International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 03, March 2017
www.ijates.com
ijates
ISSN 2348 - 7550

A STUDY ON DATA MINING TECHNIQUES

V.Arunag.Deepthi

Asst.Prof,Dept of CSEAsst.Prof, Dept of CSE
Sphoorthy EngineeringCollege,Sphoorthy EngineeringCollege,

ABSTRACT

Now a days every organization has its own data, which is bulk in amount. Consider any sector like health, finance, BPO, bank etc.All these sectors have huge data and they require the right data mining techniques which can be appropriate to various sectors. Some of the techniques involved are classification, association, clustering and outliers.

Keywords: sectors, data mining techniques.

I. INTRODUCTION

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

II. METHODOLOGIES

CLASSIFICATION TECHNIQUE:

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called

goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem.[1].

We tested each data set with four different classification tree algorithms:J48, REPTree, RandomTree and Logistical Model Trees. For each algorithm both the test options percentage splitand cross-validation were used. With percentage split, the data set is divided in a training part and a test part[2].

III. ASSOCIATION TECHNIQUE

Association rules mining are one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems. The technique is likely to be very practical in applications which use the similarity in customer buying behavior in order to make peer recommendations. Association Rules will permit you to discover rules of the kind If X then (likely) Y where X and Y can be particular items ,values, words, etc. , or conjunctions of values, items, words, etc. (e. g.,if (Car=BMW and Gender=Male and Age<20) then (Risk=High and Insurance=High)).Data patterns and models can be mined from many different kinds of databases, such as Relational Databases, Data Warehouses, Transactional Databases, and Advanced Database Systems (Object-Oriented, Relational, Spatial and Temporal, Time-Series, Multimedia, Text, Heterogeneous, Legacy, Distributed, and WWW).

Association Rule Mining Algorithms, Apriori Algorithm, FP-Growth Algorithm, Unsupervised Learning, Early Pruning, etc.

- **Apriori Algorithm:** Used to generate all frequent itemset. A Frequent itemset is an itemset whose support is greater than some user-specified minimum support.Apriori algorithm, in spite of being simple and clear, has some limitation. It is costly to handle a huge number of candidate sets.
- **FP-Growth**

Allows frequent itemset discovery without candidate itemsetgeneration. Two step approach:

(i). Step 1

Build a compact data structure called the FP-tree built using 2 passes over the data-set.

(ii). Step 2

Extracts frequent itemsets directly from the FP-tree traversal through FP-Tree.[3].

- **Unsupervised learning** :

It is a type of **machine learning** algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common **unsupervised learning** method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

- **Proposed back-track pruned algorithm**

Back-track pruned C4.5 technique is applied to construct a decision tree, with the available non-categorical attributes C1, C2 up to Cn, the categorical attribute C, and learning dataset T of instances.

Input:

1. The learning dataset 'D', and its set of training observations with respective class values.
2. Attribute list represented as A, which is the set of relevant candidate attributes.
3. Chosen splitting criteria technique.

Output: A decision tree.

IV. CLUSTERING TECHNIQUE

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

In the process of analyzing large sets of data in this step, one of the most used concepts is the aggregation of similar objects within the dataset. This method is an important one and is usually called as **clustering in data mining**.

The following are the clustering methods:

Partitioning clustering: Partitional clustering decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K ($N \geq K$) partitions of the data, with each partition representing a cluster. That is, it classifies the data into K groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. Notice that for fuzzy partitioning, a point can belong to more than one group.

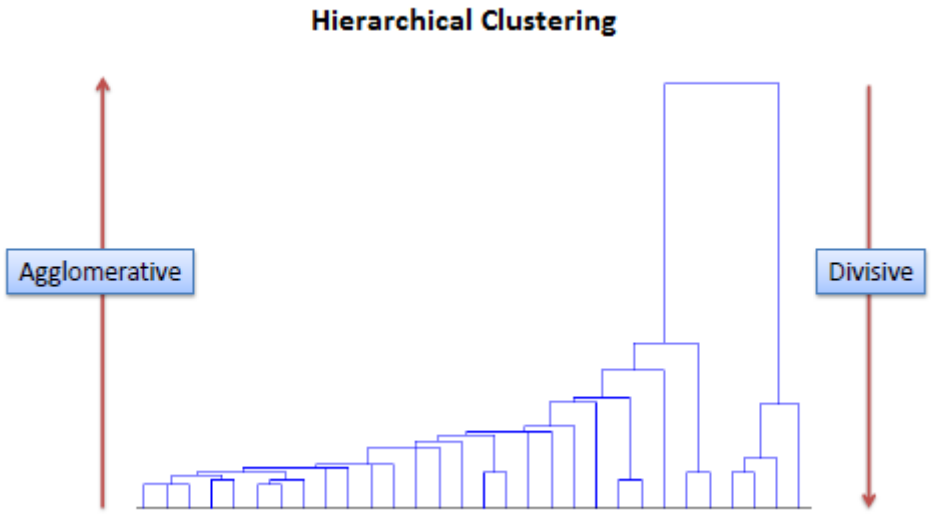
Many partitional clustering algorithms try to minimize an objective function. For example, in K-means and K-medoids the function (also referred to as the distortion function) is

