RESEARCH PAPER | Computer Science | Volume : 3 | Issue : 5 | May 2013 | ISSN - 2249-555X

Introduction of Data mining and an Analysis of Data mining Techniques

KEYWORDS: Data mining, Decision Tree, Neural Networks, Rule Induction etc.

Shah Neha K

Research Scholar, Singhania University, Pacheri Bari, Jhunjhunu

ABSTRACT *Today [1] so many data are collected by different areas like business and management, government administration, scientific and engineering, and also from environmental control via Internet for their own operations. So there is an increasing demand for efficient, effective and valuable data analysis tools. Data warehouse system provides some data analysis techniques. But it is not good enough so some additional tools required for depth analysis. I discuss another tool named 'Data Mining' and analysis of some data mining techniques that provides useful information from database.*

I. Introduction

Data mining [2] is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is also known as Knowledge Discovery in Data (KDD). Data mining uses mathematical algorithms to part the data and evaluate the probability of future events. It automatically searches large volume of data to discover pattern and trend. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining [3] is a powerful tool that can help to find patterns and relationships within our data. Data mining discovers hidden information from large databases. To ensure meaningful data mining results, we must understand our data. Figure 1 shows the Process of Data Mining.

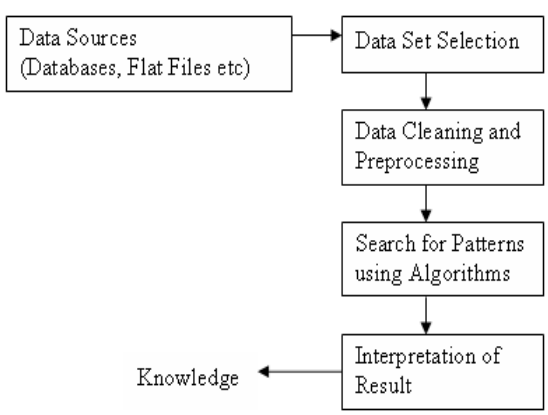

Figure 1: Data Mining Process.

Following can be done using Data mining process.

- It take out, change, and load data onto the data warehouse system.
- It store and manage the data in a database system.
- It provides data access to business analysts, information technology professionals and other persons.
- Analysis of the data can be done by application software.
- It represents the data in an understandable format, such as a graph or table.

II. Techniques used in different level of analysis

Some of the technologies that have been developed and can be used in the Data mining process are:

- Artificial neural networks
- Decision trees
- Genetic algorithms
- Nearest neighbor method
- Rule induction
- Link Analysis

Artificial neural networks

A neural network is a parallel system, capable of resolving paradigms that linear computing cannot. That means it is Non-linear predictive models that learn through training and resemble biological neural networks in structure [2].

The basic unit of an artificial neural network is node, modeled after the neurons in the brain. The other structure is the link that corresponds to the connection between neurons in the brain.

Neural networks copy the human brain. It can done this by using learning from a training of dataset and then apply the learning to generalize patterns for classification and prediction. Neural networks have high tolerance to noisy data and have the ability to classify patterns on which they have not been trained.

Advantages:

- Tasks that can not be performed by linear program are performed using neural network
- There is no matter when an element of the neural network fails, it can continue without any problem by their parallel nature.
- A neural network learns and does not need to be reprogrammed.
- It can be implemented in any application without any problem.

Disadvantages:

- The neural network needs training to operate.
- The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated.
- Large neural networks require high processing time.

Decision Tree

A decision tree is a simple inductive learning structure. Decision tree learning is a common method used in Data Mining. It is Tree-shaped structures that represent sets of decisions. The tree returns a "yes" or "no" decision for the given instance of situation.

Volume : 3 | Issue : 5 | May 2013 | ISSN - 2249-555X

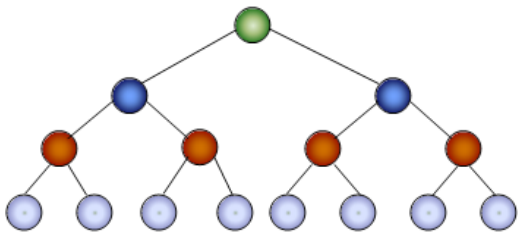

Figure 2: Decision Tree

In Decision tree each internal node denotes a test on an attribute. Each branch represents result of test and leaf node represents possible value of test.

The top most node of tree is root node. The tree can be learned by dividing the main set into subset based on attribute test. The process is repeated on each subset in a recursive manner.

