and Kennedy, D. (1995). Effectiveness of neural networks types for prediction of business failure. *Expert Systems with Applications,* 9 (4), 503-512.

[6] Bryant, S. M. (1997). A case-based reasoning approach to bankruptcy prediction modeling. *International Journal of Intelligent Systems in Accounting, Finance and Management,* 6 (3), 195-214.

[7] Cristianini, N., and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press.

[8] Deakin, E. B. (1976). Distributions of financial accounting ratios: some empirical evidence. *The Accounting Review*, 51 (1), 90-96.

[9] Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, 14 (1), 5-16.

[10] Han, I., Chandler**,** J. S., and Liang, T. P. (1996). The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods. *Expert Systems with Applications*, 10 (2), 209-221.

[11] Meyer, P. A., and Pifer, H. W. (1970). Prediction of bank failures. *The Journal of Finance,* 25 (4), 853-868.

[12] Min, J. H., and Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28 (4), 603-614.

[13] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research,* 18 (1), 109-131.

[14] Scholkopt, B., and Smola, A. J. (2000) *Statistical learning and kernel methods.* Cambridge, MA: MIT Press.

[15] Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22, 59-82.

Solution of Exam Supervisor Assignment Problem to Final Exams by Goal Programming

A. Çelik[1], H. N. Alp[2], S. Kaya[3], İ. H. Karaçizmeli[4]

[1]Harran University, Şanlıurfa, Turkey, aysecelik@harran.edu.tr
[2]Harran University, Şanlıurfa, Turkey, handenuralp@harran.edu.tr
[3]Harran University, Şanlıurfa, Turkey, serkankaya@harran.edu.tr
[4]Harran University, Şanlıurfa, Turkey, hkaracizmeli@harran.edu.tr

Abstract

In this study, 34 exam supervisors from 6 different departments of Harran University Faculty of Engineering were assigned to 170 final exams. Since this problem has more than one objective, goal programming is used as solution method. When performing these assignments, taking into account the special requests of the exam supervisors, the supervisor in each department has been assigned in an equal number of exams in its own department and in such a way that it will not be assign one after another as much as possible.

Introduction

Productivity is one of the most important requirements of our era and it is the concept that we use continuously in many sectors. It is very important to use time efficiently and to perform the work done in the most efficient way. Scheduling works for personnel in business life are also aimed at this. Scheduling is a method of decision making to keep productivity at the top when it is done in a fair and correct manner, considering the wishes of the personnel.

In the second part of the study, Goal Programming, in the third part, literature review, in the fourth part, application and in the fifth and last part, results and suggestions are given.

Goal Programming

In the studies that have been done and are being done to date, the complexity of the event makes the solution of the problem as difficult as possible. In this direction, multi- criteria decision making method is used as the most appropriate method. In solving the curret situation, it may be desirable to realize more than one goal at the same time. In this case, one of the most important multi-criteria decision-making methods, which is the most important goal programming method is preferred.

Literature Review

Varlı vd. (2017), have carried out a study with the aim of equal and fair assignment among the supervisors of midterm and final exams in Kırıkkale University Faculty of Engineering [1]. Varlı ve Eren (2017), established a goal programming model to ensure that nurses are balanced and fairly appointed as needed for shifts in a hospital in Kırıkkale [2]. Özçalıcı (2017), has developed an algorithm that prevents supervisors from being assigned to several exams at the same time and has an equal number of assignments between them. This algorithm is designed to solve even large-scale problems [3].

Application

In this study, the supervisors of the Faculty of Engineering at Harran University were assigned to the final exams for Autumn Period. Scope of the study, a total of 6 departments, 34 supervisors and 170 final exams were modeled and solved with goal programming. It is considered that each supervisor can only take exams in his / her department. In addition, it is aimed that each department should assign an equal number of examiners among their supervisors and not to take 2 consecutive exams as far as possible for each supervisor. The special requests of the supervisors were also taken into consideration. . In the Table 1 it's given the datas about problem.

