INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Comparative Study of Clustering Techniques

Prabhdip Kaur[1], Shruti Aggrwal[2]

[1]Research Scholar, [2]Assistant Professor
SGGSWU, Fatehgarh Sahib
[1]prabhdippandher@ymail.com,
[2]shruti_cse@sggswu.org

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.*

1. INTRODUCTION

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]

2. TYPES OF CLUSTERING METHODS

Number of clustering method is proposed in recent year and all clustering method are categorized into two categories: partition based clustering, hierarchical clustering. In the following section we provide the overview of all well known clustering method

2.1 Partition Based Clustering

Partition based clustering create k partition of data set with n data object. It is an iterative relocation technique is used to improve the clustering by moving up the object from one group to another. Partition based clustering is represent by centriod or mediod. [1] They use iterative way to produce the clustering. One of the disadvantages of partition based clustering is their high complexity. Even when there are a small number of objects the partition is huge. [5] The Most Efficient Algorithms Proposed under Partition Based Method are:

2.1.1 K-mean

K-mean describes that given dataset of n object divide into k cluster where k is desired number of cluster. A centriod is defined for each cluster in k-mean all data object are placed in cluster having centriod nearest to all data object. After processing all data object then k-mean centriod is calculated again and again. In each iteration centriod change their location. This process continues step by step until no centriod move. K-mean relatively scalable, efficient for processing large dataset, easy to understand and implement. Need to specify k number of cluster in advance. It is unable handle the noise or outlier or handle the cluster very different shapes [1]The complexity of k-mean clustering is O (IKN) where I denote number of iteration and k<<n.

Steps of k-mean

Select the k object as initially center
Assign each data object to cluster center
Recalculate the center of each cluster
Repeat step 2 and 3 until cluster center don't change [4]

2.1.2 PAM (Partition around Mediod)

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

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.[8] Algorithm work well with small dataset but does not work well with large dataset. [9] The computational complexity of PAM is O (IK (N-K)^2) where I is a number of iteration.

Procedure of PAM

Input dataset d
Randomly select K object from dataset G
Calculate total cost T for each pair of selected Si and non selected
For each pair if T Si<_0 then it is replaced by SK
Then find similar mediod for each non selected object
Repeat the step 2, 3, 4 until find the mediod.[8]

2.1.3 CLARA (Clustering Large Application)

CLARA is developed by Kaufman & Rousseuw in 1990.CLARA algorithm work well with several sample size of N touple in dataset. Then we apply the PAM each sample.[16]It can identify outlier and select the best mediod as output.[8]. This method takes sample of data from dataset instead of taking the full dataset. It randomly selects the data then chooses the mediod using the PAM algorithm. [10]The computational complexity of CLARA is O (K (40+K)^2+k (Ni)). [16] Clara Efficiency depends on the sample size A good clustering based on samples .It will not necessarily represent a good clustering whole data set if the sample is biased and in CLARA and if "true" mediod of the initial data are not contained in the sample, then the result is guaranteed not to be the best result

Procedure of CLARA

- Input the data set D
- Repeat N time
- Draw sample S randomly from D
- Call PAM to get mediod
- Classify entire data set to cost1--------cost k.
- Calculate the average dissimilarity from obtained cluster.[9]

2.1.4 CLARANS (Clustering Large Application Based Upon Randomized Search)

Ng and Han a new algorithm in 1994 called CLARANS. It use random search to generate neighbors by starting with arbitrary node and randomly check max-neighbors. If the neighbor represent better partition the process continue with new node otherwise local minimum is found and algorithm restart until numlocal local minima is found (value of numlocal is=2 recommended)the best node return resulting partition.[11]CLARANS take a random dynamic selection of data at each step of process. Thus the same sample set is not used throughout in the clustering process. As a result better randomization source is achieved. [12] CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasing dimensionally and draw the sample of neighbors in each step of search this is benefit of confining the search localize area.

Procedure of CLARANS

- Randomly choose k mediod
- Randomly consider the one of mediod swapped with non mediod
- If the cost of new configuration is lower repeat step 2 with new solution
- If the cost higher repeat step 2 with different non mediod object unless limit has been reached
- Compare the solution keep the best
- Return step 1 unless limit has been reached (set to the value of 2).[5]

Density Based Clustering: These algorithm group the data object according density objective function. Densities define number of object in particular neighbors of data object. In density based clustering technique the given cluster continues growing as long as the number of object in neighbors exceeds some parameter. [5]Density based clustering algorithm capable to discover the cluster of arbitrary shaped ,it provide the protection against the noise, outlier.[11] It has also good scalability.[7] It cannot cluster data sets well when there is a large difference in densities, since the minPts-ε combination cannot then be chosen appropriately for all clusters. The Most Efficient Algorithms Proposed under This Method are

2.2.1 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

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

number of object it should have.[14] 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. [17] The complexity of DBSCAN is O (N^2). 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. [14]

