DOI: http://dx.doi.org/10.26483/ijarcs.v13i2.6807

ISSN No. 0976-5697

Volume 13, No. 2, March-April 2022

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A SYSTEMATIC REVIEW ON DATA MINING METHODS AND APPLICATIONS

Archana Thakur
*Assistant Professor,
School of Computer Science & IT, Devi Ahilya University,
Indore, (Madhya Pradesh), India.

Abstract: Data mining is the practice of extracting concealed, helpful patterns and information from data. It is a novel technology that assists organizations to forecast future trends and actions, permitting them to make practical, knowledge driven decisions. The present work describes the data mining process and how it can assist decision makers to take better decisions. Practically, data mining is very fruitful for large sized organizations with huge amount of data. It also helps to augment the net profit, as a result of correct decisions taken during the right time. This paper presents the various steps taken during the data mining process and how organizations can get better answer queries from huge datasets. It also presents detailed review on data mining methods and applications.

Keywords: Data mining; classification; clustering; association; prediction.

I. INTRODUCTION

Data mining is a procedure of extracting unknown, valid and actionable information from large data sets. With the help of data mining techniques, the extracted information is used to take critical business decisions. In other words, data mining assists end users to extract fruitful, organized information from large voluminous data. Data mining is a common domain for extracting patterns from any type of large-scale data sets. The mined outcomes should be new, valid, fruitful, and understandable. Data mining also relates to the subfield of statistics termed as exploratory data analysis and subfield of artificial intelligence termed as knowledge discovery and machine learning. This paper presents a brief review on data mining process and methods. The present work explains the process of data mining and also reviews different data mining methods. It also presents data mining application domains.

II. DATA MINING PROCESS

Data mining process is a step-by-step method that cannot be finished in a single step. In other words, you cannot get the desired information easily from large voluminous data, without using the data mining methods. It is not specific to any specific industry. Basically, the data mining process has advanced from the knowledge discovery processes used largely in industry. The data mining process attempts to make big data projects to execute more efficiently. The processes of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation are to be completed in the given order. The steps involved in the data mining process are as under-

A. Understanding the business
B. Understanding the data
C. Preparing the data
D. Use of data mining model
E. Evaluation of model results
F. Deployment of model

A. *Understanding the business*

This phase focuses on development of basic understanding of the objectives and the requirements of the project. It also encompasses the current situation assessment, establishing the goals of data mining from the point of view of business. In this phase, we develop the initial plan for the project. In this phase different activities like determining different business objectives, searching the current situation, finding the data mining goal and producing the project plan exists.

B. *Understanding the data*

This phase has activities like collection of data, description of data, exploration of data, and the verification of quality of data. It basically deals with establishing the major attributes of data that contains the data structures, quality of data and recognizing any significant subsets of the data. The major functions performed under this phase are collecting initial data, describing the data, exploring the data and verifying the data.

- First, data is collected from different data sources that are available in the organization.
- Next, step is to search for the attributes of the acquired data.
- Based on the results obtained from the query, the data quality should be established. Missing data if exists should be acquired.

C. *Preparing the data*

This phase includes all the steps for constructing the final data set into the requested form. The major functions executed during this phase are selection of data, cleaning of data, integration of data and transformation of data. It is the phase where data is made production ready. The output from this phase is the final data set that can be used in modeling.

D. *Use of data mining model*

In this step, modeling techniques are selected, modeling parameters are set and assessment model is designed based on the business objectives. Once there is greater data understanding, more detailed models appropriate to the data can be applied. The various activities executed during this phase are selection of modeling technique, generate test set design, design and assess the model. For creating appropriate model, the following steps should be taken:

- Creation of a scenario to test the quality and validity of the model
- On the prepared dataset, execute the model
- All stakeholders should evaluate the model results so as to meet the objectives of data mining

E. *Evaluation of model results*

During this phase the model is validated from the data analysis point of view. The model and its steps are verified in the context of attaining business objectives. The different activities executed in this phase consist of evaluating results, reviewing the process etc. Results are evaluated as per the business objectives. A go or no-go decision is opted to shift the model in the deployment phase.

F. *Deployment of model*

During this phase the knowledge attained in the form of model is to be arranged and presented in a way that can be easily employed by the business users. This process can be simple or it may be complex like implementing the process of data mining again and again. This is the implementation phase. This phase consists of various tasks like plan deployment, plan monitoring & maintenance, produce & review the final report. Hence in this phase patterns are deployed for the desired outcome.

III. Data mining methods

The various data mining methods are discussed in this section. These methods are helpful in new knowledge discovery. The methods include classification, clustering, prediction, association rule mining, time series analysis, neural networks and summarization. Any of these methods can be used for effective decision making.

