International
SOCIAL SCIENCES STUDIES JOURNAL
SSSjournal (ISSN:2587-1587)

Economics and Administration, Tourism and Tourism Management, History, Culture, Religion, Psychology, Sociology, Fine Arts, Engineering, Architecture, Language, Literature, Educational Sciences, Pedagogy & Other Disciplines in Social Sciences

Vol:5, Issue:30	pp.960-968	2019 / February / Şubat
sssjournal.com	ISSN:2587-1587	sssjournal.info@gmail.com

Article Arrival Date (Makale Geliş Tarihi) 11/01/2019 | The Published Rel. Date (Makale Yayın Kabul Tarihi) 28/02/2019
Published Date (Makale Yayın Tarihi) 28.02.2019

DATA MINING AND STATISTICS IN DATA SCIENCE

Güner Gözde KILIÇ
Istanbul Commerce University, Institute of Science and Technology, Department of Statistics, ggteksin@gmail.com, Istanbul/TURKEY
Prof. Dr. Münevver TURANLI
Istanbul Commerce University, Institute of Science and Technology, Department of Statistics, mturanli@ticaret.edu.tr, Istanbul/TURKEY
Prof. Dr. Ünal Halit ÖZDEN
Istanbul Commerce University, Institute of Science and Technology, Department of Statistics, uozden@ticaret.edu.tr, Istanbul/TURKEY

Article Type : Research Article/ Araştırma Makalesi
Doi Number : http://dx.doi.org/10.26449/sssj.1295
Reference : Kılıç, G., G., Turanlı, M. & Özden, Ü., H. (2019). "Data Mining And Statistics İn Data Science", International Social Sciences Studies Journal, 5(30): 960-968.

ABSTRACT

In parallel with the developing technology of modern age, there has been a corresponding increase in computer domains that possess data storage function. Therefore, methods which allow storing large data gained an equally grave attention. In this study some of the most popular data analysis methods, namely data mining and statistical methods, have been investigated. The aim of this study is to exhibit the correlation between data mining and statistics. To achieve this aim, firstly data mining process has been explored. Next the need to implement statistical methods during this process has been accentuated.

Keywords: Data Science, Data Mining, Statistics, Data Preprocessing, Feature Selection

ÖZ

Günümüz teknolojisinin ilerlemesine bağlı olarak, bilgisayarların bilgi saklama kaydı yapılan alanlarında da artışlar meydana gelmektedir. Bu nedenle, büyük verilerin analiz edilmesini sağlayan yöntemler de büyük önem kazanmaktadır. Bu çalışmada, veri analizinde yaygın olarak kullanılan veri madenciliği ve istatistiksel yöntemler incelenmiştir. Bu çalışmanın amacı, veri madenciliği ile istatistik arasındaki ilişkinin ortaya konmasıdır. Bu nedenle, öncelikle veri madenciliği süreci açıklanmış daha sonra bu süreçte istatistiksel yöntemlerin gerekliliği vurgulanmıştır.

Anahtar Kelimeler: Veri Bilimi, Veri Madenciliği, İstatistik, Veri Önişleme, Öznitelik Seçimi

1. INTRODUCTION

Big data has been quite a popular concept recently. Big data is also defined as a construct in which significant and insignificant data coexist. As modern technology rapidly develops, the power of data also rises each new day and big data comes to the scene even more frequently. Thus, a rise in data dimensions leads to conditions in which classical statistical methods can fall short. In order to extract big data and the kind of information that might prove to be useful in the midst of data masses, the vitality of data analysis heightens each new day. In data analysis, the aim of statistics is to make sense of the data. Although statistical methods are widely used, data mining that is grounded on statistical methods enables us to analyze big data when we need to analyze data. In data mining processes, the widespread application of statistical methods points to the fact that statistics and data mining cannot be treated as independent entities from each other.

