*IOSR Journal of Computer Engineering (IOSRJCE)*
*ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 01-12*
*www.iosrjournals.org*

# Literature Survey On Clustering Techniques

B.G.Obula Reddy1, Dr. Maligela Ussenaiah2
*1Associate Professor, Lakireddy Balireddy College of Engineering L.B.Reddy Nagar,Mylavaram, Krishna (Dist):521 230*
*2Assistant Professor, Computer Science, Vikrama Simhapuri University Nellore Nellore, Andhra Pradesh, India*

***Abstract:*** *Clustering is the assignment of data objects (records) into groups (called clusters) so that data objects from the same cluster are more similar to each other than objects from different clusters. Clustering techniques have been discussed extensively in similarity search, Segmentation, Statistics ,Machine Learning ,Trend Analysis, Pattern Recognition and classification. Clustering methods can be classified in to i)partition methods2)Hierarchical methods,3)Density Based methods 4)Grid based methods5)Model Based methods. in this paper, i would like to give review about clustering methods by taking some example for each classification. I am also providing comparative statement by taking constraints i.e Data type, Cluster Shape, Complexity, Data Set, Measure, Advantages and Disadvantages.*

***Keywords****: clustering; Partition, Hierarchical, Density, grid, Model*

## I. Introduction

Cluster analysis is unsupervised learning method that constitutes a cornerstone of an intelligent data analysis process. It is useful for the exploration of inter-relationships among a collection of patterns, by organizing into homogeneous clusters. It is called as unsupervised learning because no a priori labeling of some patterns is available to use in categorizing others and inferring the cluster structure of the whole data. Intra-connectivity is a measure of the density. A high intra-connectivity means a good clustering arrangement because the instances grouped within the same cluster are highly dependent on each other. an Inter-connectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is advantageous because it indicates that individual clusters are largely independent of each other.

Every instance in the data set can be represented using the same set of attributes. The attributes are categorical. To stimulate a hypothesis from a given data set, a learning system requires to make assumptions about the hypothesis to be learned. These assumptions are called as biases. Since every learning algorithm uses some biases, it reacts well in some domains where its biases are appropriate while it performs poorly in other domains. The problem with clustering methods is that the interpretation of the clusters may be difficult.The algorithms will always assign the data to clusters even if there were no clusters in the data.

Cluster analysis is a difficult problem because many factors 1. effective similarity measures, 2.criterion functions, 3. algorithms are come into play in devising a well tuned clustering technique for a given clustering problems. Moreover, it is well known that no clustering method can adequately handle all sorts of cluster structures i.e shape, size and density. Sometimes the quality of the clusters that are found can be improved by preprocessing the given data. It is not uncommon to try to find noisy values and eliminate them by a preprocessing step. A common technique is to use postprocessing steps to try to fix up the clusters that have been found.

Clustering is not a recent invention, nor is its relevance to computational toxicology a recent application. Its theory, however, is often lost within black box treatments used by QSAR programs. Clustering, in the general sense, is the grouping of objects together based on their similarity, while excluding objects which are dissimilar .one of the first application of cluster analysis to drug discovery was by Harrison, who asserted that locales exist within the chemical space which favor biological activity. Consequently, these localities form clusters of structurally similar compounds. This idea that structure confers activity is also the fundamental premise of all QSAR analyses. The basic framework for compound clustering consists of three main steps: the computation of structural features, the selection of a difference metric, and the application of the clustering algorithm.

Generally clustering algorithms can be categorized into hierarchical clustering methods, partitioning clustering methods, density-based clustering methods, grid-based clustering methods, and model-based clustering methods. An excellent survey of clustering techniques can be found in (Jain et al., 1999). Section 1 deals with hierarchical methods by taking examples as BIRCH, CURE and CHAMELEON, section 2 deals with partitioning methods by taking examples as K-means Clustering, k-Medoids, Section3 deals with Density-based Clustering by taking examples as DBSCAN,OPTICS,DENCLUE, Section 4 deals with grid-based methods by taking examples as CLIQUE,STING,MAFIA,WAVE CLUSTER,O-CLUSTER,ASGC,AFR, Section 5 deals with model-based methods by taking examples as RBMN,SOM and Ensembles of Clustering Algorithms .

## II. Hierarchical methods:

A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical manner. The hierarchical methods group the data instances into a tree of clusters. There are two major methods are available under this category i.e agglomerative method, which forms the clusters in a bottom-up fashion until all data instances belong to the same cluster, divisive method, in which splits up the data set into smaller cluster in a top-down fashion until each of cluster contains only one instance. Both the divisive algorithms and agglomerative algorithms can be represented by dendrograms.

Hierarchical clustering techniques use various constraints to decide locally at each step which clusters should be joined . For agglomerative hierarchical techniques, the principle is typically to merge the "closest" pair of clusters, where "close" is defined by a specified measure of cluster closeness. There are three definitions of the closeness between two clusters i.e single link, complete link and average link. The single link similarity of two clusters is the similarity between the two most similar instances. Single link is good for handling non elliptical shapes. The complete link is less susceptible to noise and outliers and has trouble with convex shapes. The following are some of the hierarchical clustering algorithms are: Balanced Iterative Reducing and Clustering using Hierarchies – BIRCH, Clustering Using REpresentatives – CURE and CHAMELEON.

