ISSN 2320 2629
Volume 2, No.3, May - June 2013
International Journal of Information Technology Infrastructure
Available Online at http://warse.org/pdfs/2013/ijiti01232013.pdf

Review of Spatial Clustering Methods

Neethu C V
Dept. of Computer Science & Engineering
SCT College of Engineering
Trivandrum,India
neethusureshbabu@gmail.com

Mr.Subu Surendran
Associate Professor
Dept. of Computer Science & Engineering
SCT College of Engineering
Trivandrum,India

ABSTRACT

Spatial clustering can be defined as the process of grouping object with certain dimensions into groups such that objects within a group exhibits similar characteristics when compared to those which are in the other groups. It is an important part of spatial data mining since it provides certain insights into the distribution of data and characteristics of spatial clusters. Spatial clustering methods are mainly categorized into four: Hierarchical, Partitional, Density based and Grid based. All those categorizations are based on the specific criteria used in grouping similar objects. In this paper, we will introduce each of these categories and present some representative algorithms for them. In addition, we will compare these algorithms in terms of four factors such as time complexity, inputs, handling of higher dimensions and handling of irregularly shaped clusters.

Key words- Cluster centroid, Splinter group, Wavelet transformation

1. INTRODUCTION

Spatial data, also known as geospatial data or geographic information, is the data or information that identifies the geographic location of features and boundaries on earth, such as natural or constructed features, oceans, and more. Spatial data is usually stored as coordinates and topology, and is data that can be mapped. Spatial data is often accessed, manipulated or analyzed through Geographic Information Systems (GIS).This provides an effective way for displaying and information graphically. Also, various techniques are using in the further processing the spatial data for many other applications related to engineering, planning, management, transport/logistics, insurance, telecommunications, and business[1]. This techniques mainly includes the following.

- Clustering and Outlier Detection

Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible. Outlier is a data point that does not conform to the normal points characterizing the data set. Detecting outlier has important applications in data mining as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing etc.

- Association and Co-Location

When performing clustering methods on the data, we can find only characteristic rules, describing spatial objects according to their non-spatial attributes. In many situations we want to discover spatial rules that associate one or more spatial objects with others. However, one of the biggest research challenges in mining association rules is the development of methods for selecting potentially interesting rules from among the mass of all discovered rules.

- Classification

Every data object stored in a database is characterized by its attributes. Classification is a technique, which aim is to find rules that describe the partition of the database into an explicitly given set of classes. Classification is considered as predictive spatial data mining, because we first create a model according to which the whole dataset is analyzed.

- Trend-Detection

A spatial trend is a regular change of one or more non-spatial attributes when spatially moving away from a start object. Therefore, spatial trend detection is a technique for finding patterns of the attribute changes with respect to the neighborhood of some spatial object.

Important data characteristics, which affect the efficiency of clustering techniques, include the type of data (nominal, ordinal, numeric), data dimensionality (since some techniques perform better in low-dimensional spaces) and error (since some techniques are sensitive to noise) [3].

Clustering of points is a most typical task in spatial clustering, and many kinds of spatial objects clustering can be abstracted as or transformed to points clustering. This paper will discuss about various kinds of spatial clustering. In many situations, spatial objects are represented by points, such as cities distributing in a region, facilities in a city, and spatial sampling points for some research reason, e.g. ore grade samples in ore quality analysis, elevation samples in terrain analysis, land price samples in land evaluation, etc[2]. Spatial clustering is the most complicated among all the spatial data mining techniques.

Based on the technique adopted to define clusters, the clustering algorithms can be divided into four broad categories [3],[4]: Hierarchical clustering methods (AGNES, BIRCH[4] etc.), Partitional clustering algorithms(K-means, K-medoids etc.), Density-based clustering algorithms(DBSCAN DENCLUE[6] etc.), Grid based clustering algorithms (STING[5] etc.). Many of these can be adapted to or are specially tailored for spatial data.