2. DATA MINING

Data mining is the process of acquiring the big data and useful information from data masses. At the same time data mining is defined as the process of analyzing data with respect to a variety of perspectives on the classification of useful data that accumulate in data storages for data analysis, data mining algorithms and relevant information that were collected and connected together in common domains. Data mining is also referred to as; data exploration, knowledge discovery, data/pattern analysis, and data collection.

Data mining is a type of skill that integrates a number of techniques in a multitude of disciplines such as; database technology, statistics, machine learning, pattern definition, artificial neural networks, data visualization. Equally difficult as defining exact borders among these disciplines is defining the decisive borders between these domains and data mining (Hand et al., 2001).

History of data mining points out that mathematicians' interest in the 1950s on logic and computer sciences paved the way to creating concepts of artificial intelligence and machine learning. In the 1960s however, statisticians discovered new algorithms such as; regression analysis, maximum likelihood prediction and neural networks thereby took the very first steps toward data mining. In effect of developing new computer techniques and new programming languages between 1970-1990, novel algorithms such as clustering algorithms, genetic algorithms and decision trees emerged. Starting with the 1990s, first stages of data mining were formed on databases hence data storage was expanded (Ayre, 2006). In the 2000s data mining advanced and its application fields also expanded widely in due time. Likewise, currently, data mining is harnessed in every part of life and has secured its place in relevant studies as an indispensable component of data science.

In data mining, there are a number of concepts that relate to data mining analysis in order to extract knowledge amidst big data and harnessed in the knowledge-analysis process. These concepts are respectively; descriptive data mining, predictive data mining and commanding data mining.

- Descriptive data mining: It is achieved via analyzing available data and interpreting through the visualization of current status.
- Predictive data mining: It relates to making future predictions by benefiting from data from the past.
- Commanding data mining: Likewise this method also entails a prediction and at the end of prediction process, it offers novel methods that can provide better predictions.

Data mining constantly progresses in tandem with statistics, artificial intelligence, machine learning and similar techniques. Such techniques develop even further with high-performance database tools and data integration initiatives. Hence there are basic necessities of data mining. These basic necessities are; data storage, software package, methods for effective data-access, dynamism in data problems, effective algorithms, application server and deleting data. Data storage is most salient one among all the items above. Before installing data storage, it is quite significant to clearly determine objective in data usage. Failure to spot correct data and specify the framework to draw big data would obstruct data mining process (Han and Kamber, 2001).

The objective in data mining is reaching a decision on predicting future behaviors through making an analysis of past activities and to design a modeling based on these predictions (Koyuncugil, 2007)

As we dig into application fields of data mining it can be stated that the first application field is in customer relations domain; hence data mining is much frequently used in customer relations. Additionally; it is used in banking sector to specify confidential relations on financial indicators; in insurance sector it is used to identify patterns of risky customers; in other relevant sectors it is used to determine customers' choices through advertisement or tracking of credit cards. Also it is used to detect whenever new technologies emerge whether or not these new technology is fit for usage or for a specific area. Next to all of the fields above it is widely employed in health, telecommunication, astronomy, medicine, biology and sports etc.

3. DATA MINING PROCESS

Data mining is viewed as a complete process from a start to end point. Extracting through abstract studies relevant big data from the data mass in which big data and insignificant data coexist, and making data usable are also constituents of this process. Not knowing which data to work on would result in uselessness of algorithms to be used in analysis process. Therefore prior to data mining process data features should be

thoroughly examined and data should be formed to fit with the analysis. Data mining process entail stages such as data preprocessing, feature selection, model installment, success rate and model tracking.

3.1. Data Preprocessing

In data mining the biggest time consumer due to the huge size of data is stage of data preprocessing. Sampling sizes that were defined in hundreds and thousands many years ago have currently reached to millions and even billions in quantity. Data saved in database or data storage are firstly made usable through statistical methods. Making the data usable means detecting incomplete, insufficient, inconsistent, contradictory features and finding solutions for such problems. Starting analysis process before conducting any operation on millions and even billions of redundant data on the database and arriving big data based on these results would be problematic.