### 2.1 CURE:-

CURE is a hierarchical clustering algorithm, that employs the features of both the centroid based algorithms and the all point algorithms .CURE[7] obtains a data sample from the given database. The algorithm divides the data sample into groups and identifies some representative points from each group of the data sample. In the first phase, the algorithm considers a set of widely spaced points from the given datasets. In the next phase of the algorithm the selected dispersed points are moved towards the centre of the cluster by a specified value of a factor α. As a result of this process, some randomly shaped clusters are obtained from the datasets. In the process it identifies and eliminates outliers. In the next phase of the algorithm, the representative points of the clusters are checked for proximity with a threshold value and the clusters that are next to each other are grouped together to form the next set of clusters. In this hierarchical algorithm, the value of the factor α may vary between 0 and 1. The utilization of the shrinking factor alpha by the CURE overcomes the limitations of the centroid based, all-points approaches. As the representative points are moved through the clustering space, the ill effects of outliers are reduced by a greater extent. Thus the feasibility of CURE is enhanced by the shrinking factor α. The worst case time complexity of CURE is determined to be o(n2logn).

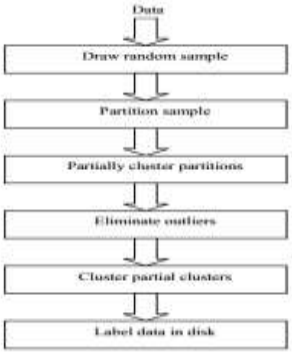



Figure overview of **CURE** implementation

A random sample of data objects is drawn from the given datasets. Partial clusters are obtained by partitioning the sample dataset and outliers are identified and removed in this stage. Final refined clusters are formed from the partial cluster set.

### 2.2 BIRCH:-

The clustering algorithm BIRCH is a main memory based algorithm, i.e., the clustering process is carried out with a memory constraint. BIRCH's incremental clustering is based on the concept of clustering feature and CF tree. A clustering feature is a triple that contains the summary of the information of each cluster. Given N d-dimensional points or objects in a cluster:{xt} where i=1,2, …, N, the Clustering feature (CF) as a vector of the cluster can be stated as,

CF=(N.LS,SS)

where N is the number of points in the cluster, LS is the linear sum on N points, i.e., $\sum_{i=1}^{N} \bar{X}_i$ , and SS is the square sum of the data points. i.e.. $\sum_{i=1}^{N} \bar{X}_i^2$

A clustering feature tree (CF tree) contains the CFs that holds the summary of clusters. A CF tree is a height balanced tree that has two parameters namely, a branching factor, B, and threshold, T. The representation of a non-leaf node can be stated as $\{CF_i, child_i\}$, where,

i = 1,2, …., B,

$child_i$ - A pointer to its ith child node

$CF_i$ - $CF$ of the subcluster represented by the I th child

The non-leaf node is provides a representation for a cluster and the contents of the node represents all of the subclusters. In the same manner a leaf-node's contents represents all of its subclusters and has to confirm to a threshold value for T. The BIRCH clustering algorithm is implemented in four phases. In phase1, the initial CF is built from the database based on the branching factor B and the threshold value T. Phase2 is an optional phase in which the initial CF tree would be educed in size to obtain a smaller CF tree. Global clustering of the data points is performed in phase3 from either the initial CF tree or the smaller tree of phase2. As has been shown in the evaluation good clusters can be obtained from phase3 of the algorithm. If it is required to improve the quality of the clusters, phase4 of the algorithm would be needed in the clustering process. The execution of Phase1 of BIRCH begins with a threshold value T. The procedure reads the entire set of data points in this phase, selects the data points based on a distance function. The selected data points are stored in the nodes of the CF tree. The data points that are closely spaced are considered to be clusters and are those selected. The data points that are widely placed are considered to be outliers and thus are discarded from clustering.In this clustering process, if the threshold limit is exceeded before the complete scan of the database, the value is increased and a much smaller tree with all the chosen data points is built. An optimum value for threshold T is necessary in order to get good quality clusters from the algorithm. If it is required to fine tune the quality of the clusters, further scans of the database is recommended through phase4 of the algorithm. Time complexity(Worst case) of this algorithm is O(n). The time needed for the execution of the algorithm varies linearly to the dataset size.

