International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-7, Issue-2, March 2019
DOI: 10.21276/ijircst.2019.7.2.4

Review Paper on Data Mining Techniques and Applications

Anshu

***Abstract*— Data mining is the process of extracting hidden and useful patterns and information from data. Data mining is a new technology that helps businesses to predict future trends and behaviors, allowing them to make proactive, knowledge driven decisions. The aim of this paper is to show the process of data mining and how it can help decision makers to make better decisions. Practically, data mining is really useful for any organization which has huge amount of data. Data mining help regular databases to perform faster. They also help to increase the profit, because of the correct decisions made with the help of data mining. This paper shows the various steps performed during the process of data mining and how it can be used by various industries to get better answers from huge amount of data.**

***Index Terms*—Data Mining, regression, Time series ,prediction, association**

I. INTRODUCTION

Data mining can be defined as the process of extracting valid, previously unknown and actionable information from large data sets. The purpose of the data mining is to use the extracted information to make crucial business decisions. So, Data Mining helps end users extract useful business information from large volume of data. This is a commonly used word for any kind of large scale data processing. The mined results should be valid, novel, useful, and understandable. Data Mining is related to the subarea of statistics called exploratory data analysis and subarea of artificial intelligence called knowledge discovery and machine learning. This paper presents a brief introduction about data mining in section one. The second section illustrates the process of data mining while the third section reviews different data mining techniques. The fourth section is committed to various application areas of Data Mining and fifth section discusses conclusion and future scope.

Manuscript received March 25, 2019

Anshu, Department of Computer Science & Engineering, G.V.M. Girls College, Sonepat, India,
+91-9991666556, (mca706@gmail.com).

II. PROCESS OF DATA MINING

Data mining process is a step by step procedure that cannot be completed in a single step. In other words, you cannot get the required information from the large volumes of data as simple as that. It is not specific to any particular industry. Basically the process has evolved from the knowledge discover processes used widely in industry. The major aim of Data Mining process is to make large data projects to run more efficiently. The processes including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation are to be completed in the given order.

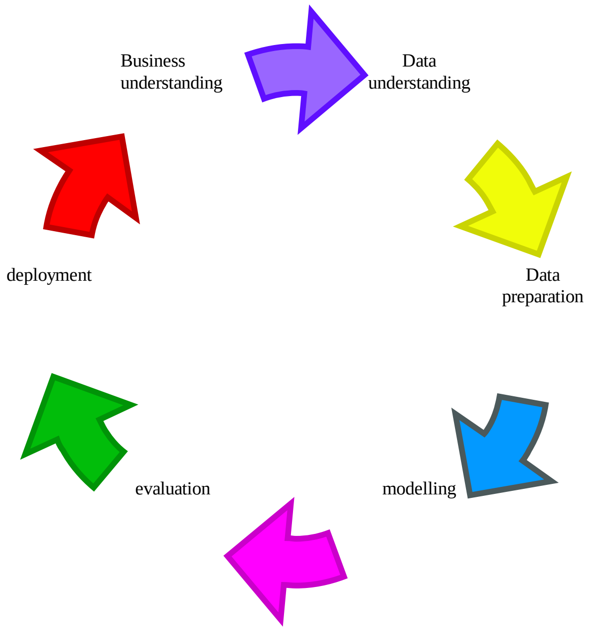

Fig 1: Process of data mining

A. *Business Understanding*

Business understanding phase focuses on understanding the project objectives and requirements, assessing the current situation, establishing data mining goals from the business point of view. In this phase we prepare the preliminary plan for the project. In this phase various activities like determining business objectives, finding current situation, determining data mining goal and producing project plan are involved.

B. *Data Understanding*

This phase includes activities like initial data collection, data description, data exploration, and the verification of data quality. It is basically concerned with establishing the main characteristics of data which includes the data structures, data quality and identifying any interesting subsets of the data. The major tasks performed under this phase are collecting initial data, describing data and exploring data and verify data.

- First, data is collected from multiple data sources available in the organization.
- Next, the step is to search for properties of acquired data.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

C. Data Preparation

This phase involves all the activities for constructing the final data set into the desired form. The main activities performed during this phase are select data, cleaning data, data integration and data transformation. In this phase, data is made production ready. The output of this phase is data set that can be used in modeling.

