A Literature Review: Data Mining Techniques, Applications & Issues

S. L. Nalawade[1], A. M. Joshi[2]

[1] Assistant Professor, Department of Computer Science, K.G.D.B.L.M., Kundal, Maharashtra, India
[2] Assistant Professor, Department of Computer Science, K.G.D.B.L.M., Kundal, Maharashtra, India

Abstract

Data mining is a process which finds useful patterns from large amount of data by turning collection of data into knowledge. The concept of data mining is center of attraction for the users because of many factors as high availability of data which needs to be converted from masses of data to useful information. Data mining as a tool was used to tackle the situation. Data mining is considered as stepping stone to procedure of knowledge discovery in databases; this is a procedure of extracting hidden information from enormous set of databases to excavate eloquent patterns and rules. This article provides an analysis of the available literature on data mining as well as some techniques, applications related to it have also been illustrated.

***Keywords**: Data Mining, Data mining techniques, Knowledge discovery in database, Knowledge base, Clustering.*

1. Introduction

Data mining involves discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. Many other terms are used for data mining, for example, knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, and information harvesting [1]. The objective of any data mining process is to build an efficient predictive or descriptive model of a large amount of data that not only best fits or explains it, but is also able to generalize to new data [2]. Based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored in either databases, data warehouses, or other information repositories.

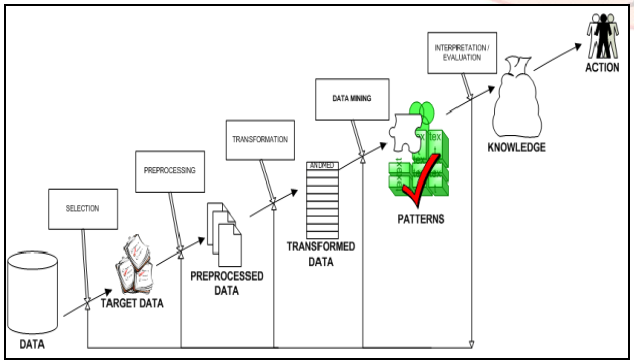

Figure 1. Knowledge Data Mining

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning –In this step, the noise and inconsistent data is removed.

Data Integration – In this step, multiple data sources are combined.

Data Selection – In this step, data relevant to the analysis task are retrieved from the database.

Data Transformation – In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Data Mining – In this step, intelligent methods are applied in order to extract data patterns.

Pattern Evaluation – In this step, data patterns are evaluated.

Knowledge Presentation – In this step, knowledge is represented to the user. [3].

2. Review of Literature

Swati N & Dr. R.V. Kulkarni(2015)[13] This paper is used to check relationship between the number of blood donors of a particular age and blood group as well as blood group of donors and disease. The purpose of this work is to analyze a data to extract knowledge of blood donor's association to aid clinical decisions in blood bank center. This study utilized real world data collected from blood

bank department of Hindratna Prakash bapu Patil Blood Bank in Sangli, Maharashtra and used Apriori algorithm for the association of donors, which can help the blood bank owner to make proper decisions faster and more accurately. Fayyad and Stolorz (1997)[14] in their paper described KDD as "generalized procedure of uncovering treasured knowledge from data with mining being one among other steps in that process that uses some algorithms for knowledge extraction process". Michael Goebel et.al (1999)[15] in their paper "A survey of data mining and knowledge discovery tools", provided an generalized view of common knowledge discovery tasks and various methodologies to resolve these. A feature classification scheme was proposed that was used to study knowledge and data mining software's. According to Rygielski et.al (2002)[16], data mining technology has added a new dimension to CRM. The data mining's' power to extract the predictive unknown information from vast datasets have found its way into the CRM to identify and evaluate valuable customers, predict the customers shopping behavior which results in helping the vendors taking proactive and knowledge based decisions. Venkatadri.M et.al (2011)[17], discuss appropriate Techniques and methodologies are needed in future to cater the needs of data mining field as it is exploring more and more complex fields so that we can explore the such complex situations where data is huge but is full of hidden information. Anand.v. saurkar et.al (2014)[18] defined data mining as "interdisciplinary field which consists of integrated databases, artificial intelligence, machine learning, statistics etc.". They defined data mining as multi-step process which comprises preparation of data for mining, mining algorithms, analysis of results and interpretation of results.

3. Techniques used in Data mining

Several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine them with example to have a good overview of them.

3.1. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in reservation systems analysis to identify in which area customers frequently make reservations. Based on this data businesses can set up corresponding reservation counters in that area to sell more tickets and make more profit.

3.2. Classification

Classification is based on machine learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Basically classification is used to categorize each item in a set of data into one of predefined set of classes or groups[4]. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay".

3.3. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. Consider library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

3.4. Prediction

It is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables [5]. For instance, prediction technique can be used in Library to predict books that need to be purchased for the future if we assume that the courses offered by a university are constant. Courses are independent variable, and books could be a dependent variable.

3.5. Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

3.6. Discrimination

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures [3].

4. Applications of Data Mining

The data mining applications in sale/marketing, finance, health care and insurance, transportation and medicine and many other sectors of day today life are remarkable. But some other distinct applications of data mining are listed below:

4.1. In Computer Security

It concentrates heavily on the use of data mining in the area of intrusion detection. The reason for this is twofold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity. This book also addresses the application of data mining to computer forensics. This is a crucial area that seeks to address the needs of law enforcement in analyzing the digital evidence [6].

4.2. In Bioinformatics

Developments in genomics and proteomics have generated a large amount of biological data in the near past. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science [7]. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Specific applications in this section of data mining are protein structure prediction, gene classification, cancer classification etc. Hence we can say that there is potential increase in the interaction between data mining and bioinformatics.

