Volume No.03, Special Issue No. 02, February 2015 ISSN (online): 2348 – 7550

A COMPARATIVE STUDY OF DIFFERENT CLUSTERING TECHNIQUE

Yogesh jain[1], Amit Kumar Nandanwar[2]

[1]CSE (Mtech Scholar), [2]Assistant Professor,

Vidya Niketan Samiti, Bhopal , (India)

ABSTRACT

Cluster analysis is process grouping the object according their similarity and dissimilarity .object can be physical or abstract. The cluster Analysis is as old as a human life and has its roots in many fields such as statistics, machine learning, biology, artificial intelligence. Cluster analysis has faced much challenge. There is several clustering method each has their own characteristics which satisfy the following criteria such as arbitrary shaped, high dimensional database, spherical shapes, domain knowledge and so on in this paper we describe the comparative study of these algorithm so user can choose particular algorithm according their need.

Keywords: Clustering, Density, Grid, Hierarchical , Model, Partition

I INTRODUCTION

A Different clustering algorithms exist in the literature review. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap, so that a method may have features from several categories. it is useful to present a relatively organized picture of the different clustering methods.

Clustering is the one of most important research area in the field of data mining. In common language clustering is division of data into different group. [1]Clustering is a process grouping the similar data into one cluster and grouping. The dissimilar data into another cluster. [2]Cluster analysis is used in wide variety of field such as- psychology, social science, biology, statics, information retrieval, machine learning and data mining.[3] Cluster analysis has not fix definition there are several working definition are commonly used. There are two main aspect of clustering which are described as below. First cluster analysis is viewed as finding only the most connected point while discarding the Background or noise point. Second it is acceptable to produce a set of cluster where the true cluster is also broken into several subcluster.[4]the main of clustering is minimize the intra class similarity and maximize inter class similarity.[1].

II TYPES OF CLUSTERING METHODS

2.1 Partitioning Methods

Given a database of n objects or data set ,a partitioning method constructs k partitions of the data, where each partition represents a each cluster(k< n). That is, it classifies the data into k groups, which together satisfy the following needs: (1) each group must contain at least one object, and (2) eachobject must belong to exactly one group.

The advantages and disadvantages of partitioning clustering methods are:

Advantages:

1. Simple and Relatively scalable.

2.Acceptable for datasets with compact spherical clusters that are well-separated.

Disadvantages:

1. Poor cluster descriptors.

3. High sensitivity to initialization phase, outliers,noise.

4. Reliance on the user to specify the number of clusters in advance..

5. Inability to deal with non-convex clusters of varying size and density.

6. Frequent entrapments into local optima.

2.1.1 PAM (Partition around Mediod)

PAM is developed by Kaufman and Rousseuw in 1987. The algorithm chooses k-mediod initially and then swaps the mediod object with non mediod as a result quality of cluster is improved. It is very robust when compare with k-mean in the presence of noise or outlier. Algorithm work well with small dataset but does not work well with large dataset. [4] The computational complexity of PAM is O (IK (N-K) 2) where I is a number of iteration[9].

Procedure Of PAM

1.Input dataset d.

2.Randomly select K object from dataset G.

3.Calculate total cost T for each pair of selected Si and non selected.

4.For each pair if T Si<_0 then it is replaced by SK.

5.Then find similar mediod for each non selected object.

6.Repeat the step 2, 3, 4 until find the mediod.[8].

2.2 Hierarchical Methods

A hierarchical method creates a hierarchical decomposition ofthe given data set of data objects. This method can be classified as being eitherdivisive or Agglomerative, based on how the hierarchical decomposition is formed.The divisive approach calledthe top-down approach,this approach starts with all of the objects in the same cluster In each successive iteration .a cluster is break up into smaller clusters, until that eventually each objectis in one cluster, or until a termination condition holds.Theagglomerative approach, also called the bottom-up approach this approach starts with each objectforming a separate group. It successively merges the objects or groups that are closeto one to another, until all of the groups are merged into one (the topmost level of thehierarchy), or until a termination condition holds. [9].

The advantages and disadvantages of hierarchical clustering methods are:

Advantages

1.Embedded flexibility regarding the level of granularity.

2.Well suited for problems involving point linkages, such that taxonomy trees.

Disadvantages

1. Inability to make corrections once the splitting and merging decision is made.
2. Lack of interpretability regarding the cluster descriptors.
3. Vagueness of termination criterion.
4. Prohibitively expensive for high dimensional and massive datasets.
5 .Severe effectiveness degradation in high dimensional spaces due to the curse of dimensionality phenomenon.

