Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey of Various Data Mining Techniques

Navjot Kaur
Assistant Professor
Computer Science Engineering
Punjab Technical University
Punjab, India

Gurpreet Kaur
Assistant Professor
Computer Applications
Punjab Technical University
Punjab, India

***Abstract**—Data mining is the process of extracting information from huge sets of data. There is a huge amount of data available in the information industry. This data is of no use until it is converted into useful information. It is necessary to analyse this huge amount of data and extract useful information from it. Extraction of information is not only the process we need to perform. Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Data mining is the process of discovering insightful, Interesting, and novel patterns, as well as descriptive, understandable and predictive models from large scale data. In this paper we overviewed different techniques in data mining. Data mining involves the techniques Association rule, Decision tree, Artificial neural networks, Classification , Genetic algorithms and Clustering.*

***Keywords**— Association rule, Decision tree, Artificial neural networks, Classification , Genetic algorithms and Clustering.*

I. INTRODUCTION

The purpose of data analysis is to discover previously unknown data characteristics, relationships, dependencies, or trends. Such discoveries then become part of the information framework on which decisions are built. A typical data analysis tool relies on end users to define the problem, select the data and initiate the appropriate data analysis to generate the information that helps model and solve problems that the end users uncover. In contrast to this traditional(reactive) BI tools, Data mining is proactive. Instead of having the end user define the problem , select the data and select the tools to analyse the data, data-mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user.

Data mining[1] is the process of analysing data from different perspectives and summarizing it into useful information patterns, associations, or relationships among all this data can provide information.

Data mining software is one of number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles. Categorize it, and summarize the relationships identified. As a advancement of information technology in various fields of human life has increased to large amount of data storage in various ways like records, documents, images, sound recordings, videos, scientific data, and many new data formats.

Data mining refers to the activities that analyse the data, uncover problems or opportunities hidden in the data relationships, form computer models based on their findings, and then use the models to predict business behaviour requiring minimal end user intervention. Therefore, the end user is able to use the system’s findings to gain knowledge that might yield competitive advantages.

Data mining describes a new breed of specialized decision support tools that automate data analysis. Data mining tools initiate analyses to create knowledge. Such knowledge can be used to address any number of business problems. For example, banks and credit companies use knowledge-based analysis to detect fraud, thereby decreasing fraudulent transactions.

II. ARCHITECTURE OF DATA MINING SYSTEM

A. *Databases and data warehouse:*

One or a set of databases, data warehouse and other kinds of information repositories and then data cleaning and data integration techniques are performed on the data.

B. *Database or data warehouse server:*

responsible for fetching the relevant data, based on user’s data mining request. It can be decouples/ loose coupled / tightly coupled with database layer.

C. *Knowledge base:*

The domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.

D. Data mining engine:

it is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis. Query languages (DMQL) based on mining primitives to access the data.

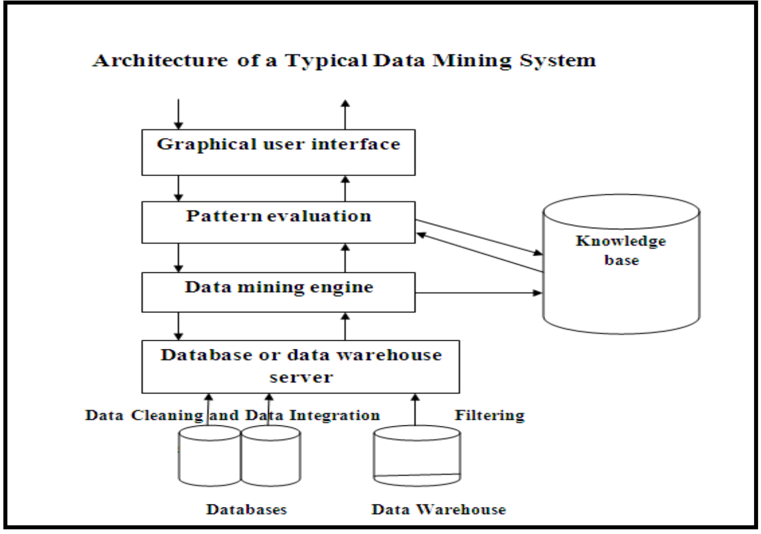
Architecture of a Typical Data Mining System
Graphical user interface
Pattern evaluation
Data mining engine
Database or data warehouse server
Knowledge base
Data Cleaning and Data Integration
Filtering
Databases
Data Warehouse

Fig. 1 Architecture of a typical data mining system

E. Pattern Evaluation

It interacts with the data mining modules so as to focus the search towards the interesting patterns. It may be integrated with mining module.

F. User interface

It communicates between user and data mining system. It allows the user to interact with the system by specifying data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. it also allows the user to browse the database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

III. DATA MINING TECHNIQUES

Data mining is an application dependent issue and different applications may require different mining technique to cope with. In general ,the kinds of knowledge which can be discovered in a database are categorized as follows.