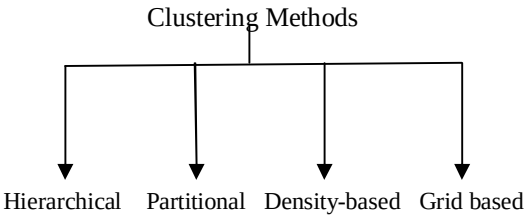

Figure 1: Categorization of spatial clustering methods

The paper is organized as follows. Section 2 introduces hierarchical clustering methods. Section 3 presents partitional clustering methods. Section 4 includes density based methods and section 5 gives grid based clustering methods and finally section 6 include comparison of different clustering algorithms.

2. CLUSTERING ALGORITHMS BASED ON HIERARCHICAL METHODS

Hierarchical clustering is a conventional clustering method which has wide variety of applications in different domains. Mainly, it is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other.

2.1 Agglomerative Nesting(Agnes)

AGNES proceeds by a series of fusions of similar objects. Initially (at step 0), all objects are apart– each object forms a small cluster by itself. At the first step, two closest or minimally dissimilar objects are merged to form a cluster[3]. Then the algorithm finds a pair of objects with minimal dissimilarity. If there are several pairs with minimal dissimilarity, the algorithm picks a pair of objects at random. Picking of pairs of objects, which are minimally dissimilar is quite straightforward, but the problem is " how to select pairs of clusters with minimal dissimilarity? " It is necessary to define a measure of dissimilarity between clusters.

Dissimilarity between clusters R and Q is defined as the average of all dissimilarities: *d(i ,j)* , where *i* is any object of R and *j* is any object of Q[3].

More formally,

$$d(R, Q) = [1/(|R||Q|)] \sum_{i \in R, j \in Q} d(i, j) \quad (1)$$

where |*R*| and |*Q*| denote the number of objects in clusters *R* and *Q* respectively.

AGNES computes Agglomerative Coefficient (AC), which measures the clustering structure of the data set. Agglomerative coefficient is defined as follows: For each object *i*, Let *d*(*i*) denote the dissimilarity of object *i* to the first cluster it is merged with, divided by the dissimilarity of the merger in the last step of the algorithm. That is,

$$AC = (\sum 1 - d(i))/n \quad (2)$$

2.2 Divisive Analysis(DIANA)

DIANA is a hierarchical clustering technique, but its main difference with the agglomerative method (AGNES) is that it constructs the hierarchy in the inverse order[3].

Initially (Step 0), there is one large cluster consisting of all n objects. At each subsequent step, the largest available cluster is split into two clusters until finally all clusters, comprise of single objects. Thus, the hierarchy is built in n-1 steps.In the first step of an agglomerative method, all possible fusions of two objects are considered leading to combinations. In the divisive method based on the same principle, there are possibilities to split the data into two clusters. This number is considerably larger than that in the case of an agglomerative method.

To avoid considering all possibilities, the algorithm proceeds as follows.

1. Find the object, which has the highest average dissimilarity to all other objects. This object initiates a new cluster– a sort of a splinter group.
2. For each object *i* outside the splinter group compute
   a. D_i = *[average d(i,j) j*ϵ $R_{splinter\ group}$ *] - [average d(i,j) j* $\epsilon R_{splinter\ group}$*]*
3. Find an object h for which the difference D_h is the largest. If D_h is positive, then h is, on the average close to the splinter group.
4. Repeat Steps 2 until all differences D_h are negative. The data set is then split into two clusters.
5. Select the cluster with the largest diameter. The diameter of a cluster is the largest dissimilarity between any two of its objects. Then divide this cluster, following steps 1-4.
6. Repeat Step 3 until all clusters contain only a single object.

Divisive Coefficient (DC): For each object j, let d(i) denote the diameter of the last cluster to which it belongs , divided by the diameter of the whole data set. The divisive coefficient (DC), given by,

$$DC = (\sum d(j))/n \quad (3)$$

indicates the strength of the clustering structure found by the algorithm.