2.2.1 BRICH (Balance Iterative Reducing and Cluster Using Hierarchies)

It is proposed by Zhang, Ramakrishna &Linvy in 1996.it is based on the idea that we don't need to keep whole touple or cluster in the main memory. In BRICH we store the triple (N, LS, and SS)N is a number of data object in cluster, LS is linear sum of number of data object & SS is sum of square of attribute value of object in cluster. These triple are called CF (clustering feature kept in tree) called CF.CF tree represent by two features these are branching factor B and threshold T[5] The computational complexity of BRICH is O (N). it can find the good clustering in single scan of data and improve the quality using few additional scan and handle the noise effectively And also achieve the scalability and compressed data may improved the performance of hierarchical algorithm.[6].

Procedure of BRICH

The data object are loaded one by one and initially CF tree is constructed and object is inserted cluster leaf entry or in sub cluster if the diameter of subcluster become larger than T then leaf node and possible other are split. If CF tree of stage 1 does not fit into the memory build the small CF tree and size of CF is controlled by parameter T. thus choosing the large value for merge sub cluster and making tree smaller Perform clustering leaf node of CF hold subcluster statics. BRICH use these statics to apply some clustering techniques k-mean and produce initially clustering Redistribution of data object using centriod of cluster .this is an optional. Which require additional scan of dataset and reassign the object their closest centriod. This phase require labeling the initially data and detecting outlier.[5] Each node in CF tree can hold limited number of entries due to its size.

There are two approaches to improving the quality of hierarchical clustering:

(1) perform careful analysis of object "linkages" at each hierarchical partitioning such as in Chameleon.

(2) Hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters and they using another clustering method such as iterative relocation, as in BIRCH.

2.3 Density-Based Methods

Other clustering methods have been developed based on the notion of density. This methods can find only spherical-shaped clusters such that circle and experience difficulty at discovering clusters of arbitrary shapes.

Their general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold that means that is for each data point within a given cluster the neighborhood of a given radius has to contain at least a minimum number of points. this method can be used to filter out noise (outliers)and discover clusters of arbitrary shape.

A advantages and disadvantages of density based clustering are:

Advantages:

1. Discovery of arbitrary-shaped clusters with varying size.

2.Resistance to noise and outliers

Disadvantages:

1.High sensitivity to the setting of input parameters

2.Poor cluster descriptors

3.Unsuitable for high-dimensional datasets because of the curse of dimensionality phenomenon.

2.3.1 DBSCAN (Density Based Spatial Clustering of Application with Noise)

This algorithm is proposed by Easter in 1996.in DBSCAN cluster is defined by the set of all point connected to their neighbors. It is the requirement of DBSCAN user specify the neighbors and mini mum number of object it should have.[7] In DBSCAN Cluster are defined by the criteria such as below:

Core point which lie interior of density based cluster and should lie within the eps (radius, threshold value).Min pts (minimum points) which are user specified parameter, border point lie within the neighbor of core point and many core point share the same border point, Noise the point which is neither a core point or nor a border point. [8] The complexity of DBSCAN is O (N2). DBSCAN find the arbitrary shaped cluster and also not much sensitive to input order every newly inserted point effect only certain point. It also provides protection against noise and outlier and we does not need to number of cluster initially. DBSCAN need to know two parameter eps and minpts but calculate eps is time consuming because eps is calculated by k-distance map but k-distance map is time consuming. [7].

Procedure Of DBSCAN Algorithm Is

1. Arbitrary select a point r.

2. Retrieve all points density-reachable from r w.r.t Eps and Minpts.

3.If r is a core point, cluster is formed.

4.If r is a border point, no points are density-5.reachable from r and DBSCAN visits the next 6.point of the database.

7.Continue the process until all of the points have been processed.[7].

2.4 Grid-Based Methods

Grid-based methods quantize the object space into a finite number of cells that form a grid structure. the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time to other methord, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

2.4.1 Wave Cluster

Wave cluster is a multi resolution clustering algorithm. It is developed by sheik holeslami in 1988. It is based on the signal processing technique (wavelet transformation) convert the spatial data into frequency domain. Each grid cell summarized information of group of point map into cell then it use the wavelet transformation the original feature space. [7] A wavelet transformation is a signal processing technique that decompose the signal into different frequency band.[8]The computational complexity of wavelet transformation is O(n) where n is number of object in data space. Wavelet transformations automatically removes outlier and discover the cluster of arbitrary shaped. It is insensitive to order of input. It can handle the data up to 20 dimensions and the large data efficiently. A prior knowledge of number of cluster is not required in wave cluster. [8]

The first step of wavelet cluster is quantized the feature space .

