Engineering and Scientific International Journal (ESIJ)
Volume 3, Issue 2, April – June 2016
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Data Mining Clustering Techniques

R.RoopRekha[#1], S.Perumal[*2]
[1]*Research Scholar, Dept of Computer Applications, Vels University, Chennai*
roopselvam@gmail.com
[2]*Head, Dept of Computer Science, Vels University, Chennai*

Abstract— Data mining is a powerful technology to extract information from the large amount of the data. Data mining is considered as one of the important field in knowledge management. Today, Data mining helps different organization focus on the data they collected based on the attitude of their customer's. For the past few years, research in data mining continues in various fields of organization and research such as Statistics, Artificial Intelligence, Pattern Recognition, Machine Learning, Business, Education, Scientific etc. This paper discuss the various concepts of data mining and its techniques.

Keywords— Data Mining; Data Base; Cluster; Prediction.

1. Introduction

1.1 Data Mining

Data mining is a technique of take out or mining facts from numerous amounts of dataset. Data mining is also referred as data or knowledge discovery. It analyze data from different perspectives and summarizes it into useful information the associations or relationships among all these data. Data mining tool is used for analyzing data. Mining allows users to analyze data from different dimensions or angle.It categorize data and summarize the relationships identified. Data collection and storage technology made it possible for organizations to store huge amounts of data at lower cost. Exploit this data to extract useful and actionable information. Data mining is the process of exploring and analyzing large amount of data to discover meaningful patterns and rules. In reality, performing data mining undergoing an entire process is essentially iterative and semi-automated and may require human interference in several key points.The two main reasons to use data mining are as follows
i) Too much data and too little information.
ii) It is essential to extract useful information from the data and to interpret the data.

1.2 Data Mining Techniques

The key techniques of data mining are
- Association
- Classification
- Clustering
- Prediction
- Sequence pattern
- Regression

A. Clustering

Clustering is a technique used in data mining that enables us to discover groups and hence identify interesting distributions and patterns in the underlying data. Clustering partitions a given data set into clusters (groups) such that the data in a cluster are more similar to each other than data in other clusters[1].

Cluster: A cluster is a set of data objects similar to one another and dissimilar to the objects in other groups.

Cluster Analysis: The main aim is to identify clusters of similar objects and to discover interesting patterns and correlations in huge data sets. It groups a set of data objects into clusters.

The similarities are identified between data depends on the features found in the data and groups similar data objects into clusters. Clustering divides a data into groups of similar objects. Clustering is a technique of unsupervised learning. Clustering group's data that share similar patterns. Clustering of data is a method by which large set of data are clustered into groups of small set of similar data.

Cluster Analysis or Clustering involves grouping similar objects in the same group (called a cluster). Each group called cluster are more similar between themselves and dissimilar to objects of other groups (clusters).The clustering technique groups data or divides a large data set into smaller data sets of some similarity. The process of data mining requires various methods such as Image Analysis, Pattern Recognition, Information Retrieval and Bioinformatics Etc.,

1.3 Methods on clustering

Clustering assigns records of similar objects into groups (called clusters) so that data objects of the same cluster are similar to one another than objects of different groups. Clustering methods have been argued extensively in Trend Analysis, similarity search, Segmentation, Pattern Recognition and classification. The clustering methods are classified into following methods
- Partioniong Method
- Grid - based Method
- Hieriarchical Method

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

• Model-based Method
• Density-based Method
• Constraint-based Method

Clusters include groups with small distances within the cluster members and more dark areas of the data space, intervals or particular statistical distributions[2]. Clustering methods for uncertain data mainly divided into two categories such as partitioning and Hierarchical approaches. Analysis similarity is the most important method using the clustering is partition and Hierarchical.

A. Partitioning Method

For 'n' data objects, the partitioning method develops k partition of data.. Each partition will represent a cluster k≤n. It classifies the data into k groups, which satisfies the following requirements:
• At least one object in each group.
• A object must belong to exactly one group not more than a group.

For a given number of 'k' partitions, the partitioning method creates an step partitioning. Then it uses the iterative relocation technique to improve the partitioning by moving data objects of one group into other.

