Data Mining System, Functionalities and Applications: A Radical Review

Dr. Poonam Chaudhary
System Programmer,
Kurukshetra University, Kurukshetra

Abstract: Data Mining is the process of locating potentially practical, interesting and previously unknown patterns from a big volume of data. It plays an important role in result orientation. Data mining can be used in each and every aspect of life. The same is similarly significant in other areas including sales/ marketing, revenue services, sports, health care and insurance etc. The said paper implies general idea of data mining system, functionalities and its applications.

Keywords: Applications, Data Mining Architecture, Data mining Challenges and Functionalities.

I. INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown valid patterns and relationships in large data set [1]. Data mining tools predict future trends and behaviors, helps organizations to take proactive knowledge-driven decision [**2].** The questions that were traditionally tedious to settle can be settled by data mining tools.

Data mining is also known as knowledge discovery in Database (KDD) and is the nontrivial extraction of implicit previously unknown and potentially useful information from data in databases. Though, databases (or KDD) are frequently treated data mining and knowledge discovery as synonyms, data mining is actually part of knowledge discovery process [3,4,5] . The following figure (Figure 1) shows the steps of knowledge discovery process in data mining.

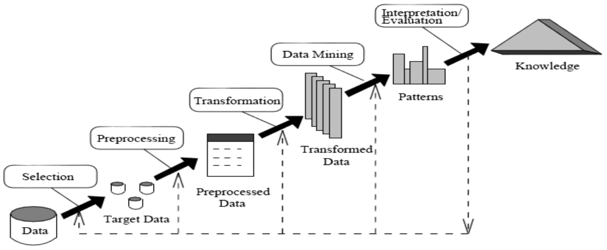


Fig 1:Data Mining is the core of Knowledge Discovery Process[4]

II. DATA MINING ARCHITECTURE

Data Mining's architecture is formed of many elements namely Data Mining Engineer / Pattern evaluation / Data Warehouse server/User Interface and Knowledge Base. The said data mining system of Architecture is presented below in figure (Fig 2)

2.1 *Knowledge Base*:

Centralized storage of Knowledge Base are used to collect the information and to evaluate the pattern.

2.2 *Data Mining Engine*:

An essential element of data mining system and consists of functional elements that perform various tasks namely clustering, classification, prediction, association and correlation analysis, characterization.

2.3 *Pattern Evaluation Module*:

The element performs interesting measures and communicates with the data mining engine module to find out interesting pattern.

