Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Applications, Issues and Techniques associated with data mining in various fields: A Survey Paper

Rajpal Dimple

Department of BCA, KLP College Rewari (123401), MD UniversityRohtak, Haryana, India
Email: dimplemalik1234@gmail.com

ABSTRACT

In this modern world we have a huge amount of information and data in various field but the major problem is extracting useful information from this huge amount of data. Data mining is the techniques which is used for selecting useful information from large amount of data. It contains various types of applications in various field but there are various issues associated with data mining in these fields such as privacy issues and security issues. In this paper we discussed various types of applications and issues associated with data mining.

KEYWORDS:KDD, SVM, KNN.

***Corresponding Author**

Rajpal Dimple

Department of BCA,

KLP College Rewari (123401),

MD UniversityRohtak, Haryana, India

Email: dimplemalik1234@gmail.com

1. INTRODUCTION

Data mining is defined as the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of statistics, machine learning and data base management systems. Apart from collecting and managing data, data mining uses tools like association, clustering, segmentation, and classification to provide objective analysis and prediction. Compared to the "traditional methods of data collection that involve manual work and interpretation", data mining utilizes state-of-the-art technology to gather and analyze data from various sources. Traditional method is "slow, expensive and highly subjective". Data mining is fast, economical, and highly objective.

Data mining refers extracting knowledge and mining from large amount of data. Sometimes data mining treated as knowledge discovery in database (KDD). KDD is an iterative process, consist a following step shown in Figure1

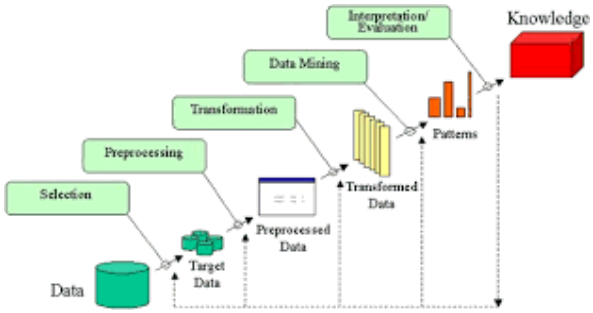

Figure1: Knowledge Data Mining

- Selection: select data from various resources where operation to be performed.
- Preprocessing: also known as data cleaning in which remove the unwanted data.
- Transformation: transform /consolidate into a new format for processing.
- Data mining: identify the desire result.
- Interpretation / evaluation: interpret the result/query to give meaningful report/information

2. APPLICATIONS OF DATA MINING

Data mining has various applicationin different fields and these are described as below:

2.1 FINANCIAL DATA ANALYSIS

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

2.2 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. Here is the list of examples of data mining in the retail industry −Design and Construction of data warehouses based on the benefits of data mining.

- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

2.3 DATA MINING IN TELECOMMUNICATION INDUSTRY

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. 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. Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

2.4 DATA MINING IN BIOLOGICAL DATA ANALYSIS

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis-

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

2.5 DATA MINING IN INTRUSION DETECTION

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

2.6 DATA MINING IN EDUCATION SECTOR

We are applying data mining in education sector then new emerging field called "Education Data Mining". Using these terms enhances the performance of student, drop out student, student behavior, which subject selected in the course. Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Use student's data to analyze their learning behavior to predict the results

2.7DATA MINING IN BANKING AND FINANCE:

Data mining has been used extensively in the banking and financial markets [11]. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction.

2.8 DATA MINING IN MARKET BASKET ANALYSIS:

These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. 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.

2.9 DATA MINING IN EARTHQUAKE PREDICTION:

Predict 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

2.10 DATA MINING IN BIOINFORMATICS:

Bioinformatics generated a large amount of biological data. 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.

2.11 DATA MINING IN AGRICULTURE:

Data mining than 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).

2.10 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.

3. DATA MINING TECHNIQUES

Data mining means collecting relevant information from unstructured data. So, it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. A descriptive model presents, in concise form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable [7]. The goal of predictive and descriptive model can be achieved using a variety of data mining techniques as shown in figure 2.

