Review Paper on Data Mining Techniques

Heena Girdher
Department of Computer Applications
Chandigarh Group of Colleges, Landran
mca.heena@gmail.com

Abstract
The data in the organizations is increasing day by day at tremendous speed. At the current stage, lack of data is no longer a problem; the problem is how to extract meaningful information (rules, regularities, patterns, constraints) from data in large databases. To extract useful information from large amount of data, data mining techniques are needed. Data mining is used as a step in the process of knowledge mining. It is also known as KDD (Knowledge discovery from data), knowledge extraction, information harvesting etc. This paper focuses on the issues of data mining, and data mining techniques such as association rules, classification, and clustering, genetic algorithm.
Keywords: Data mining, KDD, association rules, classification, clustering, genetic algorithm

Introduction

Terabytes of data is generated every day in many organizations from various fields such as www (World Wide Web), businesses, medical, science and engineering, weather data etc. This explosive growth in data makes our time truly the data age. We are drowning in data, but starving for knowledge!

Powerful techniques are needed to extract useful information from the huge amount of data. This necessity led to the development of Data mining. Data mining is used as a step in the process of knowledge mining. There are various data mining techniques that are applied to large datasets depending upon the knowledge to be mined like association rules, classification, clustering, genetic algorithm etc.

Data mining techniques helps to take better decisions in Organizations which affects the performance of the companies.

Fig 1.1 we are drowning in data, but starving for knowledge

A. Data Warehousing and Data mining

Data Warehousing and Data mining both are used as a step in the process of knowledge mining. To extract useful information from data, first data is pre-processed and then entered into data warehouse. Data warehouse contains subject-oriented, integrated, time variant and non-volatile data that is useful for decision making. Data is selected and transformed into forms that are appropriate for mining. Data mining techniques are applied on the transformed data depending upon type of knowledge interest of users. All the patterns by data mining techniques may or may not be interesting to users. The patterns are evaluated based on their interestingness measures. The knowledge mining process is shown in figure 1.2

Data Cleaning: - Data may be incomplete, inconsistent and noisy. In data cleaning step, missing values are handled by applying various techniques of data mining, outliers are removed from data and inconsistencies are corrected.

Data Integration: - Data integration involves data is gathered from multiple resources into data warehouse. Data integration is not an easy process. It involves various issues like redundancy may occur when data is merged from different resources at one place. An attribute may be derived from another attribute may cause redundancy.

Another issue is schema integration and object matching. How can the data analyst be sure that sid in one database and rno in another database refer to rollno of the student.

iii. Data selection: - In data selection, data relevant to the task is selected from datawarehouse. Data may contain relevant and irrelevant attributes. In data selection step, relevant attributes are selected and irrelevant attributes are removed from the database. Suppose a manager wants to sell CD's at a shop and he is having name, age, address, contact number, music_taste attributes of a customer. To analyse the sales of CD's name, address and contact no of a customer doesn't play any role whereas customer age and address may affect the analysis results.

Data transformation: - In data transformation, data is transformed into forms appropriate for mining. Data transformation includes the following:

Smoothing: - Smoothing involves smooth the data to remove outliers. Various techniques are used for smoothing like binning, regression and clustering.

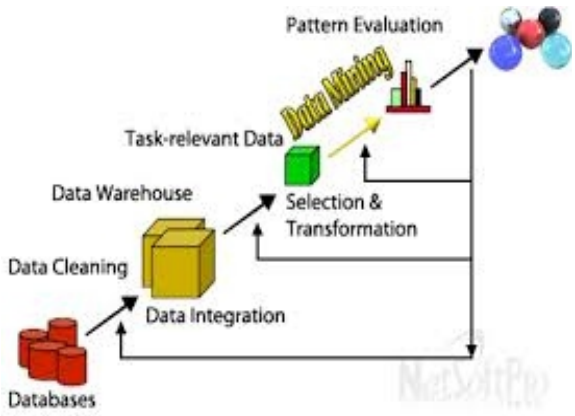
Aggregation: - Data is aggregated into forms appropriate for task. Suppose we want to compute the yearly sales of the product and we have the daily sales of the product. In this case daily sales of the product are aggregated to get the yearly sales.

