ISSN: 2321-7782 (Online)
Volume 2, Issue 11, November 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Different Data Mining Techniques-An Analysis

Sumam Sebastian[1]
M.Tech Computer and Information Science
College of Engineering
Poonjar – India

Jiby J.P[2]
Dept. Of Computer Science and Engineering
College of Engineering
Poonjar – India

Abstract: ***Data mining is a wide, interdisciplinary area of computer science that deals with the extraction of patterns from large data. Data mining is the process of discovering patterns from huge data sets using the methods of artificial intelligence, Intrusion detection and retail industry. The main goal of data mining is to find patterns and to convert it into an understandable suitable structure for further use. Data mining techniques are used to work with large datasets to discover hidden patterns and relationships that are helpful in decision making which are very essential in marketing.***

Keywords: Association Rules, Classification, Decision trees, Data Clustering, Neural Networks

I. INTRODUCTION

Data mining [1] is the process of analyzing data from different perspectives and summarizing it into important information so as to find hidden patterns from a large data set. Data mining can also be defined as the analysis step in the Knowledge Discovery in Databases process or KDD .Knowledge discovery [1] process consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, and pattern evaluation and knowledge presentation.

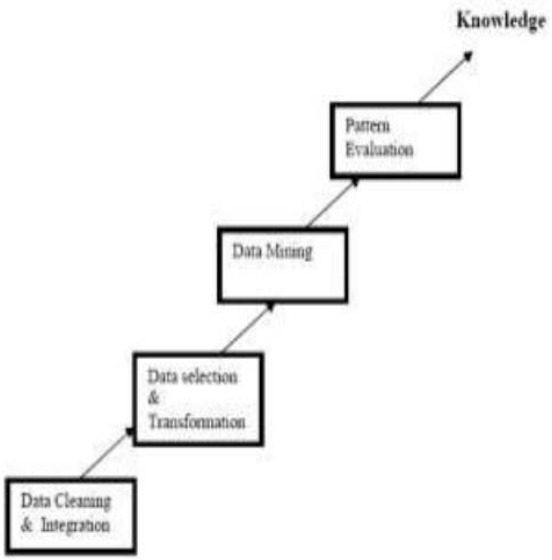

Figure 1. The steps of extracting knowledge from data

Data mining [2] refers to the nontrivial discovering of implicit, previously unknown and potentially useful information from the data in the databases. It uses techniques of machine learning, statistical and visualization to discover and present knowledge in a form which is easily understandable to us. The richness and fast evolution of the data mining discipline comes from its large variety of research areas of interest. Data mining applications uses different kind of parameters to examine the data. The main goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications.

II. DATA MINING TECHNIQUES

Data mining techniques can be basically divided as two categories

1. Classical Techniques
2. Next Generation Techniques

Classical techniques comprise of data mining techniques like Statistics, Neighborhoods and Clustering and Next generation techniques comprise of techniques like Trees, Networks and Rules etc.

Some of the techniques used in Data Mining are:

2.1 Association Rule Mining

Association [2] is one of the best-known data mining techniques. In association rule mining, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association rule mining technique is used in market basket analysis to identify a set of products that customers frequently purchase together. In DM, association rule learning is a conventional and well-researched method for determining interesting relations between attributes in large databases [2]. Association rule Mining is mainly intended to recognize strong rules from databases using different measures of support and confidence.

The preliminaries needed for performing data mining on any data are discussed below.

Let I= {I1, I2, I3….Im} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction $T \subseteq I$. Each transaction is an association with an identifier, called transaction identification (TID). Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where A $\subset I$, $B \subset I$, and $A \cup B = \varnothing$ .Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an item set. An item set that contains k items is a k-item set. The occurrence frequency of an item set is the number of transactions that contain the item set. If the relative support of an item set I satisfies a prescribed minimum support threshold, then I is a frequent item set. The association rule mining can be viewed as a two-step process:

1. Find all frequent item sets: Each of these item sets will occur at least as frequently as a predetermined minimum support count.
2. Generate strong association rules from the frequent item sets: The rules must satisfy minimum support and confidence. These rules are called strong rules.