A. *Classification*

Classification is a vital method used in data mining. It enables assignment of data instances in one of the predefined classes. The classification method falls in the category of supervised learning methods [1]. Depending upon the number of classes, the classification problem can be unary, binary or multiclass. The classification method constructs a model of training dataset containing set of example instances with known class labels. Basically, classification problem involves two steps. During the first step, a model is constructed by analyzing the instances from the training dataset consisting of a group of features. In this case for each instance present in the training dataset, the value of class label feature is known. The model is executed for the training set. If the model results in significant accuracy, then the model can be further used to classify unidentified instances [2]. The various classification methods that can be used for new knowledge discovery [3] are classification by Decision Tree induction, Bayesian classification, Neural Network, SVM and classification based on Associations, Naïve Bayesian method, Logistic Regression, Classification and Regression Tree etc. Classification methods can be efficiently employed in various applications like precision agriculture, credit-card fraud recognition systems, malware identification, computer vision etc.

B. *Clustering*

The clustering method involves arranging data into groups called as clusters so that the data objects that are of similar type are put together in the same cluster. There exist multiple ways to categorize the data objects. Clustering falls under the category of unsupervised learning in which there are no class labels provided. Instead, the data instances are grouped based upon how comparable they are with other instances. Partitioning methods, Density-based methods, Hierarchical-agglomerative methods, Grid-based methods are the different clustering methods that can be employed for effective decision making [3].

C. *Prediction*

This method depicts how certain features or attributes present within the data will behave in future. For example, the interests of products purchased by customers can be predicted by analyzing the purchase transactions of customers. Regression is one of the famous methods that is used to map a data item to a real-valued predictor variable [4]. The relationship between one or more independent variables and dependent variables can be analyzed with the help of a regression model. The prediction models fall under the category of continuous valued functions. These functions are exercised to forecast the missing or unavailable numeric values of data rather than the labels of classes. Prediction also includes the recognition of distribution trends based upon the data available. Regression analysis has evolved from statistics that is commonly used for numeric forecast [3]. The famous regression methods are Linear-Regression, Multivariate -Linear-Regression, Nonlinear-Regression, & Multivariate-Nonlinear-Regression.

D. *Asoociation rule*

The rules of association and correlation are employed to recognize the commonly used items from the big datasets. The rules of association correlate the presence of a group of items with additional range of values for another group of variables. Association attempts to determine patterns in data which are based upon relationships amongst items belonging to the same transaction. It is also called as “relation technique”. This method of data mining is fruitful is performing market-based analysis. It is used to recognize a group, or groups of products that customers frequently buy at the same time [6]. This method assists businesses to take certain decisions, such as customer shopping, design of various catalogs, cross-marketing behavior- analysis [5] etc. Association rule mining can be applied on certain types of problems for example, a customer does conditional purchasing like whenever he or she purchases a television set he or she also purchases another electronic gadget such as a radio. The different categories of association rule methods [3] are Quantitative association rule, Multilevel association rule, Multidimensional association rule etc.

E. *Neural network*

Neural network is a type of nonlinear predictive model. It resembles a biological neuron. It learns through training cycles. It offers projections and attempt to answer "what if type of questions". These models are appropriate for continuous valued inputs & outputs [3]. For example, a neural network model can be trained with symptoms in order to recognize certain disease(s). These models are best for recognizing the patterns or trends in data. These models are appropriate for prediction or forecasting problems.

F. *Time-series analysis*

Time-series analysis is the method of employing statistical methods. It is mostly used to notice the similarities or likeness within the positions of a time-series of data, which is a sequence of data collected during regular time intervals such as daily sales etc. It is a forecasting method. It uses a model to make predictions (forecasts) for futuristic events based upon some known past events [7]. For example, stock market analysis uses time-series analysis.

G. *Summarization*

Summarization is in simple terms abstraction of data. It is achieved by recognizing the features or attributes like customer name, customer date of birth, customer address, customer mobile number etc. that have distinct values. Mining operation can be performed either by eliminating redundant or inconsistent values or selecting a subset of features from them or performing a roll up operation [8]. Also, a user can apply some standard statistical method on data to represent its summary. For example, a long-distance marathon can be summarized in marathoner, speed, total time.

IV. APPLICATIONS OF DATA MINING

Data mining methods can be applied on many business areas for a variety of decision making. Different organizations have adopted data mining methods because of quick access of data and important information from a large amount of data. Some of the vital applications of data mining are as under-

A. *Data mining methods in engineering and science*

Data mining is widely used in the area of engineering and science for example in bioinformatics, medicine, genetics, electrical engineering and education field etc. Hence data mining is a multidisciplinary domain. One of the vital applications of data mining is in the field of study on human genetics, where the major focus is to recognize the relationship mapping amid the inter individual variation in human DNA sequences and changeability in disease vulnerability. It is very helpful in identifying, preventing and curing the diseases.

B. *Data mining methods in finacial and banking sectors*

Data mining is extensively useful in financial and banking sectors. In the field of banking, data mining is used to forecast credit card fraud, to approximate various risks, to examine the recent developments and profitability. Various data mining methods like distributed data mining have been researched, modeled and developed to assist in credit card fraud identification. With the help of data mining banks can detect hidden correlations among various financial indicators and can recognize stock trading rules with the help of historical market data.

