A Review : Data Mining

Prof. N. B. Mapari, Raseshwari S. Joshi, Shivani D. Supe
Department of Information Technology, Anuradha Engineering College, Chikhli, Buldana, Maharashtra, India

ABSTRACT

Data mining is a field of intersection of computer science and statistics used to discover patterns in the information bank. The main aim of the data mining process is to extract the useful information from the dossier of data and mold it into an understandable structure for future use. There are different process and techniques used to carry out data mining successfully. In this paper, we provide an overview of common knowledge discovery tasks and approaches to solve these tasks. We propose a feature classification scheme that can be used to study knowledge and data mining software.

Keywords : Data Mining, Knowledge Discovery Database, Data Warehouse, Data Sources, Architecture

I. INTRODUCTION

The enormous usage of computers has provided a huge amount of data for one's disposal. Because of the spiraling amount of data, experts have been facing challenges in extracting useful and meaningful information from it. This has lead to data mining.

Data mining is a non trivial process of extraction of information which is hidden, previously unknown and is potentially useful, from large databases. Data mining can also be explained as finding the correlations in a large relational database based on the different depth of angles we analyze it. It is a powerful tool with high potential that helps the organizations or companies to increase their sales and gain more profit from the information about the dealings of their customers.

Data mining provides us with the useful information that queries and reports are not able to provide us efficiently. The information that is extracted by the data mining etiquette is not explicitly available in the database, whereas database application only projects the information that is available in the info bank with a restricted manipulation capacity. So data mining is best described as knowledge unearthing in databases.

II. DATA WAREHOUSE

With the development of the world there are advances in data confiscating, processing power, data transmit and repository capabilities which are allowing different companies or organizations to centralize their data by integrating various databases which is known as data warehouse. [4]. All the data about customers and potential customers is stored over here. [5].In a data warehouse the data is being stored in specific categories. This maximizes the access of data by the users. This kind of data storage leads to make retrieval, interpretation, sorting tasks more convenient for the user. A data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which allows the strategic use of data.

Data Warehouse is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users for analysis.

Fig 1. Data Warehousing and Data Mining

III. KNOWLEDGE DISCOVERY IN DATABASE

Data mining is composed of seven phases, the first four phases are used for data preprocessing that is data is prepared in a format for further use and the rest three are used to work on the data so formed to retrieve the hidden information.

Data cleaning is use to remove all the noise and other inconsistent data from the input database. Data integration is used to integrate the data as data can be inputted from various sources. Data warehouse is a place where all this cleaned and integrated data is kept. Data selection phase selects the data which is best suited for data mining task. Data transformation transforms the data into a format suitable for data mining. Data mining phase use to employ intelligent methods on the data to generate the knowledge or patterns. These patterns are evaluated in the succeeding phase that is the patterns evaluation phase and in the last phase the knowledge is presented in a user friendly format.