Attribute construction: - where relevant attributes are selected from the database and irrelevant attributes are removed from the database.

Data mining: - An essential process where various techniques are applied to get the data patterns.

Pattern evaluation: - A data mining system may generate thousand of patterns but all the patterns are not of interest to a user. In Pattern evaluation, interesting patterns are analysed by applying interestingness measures support and confidence.

Knowledge representation: - It is the last phase of the knowledge discovery process where mined knowledge is presented to the users by using visualisation techniques.



Issues in Data mining

The following diagram describes the various issues in data mining.

Ability to handle different kinds of data- The database may contain different kinds of data like web data, multimedia data, hypertext data, spatial data, and temporal data. Data mining techniques should be able to handle all kinds of data. Generally a data mining system is designed to handle only one kind of data.

Presentation of Patterns- Once the patterns are discovered, they must be represented in user understandable form.

Scalability- The speed of data mining algorithms should be predictable and acceptable as the size of database increases.

Mining at different abstraction levels- The knowledge interest of the various users may vary. One kind of knowledge may be interesting for one but not for another. The data mining system should be able to mine knowledge at different abstraction level.

Data Security- Data security has major concern in data mining. As we mine knowledge at different levels, it may breach the security of user's data. For example- By mining knowledge at different abstraction levels, we may be able to find the personality traits of a customer like buying habits, areas of interest etc.

Data mining techniques

Association rules:-Association and correlation analysis is used to find frequent item sets among items in a large dataset. Frequent patterns are those patterns that appear frequently in a dataset. For example- Suppose that the customer who purchase burger at McDonalds' also tend to buy cold drink at the same time is represented by the rule:
burger==> cold drink [support=2%, confidence=60%]
Association rules are used to find customer buying habits that the customer who buys burger will purchase cold drink also. The organization may put these items in a combo to increase their sales. This is why; association rules help us to take better decisions.

Classification

Classification is one of the most widely used data mining technique. Classification involves two steps-learning step and training step. In the learning step, a model or classification rules are created based on the training data by classification algorithm. In training step, the accuracy of the rules is checked based on the test data. If the accuracy of the rules is acceptable, the rules can be applied for the classification of new data sets.
Decision tree induction is one of the approaches that are used for classification. It is a flow chart like structure that represents set of conditions. A tree is built based on the training data and classes of new data are determined from the tree. A typical decision tree is shown in the figure 4.1. It predicts whether a customer will buy the computer or not.

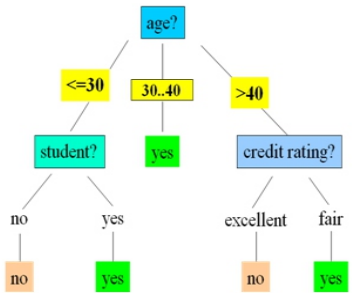


Fig 3.1: Decision tree induction

Classification by back propagation: - Back propagation is a neural network learning algorithm

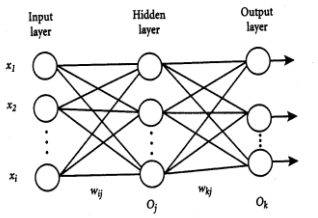


Fig 3.2: neural network

A neural network: A set of connected input/output units where each connection has a **weight** associated with it.

During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples.

Clustering- Clustering divides the large data set into a group of data sets such that data items with in a cluster are similar to each other in some way and dissimilar to data items in another cluster. Clustering has various applications in different areas like detecting outliers, handwritten character recognition, image pattern recognition, Web search etc. Credit card fraud can be detected by determining unusual behaviour of customers like infrequent and expensive transactions.

K-means clustering algorithm: k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.
The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

Step 1: Input the value of k that indicates how many clusters you want to create. And dataset containing n objects.

Step 2: arbitrarily choose k objects as initial cluster centers.