When applying AGNES and DIANA to the spatial data, each object in the data set is replaced by spatial points. These spatial points are similar to those in is the 2D plane with x and y coordinates. The main difference is that spatial points are characterized by latitude and longitude. The latitude is the location of a place on the earth, north or south of the equator and longitude is the east – west measurement of position on the earth, measured from a plane running through polar axis. Here, the similarity between objects is expressed in terms of the Euclidian distance between two points in the n-dimension, $P(p_1,p_2,p_3,\ldots,p_n)$ and $Q(q_1,q_2,q_3,\ldots,q_n)$, which is given by,

$$d(P,Q)=\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_{n1})^2} \quad (4)$$

The algorithm proceeds in the similar manner as described earlier. But, the main disadvantages are the following:

- These clustering methods are these can be used to discover only well separated isotropic clusters.
- These methods does not usually consider the attributes of the spatial objects.
- These are not well designed to handle spatial data.

3. CLUSTERING ALGORITHMS FOR PARTITIONAL CLUSTERING

Partitional clustering is another kind of conventional clustering method which decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K ($N \geq K$) partitions of the data, with each partition representing a cluster[5]. That is, it classifies the data into K groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. Notice that for fuzzy partitioning, a point can belong to more than one group.

Many partitional clustering algorithms try to minimize an objective function. For example, in K-means and K-medoids the function (also referred to as the distortion function) is,

$$\sum_{i=1}^{K} \sum_{j=1}^{|Ci|} \text{Dist}(xj, \text{center}(i)) \qquad (5)$$

where $|C_i|$ is the number of points in cluster i, Dist(x j , center(i)) is the distance between point x_j and center i.

3.1 K-Means Algorithm

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point, re-calculate k new centroids of the clusters resulting from the previous step. After having these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. This procedure has to repeat until there is no change in the sets of centroids between two iterations.In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is given by,

$$O = \sum_{j=1}^{k} \sum_{i=1}^{n} ||X_i^{(j)} - C_j||^2 \qquad (6)$$

where $|x_i^{(j)}-C_j|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre C_j , is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect.

3.2 K-Medoid Algorithm

The *k*-medoids algorithm is a clustering algorithm related to the *k*-means algorithm. Both the *k*-means and *k*-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses datapoints as centers (medoids or exemplars) and works with an arbitrary matrix of distances between data points. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster.In some applications, we need each center to be one of the points itself. This is where K-medoid comes in- an algorithm similar to K-means,except

when fitting the centers $C_1,C_2,\ldots,C_k$, restricts our attention to points.

The important step of the algorithm are the following:

Initial guess for centers $C_1,C_2,\ldots C_K$(eg. Randomly select k points from $X_1,X_2,\ldots,X_n$).

1. Mininmize over C: for each i=1,2,…,n, find the cluster center C_k closest to Xi and let C(i) = k.
2. Mininmize over $C_1,C_2,\ldots C_K$: for each k=1,…,K, $C_k=X_k$*, the medoid of points in cluster k. ie, the point Xi in the cluster k that mininmizes $\sum_{c(j)=k} ||Xj - Xi||2$.

Stops when within cluster variation doesn't change.

Even though these algorithms can be used for spatial clustering, but the main disadvantage is that these methods are sensitive to the noise and center point. Another disadvantage is that these can detect only spherical clusters. So, we prefer density based clustering methods for spatial related applications than hierarchical and partitional. Next section discuss some of the well known density based clustering algorithms and their characteristics.

4. CLUSTERING ALGORITHMS BASED ON DENSITY BASED METHODS

Density-based approaches apply a local cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed. The important algorithms in this category includes DBSCAN,DENCLUE,OPTICS etc.