D. Modeling

In Data Modeling step, we select modeling techniques, modeling parameters and assess the model created based on the business objectives. Once greater data understanding is gained (often through pattern recognition triggered by viewing model output), more detailed models suitable to the data can be applied. The various activities performed during this phase are selecting modeling technique, generate test design, build model and assess model. For creating suitable model following steps are taken:

- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

E. Evaluation

This phase validates the model from the data analysis point of view. In this step the model and the steps in modeling are verified within the context of achieving the business objectives. The various activities performed during this phase include evaluating results, reviewing process. Evaluation results should be evaluated according to the business objectives. A go or no-go decision is taken to move the model in the deployment phase.

F. Deployment

In this phase the knowledge obtained in the form of model is to be organized and presented in such a form that can be used by the business users. This process can be as simple like generating report or it may be complex as implementing the repeatable data mining process again and again. This is the execution phase. The various tasks involved in this phase are plan deployment, plan monitoring and maintenance, produce and review the final report. So, in the deployment phase, Patterns are deployed for desired outcome.

III. DATA MINING TECHNIQUES

A. Classification

Classification techniques are used to classify data records into one among a set of predefined classes. They work by constructing a model of a training dataset consisting of example records with known class labels. Classification is a supervised learning method [3]. Data classification is two-step process. In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. If the accuracy of the model is acceptable then the model can be used to classify the unknown tuples [4]. Various types of classification models can be used like Classification by decision tree induction, Bayesian Classification, Neural Networks, Support Vector Machines (SVM), and Classification Based on Associations etc [14]. Classification techniques have been used in numerous applications ranging from spam detection, credit card fraud detection, speech recognition and computer vision.

B. Clustering

Clustering can be defined as the task of organizing data into groups known as clusters such that the data objects that are similar to each other are put in the same cluster. There is no one correct basis of clustering, there could be many different ways to categorize data objects. Clustering is a form of unsupervised learning in which no class labels are provided. Instead, data records need to be grouped based on how similar they are with other records. For example, clustering can be used for profile generation for target marketing where previous response to mailing campaigns can be used to generate a profile of people who responded and this can be used to predict response and filter mailing lists to achieve the best response. Various clustering methods can be employed like Partitioning Methods, Hierarchical Agglomerative methods, Density based methods, Grid-based methods etc [14].

C. Predication

This technique show how certain attributes within the data will behave in future. For example, on the basis of analysis of buying transactions by customers. Regression is used to map a data item to a real valued prediction variable [15]. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. Basically prediction models are continuous valued functions that are used to predict missing or unavailable numerical data values rather than class labels. Prediction also encompasses the identification of distribution trends based on the available data. Regression analysis is a statistical methodology that is most often used for numeric prediction. Various types of regression methods are used like Linear Regression, Multivariate Linear Regression, Nonlinear Regression, and Multivariate

Nonlinear Regression [14].

D. Association Rule

Association and correlation are used to identify the frequently used items from the large data set. Association rules correlate the presence of a set of items with another range of values for another set of variables. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Because of its nature, association is sometimes referred to as “relation technique”. This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time [18]. This type of technique helps businesses to make certain decisions, such as catalog design, cross-marketing, and customer shopping behavior analysis [17]. For example, whenever a customer buys audio equipment, he or she also buys another electronic gadget such as memory chip Various Types of association rule are used like Multilevel association rule, Multidimensional association rule, Quantitative association rule etc[14].

E. Neural Networks

Neural network is a nonlinear predictive model that learns through training and resembles biological structure. Neural networks provide projections given new situations of interest and answers “what if questions”. These are well suited for continuous valued inputs and outputs. For example a neural network can be trained to identify the risk of any disease from a number of factors Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs [14].

F. Time Series Analysis

Time series analysis is the process of using statistical techniques to detect the similarities within positions of a time series of data, which is a sequence of data taken at regular intervals such as daily sales etc. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events [19]. For example stock market. This technique

G. Summarization

Summarization is abstraction of data. It is obtained by identifying attributes such as customer name, address etc that have too many distinct values and either removing them or performing a roll up operation. Also, we can apply standard statistics on data to represent its summary. For example, long distance race can be summarized total minutes, seconds and height. Association Rule: Association is the most popular data mining techniques and fined most frequent item set. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Because of its nature, association is sometimes referred to as “relation technique”. This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time [20].