4.3. In Telecommunications Industry

The telecommunications industry was one of the first to adopt data mining technology. This is most likely because telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. Telecommunication companies utilize data mining to improve their marketing efforts, identify fraud, and better manage their telecommunication networks [8]. However, these companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events—such as customer fraud and network failures—in real-time.

4.4. In Customer Relationship Management

CRM can be defined as the process of predicting customer behavior and selecting actions to influence that behavior for the benefit of the company [9]. What marketers want is nothing but "Increasing customer revenue and customer profitability and keeping the customers for a longer period of time." The solution is to apply data mining. Data mining techniques can be of immense help to the organization in solving business problems by: Finding patterns, associations and correlations which are hidden in the business information stored in the databases.

4.5. In Banking

Apart from execution of business processes, the creation of knowledge base and its utilization for the benefit of the organization is becoming a strategy tool to compete. The banking sector has started realizing the need of the techniques like data mining which can help them to compete in the market. Since 1990's the whole concept of banking has been shifted to centralized databases, online transaction sand ATM's all over the world, which has made banking system technically strong and more customer oriented. In the present day environment, the huge amount of electronic data is being maintained by banks around the globe. The huge size of these data bases makes it impossible for the organizations to analyze these data bases and to retrieve useful information as per the need of the decision makers [10, 11]. In today's global

competition for banks to survive should adopt a better Customer Relationship Management. Data mining is backbone for finding unrevealed information about most profitable customers, standalone customers, customers likely to leave and Customer Relationship Development. So use of Associations with Apriori Algorithm is best practices here [12].

4.6. In Healthcare

Data mining has been used by many organizations and in healthcare; data mining is becoming increasingly popular. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, evaluation of treatment effectiveness, management of healthcare, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services, for quality control and maintenance scheduling. It is useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management. Recent technologies are used in medical field to enhance the medical services in cost effective manner. In this paper the different relevant data mining tools used for Healthcare are reviewed and proposes a data model for monitoring individual's information for population based health care management [13].

5. Issues In Data Mining

Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way [19].

 Security and Social Issue: Security is an important issue with any data collection that is intended to be shared. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

 Data integrity: Data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

 Mining Methodology: An important technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment.

 Cost: Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive [19].

 Data source issues: There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem.

6. Conclusion

Data mining is concerned with extracting useful rules or interesting patterns from the bulk amount of data collected through various sources. There are many data mining techniques which can be used to perform the job efficiently. It is to be noted that a single technique cannot be used for all types of data because depending on the type of data, appropriate technique is available for extraction of information. Sometimes hybrid techniques are more useful instead of a single technique. The paper

presented a revision of literature vis-à-vis data mining, a technique used to ascertain hidden and useful patterns from vast amount of datasets. These discovered trends help originations to predict the future behaviour of customers or products. This study gives the idea about various data mining techniques, different methods, different processes and issues related to data mining.

7. References

[1] H. Jiawei and M. Kamber, Data Mining: Concepts and Techniques,Morgan Kaufmann, 2011.

[2] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A.C. Coello, "A survey of multi objective evolutionary algorithms for data mining: part I," IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, pp. 4–19, 2014.

[3] Osmar R. Zaiane, "Principles of Knowledge Discovery inDatabases", *CMPUT690*, University of Alberia.

[4] Alex Berson, Stephen Smith, and Kurt Thearling, "*Building Data Mining Applications for CRM".*

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

[6] Barbara, Daniel; Jajodia, Sushil (Eds.), "*Applications of DataMining in Computer Security*"

[7] Khalid Raza, "Application of Data Mining in Bioinformatics", *"Indian Journal of Computer Science and Engineering",* .Vol 1 No 2, 114-118

[8] Gary M. Weiss, "Data Mining in the Telecommunications Industry", *Fordham University, USA.*

[9] R.K. Mittal, Rajeev Kumar, *"E-CRM In Indian Banks- An Overview",* Delhi, Business Review

[10] S.R. Mittal, "Report of Committee on Internet Banking (2001)", Constituted by Reserve bank of India, Chairman of the Committee

[11] Rajanish Dass, "*Data Mining in Banking and Finance: A Note for Bankers*", Indian Institute of Management Ahmadabad International Journal of Engineering Research & Technology (IJERT)

[12] Deelip B. Desai, Dr. R.V. Kulkarni, "A Review: *Application of Data Mining Tools in CRM for Selected Banks*", International Journal of Computer Science and Information Technology, Vol.4(2) (2013), 199-201

[13] S. L. Nalawade, Dr. R.V. Kulkarni, "*Application of Data Mining in Health Care*", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, *(2013): 6.14 | Impact Factor (2015): 6.391*

[14]. Fayyad, Usama, and Paul Stolorz. "Data mining and KDD: Promise and challenges." Future generation computer systems 13.2 (1997): 99-115.

[15]. Goebel, Michael, and Le Gruenwald. "A survey of data mining and knowledge discovery software tools."ACM SIGKDD explorations newsletter 1.1 (1999): 20-33.

[16]. Rygielski, Chris, Jyun-Cheng Wang, and David C.Yen. "Data mining techniques for customer relationship management." Technology in society 24.4 (2002): 483-502.

[17]. Venkatadri, M., and Lokanatha C. Reddy. "A review on data mining from past to the future." International Journal of Computer Applications 15.7 (2011): 19-22.

[18]. Saurkar, Anand V., et al. "A Review Paper on Various Data Mining Techniques." International Journal of Advanced Research in Computer Science and Software Engineering 4.4 (2014).

[19] Sukhdev Singh Ghuman, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 1401-1406 © *2014, IJCSMC All Rights Reserved 1405*