IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-17, May 2019 E-ISSN: 2347-2693

Clustering Techniques and Hierarchical Distance Measure in Datamining

M. Angelin Rosy[1*], D. Shyamala[2], M. Felix Xavier Muthu[3]

[1,2,]Department of MCA, Er.Perumal Manimekalai College of Engineering, Anna University, Hosur, India
[3]Dept. of Mechanical Engineering, St.Xavier's Catholic College of Engineering, Anna University, Nagercoil, India

Corresponding Author: angel_rosym@yahoo.co.in, Tel- 9944579754

Available online at: www.ijcseonline.org

***Abstract*—**Data mining is extracting information from huge set of data. Clustering is a process of organizing object into unknown group. it deals with finding a structure in a collection of unlabeled data. Similar objects are grouped in one cluster and dissimilar are grouped in another cluster. The documents clustering will aims to group in unsupervised way. Clustering analysis is one of the main logical methods in data mining. Which focuses on the current popular and commonly used k-means algorithm? Clustering can be classified into partition method, hierarchical method, density based method, grid based method, and model based method. In hierarchical method are based on different distance measures. In each type calculate the distance between each data objects and all cluster centers .this paper provides a broad survey of the most basic techniques and identifies.

***Keywords*—**Data mining, Clustering techniques, K-means algorithm ,Hierarchical method, Partition method.

I. INTRODUCTION

Data mining is a collection of technique for efficient automated discovery of previously unknown novel, valid useful and understandable pattern in large database. Data mining is a process of extracting information. Clustering is a grouping data into classes. Data mining include pattern, association rules, changes anomalies and significant structure from large amount of data. Such has xml, rdbms, and data ware house, knowledge discovery (KDD) .it helps to extract patterns and make hypothesis from the raw data. Clustering is a process of organizing objects into unknown groups. Clustering algorithm is used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. In data mining two types of learning set are used supervised learning and unsupervised learning. It is important understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification).

A. Supervise Learning:
Classification is called the supervise learning. Classification is ordering of objects or dividing the objects into predefines classes (known class). The classes are determine before examine the data. Supervised models are decision tree, neural network, rule based method.

B. Unsupervise Learning:
Clustering is called the unsupervised learning. Clustering is the process of organizing of objects or dividing the objects into not predefines class (unknown group).

II. LITERATURE REVIEW

1) Duran and Odell5 also provide a brief exposition of cluster analysis. Theirs, however, is more mathematically detailed than Everitt's with little reference to practical problems or detailed examples.
2) Tryon and Bailey1l devote their book on cluster analysis to a comprehensive description of their factor analysis/cluster analysis package called BC TRY. Three applications of cluster analysis are used repeatedly in the book and all of the data are from Tryon's field of interest psychology.

3) Hartigan8 provides a treatment of modern clustering theory from the statistical point of view. His book contains detailed discussions of a variety of algorithms and their application to real data sets ranging from medical and biological data to political data.

III. CLUSTERING

Clustering involve combination data into class. Clustering is process of organize object into unknown group. Clustering is an unsupervised learning. Similar objects are grouped in one cluster and dissimilar are grouped in another cluster. Cluster is an ordered list of data which have familiar characteristics

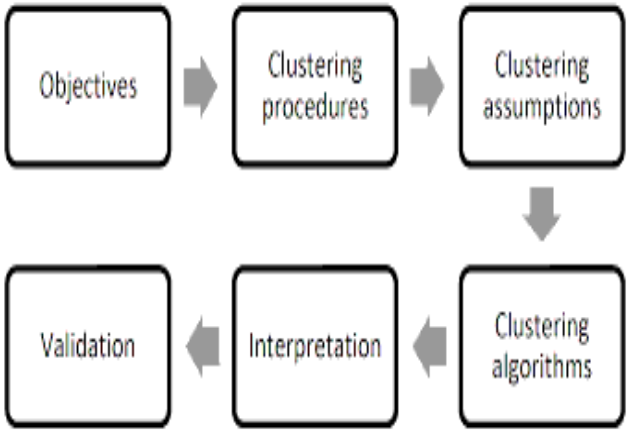

Figure 1: Stages of clustering

IV. CLUSTERING APPLICATIONS

- Business
- Medicine
- Banking &finance
- Social network analysis
- Image segmentation
- Data mining
- Earth quick

The clustering is the classified into the many techniques. Most similar data grouped into cluster.

V. CLUSTERING TECHINIQUES

1. Partition method
   - k-means algorithm
   - expectation and maximization method
2. Hierarchical method
   - agglomerative approach
   - divisive approach
3. Density based method
4. Grid based method
5. Model based method

The parallel between the objects is designed by the use of a similarity function. It is largely useful for organize papers, to recover improvement and support browsing.

