Comparative Analysis of Data Mining Algorithms

Vertika Joshi

Lecturer, Amrapali Institute, Haldwani (Uttarakhand)

Abstract- Data mining also known as knowledge discovery in databases (KDD) was discovered in 1990's by Petr Hajek as an analytical process designed to explore usually large, complex databases in search of consistent patterns between variables to make future predictions in any sector. Present paper will study the algorithms based on data mining concepts and it will analyze the various features and limitations, further this papers will analyze the advantages and disadvantages for the data mining algorithms, and this paper conclude the various aspects of data mining algorithms.

Keywords: Data mining, Association rule learning, Clustering, Classification, Regression Privacy Issues

---♦---

1.1 INTRODUCTION

Since 1960's the database and the information technology has been evolved from the earlier file processing system to powerful database systems. In 1970's the

Database systems have progressed from early hierarchical network to the development of relational database systems. In 1980's the evolutionary step was of data access. In 1990's data warehousing which helps in decision support was the evolutionary step. the evolutionary step decision support was the evolutionary step. Data mining is the evolution of an emerging field with long history. Statistics is the foundation of data mining technology on which the data mining is built. Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning and other areas. The goal of this paper is to get the idea of the concept data mining and comparative analysis of the data mining algorithms.

1.2 BACKGROUND

Generally data mining also referred to as KDD(Knowledge Discovery in Databases)

Is the process of analyzing data from different perspectives and summarizing it into useful information.

Data mining is the search for new valuable and nontrivial information in large volumes of data It is cooperative efforts of humans and computers. The Data ware house is a collection of integrated subject oriented databases designed to support the decision-support factions (DSF).Typically data warehouses are huge, storing billions of records. In many instances, an organizations may have

several local or Although, the existence of a data warehouse is not a prerequisite for Data mining in practice, the task of data mining, is especially for some

Figure1.1 Data mining a tool for innovation.

Large companies is made a lot easier by having access to a data warehouse.

We are data rich, but information poor.

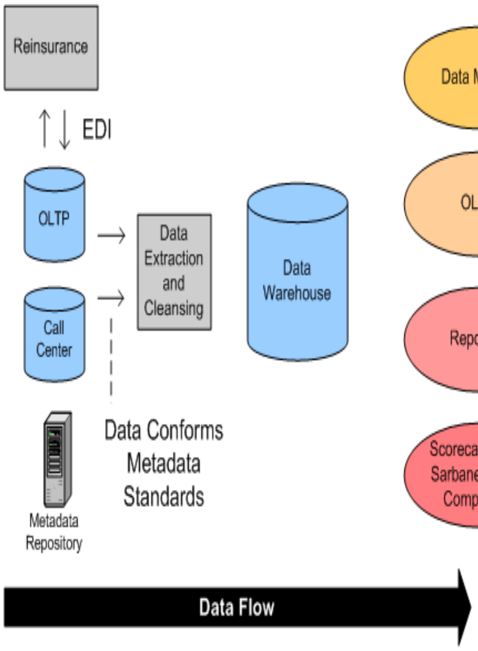
Reinsurance
EDI
OLTP
Call Center
Data Extraction and Cleansing
Data Warehouse
Data Conforms Metadata Standards
Metadata Repository
Data Mining
OLAP
Reporting
Scorecarding for Sarbanes-Oxley Compliance
Data Flow

Figure 1.2 Data ware house keeping a record of historical data

1.3 DATA MINING ALGORITHMS

Before data mining is set the target data set must be assembled .A common source for data is a data mart or data ware house. The Data mining algorithms is the mechanism that creates a data mining model. **Association Rule Learning:** It searches for relationship between the variables. It focuses on data bases of transactions. The association rules are useful for discovering regularities between products in large scale transaction data recorded by point -of-sale (POS) systems in supermarkets**.** **Market basket analysis** has also been used to identify the purchase patterns of the alpha consumers. **Market Basket Analysis can be considered the best as it identifies customers purchasing habits.**

Market basket analysis might tell a retailer that customers often purchase shampoo and conditioner together, so putting both items on promotion at the same time would not create a significant increase in profit, while a promotion involving just one of the items would likely drive sales of the other.