In data mining, preprocessing is the most time consuming stage because researchers lack knowledge on how to transfer analyzable data on the database. In data entries there may be incomplete, redundant or repetitive data on a database. In data mining applications, preprocessing stage constitutes 80% of the analysis (Piramuthu, 2003). Hence data preprocessing stage is the most salient stage of data mining.

Each data-analysis process starts with collecting, describing and deleting of new data sets. After this process data can be analyzed and results can be gathered (Dasu and Johnson, 2003). Data preprocessing step consists of stages of data collecting, deleting, connecting, transforming and reducing.

Data collecting is identifying required data for the problem to solve and locating essential resources to collect those data (Akpinar, 2000).

Data deleting is detecting incomplete data to be complemented; regulating noises in order to identify contradictory values or anomalies and correcting any inconsistencies in the data. Data deleting is quite a popular stage in data mining. Inconsistent, incomplete and noisy data are frequent features in databases. Some other problems are data noise due to incorrect use of data collection tools; data inconsistency due to deleting wrong data; data incompleteness due to accepting data as redundant at the time it was entered and problem because of failing to record. In data-deleting stage in order to complement incomplete values it is suggested to exclude incomplete values from the analysis and in place of incomplete values it is possible to use mean variable and mean value of the variables could be employed for the samples in the same class (Roiger and Geatz, 2003). One of the reasons behind the popularity of data deleting technique is that in the database there is variance and random error or in other words there is measurable variable of noisy data. Histogram, clustering analysis and regression are some of the applicable techniques used in the identification of noisy data.

In data-connecting stage, basically, it is aimed to connect data in different databases under one cluster. It is likely to cause diagram-connection errors during the stage of connecting data in different databases under one cluster. Meta data is applied in order to avoid such errors.

Data is transformed to fit with the analysis after performing data transformation. Data transformation is a structure that entails one or a few of the procedures such as connecting, correcting and normalizing.

In data preprocessing stage another salient feature that emerges is data reduction which points to extracting smaller-size data from bigger-size data. By this way it can be easier to apply data- mining techniques on the reduced data set. Deriving this variable from a different table and inconsistencies in the variable may cause redundancy. These redundancies can be analyzed via correlation coefficient and in case correlation coefficient is measured to be higher, reduction is possible by deleting the variable from the database.

3.2. Feature Selection

Feature selection relates to the entirety of procedures implemented to nullify dimension reduction that is a serious problem in data mining. In feature selection the aim is to select the best subset that can represent raw data set. In general, during feature-selection procedures in machine learning system, the features that can optimize learning are chosen while features that negatively affect learning are deleted and it is also aimed to multiply the quantity of features in the data set.

Feature-selection methods are bifurcated as statistics-based and classification-based methods. In statistical methods it is often the case to obtain data set through statistical distributions. In classification-based methods however, it is essential to select the top-achiever features according to classification results of features but eliminate other features.

Within statistics-based and classification-based feature selection methods, the most widespread ones are entropy criteria, formal similarity, principal components analysis and Fisher's criteria. In entropy criteria, selection is made through the assistance of knowledge on the feature. Feature is viewed as a distribution to measure its entropy and it is considered that as entropy increases, classes could be discriminated more conveniently. In formal similarity based method, if features emerge as forms with similar shapes, it is believed that distributions are also similar to one another. If that happens, it is aimed to lower noise through Euclidean distance instead of deleting one of the features. Fisher's criteria is one of the algorithms that can increase class discrimination and allow the distance between class centers to be maximum. Via feature transformation, principal components analysis can delete the features with similar attributes so that it can be feasible to transmit from many features to fewer features.