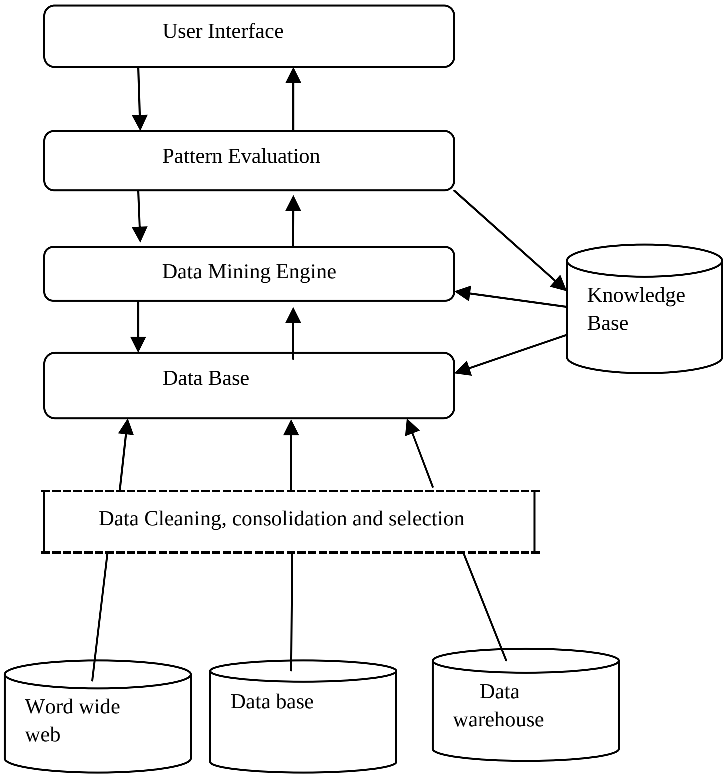


Fig.2 : Data mining system architecture

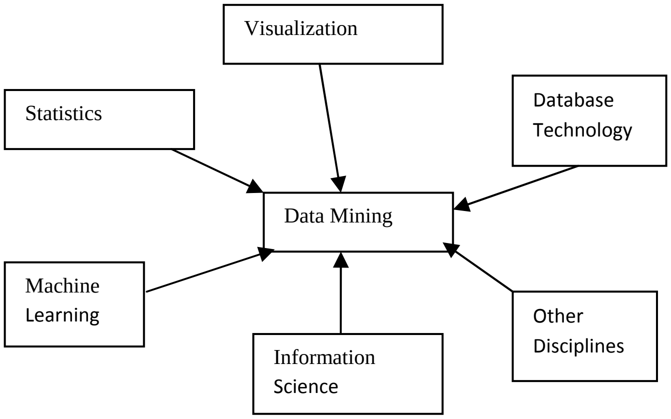
1.4 *User Interface*:

User interface module interacts between user and data exploring system. It allows the subscriber to do interaction with the system by explaining his query and simultaneously by identifying information in order to help in search and to carry out exploratory data mining based on the intermediate data mining results.

III. DATA MINING SYSTEM CLASSIFICATION:

Following criteria is evolved by classifying the said data mining system:-

1. Visualization
2. Data Base Technology
3. Machine Learning
4. Information Science
5. Other Disciplines



3.*1 Some Other Classification Criteria*:

Data Mining System can be divided on the basis of other criteria's that are mentioned below:

3.1.1. *Classification of data mining system according to the type of data sources mined:*

This mode depends upon the type of data used such as text data, multimedia data, World Wide Web, spatial data and time series data etc.

3.1.2. *Classification according to kind of data bases mined:*

This classification is based on the kind of database excavation which is relational database, transactional database, data warehouse, object-oriented database etc.

3.1.3 *Classification according to kind of knowledge mined:*

This division is according to the kind of knowledge discovered in data mining and its functionalities, such as clustering, prediction, Association and correlation analysis, discrimination, outlier analysis, characterization etc.

3.1.4 *Classification bases on the type of Techniques used:*

This categorization is according to the type of techniques utilized such as genetic algorithms, learning of machine, neural networks, oriented database or data ware houses–oriented, Statistics, and visualization etc.

IV. DATA MINING FUNCTIONALITIES:

The said functionalities are measured to perceive the type of patterns to be found in data mining tasks, Data Mining tasks can be categorized in to two categories.

4.*1Descriptive Task:*

These tasks present the general properties of data stored in database. The descriptive tasks are used to find out patterns in data i.e. cluster, correlation, trends and anomalies etc.

4.2 *Predictive Tasks*:

Predictive data mining tasks predict the value of one attribute on the bases of values of other attributes, which is known as target or dependent variable and the attributes used for making the prediction are known as independent variables.

Data mining functionalities are described as follows:-

4.3 *Prediction:*

Predictive model determined the future outcome rather than present behavior. The predictive attribute of a predictive model can be geometric or categorical. It engross the ruling of set of characteristics relevant to the attribute of interest and predicting the value distribution based on the set of data similar to the selected object (S) for example one may predict the kind of disease based on the symptoms of patient.

4.4 *Classification*:

Classification is used to builds models from data with predefined classes as the model is used to classify new instance whose classification is not known. The instances used to create the model are known as training data. A decision tree or set of classification rules is based on such type of mechanism of classification which can be retrieved for identification of future data for example one may classify the employee's potential salary on the bases of salary classification of similar employees in the company.

4.5 *Clustering:*

Clustering is the process of partitioning a set of object or data in a same group called a cluster. These objects are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering is used in many fields, including machine learning, patterns recognition, bioinformatics, image analysis and information retrieval.