1) Mining Association Rules
2) Classification
3) Decision Trees
4) Artificial Neural Networks
5) Genetic Algorithms
6) Clustering

A. Mining Association Rules

Mining association rules in transactional and relational databases has recently attracted a lot of attention in database communities .An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database.

Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.

An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.[1]

B. Classification

Data mine tools have to infer a model from the database and in the case of supervised learning, this requires the user to define one or more classes. The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes whereas the remaining attributes are called as predicting attributes. A combination of values for the predicted attributes defines a class.

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

1) Classification of credit applicants as low, medium or high risk
2) Classification of mushrooms as edible or poisonous
3) Determination of which home telephone lines are used for internet access.[1]

C. Decision Trees

Decision trees are one of the powerful tool for classification and prediction . The strength of decision trees is due to the fact that, decision trees represent rules. Rules can readily be expressed so that human can understand them or even directly used in database access language like SQL so that records falling into a particular category may be retrieved.

A decision tree is a predictive modeling technique used to classify, cluster and predict tasks.It uses “divide – and - conquer” technique to split the problem search space into subsets.

For example, in marketing one has described the customer segments into marketing professionals, so that the can utilize this knowledge in launching a successful marketing campaign. These domain experts must recognize this discovered knowledge , and for this we need good descriptions. There are a variety of algorithms for building decision trees that share the desirable quality of interpretability.

Decision tree is an approach to learn knowledge on classification. The main requirements to do mining with decision tree are:

1) *Attribute- value description:* Objects must be expressible in terms of a fixed collection of properties.
2) *Predefined Classes***:** The categories to which examples are to be assigned must have been established beforehand.
3) *Discrete Classes:* An object does or does not belong to a particular class and there must be more objects than classes.
4) *Sufficient data:* Hundreds or thousands of training cases.

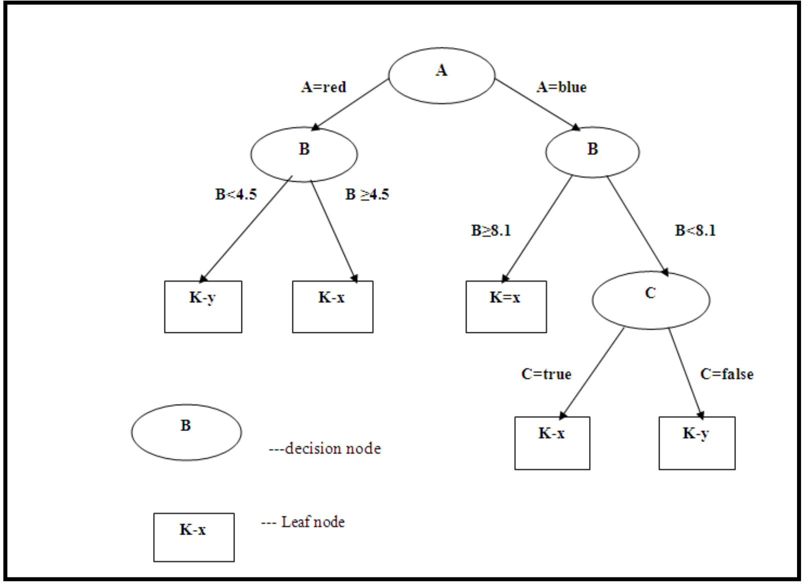

Fig. 2 Decision tree

D. Artificial Neural Network(ANN)

A successful approach to modeling non- linear relationships has been so called neural networks. Neural Networks are analytic technique modeled after the processes of learning in cognitive system and the neurological functions of the brain and capable of predicting new observations from other observations after executing a process of so called learning from existing data.

The first step is to design a specific network architecture .The size and structure of the network needs to match the nature of the investigated phenomenon. Because the latter is obviously not known very well at this early stage, this task is not easy and often involves multiple "trials and errors." The new network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions.

Neural Network techniques can also be used as a component of analyses designed to build explanatory models because *Neural Networks* can help explore data sets in search for relevant variables or groups of variables; the results of such explorations can then facilitate the process of model building. Moreover, now there is neural network software that uses sophisticated algorithms to search for the most relevant input variables, thus potentially contributing directly to the model building process. One of the major advantages of *neural networks* is that, theoretically, they are capable of approximating any continuous function, and thus the researcher does not need to have any hypotheses about the underlying model, or even to some extent, which variables matter. An important disadvantage, however, is that the final solution depends on the initial conditions of the network, and, as stated before, it is virtually impossible to "interpret" the solution in traditional, analytic terms, such as those used to build theories that explain phenomena.