About their

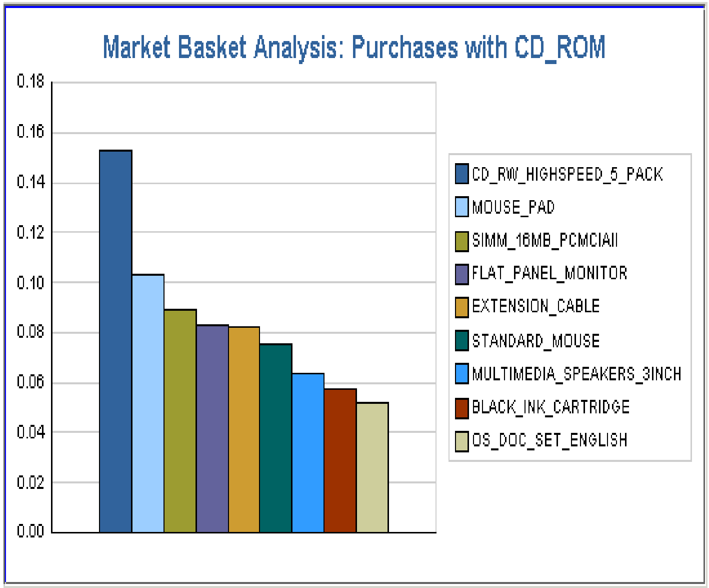

Figure 1.3 Market basket analyses of computer equipments

Market basket analysis can be used to divide customers into groups. A company could look at what other items people purchase along with eggs, and classify them as baking a cake (if they're buying eggs along with flour and sugar) or making omelets (if they're buying eggs along with bacon and cheese). This identification could then be used to drive other programs.

IMPACT OF ASSOCIATION RULE LEARNING

These can direct marketers by providing them with useful and accurate trends out customer's purchasing behavior Based on these trends; marketers can direct their marketing attentions to their customers with more precision.

Retail stores can also benefit from data mining in similar ways.

For Example the store managers can arrange shelves, stock certain items, or provide a certain discount that will attract their customers.

2.) Clustering: Clustering analysis finds clusters of objects that are similar in some sense to one another like members of other clusters. **For Example,** a group of diners sharing the same table in a restaurant may

be regarded as a cluster of people.

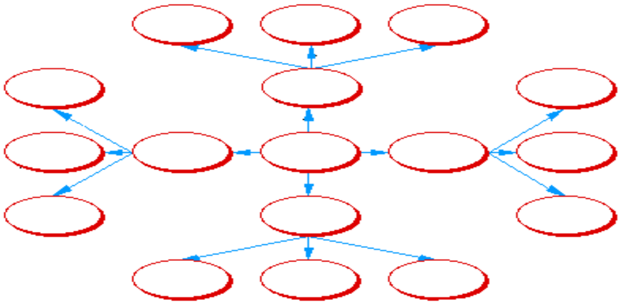

Figure 1.4 Association of similar objects

IMPACT OF CLUSTERING

Clustering can be used for anomaly detection.

For Example: an unwanted user types the username and password several types the maximum number of time the user can make an attempt is five. Exceeding that attempt will display a message

3.) Classification: Working with categorical data or a mixture of continuous numeric and categorical data? This technique is capable of processing a wider variety of data than regression and is growing in popularity. Popular classification techniques include

a.) Decision Trees b.) Neural Networks

Decision trees are basically used to illustrate the logic of a policy Decision trees are white boxes means they generate impel understandable rules. It is one of the best **independent variable** selection algorithms, Decision trees clearly lay out the situations so that each and every 'node' can be scrutinized. Many possible consequences can be worked through. Information available can be used to give clearer and more accurate assessments. The information put into the tree will determine the result.