4.6 *Mining Frequent patterns, Associations and correlations:*

Frequent patterns can be defined as a pattern (a set of items, subsequence, substructures, etc.) that appears intermittently in data. A intermittent item set is a set of data that occurs frequently together in a transaction data set for example, a set of items, such as table and chair. Subsequence means first of all buying a Computer system, then UPS, and thereafter a printer. This appears frequently in a shopping history data base and is called a frequent sequential pattern. Substructure as particular structural forms such as sub graphs, sub tree. If a substructure appears intermittently, it is named as a frequent structural pattern. Discovering such type of frequent pattern plays an important role in correlation mining association clustering and other data mining tasks.

4.*7Outlier Analysis:*

Outer analysis is an object in database which is significantly different from the existing data. "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"[6]. Deviants, Abnormalities, Discordant and Anomalies are also referred as outliers in data mining and statistics literature. The outlier can be diagnosed with the help of statistical tests that assume probability model for the data.

V. APPLICATIONS OF DATA MINING:

5.1 *Data mining applications in sales/ marketing*:

Data mining is the process of extracting unknown patterns from database which help in planning, organizing, managing and launching new market in a cost effective way. Data mining plays an important role in Market Basket Analysis. It gives information relevant to item sets that are purchased together, their sequence and when they were bought. This information helps business encouragement and to make it most profitable.

5.2 *Data mining applications in banking / finance*:

There are numerous fields in which data mining can be used like in financial and banking sector for credit analysis, fraudulent transactions, customer segmentation and profitability, optimizing stocks portfolios, predicting payment default, ranking investments, marketing, high risk loan applicants, cash management and forecasting operations and most profitable credit card customers and cross selling.

5.3 *Data mining applications in Health Care and Insurance*:

Insurance industry growth is completely depends on the ability of transforming data into information regarding customers, competitors and its market. The insurance industries have implemented the Data Mining successfully and have achieved tremendous competitive advantages. The data mining applications in insurance industry can be used in the form that, data mining is applied in claims analysis such as identifying the medical procedures which are claimed together. Data mining enables to forecasts the potential customers who will buy new schemes. This data mining also proactive insurance companies to detect risky customer's behavior patterns. Data mining also helps in detecting fraudulent behavior.

5.4 *Data Mining for the Retail Industry*:

Retail industry assemble huge amount of data related to sales and customer history of shopping. Retail data mining helps in analyzing client behavior, client patterns of shopping and trends which increases the quality of client service, enhance things consumption ratios, design more effective goods transportations and distribution policies achieve better customer retention and satisfaction and to minimize the cost of business.

5.5 Data mining for the Telecommunications industry:

Telecommunication industries generally generate and store large amount of high quality data, having a very huge customer base, and operate in rapidly changing and highly competitive environment. Telecommunication companies use data mining to enhance their marketing efforts to detect fraud and to betterment of their telecommunication networks.

5.6 *Data Mining Application in Higher Education*:

Data mining can be effectively used to address students and alumni challenges. Data mining facilitate organizations to use their current reporting capabilities to uncover and understand hidden patterns in huge databases. These patterns are then built into data mining models and used to predict individual behavior accurately. As a result of their insight, institutions are able to allocate resources and staff effeciently.This data mining can provide an entity the information necessary to take action before a student drops out, or to efficiently allocate resource with an accurate estimate of how many students will take a particular course.

5.7 Data mining for instruction Detection:

Instructions are the set of actions that threatens the availability and integrity of a network resource. Network instruction detection has been considered to be one of the most promising method for defending complex and dynamic intrusion behaviors.

Intrusion detection techniques using data mining have attracted more and more interests in recent years. Data mining techniques used for intrusion detection are frequent modalities for mining, classification, clustering and mining data streams etc. Fields where data mining technology can be applied for instruction detection are development of data mining algorithms for instruction detection, aggregation to help select and build discriminating attributes, Association and Correlation analysis, Analysis of stream data, Visualization, Distributed data mining and Querying tools.