The second step of wavelet cluster algorithm applied discrete wavelet transformation on quantized space

The third step of wavelet cluster algorithm level the unit in feature space that are include in cluster..

2.5 Model-Based Methods

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. This algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a road of automatically determining the number of clusters based on standard statistics, taking “noise” or outliers into account and thus yielding robust clustering methods.

2.5.1 Self-Organizing Feature Maps (Soms)

Self organizing feature maps are one of the popular neural networkmethods for cluster analysis. They are sometimes referred to as Kohonen self-organizing feature maps, after their creator, TeuvoKohonon, or as topologically ordered maps.SOMs’ goal is to represent all points in a high-dimensional source space by points in a low-dimensional (usually 2-D or 3-D) target space, such that the proximity relationships and distance are preserved as much as possible. The method is particularly useful when a nonlinear mapping is inherent in the problem itself.SOMs can also be viewed as a modifier version of k-means clustering, in which the cluster centers tend to lie in a low-dimensional manifold in the feature or attribute space.With SOMs, clustering is performed by having several units competing for the current object. The unit whose weight vector is closest to the current object becomes the winning or active unit. So as to move even closer to the input object the weights of the winning unit are adjusted as well as those of its nearest neighbors. SOMs assume that there is some topology or ordering among the input objects and that the units will eventually take on this structure in space. The organization of unit is said to form a feature map.

SOMs are believed to resemble processing that can occur in the brain and are useful for visualizing high-dimensional data in 2-D or 3-D space.

III COMPARISON OF DIFFERENT CLUSTERING TECHNIQUES

Sr .No	"Clustering Technique"	Proposed by	Data set	Cluster shape	Input parameter	"Outlier Handling"	Complexity	Measure
1	Partitioning Methods (PAM)	Kaufman& Rousseuw	Small	Arbitrary	No of cluster	No, Detect outlier	O(IK(n-k)2	Medoid
2	Hierarchic al Methods (BIRCH)	Zhang,Ramakrishnan& Linvy	Large	Spherical	branching factor B,threshold T (max. diameter of sub cluster)	Yes	O(n)	Feature Tree
3	Density based clustering algorith (DBSCAN)	Martin Ester,Hans-Peter Kriegel&XiaoweiXu	High Dimensional	Arbitrary	a) radius b) minimum points	Yes	O(n log n)	Density Based
4	Grid-Based Methods (Wave Cluster)	Sheikholeslami,Gholamhosein, SurojitChatterjee&Aidong Zhang	Large	Arbitrary	No	Yes	O(n)	Wave transform
5	Model-Based Clustering Methods (SOM)	TeuvoKohonon	Low Dimensional	Arbitrary	No	Yes	O(N2)	Object Similarity

Table.1.Comparison of Different Clustering Techniques

IV CONCLUSIONS

Clustering is process of grouping the object in which similar object are placed in one group and dissimilar are placed in another group. There is several clustering method each has their own algorithm. The algorithms which satisfy the following criteria such as arbitrary shaped, high dimensional database, spherical shapes, domain knowledge and so on. Single algorithm cannot fulfill these entire requirements of clustering .so it is difficult to choose any single algorithm for specific purpose. In this paper we describe the comparison of clustering algorithm so the user choose particular algorithm according their requirement.

REFERENCES

[1] Shalini S Singh, N C Chauhan," K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, May 2011.

[2] YujieZheng,"Clustering Methods in Data Mining with its Applications in High Education", International Conference on Education Technology and Computer, vol.43,2012.

[3] Er. Arpit Gupta, Er.Ankit Gupta, Er. Amit Mishra," RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS", International Journal of Advance Technology & Engineering Research, Vol. 1, Issue 1, pp 39-47, November 2011.

[4] P. Murugavel, Dr. M. Punithavalli," Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering, Vol. 3 No. 1,pp 333-339, Jan 2011.

[5] PeriklisAndritsos," Data Clustering Techniques", pp 1-34, March 11, 2002.

[6] mariahalkidi, yannisbatistakis, michalisvazirgiannis," On Clustering Validation Techniques", Journal of Intelligent Information Systems, 17:2/3, pp 107–145, 2001.

[7] PoojaBatraNagpal,PriyankaAhlawat Mann," Comparative Study of Density based Clustering Algorithms", International Journal of Computer Applications, Volume 27– No.11,pp 44-47, August 2011.

[8] https://sites.google.com/a/kingofat.com/wiki/data-mining/cluster-analysis.

[9] PrabhdipKaur, ShrutiAggrwal ," Comparative Study of Clustering Techniques", international journal for advance research in engineering and technology,April 2013.