**2.3 CHAMELEON:-**

In agglomerative hierarchical approaches the major disadvantage is that they are based on a static, user specified inter connectivity model,either under estimates or over estimates the inter connectivity of objects and clusters. This type limitation is overcome by the algorithm CHAMELEON. CHAMELEON makes use of a sparse graph, where the nodes represent data objects; weights in the edges represent similarities between the data objects. CHAMELEON's sparse graph implementation lets it to scale to large databases in an effective manner. This implementation of sparse graph is based on the frequently used k-nearest neighbor graph representation. CHAMELEON identifies the similarity between a pair of clusters named as Ci and Cj by evaluating their relative interconnectivity RI (Ci , Cj ) and relative closeness RC (Ci, Cj ). When the values of both RI (Ci , Cj ) and RC (Ci , Cj) are high for any two of clusters, CHAMELEON's agglomerative algorithm merges those two clusters.

$$RI(C_i, C_j) = \frac{\left|EC_{\{C_i,C_j\}}\right|}{\frac{\left|EC_{C_i}\right|+\left|EC_{C_j}\right|}{2}}$$

where,

$\left|EC_{\{C_i,C_j\}}\right|$ - edge-cut of cluster containing both Ci and Cj

$\left|EC_{C_i}\right|$ - min-cut bisector indicating internal interconnectivity of cluster C

$\left|EC_{C_j}\right|$ - min-cut bisector indicating internal interconnectivity of cluster Cj

The relative closeness of two clusters Ci and Cjis stated as follows:

$$RC(C_i, C_j) = \frac{\bar{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}}$$

where,

$\bar{S}_{EC_{C_i}}$ - average edge weight of min-cut bisector of cluster Ci

$\bar{S}_{EC_{C_j}}$ - average edge weight of min-cut bisector of cluster C

$\bar{S}_{EC_{\{C_i,C_j\}}}$ - average edge weight edges connecting vertices of cluster Ci with that of cluster Cj

CHAMELEON agglomerative hierarchical approach implements the algorithm in two separate phases. In the first phase, dynamic modeling of the data objects is done by clustering these objects into subclusters. In the second phase, a dynamic modeling framework is employed on the data objects to merge the subclusters in a

hierarchical manner to get good quality cluster. The dynamic framework model can be implemented by two different methods. In the first method it is checked that the values of relative inter-connectivity and relative closeness between a pair of cluster cross a user-specified threshold value. For this purpose, these two parameters should satisfy the following conditions:

$$1) RI(C_i, C_j) \geq T_{RI}$$
$$2) RC(C_i, C_j) \geq T_{RC}$$

In the second method, CHAMELEON chooses a pair of clusters that maximizes a function that is given by,

$$RI(C_i, C_j) * RC(C_i, C_j)^{\alpha}$$

where α is user-specified parameter that takes the values between 0 and 1.

## III. Partitioning Methods:

Partitioning methods are divided into two subcategories, one is centroid and other is medoids algorithms.Centroid algorithms represent eachcluster by using the gravity centre of the instances. The medoid algorithms represents each cluster by means of the instances closest to gravity centre. The well-known centroid algorithm is the k-means. The k-means method partitions the data set into k subsets such that all points in a given subset are closest to the same centre.

In detail, it randomly selects k of the instances to represent the clusters. Based up on the selected attributes, the remaining instances are assigned to their closer centers. K-means then computes the new centers by taking the mean of all data points belonging to the same cluster. The process is iterated until there is no change in the gravity centers. If k cannot beknown ahead of time, various values of k can bee valuated until the most suitable is found. The effectiveness of this method as well as of others relies heavily on the objective function used in measuring the distance between instances.

In detail, it randomly selects k of the instances to represent the clusters. Based on selected attributes, all remaining instances are assigned to their closer centre. K-means then computes new centers by taking the mean of all data points belonging to the same cluster. The process is iterated until there is no change in gravity centers. If k cannot be known ahead of time, various values of k can bee valuated until the most suitable one is found. The effectiveness of this method and others relies heavily on the objective function used in measuring the distance between instances.

The difficulty in finding a distance measure,it works well with all types of data. There are several procedures to define the distance between instances .Generally, the k-means algorithm has the following important characteristics : 1. It is efficient in processing large data sets, 2. It terminates at a local optimum, 3. The clusters having spherical in shapes,4. It is sensitive to noise. However, there are variants of the k-means clustering procedure, which gets around this problem. Choosing the proper initial centroids is the key step of the basic K-means procedure.

The k-modes algorithm is a recent partitioning algorithm and uses the simple matching coefficient measure to deal with categorical attributes. The k-prototypes algorithm , through the definition ofa combined dissimilarity measure, prototypes algorithm further integrates the k-means and k-modes algorithms to allow for clustering instances described by mixed attributes. More recently, in another generalization of conventional k-means clustering algorithm presented. This new one applicable to ellipse-shaped data clusters as well as ball-shaped ones without dead-unit problem, and also performs correct clustering without pre-determining the exact cluster number.

Traditional clustering approaches generate partitions; each pattern belongs to one cluster. The clusters in a hard clustering are disjoint. Fuzzy clustering extends this presentation to associate each pattern with every cluster using a membership function. Larger membership values specify higher confidence in the assignment of the pattern to the cluster. widely used one algorithm is the Fuzzy C-Means (FCM) algorithm, which is based on k-means. FCM attempts to find the most characteristic point in each cluster, this can be considered as the center of the cluster, then the grade of membership for each instance in the clusters.

Other soft clustering algorithms have been developed and most of them are based on the Expectation-Maximization (EM) algorithm. They assume an underlying probability model with parameters that describe the probability that an instance belongs to a certain cluster. The procedure in this algorithm is to start with initial guesses for the mixture model parameters. These values are used to calculate the cluster probabilities for each instance. These probabilities are in turn used tore estimate the parameters and the process is repeated.