VI. CHALLENGES IN DATA MINING:

In current situation of affairs data mining research is “too”ad-hoc” and their are so many challenges to unify different data mining tasks. Some of the challenges in the area are as under:

6.1. *Scalability:*

One important challenge is mining data from huge data bases. Computer data network and satellite data can easily be of this scale but to-days technology in data mining are too slow to handle data of this scale. If data mining algorithms are efficient enough to control these massive data sets then they must be scalable. Future data mining should be a continuous, online process instead of one time tiny process. The said scalability also warrants the execution of novel data structure to access individual records in a smooth manner.

6.2 *High Dimensional Data and High Data Streams*:

One challenge is to design classifiers to control ultra high dimensional classification problems for mining vast, enormous and high dimensional data set out-of-memory, parallel and distributed algorithms, algorithm is need to be developed. The traditional data analysis techniques developed for low-dimensional data do not work for dimensional data.

6.3 *Complex and Heterogeneous Data:*

Another challenge erupted in these years is emergence of more data complex. A good system must scale the complexity from users. Previous analysis data mining method deal with the data set consisting attribute of similar type i.e. continuous or categorical due to increasing role of data mining in different areas, a need is arisen to develop techniques which can handle heterogeneous attributes. Such developed techniques for mining such complex objects ought to have taken care the relationships in data, like temporal and spatial auto-correction, graph connectivity and parent-child relationships between the components in semi-structures text and XML documents.

6.4 *Data Ownership, Security and Privacy:*

It is a big challenge to find out data for an analysis at one location or to be owned by one location or to be owned by one entity. An automatic data mining in distributed environment can develop serious issues in terms of data privacy or its security. These issues can be addressed by developing of an efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data and further to measure the impact on the modification of data values on discovered pattern's statistical significance.

6.5 *Data Distribution:*

This challenge in data mining is very important in network problems. This can be addressed by the development of distributed data mining techniques. The key challenges in distributed data mining are:

a) To minimize the amount of communication needed to perform the distributed computation.
b) To consolidate the data mining results obtains from multiple sources in a efficient manner.
c) To tackle data security issues.

VII. CONCLUSION

Since the inception data mining has achieved marvelous success so various problems emerged during the tenure have been solved by data mining techniques. Data mining technology is an application oriented technology and is having extensive applications in various fields. It also evaluates, integrates and reasons to guide the solution of practical problems and find the relationships between events. Furthermore predictions of further activities can be easily made by using the prevailing data. Instead, there is still lack of timely exchange of important topic in a society exclusively. Several efforts have been made to design the generic data mining system but no system found completely generic. However, the decisions at different stages in data mining techniques are influenced by factors like context parameters, aim of data mining, domain and details of data. These specific applications are aimed to extract explicit ideas. The results gathered from the domain specific applications are more accurate and useful.

REFERENCES

[1] Pieter Adriaans and Dolf Zan ting , "Two crows corporation, introduction to data mining and knowledge Discovery", Third Edition (Potomac MD: Two Crows Corporation, 1999); Data Mining (New York : Addison Wesly, 1996)

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", John Wiley & Sons, Inc, 2005, ISBN 0-471-66657-2.

[3] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Dunham, M. H., Sridhar, S., “Data Mining: Introductory and Advanced Topics”, Pearson Education,New Delhi, 1st Edition,2006, ISBN: 81-7758-785-4.

[4] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, American Association for Artificial Intelligence, 1996.

[5] Hawkins, Identification of outlier, Champmaqn and Hall, 1980.

[6] Bhandari, S., Sharma, T., Singh, J., “A Review: Data Mining, its Issues, Functionalities and Applications”,International Journal of Research (IJR) july 2014;1(6):ISSN: 2348-6848.

[7] http://www.wikipedia.org/wiki/data_mining

[8] http://www.google.com

[9] Yang,Q., WU, X. , “ 10 Challenging Problems in Data Mining Research”, International Journal of Information Technology & Decision Making 2006;5(4):597-604.

[10] S.P., Thakar, V.M.,”Data Mining System and Applications: A Review”,International Journal of Distributed and Parallel Systems (IJDPS) Sep,2010;1(1) DOl : 10-5121/ijdps.2010. 1103 32.