A. Partition Method:

It obtains a single level partition of objects these methods are usually based on heuristic creating local optimum solution. Each cluster has at least one object and each object belongs to only one cluster. Methods like k-mean, PAM (Partitioning around Medoids), CLARA (Clustering LARGE Applications) and the Probabilistic Clustering are comes under partitioning clustering. Partitional clustering is considered to be the mainly popular this method is efficient and early adapted for large database. This algorithm uses for large database, this algorithm uses iterative refinement method (hile climbing or greevy method) they converted to a local minimum rather than global minimum

1. Number of cluster apriori
2. User to specify starting state of the cluster

Class of clustering algorithm is also known as iterative replacement algorithm. This method is efficient and early adapted for large database. This algorithm uses iterative refinement method (hile climbing or greavy method) there are two important methods.

- k-means algorithm method
- expectation and maximization method

1. K-means algorithm

K-means algorithm is classical clustering method which is easy to implement. The data about the entire object is located in the main memory. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to locate groups in the data, with the number of groups represented by the variable K. This procedure follow a simple and easy way to classify a given data set through a certain the number of clusters (assume k clusters) fixed a priori. The main suggestion is to define k Centroid, one for each cluster. As much a7s possible far gone from each other. So, the improved choice is to place them.

- The Centroid of the K clusters, which can be used to label new data
- Labels for the preparation data (each data point is assign to a particular cluster)

K-means is a simple algorithm that has been modified to many difficulty domains. As we are going to see, it is a good applicant for extension to work with unclear feature vectors. The method is called k-means because each of the k-cluster is represented by the mean of the object [called Centroid with in it. It is also called as Centroid method at each step Centroid point is assumed to be known and each of the remaining points is allocated to the cluster whose Centroid is closed to it. Once the allocation completed the Centroid of the cluster are recomputed by using simple mean and the process is repeated until there is no change in the cluster. the k-means algorithm uses Euclidean, manhattans distance measure with compact cluster.

2. Exceptation And Maximization Method

K-mean method does not explicitly assume any probability distribution for the attribute value consist similar object. A common task in signal processing is the estimation of the parameters of a probability distribution function. Perhaps the most frequently encountered estimation problem is the estimation of the mean of a signal in noise in maximization and expectation method objects in the dataset have attribute whose value are distributed according to some unknown linear combination of simple probability distribution. K-

means method is used to minimize within group variation were as EM method is an attempt to maximize expectation of assignment.

EM method consist two iteration

- E STEP
- M STEP

1. E-STEP:

It involves estimating the probability distribution of the cluster given data.

2. M- STEP:

Involve finding the model parameters that maximize the likely would of the solution. Em method assumes all the attribute are independent random variable in a simple case of just two cluster .it object having only one single attribute ,we may assume that the distributed value vary according to normal distribution.

1) Mean and standard deviation of the normal distribution for cluster one.
2) The mean and standard deviation of the normal distribution for cluster2.
3) The probability p of a sample belonging to cluster 1 and therefore probability of belonging to cluster 2.

B. Hierarchical Method:

It obtain a nested partition of object resulting in a tree of cluster These method either start with one cluster and split into the small and small cluster start with each object individual cluster and try to merge large cluster. Hierarchical clustering algorithms have tended to be somewhat more prominent than others, perhaps because they presuppose very little in the way of data character- is tics or of a priori knowledge on the part of the analyst. This method creates a hierarchical decomposition of the agreed set of data objects. The tree of clusters also created name as dendrogram. Every cluster node contains child clusters, sibling clusters partition the points enclosed by their ordinary parent. In hierarchical clustering assign each item to a cluster such that N items then we have N clusters. Find next pair of clusters and merge them into single cluster. Compute distance between new cluster and each of previous clusters. It uses a number of greedy heuristic schemes of iterative optimization. The algorithms to be discuss in this article focus instead on closely what is necessary in order to carry out an agglomeration at any stage of the clustering: this is usually little more than the nearest neighbour points of specified points. Hierarchical cluster is classified as

- Agglomerative approach
- Divisive approach

1. Agglometitive Approach[bottom up approach]

Agglomerative approach is a bottom up approach. each object at the start is a cluster by itself and the nearby cluster are rapidly merged. Object is merged into a single large cluster. Until all object are in a single cluster or confident termination condition is satisfied. The single cluster becomes the hierarchy's root. It successively merges the groups that are close to one another, until all the data objects are in same cluster. It finds the two clusters that are closest to each other, and combines the two to forms one cluster.