A drawback of such algorithms is that they tend to be computationally more expensive. Another problem found in the previous approach is called over fitting. This problem might be caused by two reasons. one hand, a large number of clusters maybe specified. And on the other, the distributions of probabilities have too many parameters. In this process, one possible solution is to adopt a fully Bayesian approach, in which every parameter has aprior probability distribution.

How do partitioning algorithms work?

- Construct a partition of a data set containing n objects into a set of k clusters, so to minimize a criterion (e.g., sum of squared distance)
- The goal is, given a k, find a partition of k clusters that optimizes the chosen partitioning criterion
- Global optimal: exhaustively enumerate all partitions
- Heuristic methods: k-means, k-medoids, K-means Clustering etc..,

1. Pick a number (K) of cluster centers (at random)
2. Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3. Move each cluster center to the mean of its assigned items
4. Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

**2.1 K-means Clustering:**

Partitioned clustering approach

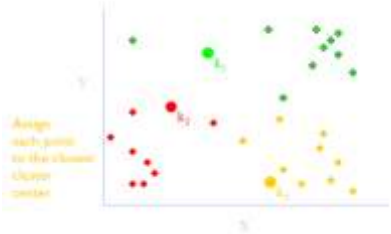
a) Each cluster is associated with a centroid(center point)
b) Each point is assigned to the cluster with the closest centroid
c) Number of clusters, K, must be specified

The basic algorithm is very simple

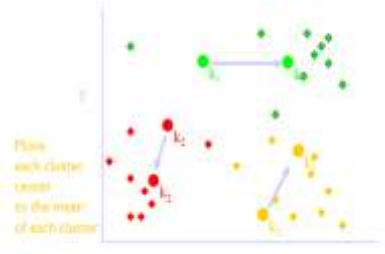
1: Select K points as the initial centroids.
2: repeat
3: from K clusters by assigning all points to the closest centroid.
4: Recompute the centroid o F each cluster.
5: Until the centroid don't change

K-means example (step 1):

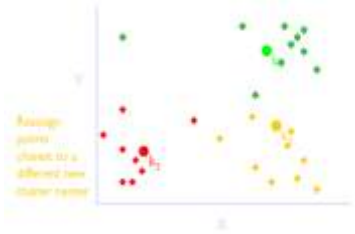
K-means example (step 2):



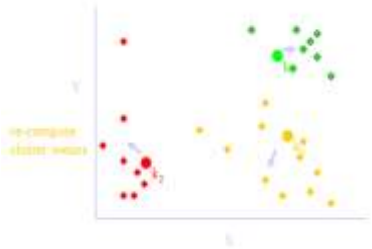
K-means example (step 3):



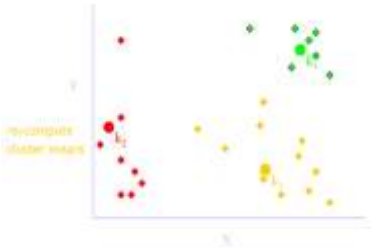
K-means example (step 4):



K-means example (step 5):



K-means example (step 6):



*2.2 k*-**Medoids:**

The k-Medoids: in k-medoids algorithm, Rather than calculate the mean of the items in each cluster, a representative item, or medoid, is chosen for each cluster at each iteration.

Medoids for each cluster are calculated by finding object *I* within the cluster that minimizes

$$\sum_{j \in C_i} d(i,j),$$

where *Ci* is the cluster containing object *i* and *d*(*i; j*) is the distance between objects *i* and *j*.

The k-medoids algorithm can be summarized as follows:

1. Choose *k* objects at random to be the initial cluster medoids.
2. Assign each object to the cluster associated with the closest medoid.
3. Recalculate the positions of the *k* medoids.
4. Repeat Steps 2 and 3 until the medoids become fixed.

Step 3 could be performed by calculating $\sum_{j \in C_i} d(i,j)$ for each object *i* from scratch at each iteration. However, many objects remain in the same cluster from one iteration of the algorithm to the next. Improvements in speed can be obtained by adjusting the sums whenever an object leaves or enters a cluster. Step 2 can also be made more efficient in terms of speed, for larger values of *k*. For each object, an array of the other objects, sorted on distance, is maintained. The closest medoid can be found by scanning through this array until a medoid is found, rather than comparing the distance of every medoid.

## IV. Density-based Clustering:

Density-based clustering algorithms try to find clusters based on density of data points in a region. The main idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well known density based clustering algorithms is the DBSCAN . DBSCAN separate data points into three classes :

• Core points. These are points that are at the interior of a cluster.
• Border points. A border point is a point that is not a core point,
• Noise points. A noise point is any point that is not a core point or a border point.

**3.1 DBSCAN: Density-Based Spatial Clustering Of Applications With Noise**

DBSCAN is a density-based clustering algorithm and an R*- Tree is implemented for the process. The basic concept of this algorithm is, in a given cluster within the specified radius of the neighborhood of every point in a cluster, there must exist a minimum number of points. The density attributed to the points in the neighborhood of a point in a cluster has to cross beyond a threshold value. Based on a distance function the shape of the neighborhood is obtained and is expressed as dist(p, q) between points p and q

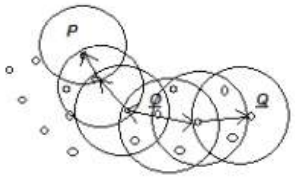



Figure. p and q are density-connected, connected by o

Density-connectivity is depicted in Figure.

