Journal of Natural Sciences and Mathematics of UT, Vol. 4, No. 7-8, 2019

UCD: 004.659-049.8
Conference paper

DATA MINING FUNCTIONALITY AND USAGE AREAS

Festim Halili[1*], Florim Idrizi[1], İsmail Bilal Devlez[2], Halim Halimi[1]

[1*] *Department of Informatics, Faculty of Natural Sciences and Mathematics, University of Tetova, RNM*

[*]*Corresponding author e-mail: festim.halili@unite.edu.mk*

Abstract

Data mining is the process of accessing information from large-scale data and mining information. With the rapid development of technologies, large data pools required analysis and modeling of large data to predict and analyze future information trends. The main purpose of the data mining process is to extract useful information from the data file and convert it into an understandable structure for later use. There are different processes and techniques used to successfully perform data mining. Data mining techniques are a result of long research and product development processes and include artificial neural networks, decision trees and genetic algorithms. In this paper, data mining technology, definition, motivation, process and architecture, data types of mining, functions and classification of data mining are examined.

Keywords: *data mining, architecture, functionality, techniques.*

1. Introduction

The rapid development of information systems and software technology has led to an increase in data and large amounts of information. This requires very large dimensions and large storage capacity. Databases also store insignificant data other than related and meaningful information. In order to process and control the required data, the manager of this data needs an application to provide control. The main point in the processing of data is the discovery of information in the database. In other words, data mining is a set of processes that include methods, techniques, tasks, models, or algorithms used to analyze huge data stored in databases.

Data mining is a very important process for extracting previously unknown and potentially useful information from large databases. It is a powerful tool that helps companies and organizations to increase sales, increase customer information and make more profits. Data mining provides useful information that queries and reports cannot provide us efficiently. Data mining is defined as information that is not visible in databases, because information obtained by data mining techniques is clearly not limited to the database.

2. Architecture is Data Mining

A. Database, data warehouse or other information store

This component is one or a series of databases, data warehouses, information tables, or other types of information stores. Data cleaning and data integration techniques can be performed on data.

B. Database or data warehouse server

The server is responsible for retrieving relevant data based on data mining requests.

C. Knowledge base

It is the field information used to direct the research or to assess the quality of the molds obtained. Includes concept hierarchies that are used to edit attributes or to edit values to different levels of abstraction.

D. Data mining engine

It is an indispensable component of the data mining system and consists of a module for tasks such as characterization, classification, association analysis, evolution and deviation analysis.

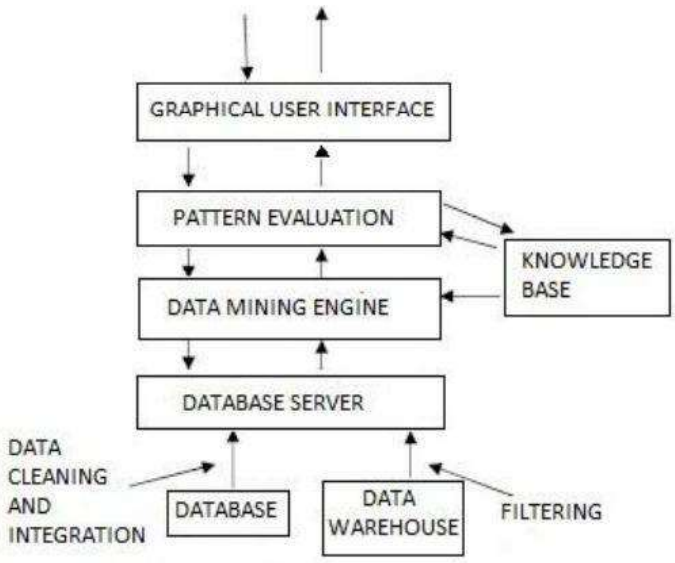

Figure 1. Architecture of data mining

E. Pattern evaluation module

This uses quirkiness criteria and interacts with data mining modules to focus research on interesting models. The model evaluation module can be integrated with the mining module depending on the application of the data mining method used, and thus effective data mining is possible.

F. Graphical user interface

This component is responsible for the communication between the user and the data mining system. It provides a user's data mining query or task and provides information to help focus the search. Based on intermediate data mining results, it enables exploration mining to interact with the system. This component also allows the user to browse database and data warehouse diagrams, upgrade data structures, evaluate problematic patterns and visualize them in different forms.

