IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-6, Issue-1 E-ISSN: 2347-2693

A Comprehensive Review on Data Mining Techniques and Applications

A. Thakur[1]

[1]School of Computer Science & IT, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India

Author's Mail Id: archana227@gmail.com

Available online at: www.ijcseonline.org

Received: 06/Jan/2018, Revised: 13/Jan/2018, Accepted: 25/Jan/2018, Published: 31/Jan/2018

***Abstract*—** Data mining is the study of mining concealed, helpful patterns and information from data. It is a new technology that helps organizations to estimate future trends and actions, allowing them to make real-world, knowledge driven decisions. The current work discusses the data mining process and how it can help the decision makers to opt for better decisions. Practically, data mining is very productive for large sized organizations with enormous amount of data. It also aids to augment the net profit, as a consequence of right decisions taken during the exact time. This paper presents the different steps taken during the data mining process and how organizations can have better answer to the queries from huge datasets. It also presents a systematic review on data mining techniques and applications.

***Keywords*—** Data Mining, Classification, Clustering, Association Rule, Neural Network.

I. INTRODUCTION

Data mining is the study of mining the unknown, valid and actionable information from big data sets. With the help of data mining techniques, the extracted information is used to take critical business decisions. In other words, data mining helps the end users to mine productive, ordered information from large capacious data. Data mining is a public domain for mining patterns from any type of large-scale data sets. The mined outcomes should be novel, valid, productive, and understandable. Data mining is also related to the subfield of statistics known as exploratory data analysis and subfield of artificial intelligence known as knowledge discovery and machine learning. This paper presents a comprehensive review on data mining process and techniques. The present work investigates the process of data mining and also reviews various data mining techniques. It also explores data mining application domains.

II. THE DATA MINING PROCESS

Data mining process is a systematic method that cannot be finished in a single step. In other words, one has to use data mining techniques to get the required information easily from large voluminous datasets. It is not specific to any particular industry. Fundamentally, the data mining process has progressed from the knowledge discovery processes used mainly in industry. The data mining process tries to make big data projects to accomplish them more efficiently. The processes of cleaning the data, data integration, selecting the data, transforming the data, data mining, pattern evaluation and knowledge representation are to be accomplished in the given order. The steps 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 concentrates upon the development of basic understanding of the objectives and the requirements of the project. It also consists of the existing situation assessment and establishing the goals of data mining from the point of view of business. In thisphase, we develop the original plan for the project. In this phase various activities like shaping different business objectives, searching the present situation, determining the data mining goal and producing the project plan exist.

B. Understanding the data

This phase has activities like data collection, data description, data exploration and the confirmation of quality of data. It essentially deals with creating the main features of data that contains the data structures, quality of data and identifying any important subsets of the data. The main functions performed under this phase are collecting original data, describing the data, exploring the data and verifying the data.

- Initially, data is collected from various data sources that are existing in the organization.
- Next step is to search for the attributes or features of the attained data.
- Based on the results attained from the query, the data quality should be recognized. Missing data if exists should be attained.

C. Preparing the data

This phase includes all the steps for creating the final data set into the requested form. The main functions performed 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 concluding data set that can be used in modeling.

D. Use of data mining model

In this step, modeling techniques are chosen, modeling parameters are set and valuation model is designed based upon the business goals. Once there is greater data understanding, more comprehensive models suitable to the data can be applied. The various activities executed during this phase are selection of modeling technique, generate test set design, design and evaluate the model. For creating suitable model, the following steps should be taken:

- Formation of a scenario to test the quality and validity of the model
- On the prepared dataset, apply the model
- All stakeholders should assess the model results so as to meet the goals of data mining

E. Evaluation of model results

During this phase the model is authenticated from the data analysis point of view. The model and its steps are confirmed in the context of achieving business goals. The various activities executed in this phase consist of assessing results, reviewing the process etc. Results are estimated as per the business goals. A go or no-go decision is chosen to shift the model in the deployment phase.

F. Deployment of model

During this phase the knowledge achieved in the form of model is to be organized and offered in a way that can be easily employed by the business users. This process can be simple or it may be difficult like executing the process of data mining again and again. This is the execution phase. This phase consists of different tasks like plan deployment, plan monitoring & maintenance, produce & review the final report. Hence in this phase patterns are deployed for the required outcome.