The DBSCAN algorithm identifies the clusters of data objects based on the density-reachability and density-connectivity of the core and border points present in a cluster. The primary operation of the algorithm can be stated as: Given the parameters Eps and MinPts, a cluster can be identified by a two phase method. It can be specified as, 1) select an arbitrary data point from the database that satisfies the core point condition as a seed 2) fetch all the data points that are density-reachable from the seed point forming a cluster including the seed.

The algorithm requires the user to know the parameters Eps and MinPts for each cluster at least one point from the corresponding cluster. Since this is not feasible for every cluster, the algorithm uses global values for these two parameters. DBSCAN begins the clustering with an arbitrary data point p and retrieves the data points that are density-reachable from p with respect to Eps and MinPts.

This approach leads to the following inferences,

1) if p is a core point this method results in a cluster that is relevant to Eps and MinPts,

2) if p is a border point, no points are density-reachable from p and the algorithm scans the next data point in the database

.

**3.2 OPTICS:**

OPTICS is a clustering algorithm that identifies the implicit clustering in a given dataset and is a density-based clustering approach. Unlike the other density-based clustering algorithm DBSCAN which depends on a global parameter setting for cluster identification, OPTICS uses a multiple number of parameter settings. In that context the OPTICS is an extended work of DBSCAN algorithm. DBSCAN algorithm requires two parameters namely, ε the radius of the neighborhood from a given representative data object and MinPts the threshold value for the occurrence of the number of data objects in the given neighborhood. OPTICS is implemented on the concept of Density-based Cluster Ordering which is an extension of DBSCAN algorithm. Density-based Cluster Ordering works on the principle that sparsely populated cluster for a higher ε value contains highly populated clusters for a lower value of ε. Multiple number of distance parameter ε have been utilized to process the data objects. OPTICS ensures good quality clustering by maintaining the order in which the data objects are processed, i.e., high density clusters are given priority over lower density clusters. The cluster information in memory consists of two values for every processed object. one is the core-distance and other is reachability distance.

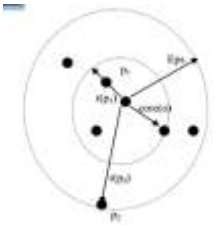


Figure. illustrates the concept of core distance and reachability distances. The reachability distances are r(p1) and r(p2) for data objects p1 and p2 respectively. The reachability distances are evaluated with respect to the Eps neighbouhood. The coredistance of a data object p is the shortest distance Eps' between p and a data object in its Eps-neighbourhood and so p is a core object with respect to Eps' if this neighbour is contained in NEps(p). Also the reachability-distance of a data object p with respect to another data object o is the shortest distance such that p is directly density-reachable from o if o is a core object. Thus OPTICS produces an ordering of the given database. Along with ordering OPTICS also stores core-distance and reachability distance of each data object, thereby resulting in better quality clusters. The OPTICS clustering algorithm provides an efficient cluster ordering with a set of ordering of the data objects with reachability-values and core-values. OPTICS implements pixel oriented visualization techniques for large multidimensional

data sets. OPTICS utilizes automatic techniques to identify start and end of cluster structures to begin with and later groups them together to determine a set of nested clusters.

**3.3 DENCLUE (DENsity-based CLUstEring) :**

It is a clustering method based on a set of density distribution functions. The method is implemented based on the following ideas: (1) the influence of each data point can be formally modeled using a mathematical function, can be called an influence function i.e., which describes the impact of a data point within its neighborhood; (2) the overall density of the data space can be modeled analytically as the sum of the influence function applied to all data points; and (3) clusters can then be determined mathematically by identifying density attractors.

## V. Grid-Based Methods:

Grid-based clustering algorithms first quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. Cells that are contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. the following are some of the grid-based clustering algorithms are: STatistical INformation Grid-based method – STING , WaveCluster, and CLustering In QUEst – CLIQUE. STING first divides the spatial area into several levels of rectangular cells in order to form a hierarchical structure.

**4.1 CLIQUE (CLUSTERING IN QUEST)**

Moreover, empirical evaluation shows that CLIQUE scales linearly with the number of instances.it has good scalability as the number of attributes is increased. The other clustering methods, WaveCluster does not require users to give the number of clusters. WaveCluster uses a wavelet transformation to transform the original feature space. In wavelet transform, Complication with an appropriate function results in a transformed space where the natural clusters in the data become obvious. It is a very powerful process, however, it is not efficient in high dimensional space.

It makes use of concepts of density and grid based methods. In first step, CLIQUE partitions the ‘ n ’ dimensional data space S into non overlapping rectangular units (grids). The units are obtained by partitioning every dimension into ξ intervals of equal length. if 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 Y is the density threshold is an another input parameter. A unit is the subspace is the intersection of an interval from each of the K attributes. The cluster is a maximal set of connected opaque units. u1, u2 are the two K-dimensional units are connected if they have a common face. The opaque units are then connected to form clusters. It uses apriori algorithm to find dense units. The opaque 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.