H. Sequence Discovery

It uncovers relationships among data [15]. This technique defines a sequential pattern of events and actions. For example, suppose a customer who buys more than twice in the first quarter of the year may be likely to buy at least once during the second quarter.

IV APPLICATIONS OF DATA MINING IN VARIOUS FIELDS

Data Mining technologies can be applied to a large variety of decision making in various business environment. Various industries adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Some of the main applications are listed below:

A. Data Mining in Science and Engineering

Data mining has been widely used in area of science and engineering like bioinformatics, genetics, medicine, education and electrical power engineering. That is why Data Mining is known as multidisciplinary technique. In the field of study on human genetics, the important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility. It is very helpful in diagnosing, preventing and treating the diseases.

B. Data Mining in Banking and Finance

Data mining has been widely used in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. Several data mining techniques like distributed data mining have been researched modeled and developed to help credit card fraud detection. Through Data Mining banks can find hidden correlations between different financial indicators and can identify stock trading rules from historical market data.

C. Data Mining in Sales and Marketing

Data Mining has been widely used in marketing field to make analysis of customer behavior based on their buying patterns like identifying products that are purchased concurrently. Also, Data mining enables businesses to determine the marketing strategies such as advertising, warehouse location etc. The ultimate goal of market analysis is finding the segmentation of customers and products so that businesses promote their most profitable products and maximize the profit. The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping [14].

D. Data Mining in Earthquake Prediction

Data Mining predicts the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth’s crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance) [22].

E. Data Mining in Telecommunication

The telecommunications field implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highly competitive environment. Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service.

F. Data Mining in Agriculture

Data mining is emerging in agriculture field for crop yield analysis a with respect to four parameters namely year, rainfall, production and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine (SVM) .

G. Data Mining in Cloud Computing

Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage .Cloud computing uses the Internet services that rely on clouds of servers to handle tasks The data mining technique in Cloud Computing to perform efficient, reliable and secure services for their users.

H. Data Mining in Retail Industry

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction.

I. Data Mining in Bio Informatics

Data Mining ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

J. Data Mining in Corporate Surveillance

Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

V CONCLUSION AND FUTURE WORK

This paper presents a detailed description of data mining techniques and applications in various fields. Data mining techniques such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Different Data Mining techniques can be used for different purposes. Each technique has its own pros and cons. In future I will work on various classifications and clustering algorithm and their significance.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] Han. J, Kamber. M, Pei. J, " Data Mining Concepts and Techniques", Third edition The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011

[4] Kabra. R, Bichkar. R, "Performance Prediction of Engineering Students using Decision Tree", International Journal of computer Applications, December ,2011

[5] VikramPudi,PRadha Krishna "Data Mining", Oxford University Press, First Edition,2009

[6] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[7] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[8] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.

[9] Tayel,Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141149,2014

[10] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," IEEE J. Quantum Electron., submitted for publication.

[11] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[12] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces (Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[13] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[14] Ramageri ," Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[15] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.

[16] Jeffrey Voasand JiaZhang, ―Cloud Computing: New Wine or Just a New Bottle? ‖, Database Systems

Journal vol. III, no. 3/2012 71IEEEInternet Computing Magazine.

[17] Yudho Giri Sucahyo, Ph. D, CISA: Introduction to Data Mining and Business Intellegence.

[18] Bianca V. D.,PhilippeBoula de Mareüil and Martine AddaDecker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and EngineeringSciences(LIMSI)".Websitewww.limsi.fr/Individu/bianca/articl e/Vieru&Boula&Madda_ParaLing07.pdf

[19] Time Series Analysis and Forecasting with Weka , http://wiki.pentaho.com/display/DATAMINING/

[20] Data mining white paper, www.ikanow.com

[21] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-BasedSystems.

[22] Venkatadri.M and Lokanatha C. Reddy, "A comparative study on decision tree classification algorithm in data mining" , International Journal Of Computer Applications In Engineering ,Technology And Sciences

[23] (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)

[24] J. Jones. (1991, May 10). Networks (2nd Ed.) [Online]. Available: http://www.atm.com

[25] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. IEEE Trans. Plasma Sci. [Online]. 21(3). pp.876—880.Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

ABOUT THE AUTHOR

Anshu, NET Qualified ,M.Tech Professional, with a rich Experience of 9 Years in the field of Teaching and Education, and current working as an Asst. Professor, Computer Science at GVM Girls college, Sonipat