3. Processing of Data Mining

Data mining consists of seven phases. The first four stages are used for data preprocessing when data is prepared in a format created for future use. The other three phases are used to work on the data generated in this way to retrieve confidential information. Data cleanup is used to remove all crowds and other inconsistent data from the input database. Data integration is used to integrate data that can be entered from a variety of sources. The data warehouse is the place where all the data is cleaned and integrated. The data selection step is to select the appropriate data for mapping. Data transformation converts data into the most appropriate format for mining. The data mining phase is to use a variety of methods on the data to produce appropriate data as patterns and information. In the next step, these patterns are evaluated and in the last step, the information is presented to the user in a suitable format.

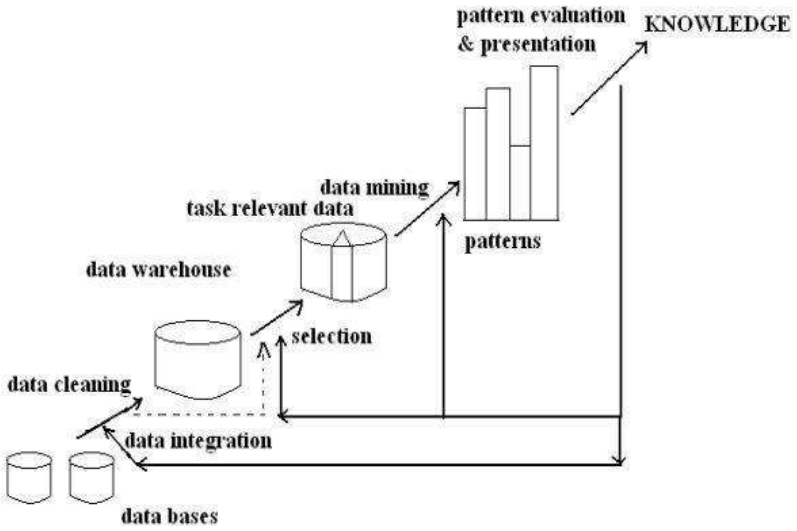

Figure 2. Data mining as a process of knowledge discovery

A. Preparing Data

The first thing to pay attention to extracting data from the database is the identification of data sources. Some databases and data warehouses may not be considered. Therefore, using sampling to simplify the data mining mechanism is necessary to format the data in a tabular or reduced dataset.

B. Data preprocessing

To improve quality, some consistency, integration, accuracy and redundancy problems need to be solved. This can be accomplished in several steps, such as cleaning, converting, reducing, and sorting data. Cleaning is to solve missing data and remove duplicate data. Conversion is to place the data in the appropriate form. Reduction is defined as eliminating unnecessary features or compression. The final step is to reduce the presentation levels of the data.

C. Data mining

This is to discover the model that represents the information. Data mining is descriptive and predictive.

D. Evaluation

Evaluation of the results obtained by different mining methods. For the accuracy of the evaluation results, the different results are compared.

E. Implementation

It is the implementation of these results by establishing a model or framework to ensure the best decision reached as a result of the evaluation phase.

4. Techniques and functionalities of Data Mining

A. Characterization

Data characterization generates the general characteristics of the objects in a target class. The data of a user-specified class are passed through a module to be subtracted at different abstraction levels.

B. Discrimination

This is basically a comparison of the general characteristics of objects between two classes, defined as the target class and the opposite class. Reveals distinctive features.

C. Association analysis

The association analysis examines the frequency of items in the database and creates frequent sets of items. Association analysis is widely used in market basket analysis.

D. Classification

Classification analysis is the regulation of data in classes. The classification, also known as supervised classification, uses given class labels to sort objects in the database. A training set associated with known class tags is used. The classification algorithm learns from the training set and forms a model. The resulting model classifies new objects.

E. Prediction

It is to estimate the data values that are not available. For some data, the class label can be estimated and linked to the classification. Once a classification model is created with the training set discussed in the classification phase, it can estimate the class label of an object, taking into account the property values of the objects and classes. The main idea is to achieve possible future values using a large number of past values.

F. Clustering

It is also called classification and is called unsupervised classification. There are different clustering operations based on the principle of maximizing similarity between objects of the same class and minimizing similarity between objects of different classes.

G. Outlier analysis

Outliers are data that cannot be included in any group after classification and clustering operations. Defining these exceptions is very important. The analysis of outliers is very important and may reveal important information in other areas.

H. Evolution and deviation analysis

Evolution analysis examines the evolutionary process in data that allows for characterization, comparison, classification and clustering of time-related data. Deviation analysis examines the condition between the measured values and the expected values and investigates the causes of deviations.