Procedure of DBSCAN algorithm is

- Arbitrary select a point r.
- Retrieve all points density-reachable from r w.r.t Eps and Minpts.
- If r is a core point, cluster is formed.
- If r is a border point, no points are density-reachable from r and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.[14]

2.2.2 DENCLUE (Density Based Clustering)

DENCLUE is proposed by Hinneburg & kein in 1998.DENCLUE is described by two factor influence, density function. Influences describe the effect of data point on neighbors. Density function which is the sum of Influence of all data point. According DENCLUE two type of cluster can be defined center defined; multicenter define. Clustering can be determined mathematically by identify density attractor. Whereas attractor is local maxima of overall densities function. Square wave function is used multicenter define cluster. Here we use two parameter σ =eps and ξ =MinPts. [14] DENCLUE has a solid mathematic foundation, it can handle outlier it use grid cell and keep information about one that do actually contain object And it also allow mathematically description of arbitrary shaped cluster in high dimension dataset.[5]

2.2.3 OPTIC (Ordering Point to Identify the Clustering Point)

It is proposed by Ankerst in 1999. The basic idea of optic is similar as DBSCAN but it is address one of the major weaknesses of DBSCAN detecting the meaning cluster of data of varying density. OPTIC use point of database of liner order so the point which are spatially closest become neighbors in ordering .additionally special distance is started for each point that represent the density. It requires the cutoff density of cluster that is no longer. The parameter e is not necessary in optic. [15]

DBSCAN and OPTIC are similar structure so they have similar computational complexity O (n log n). OPTIC detecting the meaningful cluster of data varying density .which cannot be detected by DBSCAN. [5]

2.3 Hierarchical method

Hierarchical method create hierarchical decomposition of object .they are two form Agglomerative (bottom up), divisive (top down) Agglomerative method each object initially presents a cluster of its own. Then cluster successfully merged until the desired cluster structure is obtained .Divisive method all object initially belong one cluster then cluster are divided into subcluster and subcluster are divided into their own subcluster. The result is hierarchical cluster is obtained dendogram .a cluster of data object is obtained by cutting the dendogram at desired level. [7] The major weakness of hierarchical method is that the selection of merge or split point once done cannot be undone. This problem also effect scalability of clustering. [11]. The Most Efficient Algorithms Proposed under Hierarchical Method is:

2.3.1 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

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

scalability and compressed data may improved the performance of hierarchical algorithm.[7].

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.

2.3.2 CURE (Clustering Using Representation)

CURE is proposed by Guha in 1998. Cures represent the cluster by certain number of point and then points are shrinking toward the cluster centriod. It uses the random sampling and partition clustering to handle the large database. [7]CURE work with numerical attribute with (low dimensional data)it represent the cluster by fix number of point scatter around it . [11] It is a reliable method for arbitrary shaped cluster and also takes care of outlier at level assignment stage and also insensitive to outlier. It also covers non spherical shapes. But it is sensitive to parameter such as number of representative object, shrinking factor, number of partition

Procedure of CURE Algorithm

- Draw the random sample from data set Partition this sample into equal size group each of cluster size is $\frac{n'}{p'}$ where n' is size of sample
- Cluster the point of each group we perform initial clustering until each partition has $\frac{n'}{p'}$ >1 cluster
- Eliminate outlier is two phase process first cluster are being formed until the number of cluster are being formed second if the outlier are sampled together during sampling phase
- In it point shrinking toward center replace by other point closest to center by shrinking factor
- Represent the data with corresponding label 1.[5]

CHAMELEON: Chameleon is a hierarchical clustering algorithm. It is developed by Karypis in 1999.in chameleon cluster similarity is obtained based on how well connected object with in cluster. Two clusters are merged if their interconnectivity is high and they are close together. Chameleon has a greater power for discovering the arbitrary shaped cluster of high quality than several well-known algorithms such as BIRCH and density based DBSCAN. The computational complexity of chameleon is O (n^2) time for n objects in the worst case.[18]

2.4 Grid Based Clustering

Grid based technique divide the data space into finite number of cell and then clustering method is applied in these segment. The advantage of grid based method is fast processing time which depends upon the number of cell in each dimension in quantized space. [7]

The Most Efficient Algorithms Proposed under Grid Based Method is

2.4.1 STING (Statical Information Grid Based Method)

STING is developed by wang Wei, joing yang, Richard mountz in 1997. It divides the spatial area into rectangular cell using hierarchical cell. STING goes through the dataset and compute statical parameter such as (mean, variance, and minimum, maximum)and each numerical feature of object within the cell then STING generate hierarchical structure of grid so it represent hierarchical clustering information at different level .[7]