On the other hand feature selection algorithms are analyzed in relevant literature below three main headings such as filter model, spiral model and embedded model. In terms of its advantage filter model is independent from classification algorithm and also faster although as for classification, its success rate is remarkably lower. In feature-selection algorithms that are filter modeled, statistical hypotheses have become more predominant. In spiral modeled algorithms, features that are systematically selected are changed to compute its classification success rate and this rate is harnessed to make a decision between selection or deletion procedures. Embedded modeled feature-selection algorithms are embedded within a classification algorithm. For spiral modeled feature-selection algorithms genetic algorithms are given as specimens while for embedded feature-selection algorithms, decision trees are exemplified.

Starting an analysis process after feature selection can offer a great number of benefits and advantages which can be listed as; (Ladha and Deepa, 2011).

- It minimizes the dimension of property set and elevates algorithm speed,
- It eliminates irrelevant and noisy data,
- It improves data quality,
- It can describe, visualize and simplify data set in a simpler way,
- It enables saving resource for data collection process required to forge data set,
- It decreases the size of storage that is required to store data,
- It elevates the success rate of obtained model.

Feature-selection procedure can be applied in a range of fields such as; emotion analysis, face recognition, disease diagnosis and classification of social networks.

3.3. Model Installment and Success Rate

In order to find the most optimal model, it is essential to install and test as many models as possible. Similarly, model installment process can vary with respect to models that employ controlled and uncontrolled learning.

In controlled learning model, relevant classes are predetermined and classifications are based on specific criteria. Different examples are offered for each class. Here the aim is to locate properties of each class and to describe controlling sentences that can match with these properties. When learning process stops, new examples are applied to controlling sentences that were defined and the model decides which classes these new examples best fit with.

In uncontrolled learning on the other hand it is aimed to observe relevant examples and to describe classes with the help of similarities between examples.

In controlled learning, data is formed on the basis of selected algorithm. In the first stage machine learning is applied to some parts of the data but the rest of the data receives test procedures to assure the validity of model. In general, 70% of data is taken as educational data and machine learning is applied on this data while 30% is taken as test data to check validity of the model. Educational data and test data percentages also vary according to the success rate of model. Educational data is taken as 80% and test data can be taken around 20%.

No matter how high success rate of the established model is, it can never be claimed that the modeling reflects the reality one hundred percent. Among the pivotal reasons that can transform a valid model into an

incorrect one are mistakes that are made in the accepted hypotheses while setting the model and the failure to employ correct data.

3.4. Model Tracking

In line with changing and developing technology of modern age, there are constant variations emergent in the systems and data; so it is necessary to continually track and reorganize established model. A useful method in tracking the model is graphics that can visualize the difference between estimated and observed variables.

4. METHODS THAT ARE EMPLOYED in DATA MINING

Methods employed in data mining are ranged as; statistical methods, memory-based methods, genetic algorithms, artificial-neural networks and decision trees.

4.1. Statistical Methods

- **Classification**

Classification is the procedure of examining the features of an object and assigning these objects to predefined classes. In general, during classification stage, each class properties are clearly determined beforehand and by separating each data from one another according to their properties they can be assigned to different groups. In data teaching stage a prediction is made to decide on which class newly-encountered data belongs to.

Classification is a two-step process; model formation and model usage. Model formation stage is referred to as; decision tree, classification rules or mathematical formulas. During model usage stage, model's accuracy range is estimated and it is accepted that educational sets and test sets derive from the same distribution. If accuracy range of the model is acceptable, then the model can be used.

- **Discriminant Analysis**

Discriminant analysis is one of the multivariate-statistics techniques that aims to predict the relations between categorical dependent variable or variables and metric independent variables (Kalayci, 2016)

Among the usage objectives of discriminant analysis are; deciding about the variable group to which a data could enter, to determine effective and ineffective variables in discriminating groups, to examine differences between mean features of two or more numbers of groups detected before the analysis stage, to identify how much variance of a dependent variable could be explained by independent variables, to identify to what percentage of data classification is correct or incorrect.

In discriminant analysis, to eliminate risk of faulty classification; variables must have multiple normal distribution; there should be not any multiple linear connection problems among independent variables and covariance matrixes should be at equal size. Results of discriminant analysis are testable results and this is a positive factor that can enhance validity and reliability of discriminant analysis.