**4.2 STING:(A STatistical INformation Grid Approach to spatial Data Mining):**

Spatial data mining is the extraction of implied knowledge, spatial relation and discovery of interesting characteristics and patterns that are not explicitly represented in the databases. STING[9] 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. Several levels of such rectangular cells represents different levels of resolution. Each cell is partitioned in to 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 the 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. The 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 data are loaded into database, parameters count,m,s, min,max of the bottom level cells are calculated directly . First, a layer is determined from which the query processing process is to start. This

layer may consist of small number of cells. 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.

**4.3 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 [4]. Most of the grid based algorithms uses uniform grids whereas MAFIA uses adaptive grids. MAFIA[10]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.

**4.4 Wave Cluster :**

Wave Cluster is a multi resolution clustering algorithm, it used to find clusters for very large spatial databases.

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 two set of units and they are considered as cluster. Corresponding to each resolution $\gamma$ of wavelet transform there would be set of clusters cr. In the next step wave cluster labels the units in the feature space that are included in the cluster.

**4.5 O-Cluster: (Orthogonal partitioning CLUSTERing)**

This clustering method combines a novel partitioning active sampling technique with an axis parallel strategy to identify continuous areas of high density in input space. O cluster is a method that builds upon the contracting projection concept introduced by optigrid. O cluster makes two major contributions 1) It uses statistical test to validate the quality of a cutting plane. 2) It can operate on a small buffer containing a random sample from the original data set. The partitions that do not have ambiguities are frozen and the data points associated with them are removed from the active buffer. O cluster operates repeatedly. It evaluates possible splitting points for all projections in a partition. it can selects the best one, and splits the data into new partitions. The algorithm continuous by searching for good cutting planes inside the newly created partitions. it can creates a hierarchical tree structure that can translates the input space into rectangular regions. The process stages are (1) Load data buffer (2) compute histograms for active partitions (3) Find "best" splitting points for active partitions (4) Flag ambiguous and "frozen" partitions (5) Split active partitions (6) Reload buffer. O cluster can functions optimally for large data sets with many records & high dimensionality.

**4.7 Adaptive Mesh Refinement (AMR) :**

Adaptive Mesh Refinement is a type of multi resolution algorithm that achieves high resolution in localized regions. Instead of using a single resolution mesh grid, AMR clustering algorithm adaptively creates different ISSN: 0975-5462 3443Ilango et. al. / International Journal of Engineering Science and Technology Vol. 2(8), 2010, 3441-3446 resolution grids based on the regional density. The algorithm considers each leaf as the centre of an individual cluster and recursively assigns the membership for the data objects located in the parent nodes until the root node is reached. AMR Clustering algorithm detect nested clusters at different levels of resolutions. AMR is a technique that starts with a coarse uniform grid covering the entire computational volume. It automatically refines certain regions by adding finer sub grids. Newly child grids are created from the connected parent grid cells whose attributes, density for instance, exceed given threshold. Refinement is performed on each grid separately and recursively until all regions are captured with the desired accuracy.

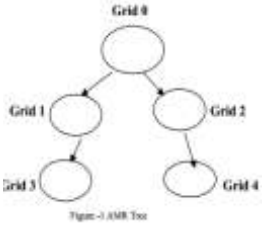



Figure-1 shows an example of AMR tree in which each tree node uses a different resolution mesh. The root grid with the coarsest granularity covers the entire domain. In which it contains two sub grids, grid 1 and grid 2. Grid 2 at level 1 also contains two sub grids discovered using a finer mesh. The deeper the node is discovered in the tree and the finer the mesh is used.

AMR clustering procedure connects the grid based and density based approaches through AMR techniques and hence preserves the advantages of both algorithms.

## VI. Model-Based Clustering Methods:

` AutoClass uses the Bayesian approach, starting from random initialization of parameters, incrementally adjusts them in an attempt to find their maximum likelihood estimates. Moreover, it is assumed that, in addition to the predictive attributes, there is hidden variable. This unobserved variable reflects the cluster membership for every case in the data set. The data-clustering problem is also an example of supervised learning from incomplete data due to the existence of such a hidden variable.

Their approach for learning has been called RBMNs. Another model based method is the SOM net. The S.O.M net can be thought of as two layers neural network. Each neuron represented by n-dimensional weight vector, m = (m1, … , mn), where n is equal to the dimension of input vectors. The neurons of the S.O.M are themselves cluster centers; but to accommodate interpretation the map units can be combined to form bigger clusters. The S.O.M is trained iteratively.

In each training step, one sample vector x from the input data set chosen randomly.The distance between it and all the weight vectors of the SOM is calculated using a distance measure. After finding the Best-Matching Unit , the weight vectors of the SOM are updated so that the Best-Matching Unit is moved closer to the input vector in the input space. The topological neighbors of the B.M.U are also treated in a similar way. The important property of the SOM is that it is very robust.

