A COMPREHENSIVE STUDY ON CLUSTERING TECHNIQUES IN DATA MINING

[1]Aravindhan,[2]Dr.D.Maruthanayagam

[1]*Research Scholar, Periyar University, Salem, Tamilnadu. India*
[2]*Assistant Professor, Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri,Tamilnadu, India.*

***Abstract*--Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. The goal of clustering is descriptive, that of classification is predictive. Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. It is widely used in biological and medical applications, computer vision, robotics, geographical data, and so on. To date, many clustering algorithms have been developed. All clustering algorithms basically can be categorized into two broad categories: partitioning and hierarchical based on the properties of generated clusters. Different algorithms proposed may follows a good features of the different methodology and thus it is difficult to categorize them with the solid boundary. The intension of this paper is to provide a categorization of some well known clustering algorithms. It also describes the clustering process and overview of the different clustering methods.**

I. INTRODUCTION

Data clustering is a method of grouping similar objects together. Thus the similar objects are clustered in the same group and dissimilar objects are clustered in different ones. An illustration example of clustering is shown in Figure 1 and 2. Data clustering is considered as an unsupervised learning technique in which objects are grouped in unknown predefined clusters. On the contrary, classification is a supervised learning in which objects are assigned to predefined classes (clusters).

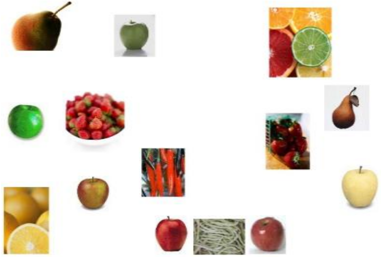

Figure 1: Clustering example -dataset contains 14 objects.

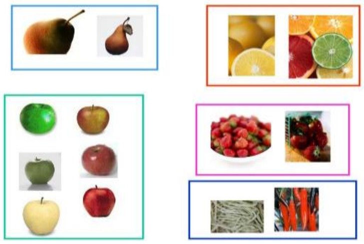

Figure 2: Clustering example - Objects are grouped into 5 clusters.

The problem of data clustering can be formulated as follows: given a dataset *D* that contains *n* objects ***x1***,***x2***,…,***xn*** (data points, records, instances, patterns, observations, items) and each data point is in a *d*-dimensional space, i.e. each 3 data point has *d* dimensions (attributes, features, variables, components). This can be expressed in a matrix format as:

$$D = \begin{pmatrix} X11\ldots\ldots X1d \\ X21\ldots\ldots X2d \\ \ldots\ldots\ldots\ldots \\ Xn1\ldots\ldots Xnd \end{pmatrix}$$

Data clustering is based on the similarity or dissimilarity (distance) measures between data points. Hence, these measures make the cluster analysis is meaningful [1]. The high quality of clustering is to obtain high intra-cluster similarity and low inter-cluster similarity as shown in Figure 3. In addition, when we use the dissimilarity (distance) concept, the latter sentence becomes: the high quality of clustering is to obtain low intra-cluster dissimilarity and high inter-cluster dissimilarity.

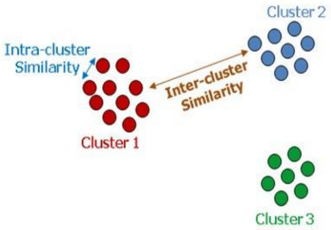

Figure 3: Inter-cluster and Intra-cluster similarities of clusters.

II. CLUSTERING PROCESS

The overall process of cluster analysis is shown in Figure 4. It involves four basic steps as explained below.

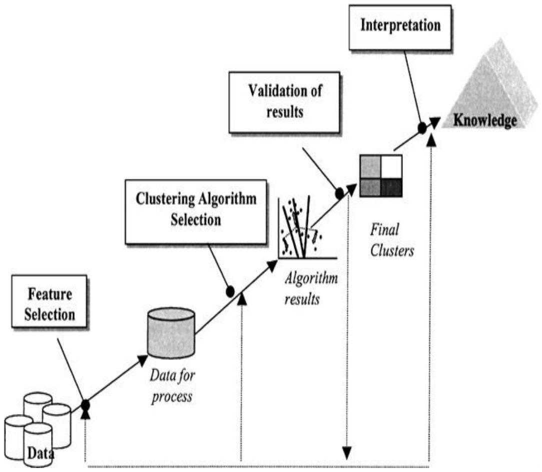

Figure 4: Steps of clustering process.