- **Clustering Analysis**

Clustering analysis is among the techniques used in the classification of grouped data. In clustering analysis it is claimed that data which are almost identical according to predetermined selection criteria cannot be classified within the same set. Clustering analysis resembles discriminant analysis which also aims to collect similar entities in same groups and also resembles factor analysis in which identical variables are collected in the same groups (Cakmak, 1999).

Distance criteria and correlation criteria are applied in clustering analysis in order to measure the existing similarities between individuals or objects. Thus data-normalcy hypothesis that is significant in other multivariate statistical analyses is less important in clustering analysis and normalcy of distance values is deemed sufficient enough (Tatlidil, 1992).

The aim in clustering analysis can be listed as; model formation, group-based prediction, hypothesis test, data research, hypothesis structuring and dimension reduction (Ball, 1970).

Clustering analysis is a significantly useful statistical method to analyze data. Since data set is very big in researches, most of the times it becomes more difficult to make data larger. In those instances data will be reduced by clustering the observations according to select criteria and upper-groups will thus be formed.

- **Regression Analysis**

Regression analysis is a statistical method used in modeling and examining the mathematical correlation between variables. In regression analysis when there is one dependent and one independent variable, simple linear-regression model is used; when there is one dependent variable and more than one independent variables, it is formed as a multiple regression model.

Aims of regression analysis are specified as; description, causality and prediction. In description stage, equations are arranged in order to describe and summarize data set. Hence regression analysis is a tool that enables developing such equations. Causality, which is another objective of regression, aims to demonstrate that there is a correlation between variables in order to describe links more effectively and correctly. In regression, an equation or a model is developed for the provided values of independent variables in order to predict dependent variable values. In short, regression analysis' mission is to learn as much as possible about the reflected environment.

- **Logistic Regression Analysis**

Logistic regression is a multivariate statistical modeling in which dependent and independent variable is discriminated. When dependent variable is a nominal variable, Least Squares Method (LSM) falls short in suggesting a prediction. In a different saying predicted variances are not minimum. LSM method suggests that variables are normally distributed and this hypothesis is achieved when dependent variable is nominal (Kalayci, 2016).

In cases when dependent variable is nominal discriminant and logistic regression models can be offered as an alternative choice for LSM technique. In discriminant analysis it is mandated that independent variables exhibit a normal distribution and covariance of independent variables are equal on each group level. Yet, in logistic regression analysis this is not an applicable scenario.

Logistic regression is one of the most popular methods in classification analyses. Because logistic regression does not demand multivariate normal distribution hypothesis, it offers an outpacing advantage in such applications. Also it has a quality of determining probabilities on class membership. One of the logistic regression hypotheses is that linear probability function and distribution of error terms can fit with the logistic distribution; thus logistic regression model can estimate parameters via maximum likelihood technique.

4.2. Memory Based Methods

Memory based methods were suggested as the onset of 1950s and due to computation advances during those years, memory usage and cost of using required technological tools, those novelties turned into a usable method. As a result of widespread usage of multiprocessor systems in particular, there has been a similar rise in the usage of memory based methods. The best example for this method is suggested as (K-Nearest Neighbor-KNN) algorithm. KNN algorithm is used in the prediction of real values for unknown sampling. Closeness is described in Euclidean distance.

4.3. Genetic Algorithms

Genetic algorithms first emerged in the 1970s and it only operates through fitness function without necessarily having some foreknowledge and hypotheses. In sum it is an optimization technique that is used to develop data mining algorithms. Genetic algorithms are applied on the data in order to reveal hidden patterns and make predictions.

Genetic algorithms generate successful results particularly in solving optimization problems in which there is no such intuitive knowledge present (Wan et al., 2016).

