Comparative Study of Data Mining Clustering Algorithms

Chethan Sharma

Department of CSE, Christ University Bangalore
Email: chethansharma89@gmail.com

Abstract— