Study Analysis of data mining Algorithms: Case Study

Deependra Dwivedi

IIMT College of Engineering, Greater Noida

Abstract: In this paper we study the Data analysis has many facets, ranging from statistics to engineering. In this paper we discussed basically data mining algorithms for data analysis. Now we can uses of cluster analysis, precedence analysis, and data mining methods are emphasized. Data mining is the best analysis method for software data.

[Deependra Dwivedi. Study Analysis of data mining Algorithms: Case Study. Researcher. 2012;4(2):16-19]. (ISSN: 1553-9865). http://www.sciencepub.net. 5

Keywords: Cluster analysis, data mining algorithms, precedence. Classification of Algorithm

Introduction:

In simple definition data mining is clustering we analyze a set of data and generate a set of grouping rules which can be used to classify future data. For example, one may classify diseases and provide the symptoms which describe each class or subclass. This has much in common with traditional work in statistics and machine learning. However, there are important new issues which arise because of the sheer size of the data. One of the important problems in data mining is the Classification-rule learning which involves finding rules that partition given data into predefined classes. In the data mining domain where millions of records and a large number of attributes are involved, the execution time of existing algorithms can become prohibitive, particularly in interactive applications. [1]

The data mining algorithm is the mechanism that creates a data mining model. To create a model, an algorithm first analyzes a set of data and looks for specific patterns and trends. The algorithm uses the results of this analysis to define the parameters of the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. [2]

The mining model that an algorithm creates can take various forms, including: [2]

- A set of rules that describe how products are grouped together in a transaction.
- A decision tree that predicts whether a particular customer will buy a product.
- A mathematical model that forecasts sales.
- A set of clusters that describe how the cases in a dataset are related.

Data mining algorithms and knowledge discovery processes are both compute and data intensive, therefore the Grid offers a computing and data management infrastructure for supporting decentralized and parallel data analysis. The opportunity of utilizing gridbased data mining systems, algorithms and applications is interesting to users wanting to analyze data distributed across geographically dispersed heterogeneous hosts. Grid-based data mining would allow corporate companies to distribute compute-intensive data analysis among a large number of remote resources. At the same time, it can lead to new algorithms and techniques that would allow organizations to mine data where it are stored. This is in contrast to the practice of having to select data and transfer it into a centralized site for mining.

Classification of Algorithm:

In Data classification one develops a description or model for each class in a database, based on the features present in a set of class-labeled training data (Figure 1). [1]

- **Statistical Algorithms** Statistical analysis systems such as SAS and SPSS have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models. Such systems have their place and will continue to be used.
- **Neural Networks** Artificial neural networks mimic the pattern-finding capacity of the human brain and hence some researchers have suggested applying Neural Network algorithms to pattern-mapping. Neural networks have been applied successfully in a few applications that involve classification.
- **Genetic algorithms** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Nearest neighbor method** 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. Sometimes called the k-nearest neighbor technique. [4]

- **Rule induction** The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization** The visual interpretation of complex relationships in multidimensional data.

Figure 1. Classification of Algorithm

Applying the Algorithm:

Selecting the best algorithm to use for a specific task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result. For example, you can use the Decision Trees algorithm not only for prediction, but also as a way to reduce the number of columns in a dataset, because the decision tree can identify columns that do not affect the final mining model. You also do not have to use algorithms independently. [3] In a single data mining solution you can use some algorithms to explore data, and then use other algorithms to predict a specific outcome based on that data. You can use multiple algorithms within one solution to perform separate tasks, for example by using a regression tree algorithm to obtain financial forecasting information, and a rule-based algorithm to perform a market basket analysis (Figure 2). [2]

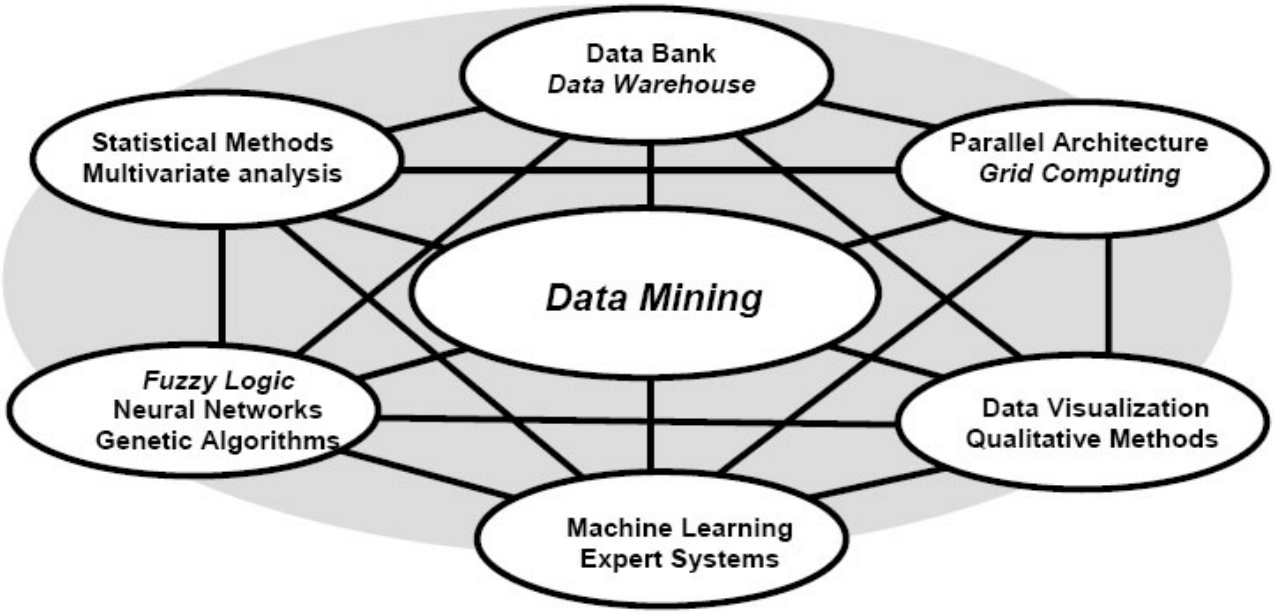

Fig. 2. Data mining process.

Data Abstraction

Many existing algorithms suggest abstracting the test data before classifying it into various classes. There are several alternatives for doing abstraction before classification: A data set can be generalized to either a minimally generalized abstraction level, an intermediate abstraction level, or a rather high abstraction level. Too low an abstraction level may result in scattered classes, bushy classification trees, and difficulty at concise semantic interpretation; whereas too high a level may result in the loss of classification accuracy. The generalization-based multi-level classification process has been implemented in the DB- Miner system.[5]

Supervised Data Mining

Supervised learning is also known as directed learning. The learning process is directed by a previously known dependent attribute or target. Directed data mining attempts to explain the behavior of the target as a function of a set of independent attributes or predictors. Supervised learning generally results in predictive models. This is in contrast to unsupervised learning where the goal is pattern detection. The building of a supervised model involves training, a process whereby the software analyzes many cases where the target value is already known. In the training process, the model "learns" the logic for making the prediction. For example, a model that seeks to identify the customers who are likely to respond to a promotion must be trained by analyzing the characteristics of many customers who are known to have responded or not responded to a promotion in the past. [6]

Conclusion:

In this paper we study the algorithm method for all algorithm uses in different ways. If we can use the algorithms then it is very useful for data allocation in mining method. It is the best method for dynamic memory allocation for business and all things whichever you use it in your future. So we can study the analysis the multiple algorithms in different ways.

References:

1. http://www-users.cs.umn.edu/~desikan/research/dataminingoverview.html#Introduction
2. http://technet.microsoft.com/en-us/library/ms175595.aspx
3. R. Agrawal, R. Srikant, ``Mining Sequential Patterns", Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.
4. R. Srikant, R. Agrawal: ``Mining Sequential Patterns: Generalizations and Performance

Improvements", Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT), Avignon, France, March 1996.
5. J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, K. Koperski, G. Liu, Y. Lu, N. Stefanovic, L. Winstone, B. Xia, O. R. Zaiane, S. Zhang, H. Zhu, `DBMiner: A System for Data Mining in Relational Databases and Data Warehouses", Proc. CASCON'97: Meeting of Minds, Toronto, Canada, November 1997.
6. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/intro_concepts.htm
7. Images taken from googleimages.com.

1/2/2012