C. *Applications of data mining in sales and marketing*

Data Mining is enormously used in marketing field to conduct analysis of customer behavior based on their purchasing patterns for example recognizing products that are purchased simultaneously. Also, data mining facilitates organizations to find out the marketing strategies like advertising, warehouse location etc. The final aim of market analysis is determining the portions of customers and products so that enterprises encourage their most profitable products and maximize the profit. The stores can use this information by accumulating these products in close nearness of each other. It helps in making these products more evident and reachable to customers at the time of shopping [3].

D. *Data mining methods in forecasting*

Data mining methods can be used to forecast earthquakes from the satellite maps. Earthquake is the unexpected movement of the Earth's crust caused by the sudden liberation of stress collected along a geologic error inside the earth's surface. Basically, there are two streams of earthquake predictions: first is forecast stream where predictions are made in advance from months to years and second is short-term prediction where predictions are made in advance from hours or days [9].

E. *Data mining applications in telecommunication systems*

The data mining applications are enormously used in telecommunication systems as these industries have large volumes of data. Telecommunication industries also have a very large customer base, and quickly changing and too much competitive environment. Data mining applications in telecommunication industry assist in recognizing the telecommunication patterns, holding fraudulent activities, optimization of resources, and enhance the quality of service.

F. *Data mining methods in agriculture*

Data mining methods are enormously used in the field of agriculture. One of the vital applications is in the field of crop yield analysis with respect to features like year, rainfall, production and area of sowing. The crop yield prediction is a vital agricultural problem that can be solved using data mining methods. Data mining methods like K Nearest Neighbor (KNN), SVM, Artificial Neural Network (ANN), K-Means are helpful in solving crop yield prediction problem.

G. *Data mining applications in cloud computing*

Data Mining applications are used tremendously in the field of cloud computing. The use of data mining applications in cloud computing permits the users to access important information from virtually integrated data warehouse that lessens the costs related to infrastructure and storage. Cloud computing employs the internet services that trust on clouds of servers to handle various tasks The use of data mining applications in cloud computing accomplishes effective, consistent and secure services for their users.

H. *Data mining applications in retail industries*

Data mining methods are enormously used in retail industries. Data mining methods assist in recognizing customer purchasing patterns and tendencies that result in enhanced quality of customer service. The use of data mining methods enables good customer satisfaction and retention.

I. Data mining applications in bioinformatics

There are many applications of data mining in bioinformatics, since it is a data-rich field. Extracting biological data assists us to mine valuable knowledge from enormous datasets collected in biology, and in other areas related to life sciences like neuroscience and medicine. Different applications of data mining in the field of bioinformatics comprise of disease recognition, determination of gene, inference of protein function, disease prediction, suggesting optimized treatment of identified diseases, forecasting protein sub-cellular location cleansing the data etc.

J. Data mining methods in surveillance

Corporate surveillance is the field of observing a person or group's conduct by a corporation. The data collected is most frequently employed for the purpose of marketing or is sold to other corporations, but is also often shared with various government agencies. It can be employed by the corporations to adapt their products required by their customers. The data can be employed for the purpose of direct marketing, for example the targeted advertisements on Yahoo and Google.

V. CONCLUSIONS AND FUTURE SCOPE

A comprehensive description of different data mining methods and applications in various fields were presented in the present work. The different data mining methods like classification, prediction, association rule mining, clustering etc., assist us in determining the various patterns to decide upon the future inclinations in businesses to expand. The different data mining methods can be employed for various purposes. Each data mining method has its own benefits and drawbacks. As a part of future work various enhancements will be suggested for classification and clustering algorithms useful in various domains.

VI. REFERENCES

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

[2] R.R Kabra, and R.S. Bichkar, "Performance Prediction of Engineering Students using Decision Tree", International Journal of computer Applications, Vol 36, Issue 11, pp. 8-12, December, 2011.

[3] B.M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering Vol. 1 No. 4, pp. 301-305, 2010.

[4] M.H. Dunham, "Data Mining, Introductory and Advanced Topics", Pearson Education, 2014.

[5] G. Parker, "Data Mining: Modules in emerging fields, CD-ROM", vol 7, 2004.

[6] K.E. DiCerbo, and K. Kidwai, "Detecting player goals from game log files," in Poster presented at the Sixth International Conference on Educational Data Mining (Memphis, TN), 2013.

[7] M. Rafiuzzaman, "Forecasting Chaotic Stock Market Data using Time Series Data Mining", International journal of computer application (0975-8887) Volume 101- Issue 10, September 2014.

[8] B. Xu, M. Recker, X. Qi, N. Flann, and L. Ye, "Clustering educational digital library usage data: a comparison of latent class analysis and k-means algorithms. J. Educ. Data Mining 5, pp. 38–68, 2013.

[9] M. Venkatadri, and L.C. Reddy, "A comparative study on decision tree classification algorithm in data mining", International Journal of Computer Applications in Engineering, Technology and Sciences, Vol 2, Issue 2, pp. 24-29, 2010.