4.1 DBSCAN

DBSCAN requires two parameters: ϵ (eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point that has not been visited[8]. This point's ϵ-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ϵ-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ϵ-neighborhood is also part of that cluster. Hence, all points that are found within the ϵ-neighborhood are added, as is their own ϵ-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

The pseudocode for the algorithm is given below[11].

```
DBSCAN(D, eps, MinPts)
  C = 0
  for each unvisited point P in dataset D
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
      mark P as NOISE
    else
      C = next cluster
      expandCluster(P, NeighborPts,
C,eps,MinPts)

  expandCluster(P, NeighborPts, C, eps, MinPts)
  add P to cluster C
  for each point P' in NeighborPts
    if P' is not visited
      mark P' as visited
      NeighborPts' = regionQuery(P', eps)
      if sizeof(NeighborPts') >= MinPts
        NeighborPts = NeighborPts joined with
 NeighborPts'
  if P' is not yet member of any cluster add P' to
cluster C
regionQuery(P, eps)
return all points within P's ϵ-neighborhood
```

4.2 DENCLUE

DENCLUE (DENsity basted CLUstEring) [10] is a generalization of partitioning, locality based and hierarchical or grid-based clustering approaches. The

algorithm models the overall point density analytically using the sum of the influence functions of the points[8][9]. Determining the density-attractors causes the clusters to be identified. DENCLUE can handle clusters of arbitrary shape using an equation based on the overall density function. The authors claim three major advantages for this method of higher-dimensional clustering[8].

- Firm mathematical base for finding arbitrary shaped clusters in high-dimensional data sets
- Good clustering properties in data sets with large amounts of noise
- Significantly faster than existing algorithms

The approach of DENCLUE is based on the concept that the influence of each data point on its neighborhood can be modeled mathematically. The mathematical function used, is called an impact function. This impact function is applied to each data point and the density of the data space is the sum of the influence function for all the data points. In DENCLUE, since many data points do not contribute to the impact function, local density functions are used. Local density functions are defined by a distance function – in this case, Euclidean distance.

The local density functions consider only data points which actually contribute to the impact function. Local maxima, or density-attractors identify clusters. These can be either center-defined clusters, similar to k-means clusters, or multi-center-defined clusters, that is a series of center-defined clusters linked by a particular path. Multi-center-defined clusters identify clusters of arbitrary shape. Clusters of arbitrary shape can also be defined mathematically. The mathematical model requires two parameters, γ and ζ. γ is a parameter which describes a threshold for the influence of a data point in the data space and ζ is a parameter which sets a threshold for determining whether a density-attractor is significant.

5. CLUSTERING ALGORITHMS BASED ON 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 In QUEst), STING (STatistical INformation Grid), Wave clustering etc.

5.1 STING(STatistical INformation Grid based method)

This 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[6]. Corresponding to different resolution, different levels of rectangular cells are arranged to form a hierarchical structure. Each cell at a higher level is partitioned into a number of cells at its immediate lower level and so on. Statistical information associated with each cell is calculated and is using to answer queries.

The root of the hierarchical structure be at level 1, then its children are level 2 etc.ie, the level 1corresponds to whole spatial area and it is divided into its children at its higher level. The following figure depicts the idea.

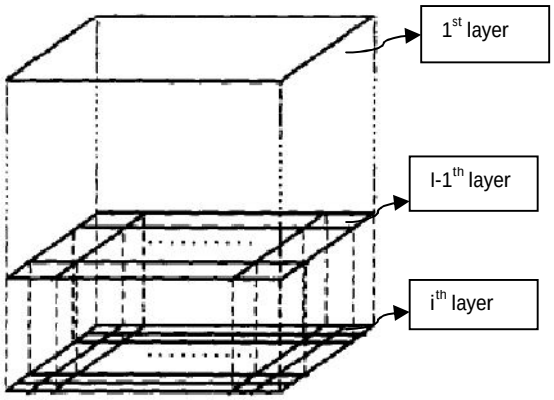

Figure 2[6]: Hierarchical arrangement of spatial data

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 that the attribute value in this cell follows.

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, check the relevancy of cell by computing confidence internal. Irrelevant cells are removed and this process is repeated until the bottom layer is reached.