$$\sum i=1K\sum j=1|Ci|Dist(xj,center(i)),\sum i=1K\sum j=1|Ci|Dist(xj,center(i)), \quad (1)$$

where $|C_i|$ is the number of points in cluster i, Dist(x_j, center(i)) is the distance between point x_j and center i. Many distance functions can be used, such as Euclidean distance and L.

- **Hierarchy clustering**:

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, *Divisive* and *Agglomerative*.



Divisive method

In this method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation.

Agglomerative method

In this method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between

each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single

cluster left.

- **Grid based clustering:**

Density-based and/or grid-based approaches are popular for mining clusters in a large multidimensional space wherein clusters are regarded as denser regions than their surroundings. In this chapter, we present some grid-based clustering algorithms.The computational complexity of most clustering algorithms is at least linearly proportional to the size of the data set. The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets.The grid-based clustering approach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points. In general, a typical grid-based clustering algorithm consists of the following five basic steps (Grabusts and Borisov, 2002):

1. Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
2. Calculating the cell density for each cell.
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

- **Model based clustering**:

The traditional clustering methods such as **hierarchical clustering** and **partitioning algorithms** (k-means and others) are heuristic and are not based on formal models.

An alternative is to use **model-based clustering**, in which, the data are considered as coming from a distribution that is mixture of two or more components (i.e. **clusters**) (Chris Fraley and Adrian E. Raftery, 2002 and 2012).

V. OUTLIER TECHNIQUE

Outlier detection is an important branch in data mining, which is the discovery of data that deviate a lot from other data patterns. D.Hawkins [1], gives definition to outlier as: An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. There are many studies have been conducted on outlier detection for large datasets. A lot of work has been done in this area of research which is detecting outliers. The early work is based on statistics ([2], [3]), and assume that a priori knowledge of distribution is known. Most of tests depended on the distribution whether or not the distribution parameters are known. In other related area dealing with detecting outliers is clustering algorithms where outliers are objects not located in clusters of a dataset, and these algorithms generate outliers as by product. Recently, researchers have proposed distance-based, density-based and connectivity-based outlier detection methods .The advantage of these methods is that, they do not have any priori knowledge about the data distribution. Inorder to detect fraud outliers are very useful

- Proximity Methods
- Projection Methods
- Extreme Value Analysis

VI. CONCLUSION

As we have various methods to knowledge or review the item. The procedures of the methods are different from one technique to other technique.Apart from these four techniques we have other techniques also.Among all these, the outlier technique is the better one in order to find the fraud or defaulters.

REFERENCES

1. Data Mining Classification- Fabriciovoznika&Leonardoviana.
2 .Data Mining Algorithms For ClassificationBsc Thesis Artificial Intelligence Author: Patrick OzerRadboud University NijmegenJanuary 2008
3. Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm 1sanjeev Rao, 2priyanka Gupta 1,2dept. Of Cse, Rimt-Maec, MandiGobindgarh, Punjab, India
4. A DATA MINING APPROACH TO PREDICT PROSPECTIVE BUSINESS SECTORS FOR LENDING IN RETAIL BANKING USING DECISION TREE Md. Rafiqul Islam1 and Md. Ahsan Habib2 Department of Information and Communication Technology MawlanaBhashani Science and Technology University, Tangail, Bangladesh.
5. From Data Mining to Knowledge Discovery in Databases, U. Fayyad, G. Piatesky-Shapiro & P. Smyth, AI Magazine, 17(3):37-54, Fall 1996.
6. The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review, 2003, 45, 167-256.
7. Learning with Labeled and Unlabeled Data, M. Seeger, University of Edinburgh (unpublished), 2002.

8 .Person Identification in Webcam Images: An Application of Semi-Supervised Learning, M. Balcan, A. Blum, P. Choi, J. lafferty, B. Pantano, M. Rwebangira, X. Zhu, *Proceedings of the 22nd ICML Workshop on Learning with Partially Classified Training Data*, 2005.

9. Wrapper-based Computation and Evaluation of Sampling Methods for Imbalanced Datasets, N. Chawla, L. Hall, and A. Joshi, in *Proceedings of the 1st International Workshop on Utility-based Data Mining*, 24-33, 2005.