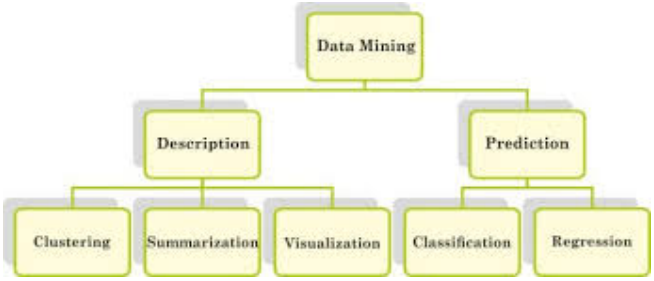

Figure2: Data Mining Model

3.1 CLASSIFICATION:

Classification based on categorical (i.e. discrete, unordered). This technique based on the supervised learning (i.e. desired output for a given input is known). It can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network and classification rule (IF- Then). for example, we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques, we can easily identify the performance of the student

3.2 REGRESSION:

Regression is used to map a data item to a real valued prediction variable [8]. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history.

3.3 TIME SERIES ANALYSIS:

Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events [9]. For example, stock market.

3.4 PREDICTION:

It is one of a data mining technique that discover the relationship between independent variables and the relationship between dependent and independent variables [4]. Prediction model based on continuous or ordered value.

3.5CLUSTERING:

Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning (i.e. desired output for a given input is not known). For example, image processing, pattern recognition, city planning.

3.6SUMMARIZATION:

Summarization is abstraction of data. It is set of relevant tasks and gives an overview of data. 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 [6].

3.7SEQUENCE DISCOVERY:

Uncovers relationships among data [8]. It is set of objects each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence. III. Data Mining Application Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical and so on.

4 ISSUES ASSOCIATED WITH DATA MINING

4.1 NOISY AND INCOMPLETE DATA

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors. Suppose a retail chain collects the email id of customers who spend more than $200 and the billing staff enters the details into their system. The person might make spelling mistakes while entering the email id which results in incorrect data. Even some customers might not be ready to disclose their email id which results in incomplete data. The data even could get altered due to system or human errors. All these result in noisy and incomplete data which makes the data mining really challenging.

4.2 DISTRIBUTED DATA

Real world data is usually stored on different platforms in distributed computing environments. It could be in databases, individual systems, or even on the Internet. It is practically very difficult to bring all the data to a centralized data repository mainly due to organizational and technical reasons. For example, different regional offices might be having their own servers to store their data whereas it will not be feasible to store all the data (millions of terabytes) from all the offices in a central server. So, data mining demands the development of tools and algorithms that enable mining of distributed data.

4.3 COMPLEX DATA

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

4.4 PERFORMANCE ISSUES

The performance of the data mining system mainly depends on the efficiency of algorithms and techniques used. If the algorithms and techniques designed are not up to the mark, then it will affect the performance of the data mining process adversely.

4.5 INCORPORATION OF BACKGROUND KNOWLEDGE

If background knowledge can be incorporated, more reliable and accurate data mining solutions can be found. Descriptive tasks can come up with more useful findings and predictive tasks can make more accurate predictions. But collecting and incorporating background knowledge is a complex process.

4.6 DATA VISUALIZATION

Data visualization is a very importance process in data mining because it is the main process that displays the output in a presentable manner to the user. The information extracted should convey the exact meaning of what it actually intends to convey. But many times, it is really difficult to represent the information in an accurate and easy-to-understand way to the end user. The input data and output information being really complex, very effective and successful data visualization techniques need to be applied to make it successful.

4.7 DATA PRIVACY AND SECURITY

Data mining normally leads to serious issues in terms of data security, privacy and governance. For example, when a retailer analyzes the purchase details, it reveals information about buying habits and preferences of customers without their permission.

4.8 A SKILLED PERSON FOR DATA MINING

Generally, tools present for data Mining are very powerful. But they require a very skilled specialist person to prepare the data and understand the output. As data Mining brings out the different patterns and relationships whose patterns significance and validity must be made by the user. So, a skilled person is a must.

4.9 PRIVACY ISSUES

As data mining collects information about people that are using some market-based techniques and information technology. And these data mining process involves several numbers of factors. But while involving those factors, this system violates the privacy of its user. That is why it lacks in the matters of safety and security of its users. Eventually, it creates miscommunication between people.

4.10 SECURITY ISSUES

As huge data is being collected in data mining systems, some of this data which is very critical might be hacked by hackers as happened with many big companies like Ford Motors, Sony etc.

4.11 ADDITIONAL IRRELEVANT INFORMATION GATHERED

The main functions of the systems create a relevant space for beneficial information. Although, there is a problem with this information collection that the collection of information process can be little overwhelming for all. Therefore, it is very much essential to maintain a minimum level of limit for all the data mining techniques.

4.12 MISUSE OF INFORMATION

In data mining system, the possibility of safety and security measure are reallyminimal. And that is why some can misuse this information to harm others in their own way. Therefore, this data mining system needs to change its course of working so that it can reduce the ratio of misuse of information through the mining process.