The main advantages of this method are the following[10]:

- It is query independent method since the statistical information exists independently of queries.
- The computational complexity is O(K) where K is the number of grid cells at the lowest level. Usually, K<<N where N is the number of objects.
- Query processing algorithms using this structure are trivial to parallelize.
- When data is updated, no need to recompute all information in the cell hierarchy. Instead, perform an incremental update.

5.2 Wave Clustering

Wave Cluster is a multi resolution clustering algorithm. It is used to find clusters in very large spatial databases [10].

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. The important steps of the algorithm are given below.

1. Quantize the feature space, then assign objects to the units.
2. Apply wavelet transform on the feature space.
3. Find connected components (clusters) in the sub bands of transformed feature space , at different levels.
4. Assign labels to the units.
5. Make look up table.
6. Map the objects to the clusters

Wave cluster connects the components in two 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.

The main advantages of this clustering method are the following:

- Wavelet transformation can automatically result in the removal of outliers
- Multi resolution properly of wavelet transformation can help in the detection of clusters at varying levels of accuracy based
- Wavelet based clustering is very fast with a computational complexity of o(n) where n is the number of objects in the database
- Discovers clusters with arbitrary shapes
- It is insensitive to the order of input.
- It can handle data up to 20 dimensions
- If can handle any large spatial database efficiently.

5.3 CLIQUE(CLustering In QUEst)

The CLIQUE algorithm integrates density based and grid based clustering unlike other clustering algorithms described earlier, CLIQUE is able to discover clusters in the subspace of the data. It is useful for clustering high dimensional data which are usually very sparse and do not form clusters in the full dimensional space[3].

In CLIQUE, the data space is partitioned into non-overlapping rectangular units by equal space partition along each dimension. A unit is dense if the fraction of total data points contained in it exceeds an input model parameter.

CLIQUE performs multidimensional clustering by moving from the lower dimensional space to higher . When searches for dense units at the k-dimensional space, CLIQUE make use of information that is obtained from clustering at the (k-1) dimensional space to prune off unnecessary search.

This is done by observing the Apriori property used in association rule mining. In general, property employs prior knowledge of the items in the search space so that portions of the space can be pruned. The property adapted for the CLIQUE states the following: If a k-dimensional unit is dense, then so are its projections in the (k-1) dimensional space. That is, given a k-dimensional candidate dense unit, check its (k-1) th projection units and find any that is not dense, then we know that the kth dimensional unit cannot be dense either. Therefore, we generate the potential candidate dense units in the k-dimensional space from the dense units found in the (k-1) dimensional space. It illustrated in the following figure. In general, the resulting space searched is much smaller the original one. The dense units that are then examined to determine clusters.

Having found clusters, CLIQUE generates a minimal description for each cluster as follows: For each cluster, it determines the maximal region that covers the cluster of connected dense units. It then determines a minimal cover for each cluster[7].

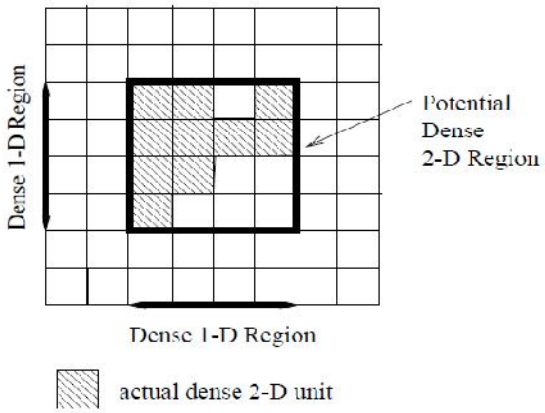

Figure 3: Determining potential region with dense units

CLIQUE automatically finds subspace of highest dimensionality such that high density clusters exist in those subspaces. It is intensive to the order of input tuples and does not presume any canonical distribution. It scales linearly with the size of input and has good scalability as the number of dimensions in the data is increased. However, the accuracy of clustering results may be degraded at the expense of the simplicity of the method.