Genetic algorithms entail specific advantages as well as disadvantages. Among the advantages are; ability to perform with discrete and indiscrete variable, not needing derivative computations and operating with quite a number of variables (Haupt, 2004). Choosing certain parameters, lack of a method that explains when to terminate the study, and operating the multiple fitness functions many times are among its disadvantages (Sivanandam and Deepa, 2008).

4.4. Artificial Neural Networks

Artificial neural networks are an artificial intelligence technique that was developed by mimicking the working of human mind. Artificial intelligence on the other hand is to endow the machines with a number

of human attributes such as thinking, analyzing, comparing and deciding just as humans do. Hence by making use of artificial neural networks the aim is to train, educate machines who can make decisions as humans.

The earliest studies on artificial neural networks emerged as modeling of neurons and application of models on computer systems. Unlike statistical methods artificial neural networks cannot provide a parametric model hypothesis of the data. This finding reveals that application fields of artificial neural networks are wider and as opposed to memory based methods they do not need a memory need and operation.

4.5. Decision Trees

Decision trees are among the most pervasive data mining classification methods. In general, classification and regression methods are widely used in solving the problems.

On the other hand, decision tree methods belong to a classification method which is formed via conditional probabilities and of which results, costs and events are depicted as a tree (Geetha and Nasira, 2014)

In decision trees, if dependent variable is categorical, tree is named as a classification tree; if dependent variable is continuous then the tree is named as a regression tree.

In order to easily draw a decision tree from data sets, a number of decision tree algorithms have been developed and these algorithms generically aim to minimize the error while they also seek to structure the form of an optimal decision tree.

Since decision trees are easy to comprehend and interpret and can offer advantages for decision-makers,they are still much more popular despite the presence of other methods such as artificial neural networks (Chien and Chen, 2008). On the other hand; simple structure of decision trees, convenience and comprehensibility of established classification model, offering a favorable structure for knowledge discovery, the shorter time to structure and the nonparametric feature of decision trees method in comparison to other decision trees, the most popular application is classification techniques among a variety of decision trees (Gehrke, 2003).

5. CORRELATION BETWEEN DATA MINING and STATISTICS

Data mining and statistics are closely correlated in a myriad of aspects (Zhao Chung-Mei and Luan, 2006). Common point between data mining and statistics is “learning from data” or “transforming the data into knowledge” (Kuonen, 2004).

In particular, statistical techniques are commonly used in stages such as data reducing, data modeling, analyzing the relations between variables and generalizing results that were extracted from the sampling for the data mass.

Analyzing the relations between variables, measuring mean values, standard deviations and making future predictions are all possible thanks to science of statistics. All in all, in every area where there is data, it is a must to make use of statistics.

Among the commonalities of data mining and statistical methods there is; focusing on data analysis, identifying effective factors behind an event and estimating potential future events in a much better way.

Data mining is an application that builds upon statistical methods. In the literature of statistics data mining is grouped below the heading of multivariate statistical methods and it is assumed that in general data derives from a parametric model.

Data mining and statistical methods also possess dissimilar aspects. Generalizability, hypothesis test, theory’s role and confidence level are a few of these distinctions (Ganesh, 2002). In data mining, deduction is the case but in statistics induction is applied. Statistics pays interest to the sampling, hence since it is essential to generalize results to the universe, in statistics theory and hypothesis tests bear significance. On the other hand in data mining since it matters to collect detailed, very specific and local knowledge; theory and hypothesis tests have no significant meaning.

In statistics, data set would not be as huge as data set used in data mining so before starting analysis process in data mining it is suggested to preprocess the data. Nevertheless in statistic studies, analysis is usually started before preprocessing the data. In data mining preprocessing stage takes too much time but it

enables to perform a higher quality analysis with quality data in data mining. This distinction is the most important factor in discriminating data mining and statistics.

6. CONCLUSION

As a result of the rise in data dimensions in our age, statistical methods have failed to be sufficient on their own. Data mining that emerged as a response to such developments can provide benefits in the analysis stage of big data dimensions. Just as the trend in globe, there is a corresponding interest and attention toward data mining in Turkey. In addition to statistics which already has a wide field of application, the usage area of data mining is also expanded.