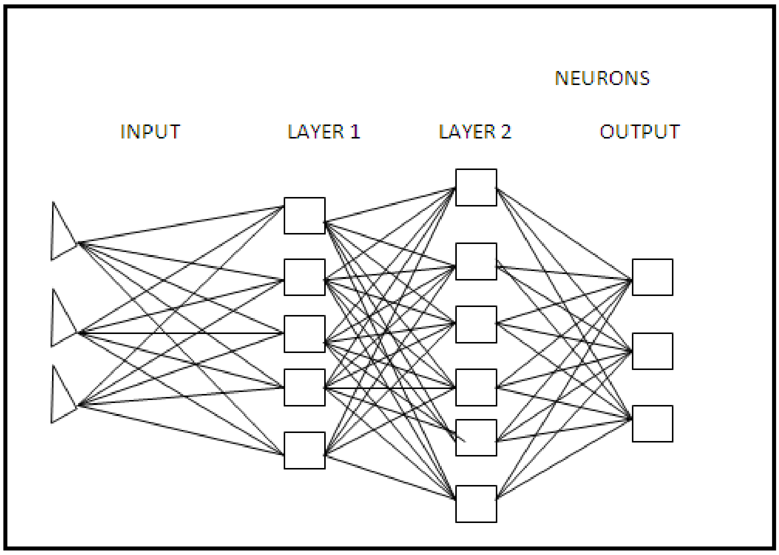

Fig. 2 Structure of ANN

E. Genetic Algorithms

Genetic algorithms are search methods which have been inspired by the process of biological evolution. In data mining , GAs' can be used either to optimize parameters for other kinds of data mining algorithms or to discover knowledge by itself. In Genetic algorithm the goal is to reach accurate and comprehensible knowledge. Hence, the user can understand the system results and combine them with his/her knowledge to make well informed decision. Genetic algorithms encode IF-THEN classification rules similarly to the rules discovered by data mining algorithms.

A Simple Genetic Algorithm

1. Start with a randomly generated population of n chromosomes.
2. Calculate the fitness f(x) of each chromosome x in the population.
3. Repeat the following steps until n offspring have been created.
   a. Randomly select a pair of parent chromosomes from the current population.
   b. Cross the pair at a randomly chosen point to form two offspring.
   c. Randomly mutate the two offspring and add the resulting chromosomes to the population.
   d. Calculate the fitness of the resulting chromosomes.
4. Let the n fittest chromosomes survive to the next generation.
5. Go to step 3(repeat for 50 generations).

F. Clustering

Clustering is similar to classification in grouping but unlike classification, the groups are not predefined. Instead , the grouping is done according to similarities between data. These groups are called clusters. A term similar to clustering is database segmentation. Where like tuples in a database are groups together.This is done to partition the database into componenets that then give the user a more general view of the data.

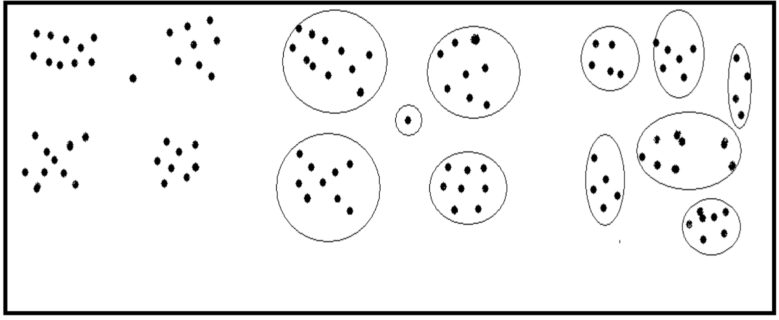

Fig. 3 Different clustering attributes

IV. CONCLUSIONS

Data mining is to discover or extract knowledge or data from large amount of database. In this paper, we briefly reviewed the concept of data mining, various data mining techniques and architecture of data mining . It would be helpful to researchers to focus on the various issues and challenges of data mining. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields.

ACKNOWLEDGMENT

The author would like to thank Amanpreet , Manpreet , Karishma Bhardwaj and the anonymous reviewers for their valuable comments.

REFERENCES

[1] A.V. Saurkar, V. Bhujade, P. Bhagat, A. Khaparde," *A Review Paper on Various Data Mining Techniques*" International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4 , pp. 98-101, April 2014.

[2] T. Hilage,&R. V. Kulkarni**,** "*Review of Literature on Data Mining*" International Journal of Research and Reviews in Applied Sciences vol.10, pp. 107-114, January 2012.

[3] M. Sharma , "*Data Mining: A Literature Survey*" International Journal of Emerging Research in Management &Technology, vol.3, pp. 1-4, February 2014

[4] S.H Liao , P.H Chu, P.Y Hsiao, " *Data Mining Techniques and Applications – A decade review from 2000 to 2011*" ELSEVIER Expert Systems with Applications, vol.39, pp. 11303-11311, 2012.

[5] S. J. Lee, K. Siau, "*A Review of Data Mining Techniques*" Industrial Management and Data Systems, vol.101, pp.41-46, 2001.