Step 3: Assign each object to the cluster which is close to the mean value of the objects in the cluster.

Step 4: Update the cluster means.

Step 5: Repeat this process until there is no change in cluster.

K-medoids clustering algorithm- k-means algorithm is sensitive to noisy data so we may get inaccurate results if data contains noise. K-medoid algorithm overcomes the disadvantage of k-means algorithm.
Step 1: Input the value of k that indicates how many clusters you want to create. And dataset containing n objects.

Step 2: arbitrarily choose k objects as initial representative objects.

Step 3: Assign each object to the cluster which is close to the representative object.

Step 4: While the cost of the configuration decreases:

1. For each medoid *m*, for each non-medoid data point *o*:

2. Swap *m* and *o*, recompute the cost (sum of distances of points to their medoid)

3. If the total cost of the configuration increased in the previous step, undo the swap

Hierarchical algorithm Hierarchical method group the data objects into a hierarchy of clusters. Hierarchical method can be either agglomerative and divisive.
Agglomerative method works by having each object has its own cluster and merge the clusters into larger cluster until we get all the objects in a single cluster.

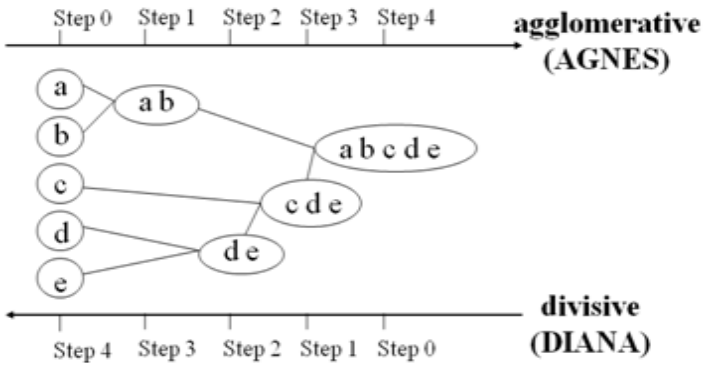


Fig 3.3: Hierarchical clustering

Divisive method works by having all the objects in a single cluster and divide the cluster into small subclusters until we get a cluster containing one object.

Genetic algorithm Genetic algorithm is based on the theory of natural evolution and genetics. This heuristic is used to generate solutions to optimization problems. A population of rules, each representing a solution to a problem, is initially created at random. Then pair of solutions (usually the better solutions are parents) are combined to produce better solutions for next generation by applying crossover and mutation operation.

Mutation process is used to modify the genetic structure of some members of each new generation. The process is repeated until optimal solution is found.
Crossover is a genetic operator that involves producing a child solution from more than one parent solutions.

Conclusion
Data mining is used to discover meaningful information from large amount of data. Organizations have realized the importance of data mining in their decision making process. The customer buying patterns can be found by applying association rules to increase sales. Credit card fraud can be detected by analysing unusual behaviour of the customer like infrequent transactions. Data mining is used in almost every aspect of life. Although progresses are continuously been made in the DM field, still many issues are there which is needed to be resolved and research is going on in this context.

References

[1] Aggarwal, R., Imielinski, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases, Paper presented at the ACM SIGMOD May.

[2] Fayyad, U., Piatetsky-Shapiro, G. And Smyth, P.(1996), "From data mining to knowledge Discovery: an overview", in Fayyad, U., Piatestsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA.

[3] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers

[4] Piatetsky-Shapiro, G. And Frawley, W.J.(1991), Knowledge Discovery in Database, AAAI/MIT Press.

[5] Savasere, A., Omiecinski, E. And Navathe, S.(1995), An Effective Algorithm for Mining Association Rules in Large Databases, paper presented at the 21st International Conference, Very Large Data Bases, September.

[6] Sang Jun Lee, Keng Siau (2001). A review of data mining techniques, Industrial Management & Data Systems.

[7] Technology Forecast: 1997 (1997), Price Warehouse World Technology Center, Menlo Park, CA