III. DATA MINING TECHNIQUES

The different data mining techniques are discussed in this section. These techniques are supportive in innovative knowledge discovery. The techniques include classification, clustering, prediction, association rule mining, time series analysis, neural networks and summarization. Any of these techniques can be used for effective decision making. The techniques are described as under –

A. Classification

Classification is a vital technique used in data mining. It allows assignment of data instances in one of the predefined classes. The classification technique falls in the category of supervised learning [1]. Based on the number of classes, the classification problem can be unary, binary or multiclass. The classification technique creates a model of training dataset encompassing set of sample instances with known class labels. Essentially, classification problem includes two steps. During the first step, a model is created by examining 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 recognized. The model is executed for the training set. If the model results in substantial accuracy, then the model can be further used to classify the instances with unknown class label [2]. The various classification techniques that can be used for novel 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 techniques can be competently employed in numerous applications like precision agriculture, credit-card fraud recognition systems, malware identification, computer vision etc.

B. Clustering

The clustering technique consists of organizing data into groups known as clusters so that the data objects that are of comparable types are put together in the same cluster. There exist numerous ways to group the data objects. Clustering falls in 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 techniques, Density-based techniques, Hierarchical-agglomerative techniques, Grid-based techniques are the various clustering techniques that can be employed for efficient decision making [3].

C. Prediction

This technique portrays how certain features or attributes present within the data will behave in future. For example, the interests of products purchased by customers can be forecasted by investigating the purchase transactions of customers. Regression is one of the well-known techniques that is used to map a data item to a real-valued predictor variable [4]. The relationship amongst one or more independent variables and dependent variables can be examined using a regression model. The prediction models are in 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 comprises of the identification of distribution trends based upon the available data. Regression analysis has evolved from statistics that is usually used for numeric forecast [3]. The famous regression techniques are Linear-Regression, Multivariate Linear-Regression, Nonlinear-Regression, & Multivariate-Nonlinear-Regression.

D. Association rule

The rules of association and correlation are employed to recognize the normally used items from the big datasets. The rules of association relate the presence of a group of items with supplementary range of values for another group of variables. Association attempts to find out the patterns in datawhich are based upon relationships between items belonging to the same transaction. It is also called as "relation technique". This technique of data mining is productive in performing market-based analysis. It is used to identify a group, or groups of products that customers normally buy at the same time [6]. This technique helps industries to take certain decisions, such as customer shopping, design of different catalogues, cross-marketing behavior- analysis [5] etc. Association rule mining can be applied on different types of problems for example, a customer does conditional purchasing like whenever he or she purchases a television set he or she also buys another electronic gadget such as a radio. The various categories of association rule techniques [3] are Quantitative association rule, Multilevel association rule, Multidimensional association rule etc.

E. Neural network

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

F. Time series analysis

Time-series analysis is the technique of using statistical techniques. It is typically used to notice the similarities or likeness within the positions of a time-series of data, which is a sequence of data collected throughout regular time intervals for example daily sales etc. It is a predicting technique. It employs a model to make predictions (forecasts) for futuristic events based upon some known past events [7]. For example, stock market analysis employs time-series analysis.

G. Summarization

Summarization is in simple terms abstraction of data. It is achieved by knowing the features or attributes like customer name, customer date of birth, customer address, customer mobile number etc. that have different values. Mining operation can be performed either by removing redundant or inconsistent values or by selecting a subset of features from them or by executing a roll up operation [8]. Also, a user can apply some standard statistical technique 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 TECHNIQUES

Data mining techniques can be used in many business areas for a variety of decision making. Different organizations have approved data mining techniques because ofquick access of data and significant information from a large amount of data. Some of the important applications of data mining are discussed as under-

A. Data mining techniques in engineering and science

Data mining is extensively 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 identify the relationship mapping amid the inter individual variation in human DNA sequences and changeability in disease susceptibility. It is very supportive in recognizing, avoiding and curing the diseases.