Table 1. Datas on the problem

Department	Supervisor Code (i)	Exam Code (j)	Required Assignment	Average Assignment for Each Supervisor
Computer	1,2,3,4,5	1,2,.....,27	52	4 and 5.supervisors 5, others 14
Environment	6,7,8,9	28,29,.....,55	32	8
Electric- Electronics	10,11,12,13,14,15,16	56,57,.....,86	60	9
Map	17,18,19	87,88,.....,97	20	7
Civil	20,21,22,23,24,25,26	98,99,.....,128	115	16
Machine	27,28,29,30,31,32,33,34	129,130,.....,170	97	33 ve 34.supervisors 7, others 14

Since the 4th and 5th supervisors in the computer department and the 33rd and 34th supervisors in the machine department are lecturers, they have fewer appointments than the research assistants. When writing the exam codes, numbering was made for each department considering the priority order in the exam program. For example, in the computer department, 1 code is given to the first exam during the exam week and 27 code is given to the last exam.

Indices

i : supervisors (i=1,2,.....,34)

j : exams (j=1,2,.....,170)

Parameters

y_{ij} $\begin{cases} 1, i.supervisor\ can\ be\ assigned\ to\ the\ j.exam \\ 0, i.supervisor\ can't\ be\ assigned\ to\ the\ j.exam \end{cases}$

t_j : Number of supervisors required for j.exam.

Decision Variables

x_{ij} $\begin{cases} 1, i.supervisor\ assigned\ to\ the\ j.exam \\ 0, \qquad\qquad other\ events \end{cases}$,

$d1_{ij}^{+}$:Positive deviation from the goal of assigning an equal number of exams for i.supervisor

$d1_{ij}^{-}$:Negative deviation from the goal of assigning an equal number of exams for i.supervisor

$d2_{ij}^{+}$: Positive deviation from the goal of not taking 2 consecutive exams for i. supervisor

$d2_{ij}^{-}$: Negative deviation from the goal of not taking 2 consecutive exams for i. supervisor

Equation (1) aims to minimize deviation from the specified targets as much as possible.

$$MinZ = \sum_{i\epsilon I} \sum_{j\in J} d1_{ij}^{+} + d1_{ij}^{-} + d2_{ij}^{+} \qquad \text{(Eq.1)}$$

Goal 1: To conduct an equal number of exams between the supervisors of each department, Equation (2) and (3) computer, equation (4) environment, equation (5) electrical-electronics, equation (6) map, equation (7) construction, equations (8) and (9) targeted total for machine supervisors shows the number of assignments.

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 14 \qquad \text{i=1,2,3} \qquad \text{(Eq. 2)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 5 \qquad \text{i=4,5} \qquad \text{(Eq. 3)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 8 \qquad \text{i=6,7,8,9} \qquad \text{(Eq. 4)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 9 \qquad \text{i=10,11,12,13,14,15,16} \qquad \text{(Eq. 5)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 7 \qquad \text{i=17,18,19} \qquad \text{(Eq. 6)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 16 \qquad \text{i=20,21,22,23,24,25,26} \qquad \text{(Eq. 7)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 17 \qquad \text{i=27,28,29,30,31,32} \qquad \text{(Eq. 8)}$$

$$\sum_j x_{ij} - d1_{ij}^{+} + d1_{ij}^{-} = 7 \qquad \text{i=33,34} \qquad \text{(Eq. 9)}$$

Goal 2: Supervisors do not take 2 consecutive exams, Equation (10) is the constraint that prevents supervisors from entering 2 consecutive examinations as far as possible. 7. supervisor is exempt because she/he wanted to repeatedly enter some exams.

$$x_{ij} + x_{ij+1} - d2_{ij}^{+} + d2_{ij}^{-} = 1 \qquad \forall \text{i} \in \text{I}/\{7\},\ \forall j \in J/\{170\} \qquad \text{(Eq. 10)}$$

Equation (11) allows each supervisor to be assigned to the exam only in his or her department.

$$x_{ij} \le y_{ij} \qquad \forall \text{i,j} \qquad \text{(Eq. 11)}$$

Equation (12) provides the number of supervisors required for each exam.

$$\sum_i x_{ij} = t_j \qquad \forall \text{j} \qquad \text{(Eq. 12)}$$

Equation (13) shows that each supervisor can be assigned to a maximum of one more exam from the specified target. Equation (14) shows that no less than one exam assignment can be made.