5. CONCLUSION

This paper provides a general idea of data mining, data mining techniques and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. These applications use classification, Prediction, clustering, Association techniques and so on. Hopefully in future work we review various classifications and clustering algorithm and its significance's.

REFERENCES

1. Yongjian Fu "data mining: task, techniques and application"
2. Er. Romy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance: A Case Study" International Journal of Computer Science and Management Research October 2012; 1(3)
3. J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.
4. Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT),November- 2012;1(9).
5. Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, 2011; 2(6).
6. Data mining white paper, www.ikanow.com

7. Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Nov-2013;2.
8. Dr. M.H. Dunham, "Data Mining, Introductory and Advanced Topics", 2002Prentice Hall.
9. Time Series Analysis and Forecasting with Weka,.
10. Umamaheswari. K, S. Niraimathi "A Study on Student Data Analysis Using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, August 2013; 3(8).
11. Industry Application of datmining,http://www.pearsonhighered.com/samplechapter/0130862711.pdf
12. David L Olson, Dursun Delen"Advance data mining techniques" springer 2008
13. G. V. Otari, Dr. R. V. Kulkarni, "A Review of Application of Data Mining in Earthquake Prediction" G. V. Otari et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, 2012,3570-3574
14. DRamesh, B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering, September 2013; 2(9)
15. Ruxandra-Ştefania PETRE, "Data mining in Cloud Computing" Database Systems Journal vol. III, no. 3/2012
16. BhagyashreeAmbulkar and VaishaliBorkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012, Link http://research.ijcaonline.org/ncrtc/number6/mpginmc1047.p
17. Arun K Pujari, Data Mining Techniques, University Press, 2013.
18. Christos N. Moridis and Anastasios A. Econo-mides "Mood Recognition during Online Self-
19. Assessment Tests" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL.2, NO. 1, JANUARY MARCH 2009
20. Eric Hsueh-Chan Lu, Wang-Chien Lee, Member, IEEE, and Vincent S. Tseng Member, IEEE," A Framework for Personal Mobile Commerce Pattern Mining and Prediction", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 24, NO. 5, MAY 2012.
21. H. Kargupta and A. Joshi, "Data Mining to Go: Ubiquitous KDD for Mobile and Distributed Environments", KDD-2001, San Francisco, August 2001.
22. J. Han, V.S. Lakshmanan and R T Ng, "Constraint-based, Multidimensional Data Mining", COMPUTER (Special issue on Data Mining), 1999; 32(8): 45-50

23. Jiawei Han and MichelineKamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2003.
24. KasunWickramaratna, Student Member,IEEE ,MiroslavKubat, Senior Member, IEEE, and Kamal Premaratne, Senior Member, IEEE," Predicting Missing Items in Shopping Carts", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, JULY 2009; 21(7).
25. Luigi Lancieri, Member, IEEE, and Nicolas Durand "Internet User Behavior: Compared
26. Study of the Access Traces and Application tothe Discovery of Communities" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, JANUARY 2006; 36(1).
27. Manuel Fogue, PiedadGarrido, Member, IEEE, Francisco J. Martinez, Member, IEEE, Juan-Carlos Cano, Carlos T. Calafate, and Pietro Manzoni, Member, IEEE," A System for Automatic Notification and Severity Estimation of Automotive Accidents", IEEETRANSACTIONS ON MOBILECOMPUTING, MAY 2014; 13(5).
28. Maya Nayak and JnanaRanjanTripathy: Pattern Classification Using NeuroFuzzyand Support Vector Machine (SVM) – A Comparative Study", International Journal of Advanced Research in Computer and Communication Engineering 13 May 2013; 2(5). N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement", International Journal of Advanced Research in Computer Science and Software Engineering, March 2016; 6(3).
29. PaškoKonjevoda and Nikola Štambuk, "Open-Source Tools for Data Mining in SocialScience," Theoretical and Methodological Approaches to Social Sciences and Knowledge Management, 163-176 .
30. Shangguang Wang, Member, IEEE, ZibinZheng, Member, IEEE, ZhengpingWu,Member, IEEE, Fangchun Yang, Member, IEEE, Michael R. Lyu, Fellow, IE.
31. https://docs.oracle.com/cd/B13789_01/datamine.101/b10698/3predict.htm
32. https://www.data-mine.com/white-papers-articles/data-mining-model-types/
33. https://www.guru99.com/data-mining-tutorial.html