2. Divisive Approach[top down approach]

Divisive approach is a top down approach. A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. All objects are put it in a single cluster. Then rapidly perform splitting of cluster .and resulting smaller and smaller cluster. Until stopping criteria is met. Then successively splits resulting clusters until only clusters of individual objects remember. There are two types

1. Monothetic
2. polythetic

Hierarchical cluster based on different distance measures are used

- Single link algorithm
- Complete link algorithm
- Centroid link algorithm
- Average link algorithm
- Ward's minimum variance algorithm

A. Single Link Algorithm:

It determines the distance between two clusters has the minimum of the distance between all pair of points nearest neighbour.

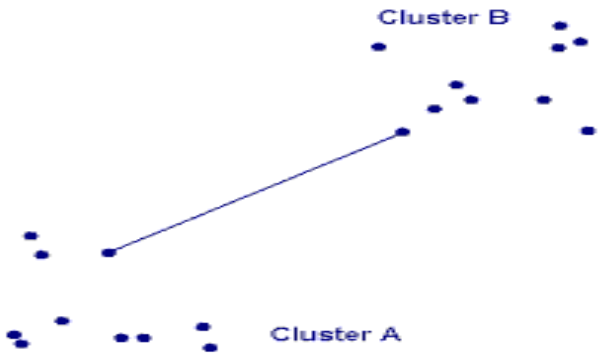

Figure2: Single linkage cluster

The dissimilarity between 2 clusters is the least variation between members of the two clusters. This method produces long chains which form loose, untidy clusters.

B. Complete Link Algorithm:

The distance between two cluster is define has a maximum of the pair wise distance. Therefore must be computed the largest chosen. Furthest neighbour

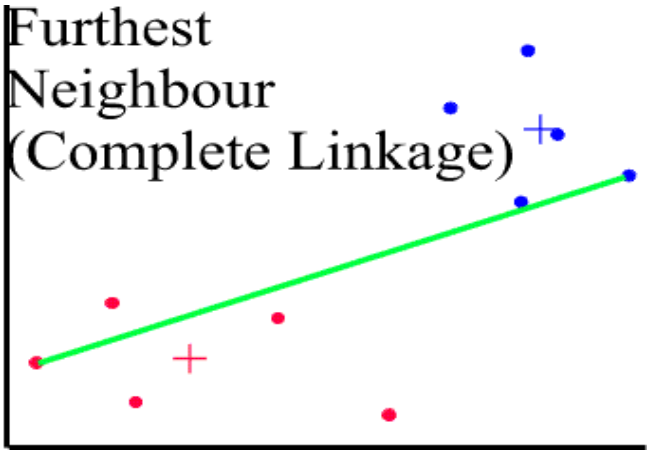

Figure 3: complete linkage cluster

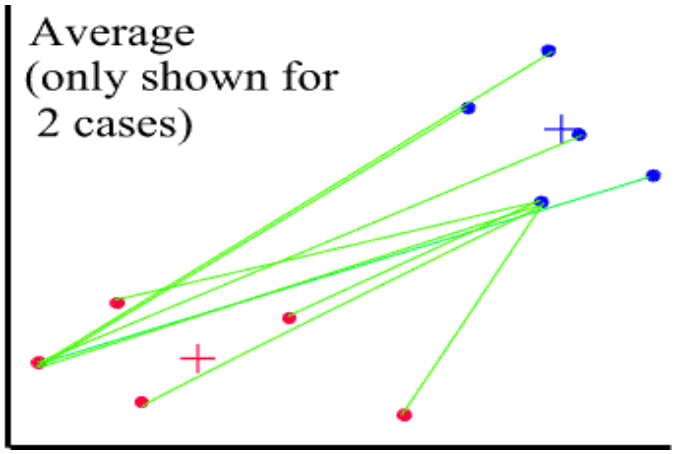

Figure5: Average linkage clustering

This variation uses the group centroid as the average. The centroid is defined as the center of a cloud of points.

C. Centroid Algorithm:

In this algorithm distance between two cluster is determines has the distance between centriod of the cluster. it computed the distance between two cluster has the distance between the average point of the two cluster

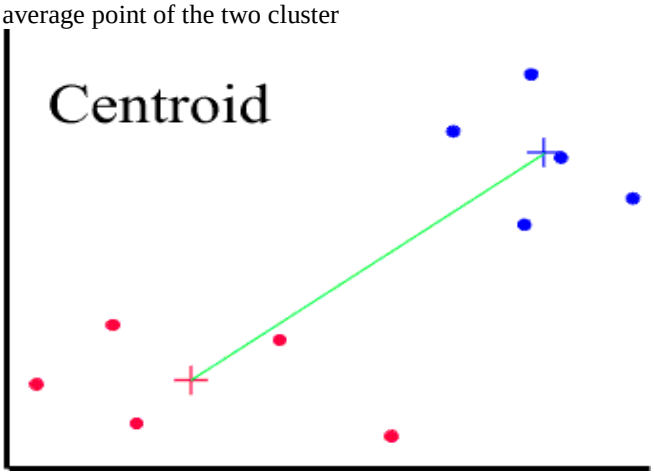

Figure 4: Centroid clustering

And we use the Euclidian distance to complex the Centroid. (Maximization of method) The Centroid method uses the Centroid (center of the group of cases) to determine the average distance between clusters of cases. Greatest similarity is between two members of cluster.

