SURVEY OF CLUSTERING TECHNIQUES FOR DATA MINING

O.Bhaskaru[1], J.Raghunath[2]
[1]*Gates Institute of technology, Gooty, Anantapur, Andhra Pradesh, India*
[2]*Gates Institute of technology, Gooty, Anantapur, Andhra Pradesh, India*
(E-mail: bha_jesus@yahoo.co.in, raghu.jangam@gmail.com)

***Abstract*—**The main aim of this review paper is to provide a comprehensive review of different clustering techniques in data mining. Clustering is used data mining technique in which a group of similar objects is combined together to form clusters, these clusters are different from the objects in another clusters. This paper describes some clusterization techniques like, hierarchical technique, grid-based technique, partitional technique, density-based technique, and their algorithms. The grid-based clustering approach uses multi resolution grid data structure. Partitional method divides the data set into objects based on some similarity criterion, hierarchical method creates a hierarchy between clusters by combining the data objects into clusters, and then these clusters are further combined together to form large clusters and so on, density based method are used to separate the high dense clusters from low dense clusters. (Abstract)

Keywords*——Cluster, Clustering methods, Classifications (key words)*

I. INTRODUCTION

We provide a comprehensive review of different clustering techniques in data mining. Clustering refers to the division of data into groups of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups. A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Data Clustering is one of the challenging mining techniques exploited in the knowledge discovery process. Clustering huge amounts of data is a difficult task since the goal is to find a suitable partition in a unsupervised way (i.e. without any prior knowledge) trying to maximize the similarity of objects belonging to the same cluster and minimizing the similarity among objects in different clusters. Many different clustering techniques have been defined in order to solve the problem from different perspective, i.e. partition based clustering, density based clustering, hierarchical methods and grid-based methods etc. In this paper we represent a survey of recent clustering approaches for data mining research.

II. GENERAL TYPES OF CLUSTERS

2.1. Density-Based Clusters

A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density. Used when the clusters are intertwined or irregular, and when noise and outliers are present.

2.2. Well-Separated Clusters

If the clusters are sufficiently well separated, then any clustering method performs well. A cluster is a set of node such that any node in a cluster is closer to every other node in the cluster then to any node not in the cluster.

2.3. Center-Based Clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any cluster other then it. The center of a cluster is often called as centroid, the average of all the points in the cluster, or a mediod, the most "representative" point of a cluster.

2.4 Contiguous Clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

2.5 Shared Property or Conceptual Clusters

Finds clusters that share some common property or represent a particular concept.

2.6 Grid Based Clustering

Grid-based clustering is used where the data space is divided into finite number of cells which forms the grid structure and performs clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest processing time that depends only on the size of the grid not on the data. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. All these methods use a uniform grid mesh to cover the whole problem.

III. CLASSIFICATION OF CLUSTERING

Clustering is the main task of Data Mining. And it is done by the number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density and Grid based algorithms.

Hierarchical Clustering

Hierarchical clustering is a clustering technique in which the similar dataset is divided by constructing a hierarchy of clusters. This method is based on the connectivity approach. This hierarchy is created using two algorithms which are: Agglomerative and Divisive.

Agglomerative - The method starts with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other thus merged together to form the next largest cluster. The merging thus continues until a hierarchy of clusters is constructed with just a single cluster comprising all the records at the top of the hierarchy.

Divisive – The technique take the opposite approach from agglomerative techniques. They start with all the records in one cluster and then split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

Partitioning clustering

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and the Probabilistic Clustering. We are discussing the k-mean algorithm as: In k-means algorithm, a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition „n" observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Grid based clustering

The grid based clustering uses a multi resolution grid data structure. It is used for building clusters in a large multidimensional space wherein clusters are regarded as denser regions than their surroundings. This method partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. It 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. The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. Grid based methods help in expressing the data at varied level of detail based on all the attributes that have been selected as dimensional attributes. In this approach representation of cluster data is done in a more meaningful manner. A typical grid-based clustering algorithm consists of the following five basic steps:

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 STING (Statistical Information Grid based) and Wave Cluster are examples of grid based clustering.