In it as user so higher in the structure static are being summarized from lower level.[5] There are several level of such rectangular cell according different level of resolution .each cell is at higher level is position to form child cell at lower level. A cell in level I corresponding of union of children it. Each cell (expect the leave) 4 child each. Child corresponding to one parent level. Each cell there is independent and dependent parameter. Irrelevant cell are removed and process is repeated until bottom layer is reached. [16] STING is highly scalable. It

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

handle the shape of cluster have vertically and horizontally boundaries and New object are easily inserted. When merging the grid cell from a cluster children are not properly merged because they are corresponding to dedicate parent. [5]

2.4.2 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.[17]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. [17]

Procedure of wave Cluster

- 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.[16]

3. COMPARATIVE STUDIES

Clustering is challenging tasks of data mining than classification .large number of clustering algorithm is proposed tell now. Each algorithm handle some specific issue .single algorithm cannot fulfill the all requirement .so it is difficult to choose single algorithm for specific purpose. A comparative study of different clustering algorithm proposed under this method. So user can choose particular algorithm according their requirement.

Table 3.1 Show the Comparative Study of Various Clustering Algorithm

Sr. No	Name	Proposed by	Year	Type of data	Data set	Cluster shape	Input parameter	Outlier Handling	Complexity
1	k-mean		1995	numerical	Large	Spherical	No of cluster	No, Detect outlier	O(IKn)
2	PAM	Kaufman& Rousseuw	1990	numerical	Small	Arbitrary	No of cluster	No, Detect outlier	$O(IK(n-k)^2$
3	CLARA	Kaufman& Rousseuw	1990	numerical	Sample	Arbitrary	No of cluster	No, Detect outlier	$O(ks^2+k(n-k))$ s is a size of sample
4	CLARANS	Ng Raymond T. & Jiawei Han	1994	numerical	Sample	Arbitrary	No of cluster	No, Detect outlier	$O(n^{2)}$
5	DBSCAN	Martin Ester, Hans-Peter Kriegel & XiaoweiXu	1996	numerical	High Dimensional	Arbitrary	a) radius b) minimum points	Yes	O(n log n)
6	DENCLUE	Hinneburg & Keim	1998	numerical	High Dimensional	Arbitrary	density parameter, noise threshold	Yes	$O(n^2)$

7	OPTICS	Ankerst	1999	numerical	High Dimensional	Arbitrary	density threshold	Yes	O(n log n)
8	BIRCH	Zhang, Ramakrishnan & Linvy	1996	numerical	Large	Spherical	branching factor B, threshold T (max. diameter of sub cluster)	Yes	O(n)
9	CURE	Guha	1998	numerical	Two Dimensional	Arbitrary	Min: Similarity	Yes	$O(n^2 \log n)$
10	CHAME LEON	Karypis	1999	Discrete	Small	Arbitrary	Min. Similarity	Yes	$O(n^2)$
11	STING	Wang Wei, Jiong Yang & Richard Muntz	1997	numerical	Any size	Rectangular	Statical	Yes	O(k)
12	Wave Cluster	Sheikholeslami, Gholamhosein, Surojit Chatterjee & Aidong Zhang	1998	numerical	Large	Arbitrary	No	Yes	O(n)

4. 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] Yujie Zheng,"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] LiorRokach,OdedMaimon,"CLUSTERING METHODS", DATA MINING AND KNOWLEDGEDISCOVERYHANDBOOK.

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

[6] M.D. Boomija,"COMPARISON OF PARTITION BASED CLUSTERING ALGORITHMS", Vol – 1, No.4, pp. 18-21, Oct – Dec 2008.

[7] MARIA HALKIDI, YANNIS BATISTAKIS, MICHALIS VAZIRGIANNIS," On Clustering Validation Techniques", Journal of Intelligent Information Systems, 17:2/3, pp 107–145, 2001

[9] **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**

[10] **Deepak Soni, Asst. Prof Naveen Jha, Deepak Sinwar," Discovery of Outlier from Database using different Clustering Algorithms", Vol. 1, No. 6,pp 388-391, (Sep 2012).**

[11] **S.Vijayarani, S.Nithya," An Efficient Clustering Algorithm for Outlier Detection", International Journal of Computer Applications, Volume 32– No.7, pp22-27, October 2011**

[12] **Pavel Berkhin," Survey of Clustering Data Mining Techniques". David Breitkreutz, Kate Casey," Clusterers: a Comparison of Partitioning and Density-Based Algorithms and a Discussion of Optimisations".**

[13] **Periklis Andritsos," Data Clustering Techniques", pp 1-34, March 11, 2002.**

[14] **Pooja Batra Nagpal,Priyanka Ahlawat Mann," Comparative Study of Density based Clustering Algorithms", International Journal of Computer Applications, Volume 27– No.11,pp 44-47, August 2011.**

[15] **Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining", Vol. 2, Issue 3, pp.1379-1384, May-Jun 2012.**

[16] **MR ILANGO, Dr V MOHAN," A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, Vol. 2(8), pp 3441-3446, 2010.**

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