5. Benefits, usage areas and problems of Data Mining

A. Benefits and Usage Areas of Data Mining

Implementation of data mining technology has many benefits. Some of these are:

- Provides automation support for available software and hardware.
- Increases efficiency and helps save budget.
- Medicine: Supports different treatment methods for different diseases.
- Law enforcer: By examining trends in behavioral patterns, it can help to identify and suspend criminal suspects.
- Marketing: It helps customers to determine the products they want to buy in advance. By revealing the facts in the database, it uncovers previously unknown elements about customers.
- Banking: Assists financial institutions in credit assessment and measurement.
- Research: Data analysis is done much faster.

B. Problems of Data Mining

1. Security and social issues

Security is an important matter for all data intended to be used in shared and strategic decision making. Security is a controversial matter because some of this data is confidential and there is the possibility of illegal access to information. Data mining may disclose information about individuals or groups that may be against privacy policies. Due to competition, a significant portion of the data obtained from confidential information may be stored and other information may be widely distributed and used in an uncontrolled manner.

2. User interface issues

The information discovered by data mining tools is useful as long as it is understandable by the user. User interfaces and visualization are very important for the user to focus on mining tasks and to depict information from different perspectives.

3. Mining methodology problems

The size of the search field for data mining techniques is more decisive than data size. The size of the search field usually depends on the number of dimensions in the domain. When the number of measures increases, the search area expands exponentially. This is observed as the most urgent problem in terms of the performance of some data mining techniques.

4. Performance issues

Many artificial intelligence and statistical methods are available for data analysis and interpretation and are generally not developed for data mining. When processing huge data, the efficiency of data mining methods is discussed.

6. Knowledge Management

A. Big Data

Big data in its simplest sense means that data is too big for computers to handle. Therefore, the size of large data is constantly increasing. That is, the amount that a computer farm can process is constantly being enhanced by technological innovations such as processors, the software world, connection speeds and the like, so the maximum data definition that can be processed is constantly increasing.

It is possible to make basic 3 definitions for large data:

- The size of the data to be processed (if we are talking about a dimension above the hardware limits, we can say that it is big data).
- Structure of data to be processed.
- The complexity of the result to be processed over the data to be processed.

The concept of large data is not only used in the literature as processing capacity. It also refers to the suitability of the data for processing. For example, let's want to process newspaper news. The number of this news is an important parameter for the processing we are going to do and if there is any news articles above that we can process, we can define this news source as big data. But let's say that half of the data we can process (that is, within our capacity) is a data source that we cannot process because of its complexity or irregularity, in which case we can call this data source big data. In other words, the critical point is the data sources that are always above the processing capacity that computers can access.

Another situation can be as follows. For example, we want to process articles written on Facebook. Let's say we find associated with Turkey and a list of words which we have a simple solution for it and Facebook messages dedicating of the words in this list. We can say that we can handle big data on this problem, but if we want to change the problem a little and follow the relationship of friends on the same data, for example, we will have much less capacity for this problem. Therefore, the third dimension in the definition of large data is the aim to be reached.

For large data it is also a blanket term (blanket term). Blankets are often used to refer to a group of sub-concepts related to each other.

In other words, large data studies actually consist of several subgroups.

1. *Problem of storage of data:*

   - Hardware to store data
   - Lack of database solutions to store data

2. *Problem of data processing:*
   - Memory problems during processing (RAM limits)
   - Time problems during processing (critical role of time in continuous real-time applications)

3. *Data structuring problems (eg. indexing problem in a search engine)*

Based on these informations, the big data concept can be examined in 3 different sizes.

- Volume: Area covered by data
- Velocity: Change or accumulation rate of data
- Variety: Variety of sources of data (email, Facebook, videos, pictures, sound recordings etc.).

It is also possible to add two additional dimensions to the above dimensions.

- Variability: Change in data. For example, a topic with a trend in social networks may change after a short time.
- Complexity: The complexity of processing data.

For instance, Twitter users have a difference in complexity between the follow-up of their Facebook friends and the follow-up of their youtube followers.