Although data mining concept seems to be a distinct concept, it is a fact its roots are grounded on statistical methods. Learning from data, acquiring big knowledge from big data, exploring the correlations are just a few stages harnessed in statistics or data mining. Nevertheless, in terms of specific features, data mining differentiates from statistical methods. Even though data mining concept has become quite a popular one statistical methods are still the most frequently employed analysis method in this age. It was thus aimed in this study to examine similarities and differences between data mining and statistics and it was then clearly demonstrated that in each stage of data mining, there is need for statistics.

REFERENCES

Akpinar, H. (2000). "Knowledge discovery and Data mining in databases", İstanbul University Faculty of Management Journal, Vol. 29, No. 1/April, p. 1–22.

Ball, G.H. (1970). Classification Analysis, Menlo Park, Calif.: Standford Research Institute.

Chien, C. F., Chen. L. F. (2008). "Data Mining to Improve Personnel Selection and Enhance

Human Capital: A Case Study in High-Technology Industry," Expert Systems

with Applications, vol. 34, 280-290.

Cakmak, Z. (1999). "Validity Problem in Clustering Analysis and Evaluating Clustering Results", Dumlupınar University Social Sciences Journal, No:3, Nov., p.187-205.

Dasu, T. and Johnson, T. (2003). Exploratory Data Mining and Data Delete, John Wiley & Sons Publication, New Jersey, USA.

Ganesh, S. (2002). "Data Mining: Should it be included in the 'Statistics' curriculum?", The Sixth International Conference on Teaching Statistics, Cape Town, South Africa, 7–12 July.

Geetha, A. and Nasira, G.M. (2014). "Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction" IEEE International Conference on Computational Intelligence and Computing Research.

Gehrke, J. (2003). "Decision Trees", The Handbook of Data Mining", Editor: Nong Ye,

Lawrence Erlbaum Associates Publishers, London, 149-175.

Han, J. and Kamber, M. (2001). "Data mining concepts and techniques", Morgan Kaufmann Publishers, Tokyo, 30-33.

Hand, D., Mannila, H. and SMYTH, P. (2001). Principles of Data Mining, MIT, USA, 546p.

Haupt, R. L. and Haupt, S. E. (2004). Practical Genetic Algorithms, New Jersey, John Wiley & Sons.

Kalayci, S. (2016). "SPSS Applied Multivarite Statistical Techniques".

Koyuncugil, A. S. (2007). "Data mining and its Application on Capital Markets", Capital Market Board Research Report, Research Office.

Kuonen, D. (2004). "Data Mining and Statistics: What is the Connection?", The Data Administration Newsletter).

Ladha, L., Deepa, T. (2011). Feature Selection Methods and Algorithms, International Journal on Computer Science and Engineering, 3(5), 1787-1797.

Lori Bowen Ayre. (2006). Data Mining for Information Professionals.

Piramuthu, S. (2003). "Evaluating Feature Selection Methods for Learning in Data Mining Applications" European Journal of Operational Research, Article in Press, pp.1-11.

Roiger, R.J. and Geatz, M.W. (2003). Data Mining a Tutorial-Based Primer, Addison Wesley, USA, 350p.

Sivanandam, S.N. and Deepa, S.N. (2008). Introduction to Genetic Algorithms, New York, Springer.

Tatlidil, H. (1992). Applied Multi-Variable Statistical Analysis H.U. Faculty of Science Department of Statistics, Ankara: p.252.

Wan, Y., Wang, M., Ye, Z. and Lai, X. (2016). "A Feature Selection Method Based on Modified Binary Coded Ant Colony Optimization Algorithm" Applied Soft Computing, 49:248-258.

Zhao Chung-Mei and Luan, J. (2006). "Data Mining: Going Beyond Traditional Statistics",

New Directions for Institutional Research, No. 131, pp. 7–16.