The clustering process may result in different partitioning of a data set, depending on the specific criterion used for clustering. Thus, there is a need of preprocessing before we assume a clustering task in a data set. The basic steps to develop clustering process are presented in figure 4 and can be summarized as follows:

- ***Feature selection:*** The goal is to select properly the features on which clustering is to be performed so as to encode as much information as possible concerning the task of our interest. Thus, preprocessing of data may be necessary prior to their utilization in clustering task.
- ***Clustering algorithm.*** This step refers to the choice of an algorithm those results in the definition of a good clustering scheme for a data set. Aproximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set.
    i. ***Proximity measure*** is a measure that quantifies how “similar” two data points (i.e. feature vectors) are. In most of the cases we have to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others.
    ii. ***Clustering criterion:*** In this step, we have to define the clustering criterion, which can be expressed via a cost function or some other type of rules. We should stress that we have to take into account the type of clusters that are expected to occur in the data set. Thus, we may define a “good” clustering criterion, leading to a partitioning that fits well the data set.
- ***Validation of the results:*** The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering methods, the final partition of data requires some kind of evaluation in most applications.
- ***Interpretation of the result:*** In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.

III. CATEGORIZATION OF CLUSTERING METHODS

There is some difference between clustering method and clustering algorithm [2]. A clustering method is a general strategy applied to solve a clustering problem, whereas a clustering algorithm is simply an instance of a method. As mentioned earlier no algorithm exist to satisfy all the requirements of clustering and therefore large numbers of clustering methods proposed till date, each with a particular intension like application or data types or to fulfill a specific requirement. All clustering algorithms basically can be categorized into two broad categories: partitioning and hierarchical, based on the properties of generated clusters [2] [3]. Different algorithms proposed may follows a good features of the different methodology and thus it is difficult to categorize them with the solid boundary. Though we have tried to provide as much clarity as possible, there is still a scope of variation. The overview of each categorization is discussed below.

A. Hierarchical Methods

As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition can be represented by a tree structure diagram called as a *dendrogram*; whose root node represents the whole dataset and each leaf node is a single object of the dataset. The clustering results can be obtained by cutting the dendrogram at different level. There are two general approaches for the hierarchical method: **agglomerative** (bottom-up) and **divisive** (top down) [4] [5].

- **Agglomerative method** starts with n leaf nodes(n clusters) that is by considering each object in the dataset as a single node(cluster) and in successive steps apply merge operation to reach to root node, which is a cluster containing all data objects. The merge operation is based on the distance between two clusters. There are three different notions of distance: single link, average link, complete link.
- **Divisive method** is opposite to agglomerative, starts with a root node that is considering all data objects into a single cluster, and in successive steps tries to divide the dataset until reaches to a leaf node containing a single object. For a dataset having n objects there is 2n-1 – 1 possible two-subset divisions, which is very expensive in computation. Two divisive clustering algorithms, **DIANA** and **MONA** [6] [7].

B. Partitioning Methods

As the name suggest, the partitioning methods, in general creates k partitions of the datasets with n objects, each partition represent a cluster, where k<= n. It tries to divide the data into subset or partition based on some evaluation criteria. As checking of all possible partition is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization [8].

- **Relocation based:** One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found, can be known as a probabilistic models or simply model based clustering. Here, a model assumes that the data comes from a mixture of several populations whose distributions and priors we want to find. The representative algorithms are **EM**, **SNOB**, **AUTOCLASS** and **MCLUST** [6] [5]. The other approach to partition is based on the objective function, in which the instead of pair-wise computations of the proximity measures, unique cluster representatives are constructed. Depending on how representatives are constructed iterative partitioning algorithms are divided into **k-means** and **k-mediods** [6] [7]. The partitioning algorithm in which each cluster is represented by the gravity of the centre is known as k-means algorithms. The one most efficient algorithm proposed under this scheme is named as k-means only. From the invention of k-means to till date large number of variations had been proposed, some of them can be listed as, ISODATA, Forgy, bisecting k-means, x-means, kernel k-means and so on[5][6]. The partitioning algorithm in which cluster is represented by one of the objects located near its centre is called as a k-mediods. PAM, CLARA and CLARANS are three main algorithms proposed under the k-mediod method [5].
- **Grid Based:** As the name suggest, grid based clustering methods uses a multidimensional grid data structure. It divides the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The representative algorithms based on this method are: **STING**, **Wave Cluster**, and **CLIQUE** [9].
- **Density Based:** This method has been developed based on the notion of density that is the no of objects in the given cluster, in this context. The general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold; 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. The density bases algorithms can further classified as: density based on connectivity of points and based on density function. The main representative algorithms in the former are DBSCAN and its extensions, OPTICS, DBCLASD, whereas under the latter category are DENCLUE and SNN [6][9].