Apriori Algorithm

Apriori[3]is an algorithm implemented by R. Agarwal and R. Srikant in 1994 for mining item sets for association rules. The algorithm is named so based on the fact that the algorithm uses prior knowledge of frequent item set properties for mining. Apriori [4] is a representative of the candidate generation approach. It generates length (k+1) candidate item sets based on length (k) frequent item sets. The frequency of item sets is defined as the counting occurrence of item in transactions.Apriori algorithm suffers has some drawback in spite of being clear and simple. The main limitation is large wasting of time to hold a vast number of candidate sets with much frequent item sets, low minimum support. Apriori is having very low and inefficiency when memory capacity is limited with large number of transactions [4]. The main drawbacks of the association rule algorithms are the following:

» Obtaining non interesting rules

» Huge number of discovered rules

» Low algorithm performance

Association rule mining finds set of all subsets of attributes that frequently occur in the database records. This algorithm is applicable in business decision making process and in cross validation. Apriori is the most influential ARM algorithm. Efficiency of algorithm can be improved by using hash based technique.

KeyFeatures: Association rule mining uses unsupervised learning method, it helps to identifiE strong rules, and it handles Web usage mining, intrusion detection.

2.2 Clustering

Clustering [5] is the subject of active research in several fields like statistics, pattern recognition, machine learning etc. Data Clustering is unsupervised and statistical data analysis technique. It is used to classify the same data into a homogeneous group. It is used to operate on a large data-set to discover hidden pattern and relationship helps to make decision quickly and efficiently. In a word, Cluster analysis [5] is used to segment a large set of data into subsets called clusters. Each cluster is collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters.The k-means algorithm [6] is very widely used to produce clustering of data, due to its simplicity and speed. K-mean clustering is widely used to minimize squared distance between features values of two points reside in the same cluster. K-means is an old and widely used technique in clustering method. Here, k-means is applied to the processed data to get valuable information. This algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

The pseudo-code of k-means clustering is given below.

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or

- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters

- Each record is assigned to the nearest cluster using a measure of distance.

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

Advantages of K-means clustering

» K-means clustering is simple and flexible.

» K-mean clustering algorithm is easy to understand and to implement.

» Disadvantages of K-means clustering

» In K-means clustering user need to specify the number of cluster as proir [7].

» K-means clustering algorithm's performance depends on an initial centroids that why the algorithm doesn't have guarantee for optimal solution [7].

K means clustering has less computational cost when compared to Fuzzy k means clustering and takes less amount of time for particular change of data in database. This algorithm never waits time for handling outliers.

KeyFeatures:K means clustering uses unsupervised *Feature* Selection as learning method, uses partitioned Clustering Technique and it handles business applications.

2.3 Decision Tree

Decision tree [8] learning uses a decision tree as a predictive model which maps observations about an item to obtain conclusions about the item's target value. It is a predictive modeling approaches used in data mining and machine learning. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. A decision tree method is a flow-chart-like tree structure, where each internal node is represented by rectangles and leaf nodes by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it. A decision tree is constructed from a training set, which consists of data tuples.Each tuple is completely described by a set of attributes and a class label. Attributes can have discrete or continuous values. Decision trees are used to classify the data tuples whose class label is unknown. Based on the tuple's attribute value, the path from root to a leaf can be followed. The class of the leaf is the class predicted by decision tree for that tuple[8].The different commonly used decision tree algorithms are

» ID3 algorithm introduced by J. R. Quinlan is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes.

» C4.5, the most popular algorithm, is a successor of ID3.

» CART algorithm, which was proposed by Breiman, is conceptually is same as that of ID3.

» CHAID uses Chi square contingency test.

Decision Tree Induction

The task of constructing a tree from the training set has been called tree induction.

Algorithm:

1) Create a node N.

2) If all the tuples in the partition are of the same class then return N as a leaf node labeled with that class.

3) If attributes list is empty then return N as a leaf node labeled with the most common class in samples.

4) identify the splitting attribute so that resulting partitions at each branch are as pure as possible.

5) Label node N with splitting criterion which serves as test at that node.

6) If splitting attribute is discrete valued then remove splitting attribute from attribute list.

7) Let Pi be the partitions created based on the i outcomes on splitting criterion.

8) If any Pi is empty then attach a leaf with the majority class in the partition to node N.

9) Else recursively apply the complete process on each partition.

10) Return N.

Advantages of using decision tree in data mining are

» Decision trees implicitly perform variable screening or feature selection.
» Decision trees require relatively little effort from users for data preparation.
» Nonlinear relationships between parameters do not affect tree performance.
» The best feature of using trees for analytics - easy to interpret and explain to executives.