IMPACT OF DECISION TREES

The following is an example of objects that describe the weather at a given time. The objects contain information on the outlook, humidity etc..**Some objects are positive examples denoted by P and others are negative I. e. N .**Classification in this case the construction of a tree structure, illustrated in the following diagram, which can be used to classify all the objects correctly.

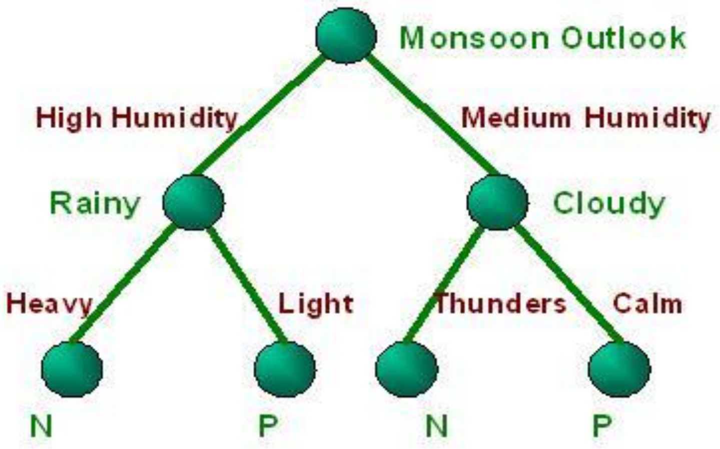

Figure 1.5 Decision tree

NEURAL NETWORKS:

Neural networks are an approach to computing that involves developing mathematical structures with the ability to learn. A neural network is a collection of interconnected

Simple processing elements or neurons. Neural networks are potentially useful for

Studying the complex relationships between inputs and out puts of a system (White, 1990). The methods are the result of academic investigations to model nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze.

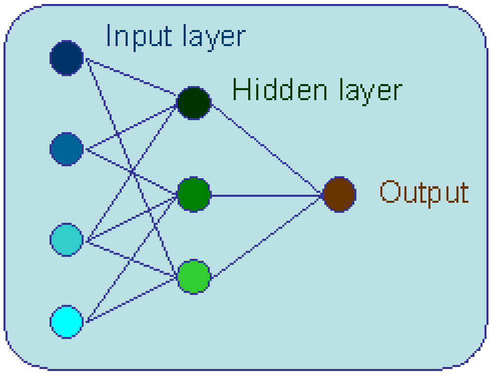

Figure 1.6 Neural Networks

IMPACT OF NEURAL NETWORKS

Neural networks have been applied to solve hard real world problem such as financial forecasting.

Many of the patterns finding algorithms such as decision tree, classification rules and clustering techniques that are frequently used in data mining have been developed in machine learning research community. Frequent pattern and association rule mining is one of the few exceptions to this tradition

4.) Regression Analysis: Regression is basically a data mining application that predicts a number, regression analysis is widely used for prediction and forecasting, a regression always begins with a data set in which the target values are known.

IMPACT OF REGRESSION ANALYSIS

Regression analysis is an opportunity to specify hypothesis concerning the nature of effects. **For example**, a regression model that predicts a used car values could be developed based on observed data for many used cars over a period of time. In addition to the value, the data might track the age of the cars, regression algorithm estimates the value of target as a function of predictors for each case in the build data.

1.4 APPLICATION AREAS OF THE DATA MINING ALGORITHMS

Association rules are applied today in many areas **the market basket analysis** can be used as the basis for decision making activities such as promotional pricing. In many areas from market basket analysis association rules are employed today in many application areas including **Web usage mining, intrusion detection and bioinformatics**.

Cluster analysis is widely used in market research when working with multivariate data from surveys market researchers used clusters to generally partition the population of consumers and to better understand the relationship between different groups of consumers. It helps in **product positioning, new product development**. In social network analysis clustering may be used to recognize communities within large groups of people. In the process of intelligent grouping of the files and websites clustering may be used to create a more relevant set of search results compared to normal search engines like **Google.**

Classification application can be used to identify loan applicants as low, medium or high credit risks**. For example,** by examining previous customers with similar attributes, a bank can estimated the level of risk associated with each given loan. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. Although the data mining technique is not a 100% accurate in its prediction about fraudulent charges, it does help the credit card issuers reduce their losses. The regression models can be used to predict the value of a car, based on how much kilometers the car has run, music system, time of purchase etc.