The outlier can be easily detected from the map, since its distance in input space from other units are large. The S.O.M can deal with missing data values. Many applications require the clustering of large amounts of high dimensional data sets. However, most automated clustering techniques do not work effectively and/or efficiently on high dimensional data.

## VII. Ensembles of Clustering Algorithms:

The theoretical foundation of combining multiple clustering algorithms is still in early stages. combining multiple clustering algorithms is a more challenging problem than combining multiple classifiers.The reason is that impede the study of clustering combination has been identified as various clustering algorithms produce different results.The main reason is due to different clustering field, combining the clustering results directly with integration rules such as sum, product, median . Cluster ensembles can be formed in different ways.i.e.,(1) the use of a number of different clustering techniques . (2) The use of a single technique many times with different initial conditions. (3) The usage of different partial subsets of features or patterns. In a split-and-merge strategy is followed.

The first step is to decompose complex data into small compact clusters. The K-means algorithm works this purpose. Data partitions present in these clusterings are mapped into a new similarity matrix between patterns based on a voting mechanism. This matrix is independent of data sparseness.it is used to extract the natural clusters using the single link algorithm. Recently, the idea of combining multiple different clustering algorithms of a set of data patterns based on a Weighted Shared nearest neighbors Graph WSnnG is introduced in.

### comparative statement of various clustering techniques:

| Clustering Technique | Examples | Data Type | Cluster Shape | Complexity | Data Set | Measure | Advantages | Disadvantages |
|---|---|---|---|---|---|---|---|---|
| | CURE | Numerical | Arbitrary | $O(N^2)$ | Large | Similarity Measure | Attempt to address the scalability problem and improve the quality of clustering results | These methods do not scale well with the number of data objects |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Hierarchical Methods** | BIRICH | Numerical | Spherical | O(N)(Time) | Large | Feature Tree | 1.Suitable large scale data sets in main memory 2.Minimize number of scans 3.I/O costs are very low | Suffers from identifying only convex or spherical clusters of uniform size |
| | CHAMELEON | Discrete | Arbitrary | **$O(N^2)$** | Large | Similarity Measure | 1.Hierarchical clustering come at the cost of lower efficiency. 2.CHAMELEON strongly relies on graph partitioning implemented in the library HMETIS | issue of scaling to large data sets that cannot fit in the main memory |
| | ROCK | Mixed | Graph | **$O(KN^2)$** | Small size | Similarity Measure | 1. ROCK employs link but not distance when merging points 2. It also introduces some global property and thus provides better quality | The algorithm can breakdown if the choice of parameters in the static model is incorrect with respect to the data set being clustered, or if the model is not adequate to capture the characteristics of clusters |
| **Partitioning Methods** | K-means Method | Numerical | Spherical | O(NKD)(Time)<br>O(N+K)(Space) | Large | Mean | 1. K-means is relatively scalable and efficient in processing large data sets | 1. Different sized clusters 2. Clusters of different Densities 3. Non-globular clusters<br>4. Wrong number of Clusters 5. Outliers and empty Clusters |
| | K-Medoids Method | Numerical | Arbitrary | O(TKN) | Large | Medoid | 1.K-Medoids method is more robust than k-Means in the presence of noise and outliers 2. K-Medoids algorithm seems to perform better for large data sets | 1.K-Medoids is more costly that the k-Means method 2. Like k-means, k-medoids requires the user to specify k 3. It does not scale well for large data sets |
| | CLARA | Numerical | Arbitrary | O(K(40+K2)+-K<br>(N-K))+ | Sample | Medoid | 1.Deals with larger data sets than *PAM*<br>2.More efficient and scalable than both PAM | 1. The best k medoids may not be selected during the sampling process, in this case, CLARA will never find the best clustering 2. If the sampling is biased we cannot have a good clustering 3. Trade off-of efficiency |
| | CLARANS | Numerical | Arbitrary | Quadratic in total performance | Sample | Medoid | 1. Experiments show that CLARANS is more effective than both PAM and CLARA 2. Handles outliers | 1.The computational complexity of LARANS is **O(n2)**, where n is the number of objects 2. The clustering quality depends on the sampling method |
| Density based clustering algorithm | DBSCAN | Numerical | Arbitrary | O(NLogN)(Time) | High Dimensional | Density Based | 1. DBSCAN Algorithm perform efficiently for low dimensional data. 2. DBSCAN is highly sensitive to user parameters MinPts and Eps | 1. The data set can't be sampled as sampling would effect the density measures 2.The algorithm is not partitionable for multi processing systems. |
| | OPTICS | Numerical | Arbitrary | O(NLOGN) | Low Dimensional | Density Based | 1.it can discovers the clustering groups with irregular shape, uncertain amount of noise 2. it can discovers high density data included in low density group 3.final clustering structure are incentive to parameters. | 1.Expect some kind of density drop to detect cluster borders 2.less sensitive to outliers |
| | DENCLUE | Numerical | Arbitrary | **$O(N^2)$** | Low Dimensional | Density Based | 1 it has a solid mathematical foundation and generalizes various clustering methods 2. it has good clustering properties for data sets with large amounts of noise | less sensitive to outliers |
| **Grid-Based Methods** | CLIQUE | Mixed | Arbitrary | **$O(C^k+mk)k$ – Highest Dimensionality m- number of input points C-number of clusters** | High Dimensional | cosine similarity, the Jacquard index | 1.It *automatically* finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces 2. It is quite efficient 3. It is *insensitive* to the order of records in input and does not presume some canonical data distribution | The accuracy of the clustering result may be degraded at the expense of simplicity of the simplicity of this method |
| | STING | Numerical | Rectangular | **O(K)**<br>**K – Number** | Any size | Statistical | 1.The grid-based computation is *query-independent* 2. the grid structure facilitates | All Cluster boundaries are either horizontal or vertical, and no diagonal boundary is |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **of Cells at bottom Layer** | | | parallel processing and incremental updating 3. the method's efficiency is a major advantage 4. Very efficient | selected |
| | MAFIA | **Numerical** | Arbitrary | $O(c^{k}+N/p^{B}k'\gamma+\alpha Spk')$ | Large | | 1.Praposes adaptive grids for fast sub space clustering. 2.Introduces scalable parallel frame work on a shared – nothig architecture to handle massive data sets. | Gains on higher dimensional data and larger data sets has been observed to be even more dramatic |
| | Wave Cluster | **Spatial Data** | Arbitrary | O(N) n – number of objects | Low Dimensional | Wave transform | 1.It provides unsupervised clustering 2. The multiresolution property of wavelet transformations can help detect clusters at varying levels of accuracy | |
| | O-Cluster | **Categorical and Numerical Data** | | O(N x d) | Large scale high Dimensional | | 1. Good accuracy and scalability. 2.It is robust to noise. 3. Automatically detects the number of clusters in the data, 4.Successfully operate with limited memory resources | O- clustering algorithms encounter serious scalability and/or accuracy related problems when used on data sets with a large number of records and/or dimensions |
| **Model-Based Clustering Methods** | RBMNs | | O(nN ˆL) | | Syntactic data sets | | 1.Improves classification accuracy. 2. More flexible representation scheme than BNs | 1. CBBN classifiers perform significantly better than RBMN classifiers.2. RBMNs rely on restricted partitioning of domain knowledge using a DT induction algorithm which may not yield the best partitioning of domain knowledge. |
| | SOM | **Numerical** | Arbitrary | **$O(N^2)$** | Low Dimensional | Object Similarity | 1..Different kinds of distance measures and joining criteria can be utilized to form big cluster. | The S.O.M generates a sub-optimal weights are not chosen properly |