The quality of clustering produced by this method is directly related to the granularity of the bottom most layers, approaching the result of DBSCAN as granularity reaches zero. It explores statistical information stored in grid cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure: each cell at high level is partitioned to form a number of cells at the next lower level. Statistical information regarding the attributes in each grid cell is pre-computed and stored. CLIQUE was the first algorithm proposed for dimension –growth subspace clustering in high dimensional space. Wave Cluster does not require users to give the number of clusters applicable to low dimensional space. It uses a wavelet transformation to transform the original feature space resulting in a transformed space where the natural clusters in the data become distinguishable.

Advantages
1. Fast processing time.
2. Independent of the number of data objects.

Drawbacks
1. Depends only on the number of cells in each dimension in the quantized space.

Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remaining of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. There are two major approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity both measured in terms of local distribution of nearest neighbors. So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach pins density to a point in the attribute space and is explained in the sub-section Density Functions. In this, density function is used to compute the density. Overall density is modeled as the sum of the density functions of all objects. Clusters are determined by density

attractors, where density attractors are local maxima of the overall density function. The influence function can be an arbitrary one. It includes the algorithm DENCLUE.

Density Based Spatial Clustering of Applications Noise

DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is a density based clustering algorithm. In this algorithm the regions grow with sufficiently high density are known as clusters. The Eps and the Minpts are the two parameters of the DBSCAN. The basic idea of DBSCAN algorithm is that for each object of a cluster, the neighborhood of a given radius (Eps) has to contain at least a minimum number of objects (MinPts).The clustering quality of DBSCAN algorithm strongly depend on the parameters does not depend upon the database. The parameters are set by users which will considered in the computational of clusters. The users have to select the parameters properly to get the better results the reason is that the same database with different parameters; the algorithm can produce different results. However, DBSCAN algorithm uses global parameters, which are not suitable for discovering clusters with different densities, without considering different possible density, only using a given possible density of any clusters, when the densities of clusters are totally separated.

IV. CONCLUSION

Clustering is that technique of data mining which is used to extract the useful information from raw data. We can say that raw data is useless without the different clustering techniques.

V. ACKNOWLEDGMENT

I would like to thanks the Department of Computer Science & Engineering of RIMT Institutes near Floating Restaurant, Sirhind Side, Mandi Gobindgarh-147301, and Punjab, India

REFERENCES

[1] Amandeep Kaur Mann, Navneet Kaur "SURVEY PAPER ON CLUSTERING TECHNIQUES", ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[2] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya ",COMPARISON THE VARIOUS CLUSTERING ALGORITHMS OF WEKA TOOLS", ISSN 2250-2459, Volume 2, Issue 5, May 2012.

[3] Aastha Joshi, Rajneet Kaur "A REVIEW: COMPARATIVE STUDY OF VARIOUS CLUSTERING TECHNIQUES IN DATA MINING", Volume 3, Issue 3, March 2013.

[4] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim "An Efficient Clustering Algorithm for Large Databases"

[5] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.

[6] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, "OPTICS: Ordering Points To Identify the Clustering Structure" Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[7] Osama Abu Abbas "Comparisons between data clustering algorithms" [9] Preeti Baser, Dr. Jatinderkumar R. Saini "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets"

[8] Kehar Singh, Dimple Malik and Naveen Sharma "Evolving limitations in K-means algorithm in data mining and their removal" IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011

[9] Suman and Mrs.Pooja Mittal "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weak tool" IJETTCS.

[10], M Meila, D Verma, 2001. Comparison of spectral clustering algorithm. University of Washington, Technical report

[11], N. Pal, J. Bezdek, and E. Tsao, "Generalized clustering networks and Kohonen's self-organizing scheme," IEEE Trans. Neural Netw., vol. 4, no. 4, pp. 549–557, Jul. 1993.