Usage Areas of Big Data:

- In the analysis of systems, it has earned billions of dollars in detecting errors and problems and developing solutions.
- It provides economic contribution to the solution of the optimization problems of the working systems in real time. (eg. calculating the best path for moving vehicles).
- It adds economic value to enterprises with market researches (such as stock and product prices) to increase warehouse tracking or profitability.
- Supports marketing and sales with customer oriented data processing activities such as creating campaigns. For example, there are many applications such as discount vouchers, loyalty cards.
- Provides location-dependent solutions with applications developed on mobile devices (eg. making recommendations, advertisements and notifications close to the customer's location and again produced from the customer's past habits.
- In the calculation of risk, for example in the calculation of operations, operations and field risks, it provides instant and fast data in the field of insurance.
- Customer selection (the most important, the most effective, and the most risky, such as the highest purchasing power) can be done quickly.
- System abuses (viruses, malware, system attackers, terrorist organizations etc.) can be found in a short time.

B. *NORA (Non-Obvious Relationship Awareness)*

The concept is simply using multiple data sources to achieve results where the data sources alone are not sufficient for inference. At first, although it has entered the literature for the purpose of finding terrorist groups or criminal organizations, both economic and social is used for many different purposes.

Sample:

Suppose a user searches for holidays in Antalya on a travel site. This is a simple NORA application to show advertisements about the holidays in Antalya on the food site that the user has identified as his / her spouse on Facebook.

NORA application can be seen in various blacklist applications. For example, casinos are a well-known fact that they hold lists of players that are famous for counting cards or doing tricks, and that the NORA app does on these lists. Accordingly, there are practices such as prohibiting the use of persons in casinos with the same cell phone number as the cheating person (against changing names or nicknames). Players who have the potential to cheat from 18 different lists are kept up to date.

Countability is important. In the literature, the concept as countability, cardinality is of great importance in computer processing. For example, countable data such as persons, cases, procedures are always preferred. It is possible to obtain and process vector over these data. On the other hand, unqualified values (such as how much a person may want to make a holiday, or how high an organization is), always give rise to problems such as a higher load in processing and, in some cases, no results.

Personality rights are important but not an obstacle. In the process of processing personal information, it is seen that most users are not sensitive about personal information (such as summary functions) as well as ways of not violating their personal rights (for example, only if a game written

for this purpose will clearly/surely take all your information in the end-user agreement and use it for all purposes. it was seen to confirm without reading it).

It is important to remember that malicious people may also be smart. Generally, the methods and inferences applied are recognized and prevented by malicious people after a certain period of time. With the help of technology, this time is decreasing day by day and life is getting faster. The methods used for this are not to be explained and kept confidential, should be kept at a level that will not disturb the users according to the usage area.

C. Social Search

The section aim is to explain the concept of social search in literature. The concept is no different from other searches on the Internet. The search is done again via the web. The only obvious difference is that the search is limited or prioritized to social networks and, in particular, to the social network where the call is made.

The social graph is the most important feature of social search. A social graph is a form consisting of the contents produced by individuals and persons. For example, each person can be imagined as a node and a social figure in which each friendship relationship is considered as an edge. Such proximity is taken into account in social searches.

In the literature, the concept of social discovery, again, examines the social relations of the searcher and tries to predict the result that the person desires to see the most. For example, you may be interested in a purchase, vacation or training by your friend. In this case, sorting the results of your searches by your friends' history makes the results more preferable. For example, iTunes shows the music preferences of the person's social environment and the most preferred music tracks through the shares and sales processes.

Similarly, people search on Facebook shows results by taking into account the social environment of the person making the call.

D. Text Mining

In the simplest sense, text mining work is a data mining study that accepts the text as a data source. Another definition aims to obtain structured data from text. For example, it aims at studies such as classification, clustering, concept/entity extraction, production of granular taxonomy, sentimental analysis, document summarization, entity relationship modeling.

In order to achieve the above objectives, text mining uses methods such as information retrieval, lexical analysis, word frequency distribution, pattern recognition, tagging, information extraction, data mining and even visualization.

Text mining studies are often accompanied by natural language processing (NLP). Natural language processing studies mainly involve studies based on linguistic knowledge under artificial intelligence. Text mining studies aim to reach more statistical results. During text mining studies, feature extraction is often done by using natural language processing.

Data from a text database are primarily subjected to feature extraction. Then, a machine learning algorithm runs on the extracted properties (classification, clustering, prediction, etc.) and the resulting structured data is obtained.

Although the machine learning stage here is generally used, it is not a requirement for text mining. In some cases, the directly extracted feature may be the searched structured data. In some cases, some statistical methods can be used instead of the machine learning step.