$$\sum_j d1_{ij}^{+} \le 1 \qquad \forall \text{i} \qquad \text{(Eq. 13)}$$

$$\sum_j d1_{ij}^{-} \le 1 \qquad \forall \text{i} \qquad \text{(Eq. 14)}$$

Equations (15), (16) and (17) indicate that the 8th supervisor should take the 28th, 31st and 34th exams in the environmental engineering department. Equations (18) and (19) show that the 18th supervisor in the map engineering department cannot take the 90th and 92th exams. Equation (20) shows that at least one of the 11th and 14th supervisors should be assigned to the 61st exam in the department of electrical and electronics engineering. Equation (21) shows that the 15th supervisor should take the 70th exam. Equation (22) shows that at least one of the 12th and 16th supervisors should be assigned to the 71st exam. Equations (23), (24), (25) and (26) indicate that the 20th supervisor in the civil engineering department cannot take the 117, 118, 119 and 120 exams. Equations (27), (28), (29) and (30) indicate that the 7th supervisor in the environmental engineering department wants to take the 47th, 48th, 49th and 50th exams one after the other.

$$x_{8\,28} = 1 \qquad \text{(Eq. 15)}$$

$$x_{8\,31} = 1 \qquad \text{(Eq. 16)}$$

$$x_{8\,34} = 1 \qquad \text{(Eq. 17)}$$

$$x_{18\,90} = 0 \qquad \text{(Eq. 18)}$$

$$x_{18\,92} = 0 \qquad \text{(Eq. 19)}$$

$$x_{11\,61} + x_{14\,61} \geq 1 \quad \text{(Eq. 20)}$$
$$x_{15\,70} = 1 \quad \text{(Eq. 21)}$$
$$x_{12\,71} + x_{16\,71} \geq 1 \quad \text{(Eq. 22)}$$
$$x_{20\,117} = 0 \quad \text{(Eq. 23)}$$
$$x_{20\,118} = 0 \quad \text{(Eq. 24)}$$
$$x_{20\,119} = 0 \quad \text{(Eq. 25)}$$
$$x_{20\,120} = 0 \quad \text{(Eq. 26)}$$
$$x_{7\,47} = 1 \quad \text{(Eq. 27)}$$
$$x_{7\,48} = 1 \quad \text{(Eq. 28)}$$
$$x_{7\,49} = 1 \quad \text{(Eq. 29)}$$
$$x_{7\,50} = 1 \quad \text{(Eq. 30)}$$

Equation (31) and (32) are sign constraints of decision variables.

$$x_{ij} \in \{1,0\} \qquad \forall i,j \qquad \text{(Eq. 31)}$$
$$d1_{ij}^{+}, d1_{ij}^{-}, d2_{ij}^{+}, d2_{ij}^{-} \geq 0 \qquad \forall i,j \qquad \text{(Eq. 32)}$$

The solution of the problem was realized by coding the mathematical model in GAMS 24.0.2 application on an 8GB Ram Memory computer with Intel (R) Core (TM) i7-6500U CPU @ 2.50 GHz. Table 2 presents the comparison of the current situation with the solution obtained.

Table 2. Comparative results (some part)

	Current State				Recommended Solution			
Supervisor	**Total Assignment**	**1.Positive deviation from target**	**1.Negative deviation from target**	**2.Positive deviation from target**	**Total Assignment**	**1.Positive deviation from target**	**1.Negative deviation from target**	**2.Positive deviation from target**
1	14	0	0	11	14	0	0	2
2	14	0	0	9	14	0	0	2
3	14	0	0	5	14	0	0	2
4	3	0	2	1	5	0	0	0
5	7	2	0	0	5	0	0	0

,

Results and Suggestions

When the optimal solution obtained was compared with the current situation, the positive deviation rate from the target assignment amount for each supervisor was reduced from 22 to 3, while 86% improvement was achieved while the negative deviation rate was reduced from 24 to 5, resulting in approximately 79% improvement. The positive deviation rate of the supervisors from 2 consecutive examinations was reduced from 173 to 54, resulting in an improvement of 69%. In subsequent studies, besides the supervisor assignments can also be made of scheduling the exam or supervisor assignments can be made by considering both midterm and final exams.

References

[1] Varlı, E., Alağaş, H. M., Eren, T., and Özder, E. H. (2017). Sınav görevlisi atama probleminin hedef programlama yöntemiyle çözümü. Bilge International Journal of Science and Technology Research, 1(2), 105-118.
[2] Varlı, E., and Eren, T. (2017). Hemşire çizelgeleme problemi ve hastanede bir uygulama. Academic Platform-Journal of Engineering and Science, 5(1), 34-40.
[3] Özçalıcı, M. (2017). Sınavlara Gözetmen Atama Probleminin Çözümü İçin Takas Bazlı Bir Algoritma Önerisi. İktisadi ve İdari Bilimler Fakültesi Dergisi, 19(2), 492-506.