Data Mining Clustering Techniques : A Case Study

S. Chinmayee Krishna* & Shoban Babu Sriramoju**
*Student, S R Edu Center, Warangal
**Professor, Department of CSE, S R Engineering College, Warangal

Received: March 31, 2018 Accepted: May 04, 2018

ABSTRACT *Now a day's data is growing in a sense of size and variety. How to fetch information from the databases is a important concern. Decision making out of information is challenging these days. Many techniques have been developed for extracting. One of techniques is data clustering. In this paper, a review of several clustering techniques that are being used in Data Mining is presented. In clustering we use cluster of same type of data and current data mining clustering techniques.*

Keywords: *database, techniques, clustering, data mining etc.*

INTRODUCTION

For any organization collection of data for future reference is very important. Along with that having tools that can be used to check coming requirements, trends and problem of existing data is necessary. Data clustering is a technique that can check many data sets and each data set can contain different data types. Each data set is having different size and size of dataset is depends upon the count of objects, dimensions and different data types. In case of data clustering method internal structure of data is not known.

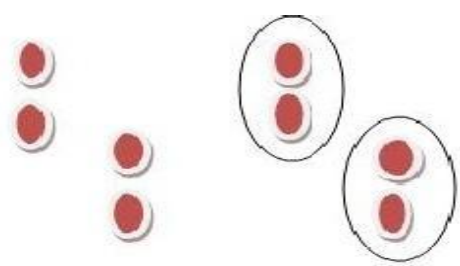
A. Data Mining Techniques

Classification, Predication, Association, neural networks are various techniques of Data Mining.

B. Clustering

Clustering is a fundamental operation in data mining. Clustering can be understood as recognition of alike classes of objects. It can discover overall division pattern and correlations among data attributes. Clustering methods have been projected and they can be mostly classified into four categories like partitioning methods, hierarchical methods, density-based methods and grid-based methods.

C. Partitioning Methods



1) Originalcluster
2) Partitioned cluster

Partitioning method simply moving instances from one cluster to another for relocation. It starts moving from initial Partitioning. In this method user will predetermined number of cluster .

To achieve global optimality in partitioned-based clustering, an complete enumeration process of all possible partitions is required. Namely, a relocation method iteratively relocates points between the k clusters. The following sections present different types of partitioning methods.

A partitioning-based clustering algorithms combinatorial optimization algorithm is the most popular class of clustering algorithms. They are also known as iterative relocation algorithms. These algorithms minimize A given clustering criterion is minimized using these algorithms by iteratively relocating data points between clusters until an optimal partition is attained. In a basic iterative algorithm, such as

K-means algorithms which are used as a solution to clustering problem. In this algorithm, a given dataset is classified into a fixed number of clusters (assume k clusters). The main idea is to define the centroids of each cluster. The centroids of each cluster are placed as far as possible from each other. In the next step, each point belonging to a given data set is taken and associated to the nearest centroid. When there exist no point and early grouping is done. A loop has been generated. Because the number of data points in any

data set is always finite and, thereby, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods. Graph-Theoretic, Error Minimization Algorithms are used for partitioningmethod.

D. Hierarchical Agglomerative (divisive)Methods

As the name specifies this method is to create a hierarchy of clusters i.e. like tree structure. Now the point is that we can create a hierarchy by using bottom up or top down approach. So hierarchy clustering method is divided into two parts Agglomerative and Divisive. Agglomerative is a bottom up approach. The process of clustering is continued till certain condition is satisfied.

Each examination starts on its own cluster and then the pairs of clusters are combined as soon as one moves up the hierarchy in case of bottom up approach.

The top down approach works as twice as faster than bottom up approach.

To increase agglomerative clustering algorithm efficiency as well as to make it suit for large data, we require the cloud computing virtualized environment]. Virtualization is a key technology used in data centers to optimize resource.

E. Density BasedMethods

Density based clustering algorithm is one of the prime methods for clustering in data mining. In this clusters are formed on the bases of density.

Density based clustering method is useful because it can find clusters in random shapes and it can handle noisy data efficiently. It is called as one scan algorithm because raw data is examined only once. In density based clustering clusters are defined as areas of higher density than remainder of the data sets. One cluster is separated from other clusters by lower density regions.