Text sources are usually written in natural language, so a column in a newspaper, a book, an article. Even websites on the internet can be seen as a text source (this is more specifically called web mining). These texts have important information about text mining. For example, the history of the article, the web site where the article is published, the author's information, but not in the text, it can be used in the text of the metadata that can be used in text mining.

In the feature extraction stage, the desired properties can be subtracted from the direct content or the header information of the texts, and the extracted properties can be processed.

Sample Text Mining application:

For example, we have 100 articles. Let us know the authors of these writings (let's say 5 different writers have 20 articles). The new letter will be 101. Finding out which of these 5 authors belong to a classic text mining practice, and in the literature the author is also known as recognition.

Here, for example, we would like to use word frequency in texts for feature extraction. In other words, we think that we can recognize our authors from the frequency of words that they use. In each text, and therefore for each writer, how often he used the word is our feature extraction stage.

Then we give the word frequencies used by the CNN algorithm, for example, to the machine learning algorithm, and let's say that we list the authors who use the most for each word in the 101st article we want to recognize. As a result, a list of possible writers comes up and we are most likely telling the author who may have written this article. This result can be considered as a meaningful and structured result for the 101st article.

7. Related works

Data mining, which has been used effectively in many areas, has become one of the most applied disciplines of our day. Although it has become more widespread with each passing year, it has become one of the most widely used methods due to its easy applicability and effective results. We can summarize the studies performed with data mining by literature review.

A. Data Mining in Agriculture

Recent technologies can provide a wealth of information about agriculture-related activities today, which then can be analyzed to find important information.

B. Surveillance / Mass surveillance

Surveillance is the observation of behaviors, activities or other changing information. Systematically collecting, processing and transferring these data to those who will act according to the results. This generally refers to the monitoring of individuals or groups by state institutions. Disease surveillance, for example, monitors the progression of a disease in society.

C. Data Mining in Engineering

In a post-graduate study conducted by Kıyas Kayaalp, short-circuit or insulation defects that may occur between the winding spirals in the three-phase asynchronous motor and the mechanical imbalance faults that may occur in the motor shaft were determined by data mining technique (Kayaalp, 2007).

Yomi Kastro created an implementation for the purpose of creating a model that estimates the error rate in new versions of software- based on its previous versions. The modifications mentioned in this application may be an innovation in the software, an algorithm change, and even a debugging change. By analyzing the type of such changes from a formal and objective point of view and including the volumetric change of the software, it is aimed to accurately estimate the error rate in the new version. By using the proposed model in this study, it was possible to shorten the test life in the software life cycle and reduce the power spent. In addition, it is possible to determine the robustness of a new software version thanks to this model. This model also helps to understand the contribution of changes in the software product, such as debugging changes, to the possibilities of error generation (Kastro, 2006).

D. Data Mining in Medical Field

A data mining application was implemented by Barış Aksoy on Cluster Analysis of Decompression Analysis. In this study, decompression sickness using different clustering algorithms and symptom and symptom lists obtained from Divers Alert Network diving injury reporting forms were classified. The results were compared with classical classification methods, new statistical classification methods and treatment results. In addition, association rules have been obtained which can help in the diagnosis. As a result, it is observed that the classes obtained by clustering methods are in conformity with the new statistical classifications and classical classifications and they are in hierarchical structure leading to mild to severe cases (Aksoy, 2009).

Mustafa Danacı, Mete Çelik and A. Erhan Akkaya conducted a study were brief information is given about the most common breast cancer among women. Then, with the help of the Xcyt pattern recognition program, general data about the tissue were obtained, and breast cancer cells were determined and diagnosed using the Weka program (Danaci 2010).

E. Data Mining in Banking and Stock Market

In this study carried out by Nihal Ata, Erengül Özkök and Uğur Karabey, after analyzing the methods of life analysis within the framework of data mining, they examined life probabilities, hazard probabilities and regression models for a data set of credit card holders. Accordingly, age, income and marital status were found to be important risk factors for customers to stop using credit cards (Ata 2008).

F. Data Mining in Commercial Area

Çağatan Taşkın and Gül Gökay Emel implemented an application in data mining with clustering approaches and Kohonen networks in the retail sector. In this application; clustering of a retail enterprise customer with Kohonen networks. The purpose of clustering analysis; The aim of the project is to provide critical insights into critical customer characteristics and importance to help the strategic marketing decisions such as market segmentation and target market selection (Taşkın and Emel, 2010).