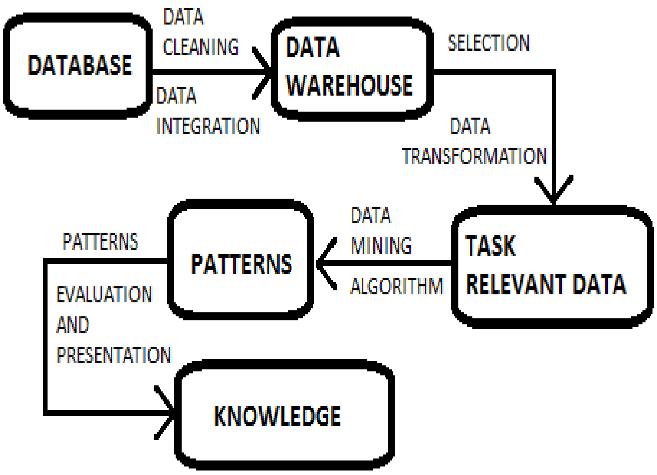

Fig 2. Steps in KDD

IV. ARCHITECTURE OF DATA MINING SYSTEM

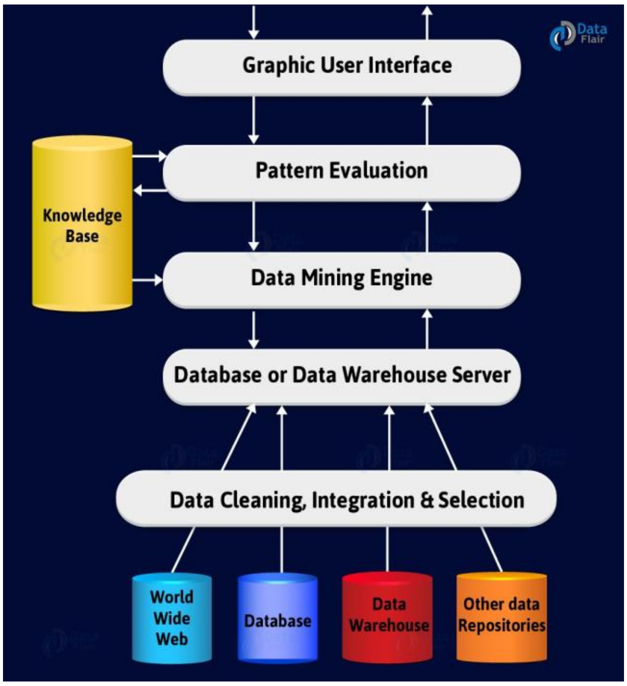

Fig 3. Data Mining system Architecture

Data mining Architecture system contains too many components. That is a data source, data warehouse server, data mining engine, and knowledge base.

- **Data Sources**

There are so many documents present. That is a database, data warehouse, World Wide Web (WWW). That are the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

- **Database or Data Warehouse Server**

The database server contains the actual data that is ready to be processed. Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.

- **Data Mining Engine**

In data mining system data mining engine is the core component. As It consists a number of modules. That we used to perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.

- **Pattern Evaluation Modules**

This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value. Also, it interacts with the data mining engine. That's main focus is to search towards interesting patterns.

- **Graphical User Interface**

We use this interface to communicate between the user and the data mining system. Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process. When the user specifies a query, this module interacts with the data mining system. Thus, displays the result in an easily understandable manner.

- **Knowledge Base**

In whole data mining process, the knowledge base is beneficial. We use it to guiding the search for the result patterns. The knowledge base might even contain user beliefs and data from user experiences. That can be useful in the process of data mining. The data mining engine might get inputs from the knowledge. That is the base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base. That is on a regular basis to get inputs and also to update it.

V. DATA MINING ALGORITHMS

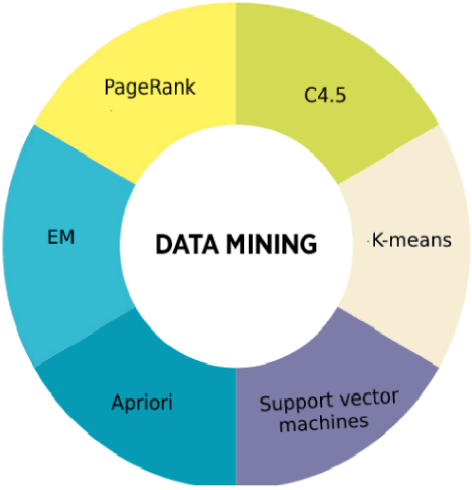

Fig 4. Data Mining Algorithm

Data mining is known as an interdisciplinary subfield of computer science and basically is a computing process of discovering patterns in large data sets. It is considered as an essential process where intelligent methods are applied in order to extract data patterns.

Given below is a list of Top Data Mining Algorithms:

A. *C4.5*

C4.5 is an algorithm that is used to generate a classifier in the form of a decision tree and has been developed by Ross Quinlan. And in order to do the same, C4.5 is given a set of data that represent things that have already been classified.

C4.5 that is often referred to as a statistical classifier is basically an extension of Quinlan's ID3 algorithm. The decision trees that are generated by C4.5 can be further used for classification. The C4.5 algorithm has also been described as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date" by the authors of the Weka machine learning software.

B. *k-means*

k-means clustering that is also known as nearest centroid classifier or The Rocchio algorithm is a method of vector quantization, that is considerably popular for cluster analysis in data mining.

k-means is used to create k groups from a set of objects just so that the members of a group are more similar. It's a well known popular cluster analysis technique used for exploring a dataset.

C. *Support vector machines*

When it comes to machine learning, support vector machines that are also known as support vector networks are basically supervised learning models that come with associated learning algorithms which then analyze data that are used for the analysis of regression and classification.

An SVM model is created that is a representation of the examples as points in space, that are further mapped so that the examples of the separate categories are then divided by a clear gap that is ought to be as wide as possible.

D. *Apriori*

Apriori is an algorithm that is used for frequent itemset mining and association rule learning overall transactional databases. The algorithm is proceeded by the identification of the individual items that are frequent in the database and then extending them to larger itemsets as long as sufficiently those item sets appear often enough in the database. These frequent itemsets that are determined by Apriori can be used for the determination of association rules which then highlight general trends.

E. EM(Expectation-Maximization)

An expectation–maximization (EM) algorithm, when it comes to statistics is an iterative method that is used to find maximum a posteriori(MAP) or maximum likelihood estimates of parameters in statistical models, that basically depends on unobserved latent variables.