IV. HIERARCHICAL METHODS

The hierarchical methods group training data into a tree of clusters. This tree structure called dendrogram (Figure 5). It represents a sequence of nested cluster which constructed top-down or bottom-up. The root of the tree represents one cluster, containing all data points, while at the leaves of the tree, there are n clusters, each containing one data point. By cutting the tree at a desired level, a clustering of the data points into disjoint groups is obtained [10].

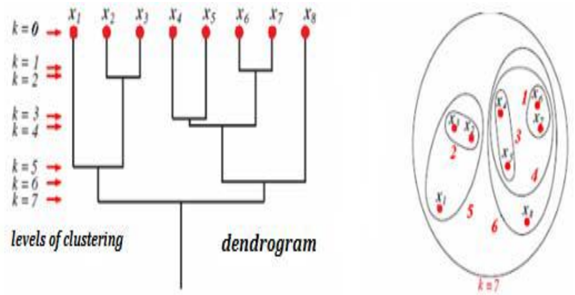

Figure 5: Tree structure of training data (dendrogram)[13].

Hierarchical clustering algorithms divide into two categories: Agglomerative and Divisive. Agglomerative clustering executes in a bottom–top fashion, which initially treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single remaining cluster. Divisive clustering, on the other hand, initially treats all the data points in one cluster and then split them gradually until the desired number of clusters is obtained. To be specific, two major steps are in order. The first one is to choose a suitable cluster to split and the second one is to determine how to split the selected cluster into two new clusters [11]. Figure 5 shows structure of clustering algorithms. Many agglomerative clustering algorithms have been proposed, such as **CURE**, **ROCK**, **CHAMELEON**, **BIRCH**, single-link, complete-link, average-link, and Leaders-Sub leaders. One representative divisive clustering algorithm is the bisecting k-means method. Figure 6 is a representation of Hierarchical clustering schemes. Agglomerative and Divisive clustering have been applied to document clustering [12]. Divisive clustering is very useful in linguistics, information retrieval, and document clustering applications [11].

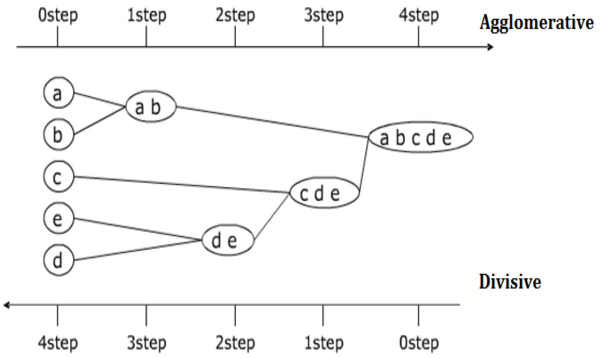

Figure 6: Application of agglomerative and divisive to a data set of five objects, {a, b, c, d, e}[13].

V. PARTITION METHODS

Partitioning techniques divides the object in multiple partitions where single partition describes cluster. The objects with in single clusters are of similar characteristics where the objects of different clusters have dissimilar characteristics in terms of dataset attributes. A distance measure is one of the feature space used to identify similarity or dissimilarity of patterns between the data objects [14]. K-mean, K-medoid are partitioning algorithms [15].