Decision trees have several disadvantages:

» Decision rules yield orthogonal hyper planes in the n-dimensional space, thus each region has the form of a hyper-rectangle. But if in reality many of these decision hyper planes are not perpendicular to the coordinates.
» There is a type of classification problem where the classification criterion has the form: a given class is supported if n out of m conditions are present. Decision trees are not the appropriate tool for modeling this type of problems.

Using Decision tree techniques for data mining provide human readable rules for classification. These algorithms are easy to interpret. They provide better accuracy by making tree construction fast. These are robust algorithms.

Key Features: Decision tree techniques uses supervised learning methods handles classification and estimation tasks, and uses recursive partitioning.

1.4 Artificial Neural Network

An artificial neural network (ANN), often called as a "neural network" (NN), is a computational model based on the biological neural networks, in other words, is a representation and emulation of human neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most of the cases an artificial neural network is an adaptive system that changes its structure based on the information that flows through the network during the learning phase. In practical terms neural networks are non-linear statistical data modeling tools [9]. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions.

Multilayer* Perceptron *(MLP)

MLP algorithm is one of the most widely used and popular neural networks. The network consists of a set of sensory elements that make up the input layer, one or more hidden layers of processing elements, and the output layer of the processing elements.

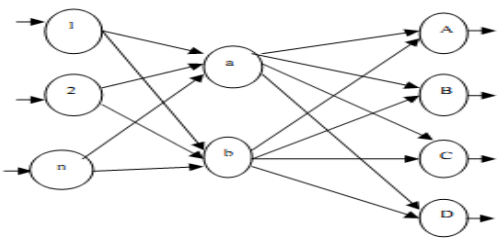

Fig 1.A Feed Forward Neural Network

MLP is especially suitable for approximating a classification function which sets the example determined by the vector attribute values into one or more classes. MLP trained with back propagation algorithm is used for data mining.

There are two phases

1. Learning phase where the network learns by modification of weights.
2. Testing phase where an unknown input is tested for proper learning of neural network.

The Back-propagation algorithm [5] is as follows.

Input

» The samples used for training

» The rate of learning - l

» A feed – forward multi layered fully connected network

Method

Network is initialized with random values of weights and biases.

Repeat till termination conditions are met

{

For each sample S in the set of training samples

{

// Propagate the inputs forward

for each hidden or output layer unit j

{

• Sum of products of weight and output of a particular node with the bias assumed is the input value for the next layer

• For each unit j, compute the output using exponential activation function.

- Back propagate the errors with the following steps
- for each unit j in the output layer
  - » Error j = Out j (1- Outj) (Tru j – Out j) ;
- for each unit j in the hidden layers, from the last to the first hidden layers
- Error j = Out j (1- Outj) Σk Error k w jk;

//Compute the error wrt the next higher layer ,k

- for each weight wt ij in network
- Increment the weight, bias values
- wtij = wtij + Δ wt ij ;
- for each bias θ j in network
- Δ θb j = (lr)Err j ;

o $\theta b_j = \theta_j + \Delta \theta b_j$;

}

}

Advantages of Neural Networks Technique

1. High Accuracy: Neural networks are able to get an approximation of complex non-linear mappings
2. Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing,noisyand incomplete data.
3. Independence from prior assumptions: Neural networks do not make prior assumptions about the distribution of the data, or about the interactions between factors.
4. Ease of maintenance: Neural networks can be updated with fresh data, making them useful for dynamic environments.
5. Neural networks can be implemented in parallel hardware
6. When an element of the neural network fails, it can continue without any problem by their parallel nature. 7. capable of producing an arbitrarily complex relationship between inputs and outputs.
7. able to analyze and organize data using its intrinsic features without any external guidance.
8. Neural Networks of various kinds can be used for clustering and prototype creation.

Disadvantages of Neural Networks Technique

1. Do not work well when there are many hundreds or thousands of input features.
2. Do not yield acceptable performance for complex problems.

Artificial neural network is an alternative to conventional classifier methods [10]. Neural networks are data driven self adaptive method with high accuracy and efficiency. The performance of neural networks can be improved by

» Designing Neural Networks using Genetic Algorithms.

» Neuro-Fuzzy Systems.