F. *PageRank(PR)*

PageRank (PR) that was named after Larry Page who is one of the founders of Google is an algorithm that is used by Google Search to rank the websites in their search engine results. PageRank, that is the first algorithm that was used by the company is not the only algorithm that is being used by Google to order search engine results, but it is the best-known way of measuring the importance of website pages.

VI. TOOLS OF DATA MINING

Data is definitely priceless. But it is not a cake walk to analyze it as greater things come at a greater cost. With the exponential growth in data, there requires a process to extract meaningful information as conclude to useful insights.

Data mining is the process where the discovery of patterns among large sets of data to transform it into effective information is performed. This technique utilizes specific algorithms, statistical analysis, artificial intelligence and database systems to juice out the information from huge datasets and convert them into an understandable form. This paper lists out 5 comprehensive data mining tools widely used in the big data industry.

A. Rapid Miner

Rapid Miner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis. It is one of the apex leading open source system for data mining. The program is written entirely in Java programming language. The program provides an option to try around with a huge number of arbitrarily nestable operators which are detailed in XML files and are made with graphical user interference of rapid miner.

B. Oracle Data Mining

It is a representative of the Oracle's Advanced Analytics Database. Market leading companies use it to maximize the potential of their data to make accurate predictions. The system works with a powerful data algorithm to target best customers. Also, it identifies both anomalies and cross-selling opportunities and enables users to apply a different predictive model based on their need. Further, it customizes customer profiles in the desired way.

C. IBM SPSS Modeler

When it comes to large-scale projects IBM SPSS Modeler turns out to be the best fit. In this modeler, text analytics and its state-of-the-art visual interface prove to be extremely valuable. It helps to generate data mining algorithms with minimal or no programming. It can be widely used in anomaly detection, Bayesian networks, CARMA, Cox regression and basic neural networks that use multilayer perceptron with back-propagation learning.

D. KNIME

Konstanz Information Miner is an open source data analysis platform. In this, you can deploy, scale and familiarize data within less than no time. In the business intelligent world, KNIME is known as the platform that helps to make predictive intelligence accessible to inexperienced users. Moreover, the data-driven innovation system helps uncover data potential. Also, it includes more than thousands of modules and ready-to-use examples and an array of integrated tools and algorithms.

E. Python

Available as a free and open source language, Python is most often compared to R for ease of use. Unlike R, Python's learning curve tends to be so short that it becomes easy to use. Many users find that they can start building datasets and doing extremely complex affinity analysis in minutes. The most common business-use case-data visualizations are straightforward as long as you are comfortable with basic programming concepts like variables, data types, functions, conditionals and loops.

VII. Data Mining Techniques

Data mining is highly effective, so long as it draws upon one or more of these techniques:

A. Tracking patterns

One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

B. Classification

Classification is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

C. Association

Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

D. Outlier detection

In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

E. *Clustering*

Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

F. *Regression*

Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

G. Prediction

Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

VIII. APPLICATIONS OF DATA MINING

Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

Here is the list of other important areas where data mining is widely used:

A. Future Healthcare

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

B. Market Basket Analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

C. Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

D. Manufacturing Engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

E. CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

F. Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

IX. ADVANTAGES OF DATA MINING

A. Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

B. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

C. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of the golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

D. Governments

Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

X. LIMITATIONS OF DATA MINING

A. Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

B. Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much per rsonal and financial information available, the credit card stolen and identity theft become a big problem.

C. Misuse of information/inaccurate information

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. In addition, data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence.

XI. CONCLUSION

This paper gives a general introduction of data mining, the process of discovering interesting knowledge from large amounts of data stored in information repositories. It also discusses background on data mining and methods to integrate uncertainty in data mining such as K-means algorithm. It is also shown that data mining technology can be used in many areas in real life including biomedical and DNA data analysis, financial data analysis, the retail industry and also in the telecommunication industry. One of the biggest challenges for data mining technology is managing the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. Future research will involve the development of new techniques for incorporating uncertainty management in data mining.

XII. REFERENCES

[1]. N.P. Gopalan and B. Sivaselan book on Data Mining techniques and trends published by Asoke K. Ghosh, PHI learning private limited.

[2]. Micheline Comber's book on Data Mining second edition published by Hill publications.

[3]. Data mining article present at wikipedia.com

[4]. Mr. Anderson's article on data mining:what is data mining?

[5]. Mr. Doug Alexander article on data mining.

[6]. Fayyad, Usama, Gregory Piatetsky-Shapiro, Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, 1996.

[7]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, London: Academic Press, 5, 2001.

[8]. Kantardzic, Mehmed, Data Mining: Concepts, Models, Methods, and Algorithms, New York: John Wiley & Sons Inc publishes, 2003.

[9]. Michael Chau, Reynold Cheng, Ben Pao, Uncertain Data Mining: A New Research Direction, Introduction, 2005.