B. Data mining techniques in financial and banking sectors

Data mining is broadly suitable in financial and banking sectors. In the field of banking, data mining is used to forecast credit card fraud, to estimate different risks, to study the recent developments and profitability. Different data mining techniques like distributed data mining have been researched, modeled and developed to help in credit card fraud detection. With the help of data mining banks can discover concealed correlations among various financial indicators and can identify stock trading rules with the help of historical market data.

C. Data mining techniques in sales and marketing

Data Mining is enormously used in marketing field to conduct study of customer behavior based on their buying patterns. For example, identifying products that are purchased concurrently. Also, data mining enables organizations to find out the marketing strategies like advertising, warehouse location etc. The final goal of market analysis is finding out the segments of customers and products so that enterprises inspire their most profitable products and maximize the profit. The stores can use this information by gathering these products in close nearness of each other. It aids in making these products more obvious and reachable to customers at the time of shopping [3].

D. Data mining techniques in forecasting

Data mining techniques can be used to predict earthquakes from the satellite maps. Fundamentally, there are two streams of earthquake predictions: first is to predict the stream where predictions are prepared in advance from months to years and second is short- term prediction where predictions are prepared in advance from hours or days [9].

E. Data mining applications in tele communication systems

The data mining applications are extremely used in telecommunication systems as these businesses have large voluminous data. Telecommunication businesses also have a very large customer base, and quickly changing and too much competitive environment. Data mining applications in telecommunication industry help in knowing the telecommunication patterns, holding fraudulent activities, optimization of resources, and improve the quality of service.

F. Data mining techniques 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 arehelpful in solving crop yield prediction problem.

G. Data mining techniques in cloud computing

Data Mining applications are used extremely in the field of cloud computing. The use of data mining applications in cloud computing allows the users to access significant information from virtually integrated data warehouse that reduces the costs related to infrastructure and storage. Cloud computing employs the internet services that relies upon clouds of servers to handle different tasks The use of data mining applications in cloud computing achieves effective, consistent and secure services for their users.

H. Data mining applications in retail industries

Data mining techniques are extremely used in retail industries. Data mining techniques help in identifying customer buying patterns and inclinations that result in improved quality of customer service. The use of data mining techniques aid in good customer satisfaction and retention.

I. Data mining applications in bioinformatics

There are multiple applications of data mining in bioinformatics, since it is a data-rich field. Mining biological data helps us to extract valuable knowledge from massive datasets collected in biology, and in other areas of life sciences like neuroscience and medicine. Different applications of data mining in the field of bioinformatics consists of disease diagnosis, determining the gene structure, inference of protein function, prediction of disease(s), suggesting optimized treatment of recognized diseases, predicting protein sub-cellular location cleansing the data etc.

J. Data mining applications in surveillance

Corporate surveillance is the field of witnessing a person or group’s conduct by a corporation. The data collected is most commonly used for the purpose of marketing or is sold to other corporations, but is also frequently shared with different government organizations. It can be used by the corporations to familiarize their products required by the customers. The data can be used for the purpose of direct marketing, for example the targeted advertisements on Yahoo and Google.

V. CONCLUSIONS AND FUTURE SCOPE

A comprehensive description of various data mining techniques and applications in numerous fields were presented in the current work. The different data mining techniques like classification, prediction, association rule mining, clustering etc., help us in determining the different patterns to decide upon the future inclinations in industries to expand. The different data mining techniques can be used for different purposes. Each data mining technique has its own advantages and disadvantages. As a part of future work different improvements will be recommended for classification and clustering techniques useful in different domains.

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.**

AUTHOR’S PROFILE

Dr. Archana Chaudhary Thakur received M.Tech. and Ph.D. from School of Computer Science & IT, Devi Ahilya University, Indore. She is working as an Assistant Professor at School of Computer Science & IT, Devi Ahilya University, Indore. She is involved in coordinating postgraduate-level training program in computer science for the university. She is guiding many M.Tech. and Ph.D. research scholars. She has published many research papers in various reputed national and international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE whose manuscripts are also available online. She has also been esteemed author and reviewer for many Elsevier journals. Her research areas include Artificial Intelligence, Machine learning, Data Mining and Soft Computing.