G. Data Mining in Education

In a study conducted by Serdar Savaş and Nursal Arici, two different teaching materials suitable for video-supported and animation-supported teaching model for web-based distance education were prepared to examine the effects of these materials on student achievement. As a result of the analysis,

it has been determined that video-supported teaching materials have a more positive effect on student achievement than animation-supported teaching materials (Savaş & Arıcı, 2009).

H. Data Mining in Telecommunications

Umman Simsek, Tugba Gürsoy by a big company operating in the telecommunications sector in Turkey, determined that tend to leave customers; it is aimed at developing customized marketing strategies for customers. In order to determine the customer profile to be separated, Decision Trees from Logistic Regression Analysis and classification techniques were used and the results of the application were presented (Gürsoy, 2010).

Selman Bozkır, S. Güzin Mazman and Ebru Akçapınar Sezer conducted a study on social networking. In this study, the current social sharing site is examined user templates on Facebook. Facebook usage period and access frequency were examined on 570 Facebook users and their results were revealed (Bozkır, 2010).

8. Conclusions

Data mining is a highly promising technique to help organizations find patterns that are hidden in their data, models that can be used to predict the behavior of customers, products, and processes. Data mining, however, should be guided by users who understand their work, data, and the overall structure of the analytical methods involved. In line with the realistic expectations, it results in rewarding results in a wide range of applications from the improvement of revenues to the reduction of costs. Different data types are stored in various repositories. It is difficult to wait effectively and efficiently to obtain good mining results from any data mining system from a data mining system. Different data and resources may require different algorithms and techniques. Currently, it focuses on the need for data mining. We made an explanation about the typical architecture of data mining and explained the steps of the data mining process. This article summarizes the functionality of data mining and defines the classification of data mining systems. We talked about the benefits of data mining technology. We also discussed some of the main issues to be addressed and talked about a few applications where data mining technology can be implemented.

References

[1]. Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. 2013 International Conference on Machine Intelligence and Research Advancement.

[2]. Aksoy, B., (2009), Cluster Analysis of Decompression Illness, Galatasaray University, Institute of Science and Engineering.

[3]. Altun, M. (2017). Veri Madenciliği ve Uygulama Alanları

[4]. Ashour A. N. Mostafa. Review of Data Mining Concept and its Techniques.

[5]. Ata, N., Özkök, E. ve Karabey, U., (2008), "Survival Data Mining: An Application To Credit Card Holders", Sigma Mühendislik ve Fen Bilimleri Dergisi, Cilt 26, No 1, 33-42.

[6]. Bharati M. Ramageri. Data Mining Techniques and Applications. Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[7]. Bozkır, A.S., Mazman, S.G., ve Sezer, E.A., (2010), "Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case", 2nd International Symposium on Information Management in a Changing World", 22-24 September, Hacettepe University, Ankara, 145-152.

[8]. Danacı, M., Çelik, M. ve Akkaya, A.E., (2010), "Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi", Akıllı Sistemlerde Yenilikler ve Uygulama Sempozyumu, 21-24 Haz. 2010, Kayseri, 9-12.

[9]. Freitas, A. A. (2013). Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media.

[10]. Gürsoy, U.T.Ş., (2010), "Customer Churn Analysis in Telecommunication Sector", İstanbul University Journal of the School of Business Administration, Cilt 39, No 1, 35-49.

[11]. Kastro, Y., (2006), A Defect Prediction Method for Software Versioning, Yüksek Lisans Tezi, Boğaziçi University, Computer Engineering.

[12]. Kayaalp, K. (2007), Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü.

[13]. Lakshmi, B. N., & Raghunandhan, G. H. (2011). A conceptual overview of data mining. 2011 National Conference on Innovations in Emerging Technology.

[14]. Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. Expert Systems with Applications, 39(12), 11303-11311.

[15]. Özekes, S.Veri Madenciliği Modelleri Ve Uygulama Alanları.

[16]. Rüdiger Wirth & Jochen Hipp. CRISP-DM: Towards a Standard Process Model for Data Mining

[17]. Savaş, S. ve Arıcı, N., (2009), "Web Tabanlı Uzaktan Eğitimde İki Farklı Öğretim Modelinin Öğrenci Başarısı Üzerindeki Etkilerinin İncelenmesi", 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS 09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 1229.

[18]. Taşkın, Ç. ve Emel, G.G., (2010), "Veri Madenciliğinde Kümeleme Yaklaşımları Ve Kohonen Ağları İle Perakendecilik Sektöründe Bir Uygulama", Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 15, No 3, 395- 409.

[19] . Wang, H., & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. Industrial Management & Data Systems, 108(5), 622–634.