D. Average Link Algorithm:

It compute the distance between two cluster has the average of all pair wise distance between the object from one cluster to another cluster .that is if there are m element in one cluster n .

In the other there is distance to be completed. Added and divided by the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster

E. Ward's Mininmum Variance Method:

Ward's minimum is the different between the total within the cluster some of the square for the two cluster separately and within the cluster some of squares resulting from merging the cluster.

$$Dn(A,B)=NaNbDc(A,B)/(Na+Nb)$$

Cluster membership is assigned by calculating the total sum of squares deviation from the mean of The principle for union is that it should create the smallest possible add to in the error sum of squares.

C. Density Based Method:

The density based method is based on assumption that clusters are high density collection of data. That is separated by low density of data. Therefore the basis call density base method is for each data point in a cluster at least minimum number of points exit within the given distance. There are two important parameter are used

1. R (size of the neighbor wood)
2. N (The minimum points in the neighbor wood)

These two parameters determine the density within the cluster and also determine which object is outlier (or) noise. (The number of concept required for density base clustering)

- Neighbour hood
- Core object
- Proximity
- Connectivity

For these methods a "neighbourhood" has to be defined and the density must be calculated according to number of substance in the neighbourhood.

D. Grid Based Method:

In this technique measures the object space into a limited number of cells that form a grid organization on which all of the operations for clustering are performed. The object space rather than the data is divided into grid .this method is based on characteristics of data and can deal with non-numeric data. It is based on clustering leaning query answering in multilevel grid structures. It will generate a minimal description of each cluster. Unlike other clustering methods, Wave Cluster does not require users to give the number clusters applicable to low dimensional space. 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.

E. Model Based Method:

Here a model based on probability distribution. the algorithm try to build clusters with a high level of similarity within them and low level of similarity between them based on mean value this algorithm minimized the error function. They optimize the fit among the data and some mathematical model. Ex: EM (Expectation and maximization), SOM (Self organizing feature map) in this model is theorized for each group to locate the best shape of data for a given model.

VI. CONCLUSION

The main objective of this paper is to cluster similar or dissimilar data sets into different groups. it is the task of arranging a set of object so that objects in the identical group are more related to each other than to those in other groups(cluster).clustering algorithm can be classified into partition-based algorithm, hierarchical-based algorithm, density-based algorithm, grid-based algorithm and model-based algorithm. Under partition method, a brief description of k-means and expectation and maximization are disused. In hierarchical clustering the agglomerative and divisive algorithms, hierarchical based on different distance measures have been described. So that less time is consumed to collect the data to form cluster. it is clear from the above mentioned data that different clustering techniques can be used or applied not only one small data but on large and large amount of data sets.

REFERENCES

[1] k.chithra, D.maheswari,"A comparitive study of various clustering algorithms in datamining" (ISSN 2320-088X),vol6,issue.8,auguest 2017

[2] Dhara patel,Ruchi modi and Ketan sarvakar, "A comparitive study of clustering datamining:techniques and research challenges" (ISSN 2278-2540)vol 3, issue.9,september 2014

[3] S.Mythili,E.Madhiya,"An analysis on clustering algorithms in datamining"(ISSN 2320-088X)vol.3,issue.1,january 2014

[4] Shivangi bhardwaj,"data mining clustering techniques-A review" (ISSDN 2320-088X)vol.6,issue.5,may 2017

[5] Amandeep kaur mann,Navneet kaur,"Survey paper on clustering techniquies"(ISSN:2278-7798)vol.2,issue.4,april2013

[6] Pradeep rai,Shubha singh"A survey of clustering techniques" (0975-8887) vol.7-no.12,october 2010

[7] Aastha joshi,Rajneet kaur"Comparitive study of various clustering techniques in datamining" (ISSN:2277 128X) vol3,issue 3,march 2013

[8] Shraddha K.popat,Emmanuel.M"Review and comparative study of clustering techniques" (ISSN:0975-9646)vol5(1),2014,805-812