K-MEAN

K-mean algorithm is one of the centroid based technique. It takes input parameter k and partition a set of n object from k clusters. The similarity between clusters is measured in regards to the mean value of the object. The random selection of k object is first step of algorithm which represents cluster mean or center. By comparing most similarity other objects are assigning to the cluster. **Algorithm [16]**: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- K:the number of clusters
- D:a data set containing n object

Output:

- A set of k clusters

Method:

- Arbitrarily choose k objects from D as the initial cluster centers.
- Repeat
- Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
- update the cluster means ,i.e., calculate the mean value of the objects for each cluster;
- Until no change;

K-MEDOID

The k-means method is based on the centroid techniques to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data [17].

To overcome the problem we used K-medoids method which is based on representative object techniques. Medoid is replaced with centroid to represent the cluster. Medoid is the most centrally located data object in a cluster. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. This process is continued until no any medoid move. As a result, k clusters are found representing a set of n data objects [15]. An algorithm for this method is given below

Algorithm [15]: PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:

- K: the number of clusters,
- D: a data set containing n objects.

Outputs:

- A set of k clusters.

Method:

- Arbitrarily choose k objects in D as the initial representative objects or seeds;
- Repeat
- Assign each remaining object to the cluster with the nearest representative object;
- Randomly select a non-representative object, Orandom.
- Compute the total cost of swapping representative object, Oj with Orandom;
- If S<0 then swap Oj with Orandom to form the new set of k representative object;
- Until no change;

VI. GRID BASED METHODS

Grid based methods quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. The main advantage of Grid based method is its fast processing time which depends on number of cells in each dimension in quantized space. In this paper, we present some of the grid based methods such as CLIQUE (CLustering InQUEst) [18], STING (STatistical INformation Grid) [19], MAFIA (Merging of Adaptive Intervals Approach to Spatial Data Mining) [20], Wave Cluster [21],to improve the processing time of objects.

CLIQUE (CLUSTERING IN QUEST)

It makes use of concepts of density and grid based methods. In the first step, CLIQUE partitions the dimensional data space S into non overlapping rectangular units (grids) [18]. The units are obtained by partitioning every dimension into ξ intervals of equal length. ξ is an input parameter, selectivity of a unit is defined as the total data points contained in it. A unit u is dense if selectivity (u) is greater than γ, where the density threshold γ is another input parameter. A unit in the subspace is the intersection of an interval from each of the K attributes. A cluster is a maximal set of connected dense units. Two K-dimensional unit's u1, u2 are connected if they have a common face. The dense units are then connected to form clusters. It uses apriori algorithm (bottom up algorithm) to find dense units. The dense units are identified by using a fact that if a K dimension unit (a1, b1)* (a2, b2) …….(ak ,bk) is dense, then any k-1 dimension unit (a1,b1) * (a2,b2)….(aik-1,bik-1) is also dense where (ai, bi) is the interval of the unit in the i th dimension.

Given a set of data points and the input parameters ξ and γ CLIQUE is able to find clusters in all subspaces of the original data space and present a minimal description of each cluster in the form of a DNF expression. Steps involved in CLIQUE is i) identification of sub spaces (dense Units) that contain cluster ii) merging of dense units to form cluster & iii) Generation of minimal description for the clusters.

STING :(A STATISTICAL INFORMATIONS GRID APPROACH TO SPATIAL DATA MINING).

Spatial data mining is the extraction of implicit knowledge, spatial relation and discovery of interesting characteristics and patterns that are not explicitly represented in the databases. (Spatial data mining has wide applications in many fields, including GIS system, image data base exploration, medical imaging etc). STING is a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells(using latitude and longitude) and employs a hierarchical structure [19].There are usually several levels of such rectangular cells corresponding to different levels of resolution. Each cell at a high level is partitioned to form child cells at lower level. A cell in level i corresponds to union of its children at level i + 1. Each cell (except the leaves) has 4 children & each child corresponds to one quadrant of the parent cell. Statistical information regarding the attributes in each grid cell (such as, mean, Standard Deviation maximum & minimum values) is pre computed and stored. Statistical parameters of higher level cells can easily be computed from the parameters of lower level cells. For each cell, there are attribute independent parameters and attribute dependant parameters.

i. Attribute independent parameter: count

ii. Attribute dependant parameters

M- Mean of all values in this cell; S- Standard deviation of all values in this cell

Min – minimum value of the attribute in this cell; Max – minimum value of the attribute in this cell; Distribution –Type of distribution the attribute value follows. Distribution types are normal, uniform exponential & none. Value of distribution may either be assigned by the user or obtained by hypothesis tests such as X2 test. When the data are loaded into the database, the parameters count, m, s, min, max of the bottom level cells are calculated directly from the data. First, a layer is determined from which the query processing process is to start. This layer may consist of small number of cells. For each cell in this layer we check the relevancy of cell by computing confidence internal. Irrelevant cells are removed and this process is repeated until the bottom layer is reached.