F. Grid-Based Methods

Grid based method is different from other as performs action on cell rather than data points. This method is having more calculation efficiency as compare to other methods that are traditionally used.

In fact, most of the grid-clustering algorithms achieve a time complexity. All clustering operations are performed in a gridded data space. Grid-based methods are most widely used in comparison to the other conventional models as they have high computational efficiency. The major difference between grid-based and other clustering methods is that all the clustering operations are performed on the segmented data space, instead of the original data objects. In grid-based clustering methods, it has to determine beforehand a proper size of the grid structure which is a major difficulty. Larger grid size can be managed by combining two or more clusters into a single cluster. In case of smaller grid size, a cluster may be divided into several sub-clusters. So, finding the suitable size of grid is a challenging issue in grid clustering methods. This problem is known as the locality of cluster. The third problem is how to select a merging condition to form efficientclusters.

G. Model-basedmethods

These methods attempt to optimize the fit between the given data and some mathematical models. Other than usual clustering which identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each grouprepresents a concept or class. The most frequently used induction methods are decision trees and neural networks. Model-based Clustering MLE (maximum likelihood estimation) is used in model-based clustering method to find the parameter inside the probability model. Since the probability function is a mixture summation of a couple of probability function, it makes the traditional method infeasible to find the maximum value. Latent variable technique is used here, relocation algorithm such as EM and Gibbs sampling are among the most popular. The criterion to split one data set into several data sets is to make the variance between the clusters maximum and inside the clustersminimum.

CONCLUSION

It has been concluded in this that data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining contains huge variety of applications in every domain of industry where the data is generated, so the data mining is considered as one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in InformationTechnology.

REFERENCES

1. *Shoban Babu Sriramoju, "Heat Diffusion Based Search for Experts on World Wide Web" in "International Journal of Science and Research", https://www.ijsr.net/archive/v6i11/v6i11.php, Volume 6, Issue 11, November 2017, 632 - 635, #ijsrnet*
2. *Dr. Shoban Babu Sriramoju, Prof. Mangesh Ingle, Prof. Ashish Mahalle "Trust and Iterative Filtering Approaches for Secure Data Collection in Wireless Sensor Networks" in "International Journal of Research in Science and Engineering" Vol 3, Issue 4, July-August 2017 [ISSN : 2394-8299].*
3. *Dr. Shoban Babu, Prof. Mangesh Ingle, Prof. Ashish Mahalle, "HLA Based solution for Packet Loss Detection in Mobile Ad Hoc Networks" in "International Journal of Research in Science and Engineering" Vol 3, Issue 4,July-August 2017 [ISSN : 2394-8299].*
4. *Shoban Babu Sriramoju, "A Framework for Keyword Based Query and Response System for Web Based Expert Search" in "International Journal of Science and Research" Index Copernicus Value(2015):78.96 [ISSN : 2319-7064].*
5. *Sriramoju Ajay Babu, Dr. S. Shoban Babu, "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].*
6. *Dr. Shoban Babu Sriramoju, Ramesh Gadde, "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].*
7. *Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2014 [ISSN : 2249-4510]*
8. *Mounica Doosetty, Keerthi Kodakandla, Ashok R, Shoban Babu Sriramoju, "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2012 [ISSN : 2249-4510]*
9. *Shoban Babu Sriramoju, "An Application for Annotating Web Search Results" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,March 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]*
10. *Shoban Babu Sriramoju, "Multi View Point Measure for Achieving Highest Intra-Cluster Similarity" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,March 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]*
11. *Shoban Babu Sriramoju, Madan Kumar Chandran, "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4, Issue 2, February 2014 [ISSN : 2277 128X]*
12. *Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao, "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications" Vol V,Issue II,Febrauary 2014 [ISSN : 2321-3469]*
13. *Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Analysis on Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management", Vol VII, Issue X, November 2014 [ISSN : 2249-4510]*
14. *Shoban Babu Sriramoju, "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2, Issue 1,January 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]*
15. *Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Processing" in "International Journal of Information Technology and Management"*
16. *Vol V, Issue I, August 2013 [ISSN : 2249-4510]*