

Data Mining Techniques & Distinct Applications: A Literature Review

Aakanksha Bhatnagar
Student (M.Tech-CSE)
Gurukul Institute of Engg. &
Tech., Kota(Raj)

Shweta P. Jadye
Lecturer (Deptt. of CSE)
R.N. Modi Engg. College,
Ranpur, Kota(Raj)

Madan Mohan Nagar
Lecturer (Deptt. of CSE)
R.N. Modi Engg. College,
Ranpur, Kota(Raj)

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 centre 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. The list of sources that generate these data is endless. Businesses worldwide generate gigantic sets of data everyday that may include stock, transactions and many more of similar kinds. So there comes the need of powerful and most importantly automatic tools for uncovering valuable slots of organized information from tremendous amount of data. Considering any social networking site or a search engine, they receive millions of queries every day. Firstly, the Database Management Systems evolved to handle the queries of similar types. Then the approach was modified to advanced Database management system, Data Warehousing and Data mining for advance data analysis and web based databases. Data mining has immensely penetrated in each and every field of day to day life.

1. Introduction

A proverb says that, "We are living in the era of information." But the reality is somewhat different, as raw data cannot be used, it needs to be refined and converted to information that is compatible with the raw data. This **Pattern evaluation**: In this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results[1].

process of transformation of data is termed as knowledge discovery from data (KDD). Data Mining is just a step that is being followed in KDD [2]. 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: It is also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the given collection.

Data integration: At this stage, multiple data sources, often heterogeneous, may be combined in a common source. Above defined are the initial steps that need to be performed before the selection process.

Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the integrated data collection.

Data transformation: It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: It is the crucial step in which clever techniques are applied to extract patterns that are potentially useful.

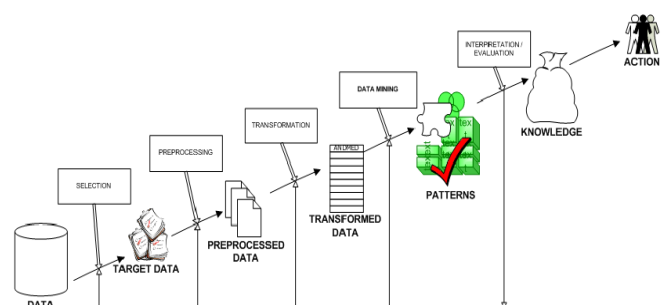


Figure 1. Knowledge Data Mining

2. 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.

2.1. 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.

2.2. 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”.

2.3. 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.

2.4. Prediction

It is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables [3]. 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.

2.5. 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.

2.6. 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 [1].

3. Applications of Data Mining

Data mining is a process that analyzes the large amount of data to find the new and hidden information that improves business efficiency. Various industries have been adopting data mining to their mission-critical business processes to gain competitive advantages. The data mining applications in sale/marketing, finance, health care and insurance, transportation and medicine and many other sectors of day to day life are remarkable. But some other distinct applications of data mining are listed below:

3.1. Applications of Data Mining 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 [5].

3.2. Application of Data Mining 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.

3.3. Data Mining in the 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 [6]. 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.

3.4. Data Mining 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 [8]. 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.

3.5. 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 transactions

and 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 [9, 10]. Data mining can be used in following ways in banking sector

- Detection of fraudulent credit card usage patterns.
- Risk management related to attribution of loans using scorecards.
- Find hidden correlations between different financial indicators.
- Find hidden correlations between different financial indicators.
- Identification of stocks trading rules from historical market data.

4. References

- [1] Osmar R. Zaiane, “Principles of Knowledge Discovery in Databases”, *CMPT690*, University of Alberta.
- [2] J. Han and M. Kamber. Data Mining, “Concepts and Techniques”, *Morgan Kaufmann*, 2000.
- [3] Bharati M. Ramageri, “Data Mining Techniques and Applications”, *Indian Journal of Computer Science and Engineering*, Vol. 1 No. 4 301-305
- [4] Alex Berson, Stephen Smith, and Kurt Thearling, “*Building Data Mining Applications for CRM*”.
- [5] Barbara, Daniel; Jajodia, Sushil (Eds.), “*Applications of Data Mining in Computer Security*”
- [6] Gary M. Weiss, “Data Mining in the Telecommunications Industry”, *Fordham University, USA*.
- [7] Khalid Raza, “Application of Data Mining in Bioinformatics”, “*Indian Journal of Computer Science and Engineering*”, .Vol 1 No 2, 114-118
- [8] R.K. Mittal, Rajeev Kumar, “*E-CRM In Indian Banks- An Overview*”, Delhi, Business Review
- [9] S.R. Mittal, “Report of Committee on Internet Banking (2001)”, Constituted by Reserve bank of India, Chairman of the Committee
- [10] Rajanish Dass, “*Data Mining in Banking and Finance: A Note for Bankers*”, Indian Institute of Management Ahmadabad