MAFIA: (MERGING OF ADAPTIVE INTERVALS APPROACH TO SPATIAL DATA MINING)

MAFIA proposes adaptive grids for fast subspace clustering and introduces a scalable parallel framework on shared nothing architecture to handle massive data sets [20]. Most of the grid based algorithms uses uniform grids whereas MAFIA uses adaptive grids. MAFIA proposes a technique for adaptive computation of the finite intervals (bins) in each dimension, which are merged to explore clusters in higher dimensions. Adaptive grid size reduces the computation and improves the clustering quality by concentrating on the portions of the data space which have more points and thus likelihood of having clusters. Performance results show MAFIA is 40 to 50 times faster than CLIQUE, due to the use of adaptive grids. MAFIA introduces parallelism to obtain a highly scalable clustering algorithm for large data sets. MAFIA proposes an adaptive interval size to partition the dimension depending on the distribution of data in the dimension. Using a histogram constructed by one pass of the data initially, MAFIA determines the minimum number of bins for a dimension. Contiguous bins with similar histogram values are combined to form larger bins. The bins and cells that have low density of data will be pruned limiting the eligible candidate dense units, thereby reducing the computation. Since the boundaries of the bins will also not be rigid, it delineates cluster boundaries more accurately in each dimension. It improves the quality of the clustering results.

WAVE CLUSTER

Wave Cluster is a multi resolution clustering algorithm. It is used to find clusters in very large spatial databases [21].Given a set of spatial objects Oi, $1\leq i \leq N$, the goal of the algorithm is to detect clusters. It first summarizes the data by imposing a multi dimensional grid structure on to the data space. The main idea is to transform the original feature by applying wavelet transform and then find the dense regions in the new space. A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub bands. The first step of the wavelet cluster algorithm is to quantize the feature space. In the second step, discrete wavelet transform is applied on the quantized feature space and hence new units are generated. Wave cluster connects the components in 2 set of units and they are considered as cluster. Corresponding to each resolution γ of wavelet transform there would be set of clusters cr, where usually at the coarser resolutions number of cluster is less? In the next step wave cluster labels the units in the feature space that are included in the cluster.

VII. DENSITY BASED METHODS

Density based clusters are defined as clusters which are differentiated from other clusters by varying densities that means a group which have dense region of objects may be surrounded by low density regions. Density based method are of two types: Density based Connectivity and Density based Functions [22].Density based Connectivity is related to training data point and DBSCAN [23] and DBCLASD [24] comes under this while Density Functions is related to data points to computing density functions defined over the underlying attribute space and DENCLUE [25] comes under this.

DBSCAN (DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)

DBSCAN[22] (Density Based Spatial Clustering of Applications with Noise) It is of Partitioned type clustering where more dense regions are considered as cluster and low dense regions are called noise. Obviously clusters are define on some criteria which is as follows

✓ ***Core:*** Core points lie in the interior of density based clusters and should lie within Eps (radius or threshold value), MinPts(minimum no of points) which are user specified parameters.

- ✓ ***Border:*** Border point lies within the neighborhood of core point and many core points may share same border point.
- ✓ ***Noise*:** The point which is neither a core point nor a border point
- ✓ ***Directly Density Reachable*:** A point r is directly density reachable from s w.r.t Eps and MinPts if a belongs to NEps(s) and |NEps (s)| >= MinPts
- ✓ ***Density Reachable:*** A point r is density reachable from r point s wrt.Eps and MinPts if there is a sequence of points r1….rn, r1 =s, rn = s such that ri+1 is directly reachable from ri.

Algorithm

Steps of algorithm of DBSCAN are as follows

- 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

DBCLASD (APPLICATION BASED CLUSTERING ALGORITHMS FOR MINING IN LARGE SPATIAL DATABASES)

DBCLASD [24] (Application Based Clustering Algorithms for Mining in Large Spatial Databases) Basically DBCLASD is an incremental approach. A point is assigned to a cluster that processed incrementally without considering the cluster. In DBCLASD cluster is defined by three properties which are as follows:

1) Expected Distribution condition NNDistSet(C) which is set of nearest neighbor of cluster C has the expected distribution with required confidence level.