The main drawback of partitioning the objects into k clusters repeatedly reallocates objects to improve the clustering. It uses an k-medoid method for each sub-set of a data stream. In order to iterative evaluation of the k-medoid algorithm[4], its objective is to maintain only the consistent good data elements ,i.e., each of which represents the cluster for the data elements.

B. Hierarchical Method

This method creates the hierarchical segregation of the given set of data objects. Thus, the decomposition of hierarchical algorithm is formed as follows:
Agglomerative: It is a 'bottom-up' approach. Each time a cluster or collection is erged with other group to shape larger ones.

Divisive: It is a 'top-down' approach. All data objects are placed in single cluster and split it up into smaller clusters.

C. Density-Based Method

The Density-based method is based on the notion of density. It allows the group to grow as long as the density in its neighborhood goes beyond some threshold level i.e. for each data point in a given cluster the radius of a given cluster must contain at least a minimum number of data points.

D. Grid-Based Method

In this the objects together form a multi-resolution grid structure. The object space is divided into fixed number of cells that create a grid formation. The major advantage of this method is fast processing time. Another advantage is dependent only on the cells in each dimension in the space.

E. Model-Based Method

A model is hypothesizing for each cluster and finds the best fit of data to the given data model. It identifies the clusters by applying the density function. This shows spatial distribution of the data points. This method serves as a way of automatically determe the number of clusters based on typical statistics considering outliers or noise into account.

F. Constraint-Based Method

It identifies the user expectation or the properties of clustering results. The constraint gives us the interactive way of communication with the clustering process. The constraints are specified by the user or the application requirement.

2. Hierarchical clustering

The Cluster analysis goal is that the objects within a group must be similar to each other and dissimilar from the objects of the other groups. The greater similarity (or homogeneity) of clustering within the group and greater difference between the groups and better or more distinct among the clustering. The hierarchical clustering is a method of cluster analysis which builds clusters in hierarchical fashions. The strategy for hierarchical clustering are of two types:
• *Agglomerative:* It is a "bottom-up" approach. Each iteration starts with one cluster and pairs of clusters are merged to get new clusters.
• *Divisive:* It is a "top- down" concept. In each time, the iterations begins with a cluster ‘A’ and splits are performed continuously as one moves down the hierarchy.

Fundamentally, the merges and splits are identified in a greedy fashion. The output of hierarchical clustering are generally displayed by using a dendrogram. The disadvantage of agglomerative clustering is it makes them too slow for large data set points.

3. Advantages of Hierarchical Clustering

The advantages of the hierarchical clustering algorithms are,
• Embedded flexibility in level of granularity.
• Easy handling of any forms of similarity or distance.
• It is applicable to any attributes types.

These advantages of hierarchical clustering leads to the cost of lower efficiency. Agglomerative hierarchical clustering presents four different algorithms,

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

- Similarity measures of a single-relink process of chaining effect,
- Complete-link process of not sensitive to outliers,
- Group-average process of Best choice for most applications,
- Centroid process of inversions can be occurred.

4. Conclusion

In this paper, various clustering algorithms and its features are discussed and analysed. Based on the results, the hierarchical clustering techniques in data mining are recognized as efficient and best for many applications in various industries.

References

[1] I.K. Ravichandra Rao (2003), "Data Mining and Clustering Techniques", DRTC Workshop on Semantic Web, Bangalore.
[2] Jiawei Han & Micheline Kamber (2006), "Data Mining: Concepts and Techniques", The Morgan Kaufmann / Elsevier India.
[3] "Clustering Uncertain Data With Possible Worlds"Peter Benjamin Volk, Frank Rosenthal, Martin Hahmann, Dirk Habich, Wolfgang Lehner, IEEE International Conference on Data Engineering.
[4] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais & S.J. Formosinho (2007), "Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering", Chemometrics and Intelligent Laboratory Systems, Vol. 87, Pp. 208–217.
[5] A.S.Aneeshkumar and Dr. C.Jothi Venkateswaran, "A novel approach for Liver disorder Classification using Data Mining Techniques", Engineering and Scientific International Journal, Volume 2, Issue 1, January - March 2015, pp.15-18.