Key Features: Artificial neural network is a tool for classification. It uses the learning method of supervised learning, and is suitable for almost all data mining tasks, and it handles pattern recognition.

***2.5* Classification**

Naive Bayes algorithm (NB)[11] is a method based on probability theory. It is a simple method based on classification method in data mining. This method solves problems based on 2 assumptions. First it assumes that with familiar classification prognostic attributes are conditionally independent. Also it supposes that process of prediction will not be affected by any hidden attributes.It is an efficient algorithm for data classification. The Bayesian Classification [11] uses supervised learning method and statistical method for classification process. It allows us to find uncertainty about the model in a sophisticated way by finding out probabilities of the results. . It can be used to solve diagnostic and predictive problems. This Classification method got its name from its author Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification gives practical learning algorithms and prior knowledge and observed data can be combined.Bayesian Classification gives us a useful perspective for understanding and evaluating many learning algorithms. It is robust to noise in input data.

Stages for Building a Bayesian Classifier

» Collect class exemplars

» Estimate the priori probabilities

» Estimate class means

» Form the covariance matrices, and find the inverse and determinant for each

» Form the discriminant function for every class

Advantages of the algorithm

» Naive Bayes classifier will meets target quicker than other models like logistic regression, so needs less training data.

» Can be used semi-supervised learning.

This approach reduces the computation cost and comparable in performance to decision trees. The yields high accuracy and sped when applied to large data bases.

Key Features: Naïve Bayesian method is simple approach with clear semantics for classification. It uses supervised learning.

III. Performance Evaluation of Data Mining Techniques

The table below is performance evaluation table for different data mining techniques like k-means clustering, decision tree technique, association rule, neural network techniques and Naive Bayesian techniques. Table considers different parameters like main feature of the technique, application area of the technique, cost of computation, accuracy, speed, execution time with respect to one processor etc.

Data mining technique	Learning method	Application area	Cost of computation	accuracy	speed	Execution time(s)(respect to one processor)
K means clustering	Unsupervised	Discovery of prediction	Less(compared to fuzzy k means clustering)	medium	Comparatively high	12.9
Decision tree	supervised	Business for future prediction	less	high	High	unavailable
Association Rule	Unsupervised	medical diagnosis	high		high	102.7
Neural Networks Technique	supervised	Stock Market Prediction	less	high	high	unavailable
Naive Bayes algorithm	supervised	solve diagnostic and predictive problems	less	High(in large database)	High(in large database)	25.1

IV. Conclusion

This paper discussed the data mining process and different techniques of data mining. Each of the methods has different criteria in extracting pattern from data set. A performance evaluation is done by considering the different feature of each technique. Artificial neural network is an emerging technique in data mining field and it can act as an alternative to all other

data mining techniques. In most cases neural networks perform as well or better than the traditional statistical techniques to which they are compared. The performance of neural networks can be improved by using Genetic Algorithms, and Neuro-Fuzzy Systems. Artificial neural networks can be applied in areas like education data mining, stock marketing etc.

References

1. Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.
2. Anwar, M. A., and Naseer Ahmed. Knowledge Mining in Supervised and Unsupervised ssessment Data of Students' Performance." 2011 2nd International Conference on Networking and Information Technology IPCSIT vol. Vol. 17. 2011.
3. Baha Sen, Emine Ucar. Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology 1 262 – 267, 2012.
4. R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.
5. Survey of Clustering Data Mining Techniques ; Pavel Berkhin ;Accrue Software, Inc.
6. J. B. MacQueen. Some method for the classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Structures, pages 281–297, 1967.
7. Garbriela derban and Grigoreta sofia moldovan, "A comparison of clustering techniques in aspect mining", Studia University, Vol LI, Number1, 2006, pp 69-78.
8. Performance Prediction of Engineering Students using Decision Trees; R. R. Kabra, R. S. Bichkar; International Journal of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011 .
9. Refaat, M. Data Preparation for Data Mining Using SAS, Elsevier, 2007.
10. NEURAL NETWORKS IN DATA MINING; 1 DR. YASHPAL SINGH, 2ALOK SINGH CHAUHAN; Journal of Theoretical and Applied Information Technology.
11. Predicting Student Performance by Using Data Mining Methods for Classification; Dorina Kabakchieva Sofia University "St. Kl. Ohridski", Sofia 1000.