2) Maximality Condition Every point that comes into neighboring of C does not fulfill condition (1).

3) Connectivity Condition Each pair (a, b) are connected through grid cell structure.

Algorithm

- Make set of candidates using region query
- If distance set of C has expected distribution then point will remain in cluster.
- Otherwise insert point in list of unsuccessful candidates.
- In the same way expand cluster and check condition
- Now list of unsuccessful candidates is again checked via condition.
- If passes then put in cluster otherwise remain in that list

There are two main concepts in DBCLASD. First one is generating candidates and candidate generation is done on the basis of region query that specifies some radius for circle query to accept candidates. Second one is testing the candidates which are done through chi square testing. Points that lie under the threshold value are considered right candidates while those lies above threshold are remain in unsuccessful candidates list. In last unsuccessful candidate list is again checked and every point go through test and points passes test are considered in cluster while left remains in unsuccessful candidates list.

DENCLUE (DENSITY BASED CLUSTERING)

DENCLUE [25] (Density based clustering) Main concepts are used here i.e. influence and density functions. Influence of each data point can be modeled as mathematical function and resulting function is called Influence Function. Influence function describes the impact of data point within its neighborhood. Second factor is Density function which is sum of influence of all data points. According to DENCLUE two types of clusters are defined i.e. centre defined and multi centre defined clusters .In centre defined cluster a density attractor. The influence function of a data objects y Є F is a function. Which is defined in terms of a basic influence function F, *F (x)* =- *F* (x, y).The density function is defined as the sum of the influence functions of all data points. DENCLUE also generalizes other clustering methods such as Density based clustering; partition based clustering, hierarchical clustering. In density based clustering DBSCAN is the example and square wave influence function is used and multicenter defined clusters are here which uses two parameter σ = Eps, c =MinPts. In partition based clustering example of k-means clustering is taken where Gaussian Influence function is discussed. Here in center defined clusters ξ=0 is taken and σ is determined. In hierarchical clustering center defined clusters hierarchy is formed for different value of σ.

Algorithm

- Take Data set in Grid whose each side is of 2σ
- Find highly densed cells i.e.
- Find out the mean of highly populated cells
- If d (mean (cl), mean (c2)) < 4a then two cubes are connected.
- Now highly populated or cubes that are connected to highly populated cells will be considered in determining clusters.
- Find Density Attractors using a Hill Climbing procedure.
- Randomly pick point r.
- Compute Local 4 σ density
- Pick another point (r+1) close to previous computed density.
- If den(r) < den(r+1) climb.
- Put points within (σ /2) of path into cluster.
- Connect the density attractor based cluster.

VIII. CONCLUSION

The clustering problem can be described as dividing a given set of objects into groups so that objects inside a group are more similar to each other than objects belonging to different groups. In this paper we described the process of clustering from the data mining point of view. We gave the properties of a “good” clustering technique and the methods used to find meaningful partitioning. At the same time, we concluded that research has emphasized numerical data sets, and the intricacies of working with large categorical databases are left to a small number of alternative techniques. We claimed that new research solutions are needed for the problem of categorical data clustering, and presented our ideas for future work.

REFERENCES

[1] G. Gan, Ch. Ma, and J. Wu, “Data Clustering: Theory, Algorithms, and Applications,”ASA-SIAM series on Statistics and Applied Probability, SIAM, 2007.

[2] A.K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A Review”, *ACM Computing Surveys*, vol. 31, pp. 264-323, Sep. 1999.

[3] P. Berkhin. (2001) “Survey of Clustering Data Mining Techniques” [Online]. Available: http://www.accure.com/products/rp_cluster_review.pdf.

[4] O. A. Abbas, “Comparisons between Data Clustering Algorithms”, *The Int. Journal of Info. Tech., vol*. 5, pp. 320-325, Jul. 2008.

[5] J. Han, M. Kamber, Data Mining, Morgan Kaufmann Publishers, 2001.