Genetic Algorithms

Genetic algorithm has been widely used in data mining applications like classification, clustering, feature selection etc. It is search algorithm. It is based on mechanics of natural selection and natural genetics [4]. It applies the "survival of the fittest" principle to Data Mining. It uses an iterative process of selection, cross-over, and mutation operators to evolve successive generations of models. At each iteration step every model competes with other by inheriting behavior from previous ones. Iterations are done until the most predictive model survives. It is differ in following ways from normal search procedure [4]:

- It is not work with parameter themselves but it is work with coding of parameter set.
- It is search from a population of points, not from a single point.
- It doesn't use derivatives or auxiliary knowledge but it is use objective function information.
- Probabilistic transitional rules are used by it instead of deterministic rule.

Nearest Neighbor method

Sometimes it is called k-nearest neighbor technique. This technique classifies each record in a dataset based on a combination of the classes of the record(s) most similar to it in a historical dataset. It means the nearest neighbor algorithm classifies a given instance based on a set of already classified instances (the training set), by calculating the distance to the nearest training case. The new instance is classified in the same class as the closest training case (i.e. the one that has the least differences/the one that is most the same).

The K-Nearest Neighbor algorithm is similar to the Nearest Neighbor algorithm, except that it looks at the closest **K** instances to the unclassified instance. Nearest Neighbor algorithm support clustering and classification matching cases internally to each other. It is used for text classification.

Figure 3: Nearest Neighbor [6]

The new Sample is large star. Here considering k=8 nearest neighbor. As per above diagram 6 of the neighbors are in the diamond class and 2 of the neighbors are in the circle class. So the new class is assigned to the diamond class.

Rule induction

Rule induction is an important technique of machine learning used for Data Mining. It is the technique which expresses the regularities hidden in data in terms of rules. This is the extraction of useful if-then rules from data based on statistical significance.

Usually rules are expressions of the form

if…then. For ex., if the customer is a male then, if he is between 35 and 40 years of ages, and his income is less than Rs. 50,000 and more than Rs. 20,000, he is likely to be driving a car that was bought as new.

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain. Data from which, rules are derived, are usually presented in the form of a table.

Link Analysis

Link Analysis is the data mining technique that addresses relationships and connections. It is based on Graph Theory, which represents relationships between different objects as edges in a graph. It can be used in so many artificial intelligence applications. Link analysis is not a specific modeling technique, so it can be used for both directed and undirected data mining.

Depending upon the types of discovery most common approaches to link analysis are [5]: Association discovery and sequence discovery. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Sequence discovery is very similar, in that a sequence is an association related over time.

Conclusion

At last I conclude that, Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is also known as Knowledge Discovery in Data (KDD).

Data Mining helps organizations and businesses in increasing the profitability, smoothening interaction with customers, detecting fraud, and improving risk management. It is also useful in applications like Banking, Insurance, Medical, E-Commerce, Airline, Telecom, and Retail.

Some of Data Mining techniques are: Artificial neural networks, Decision trees, Genetic algorithms, Nearest neighbor method, Rule induction, Link Analysis etc.

An artificial neural network is non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision tree is Tree-shaped structures that represent sets of decisions. The tree returns a "yes" or "no" decision for the given instance of situation.

Genetic algorithm is search algorithm. It is optimization techniques that use different processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Nearest neighbor method is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.

Rule induction is the extraction of useful if-then rules from data based on statistical significance.

Volume : 3 | Issue : 5 | May 2013 | ISSN - 2249-555X

Link Analysis is the data mining technique that addresses relationships and connections is based on graph theory.

REFERENCE

[1] Namita Gupta (2012) "Text Mining for Information Retrieval" Thesis at "Jaypee Institute of Information Technology, Noida". | [2] (1996) "Technology Note prepared for Management 274A "at Anderson Graduate School of Management at UCLA, Bill Placce, Retrieved from "http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm" | [3] Oracle® Data Mining Concepts | 11g Release 1 (11.1) Retrieved from "http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm" | [4] Dr. Rajani Jain (2012) "Introduction to Data Mining Techniques". | | [5] By Two Crows Corporation (2005) "Introduction to Data Mining and Knowledge Discovery" Third Edition | [6] Mohammed J. Zaki and Limsoon Wong (2003) "DATA MINING TECHNIQUES" from WSPC/Lecture Notes Series: 9in x 6in. |