6. CONCLUSION

The spatial clustering algorithms are categorized based on four factors like time complexity ,input, handling of higher dimensional data, capability to detect irregularly shaped clusters. These factors were found to be necessary for effective clustering of large spatial datasets with high dimensionality. The previous nine algorithms are addressed by how they matched these four requirements. Figure 4 summarizes the result of the comparison. It can be seen that although several of the algorithms meet some of the requirements, and some meet most, none of the algorithms as originally proposed meet all the requirements. The hierarchical clustering methods are similar in performance but consumes more time as compared to the other. The performance of partitional clustering methods like K- means and K-medoid algorithms are not well in handling irregularly shaped clusters. The density based methods and grid based methods are more suitable for handling spatial data, but when considering time complexity, grid based methods are more preferable.

Algorithm	Time Complexity	Inputs Required	Handling of Higher dimension	Handling of irregularly shaped clusters
AGNES	Atleast $O(n^2 \log n)$ where n= no. of data points	• Data set • Adjacency matrix of data points	No	Clusters are of arbitrary shape
DIANA	Atleast $O(n^2 \log n)$ where n= no. of data points	• Data set • Adjacency matrix of data points	No	Clusters are of arbitrary shape
K-MEANS	O(tknd) t: # of iterations k: # of clusters n: # of data points t: # of dimensions of data object.	• Data set • Number of clusters	Not well	Not completely
K-MEDOID	$O(k(n-k)^2)$ k: # of clusters n: # of data points	• Data set • Number of clusters	Not well	Not completely
DBSCAN	O(n logn) n: # of data points	• Data set • Two parameters	No	Not completely
DENCLUE	O(D logD) D: # of active data set	• Data set • Two parameters	Yes	Yes
STING	O(k) k: # of clusters	A-prior knowledge of data	No	Clusters are of approximate shape
Wave Clustering	O(n) n: # of data points	A-prior knowledge of data	Not well	Yes
CLIQUE	Quadratic on number of dimension.	• Data set • Two parameters	Yes	Partially

Figure 4: Comparison of different clustering algorithms.

REFFERENCES

[1]N.Sumathi,R.Geetha,Dr.S.SathiyaBhama.,"Spatial data mining –techniques,trends and its applications", Journal of Computer Applications,vol-1,2008.

[2] Limin Jiao, Yatin Liu., "Knowledge discovery by spatial clustering based on self-organizing feature map and a composite distance measure",The International Archieves of Photogrammetry, Remote sensing and Spatial Information Science, Vol.37,Part B2,Beijing 2008.

[3] J. Han,M. Kamber,A.K.H. Tang.,"Spatial Clustering in Datamining: A Survey",Geographic Datamining and Knowledge Discovery,London,2001.

[4] Sanjay Chawla,Sashi Shekhar., "Modeling Spatial Dependencies for Mining Geospatial Data",Geographic Datamining and Knowledge Discovery(GKD),2003.

[5]T. Zhang,R. Ramakrishna,M. Liuncy,"BIRCH: An efficient data clustering method for very large databases",Proc.ACM-SIGMOID International Conference on Management of Data(SIGMOID'96),1996.

[6]J. Wang,R. Yang,Muntz,"STING: A Statistical Information Grid Approach to Spatial Data mining ",International Confrence of Very Large Databases(VLDB'97),1997.

[7] Shekhar, S., Zhang, P. & Huang, Y., 2004. Trends in Spatial Data Mining. In S. Shekar, ed. Science.Minneapolis:AAAI/MIT, p. 363.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996-). "A density-based algorithm for discovering clusters in large spatial databases with noise".

[9] Sander, Jörg, Ester, Martin; Kriegel, Hans-Peter; Xu, Xiaowei (1998). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". *Data Mining and Knowledge Discovery* (Berlin: Springer-Verlag) 69–194.

[10] Pavel Berkhin ,"Survey of Clustering Data Mining Techniques" *Data Mining and Knowledge Discovery,2009.*

[11]http://en.wikipedia.org/wiki/DBSCAN