## VIII. Conclusion

In this article i provide descriptions of several clustering techniques proposed for clustering process in ad hoc wireless networks. i made review on various clustering algorithms and explained by taking some examples in each category. This review can be helpful for understanding of various clustering techniques for selection of suitable technique.

## References


[1]. Z. Huang, "Extensions to the k-means algorithm for clustering large data sets," *Data Mining and Knowledge Discovery*, vol. 2:3, pp. 283-304, 1998.

[2]. ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., and SANDER, J. 1999. OPTICS: Orderingpoints to identify clustering structure. In *Proceedings of the ACM SIGMOD Conference*,40-60 Philadelphis,PA

[3]. BRADLEY, P. and FAYYAD, U. 1998. Refining initial points for k-means clustering. In *Proceedings of the 15th ICML*, 91-99, Madison, WI

[4]. DHILLON, I., MALLELA, S., and KUMAR, R. 2002. Enhanced Word Clustering for Hierarchical Text Classification, In *Proceedings of the 8th ACM SIGKDD*, 191-200,Edmonton,canada

[5]. S. Guha, R. Rastogi, and K. Shim, 2000. ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems, vol. 25, no. 5 : 345-366.

[6]. Jiawei Han and Micheline Kamber. "Data Ware Housing and Data Mining. Concepts and Techniques", Third Edition 2007

[7]. Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: An efficient clustering algorithm for large

[8]. databases. In *Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data*, 1998.

[9]. Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. ROCK: a robust clustering algorithm for categorical attributes. In *Proc. of the 15th Int'l Conf. on Data Eng.*, 1999.

[10]. 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

[11]. 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


Mr. B.G.obula Reddy obtained his Bachelors degree in Computer Science and Engineering from J.N.T.U Anantapoor INDIA in 2005 and his Masters degree in Software Engineering from Avanthi Institute of Engineering, Makavarapalem(Village), Narsipatnam(Mandal), INDIA in 2008. He is pursuing the Doctoral degree in Computer Science and Engineering from Rayalaseema University, Karnool – INDIA.He has 9 years of teaching experience and presently working as an Assoc.Professor in the Department of Information Technology of Lakireddy Balireddy College of Engineering, Mylavaram, Andhra Pradesh, India.
Mr. B.G.obula Reddy is a member of various professional societies like IEEE and ISTE.

Dr.Maligela Ussenaiah, working as Assistant Professor of Department of Computer Science with four years of experience in , Vikrama Simhapuri Unicversity, Nellore, Andhra Pradesh -524001. He is supervising for three Doctoral thesis in the areas of computer networks, mobile wireless networks, Data warehousing & mining and image processing