[6] P. Berkhin. (2001) “Survey of Clustering Data Mining Techniques” [Online]. Available: http://www.accure.com/products/rp_cluster_review.pdf.

[7] Rui Xu, Donald C. Wunsch II, “Survey of Clustering Algorithms”, IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005.

[8] A. K. Jain, “Data Clustering: 50 Years Beyond K-Means, in Pattern Recognition Letters, vol. 31 (8), pp. 651-666, 2010.

[9] S.B. Kotsiantis, P. E. Pintelas, “Recent Advances in Clustering: A Brief Survey” WSEAS Transactions on Information Science and Applications, Vol. 1, No. 1, pp. 73–81, Citeseer, 2004.

[10] D.T. Pham and A.A. Afify, Engineering applications of clustering techniques, Intelligent Production Machines and Systems, (2006), 326-331.

[11] L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu, A fast divisive clustering algorithm using an improved discrete particle swarm optimizer, Pattern Recognition Letters, 31 (2010), 1216-1225.

[12] R. Gil-García and A. Pons-Porrata, Dynamic hierarchical algorithms for document clustering,Pattern Recognition Letters, 31 (2010), 469-477.

[13] R. Capaldo and F. Collova, Clustering: A survey, Http://uroutes.blogspot.com, (2008).

[14] A. K. Jain, M. N. Murty, and P. J. Flynn” Data clustering: a review”.ACM Computing Surveys, Vol .31No 3, pp.264–323, 1999.

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

[16] “Data Mining Concept and Techniques”, 2nd Edition, Jiawei Han, By Han Kamber.

[17] Abhishek Patel,”New Approach for K-mean and K-medoids algorithm”, International Journal of Computer Applications Technology and Research, 2013.

[18] Rakesh Agrawal, Johannes Gehrke, Dimirios Gunopulos, Prabhakar Raghavan: Automatic Subspace Clustering of High Dimensional Data.Data Mining and knowledge discovery, 11, 5-33, 2005.Springer Science + Business media, Inc. Manufactured in the Netherlands.

[19] Wei Wang, Jiong Yang, and Richard Muntz : STING : A Statistical Grid Appraoch to Spatial Data Mining : Department of Computer Science, University of California, Los Angels

[20] Sanjay Goil, Harsha Nagesh and Alok Choudhary : MAFIA: Efficient and Scalable Clustering for very large data sets : Technical Report No. CPDC – TR – 9906 – 010 ©1999 Center for Parallel and distributed Computing. June 1999

[21] Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang: WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases.

[22] A.K. Jain and R. C. Dubes, Algorithms for Clustering Data.Englewood Cliffs, NJ: Prentice-Hall, 1988.

[23] Ester M. Kriegel H.-P., Xu X.: “Knowledge Discovery in Large Spatial Databases: Focusing Techniques for efficient Class Identification”, Proc. 4th Int. Symp. on large Spatial Databases, Portland, ME, 1995, in: Lecture Notes In Computer Science, Vol. 951, Springer, 1995, pp. 67-82.

[24] XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J.1998.A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th ICDE, 324-331, Orlando, FL.[10]A. K. Jain, M. N. Murty and P. J.Flynn, Data clustering: a review, CM, 31 (1999), pp. 264–323.

[25] A. Hinneburg and D. Keim, “An efficient approach to clustering Large multimedia databases with noise,” in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining(KDD‟98), 1998, pp. 58–65.

AUTHORS PROFILE

S.Aravindhan received his M.Phil degree from Thiruvalluvar university,Vellore in the year 2012.He has received his MCA degree from Anna university,Chennai in the year 2011.He is pursuing his Ph.D degree at Periyar University, Salem, Tamilnadu, India. His areas of interest include Data Mining, Cloud Computing and Computer Networks.

Dr.D.Maruthanayagam received his Ph.D Degree from Manonmanium Sundaranar University, Tirunelveli in the year 2014. He has received his M.Phil, Degree from Bharathidasan University, Trichy in the year 2005. He has received his M.C.A Degree from Madras University, Chennai in the year 2000. He is working as Assistant Professor, Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri, Tamilnadu, India. He has 14 years of experience in academic field. He has published 1 book, 12 International Journal papers and 21 papers in National and International Conferences. His areas of interest include Grid Computing, Cloud Computing and Mobile Computing.