Data mining can assist researchers by speeding up their data analyzing process; thus allowing them more time to work on other projects.

Companies are using data mining tools and techniques to take the advantage of historical data, **specific uses of data mining include:**

- **Market segmentation -** Identify the common characteristics of customers who buy the same products from your company.
- **Customer churn -** Predict which customers are likely to leave your company and go to a competitor.

- **Fraud detection -** Identify which transactions are most likely to be fraudulent.
- **Direct marketing -** Identify which prospects should be included in a mailing list to obtain the highest response rate.
- **Interactive marketing -** Predict what each individual accessing a Web site is most likely interested in seeing.
- **Market basket analysis -** Understand what products or services are commonly purchased together..
- **Trend analysis -** Reveal the difference between typical customers this month and last.

1.5 DISADVANTAGES

Like many technologies there are that are negative things that are caused by data mining and its associated algorithm such as:

1.) Privacy issues: First of all, we need to fight back against those who steal our content for their own evil purposes without our permission. Because of the privacy issues some people do not shop on internet

2.) Security issues: although companies have a lot of personal information about us available online, they do not have sufficient systems in place to protect that information.

3.) Misuse of information/inaccurate information: Trends obtained through data mining intended to be used for marketing purpose or for some other ethical purposes may be misused.

1.6 FUTURE

The short-term, the results of data mining will be in profitable, if mundane, business related areas. Micro-marketing campaigns will explore new niches. Advertising will target potential customers with new precision.

In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to New York, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers.

The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe. There are potential dangers, though, as discussed below.

1.7 CONCLUSION

Data mining can be beneficial for businesses, governments, society as well as the individual person. However, the major flaw with data mining is that it increases the risk of privacy invasion. Currently, business organizations do not have sufficient security systems to protect the information that they obtained through data mining from unauthorized access, though the use of data mining should be restricted. In the future, when companies are willing to spend money to develop sufficient security system to protect consumer data, then the use of data mining may be supported, but still there are some questions which need to be answered.

REFERENCES:

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499

2. Ahmed S, Coenen F, Lang PH (2006) Tree-based partitioning of date for association rule mining. Know INF Cyst 10(3):315–331

3. Banerjee A, Morgue S, Dillon I, Gosh J (2005) Clustering with Bergman divergences. J Mach Learn Res 6:1705–1749

4. Bedeck JC, Chua SK, Leap D (1986) Generalized k-nearest neighbor rules. Fuzzy Sets Cyst 18(3):237– 256. http://dx.doi.org/10.1016/0165-0114 (86)90004-7

5. Bloch DA, Olsen RA, Walker MG (2002) Risk estimation for classification trees. J Compute Graph Stat 11:263–288

6. Bronchi F, Lunches C (2006) on condensed representations of constrained frequent patterns. Know INF Cyst 9(2):180–201

7. Bremen L (1968) Probability theory. Addison-Wesley, Reading. Republished (1991) in Classics of mathematics. SIAM, Philadelphia

8. Bremen L (1999) Prediction games and arcing classifiers. Neural Compute 11(7):1493–1517

9. Bremen L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

10. Bring S, Page L (1998) the anatomy of a large-scale hyper textual Web Search Engines. Computer Networks

11. Data mining by Hang Chary. (PDF)

12. Evaluating socio economic development source book2

13. Neural Networks for the Analysis and Forecasting Of Advertising and Promotion Impact Hein-Lee Pooh, Jingtao Yao* and Toe Jas˘ic

14. National University of Singapore, Singapore (PDF) SURVEY PAPER

15. Top 10 algorithms in data mining XindongWu · Vipin Kumar · J. Ross Quinlan · Joy deep Gosh · Qing Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng· Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg

16. Data Mining by Doug Alexander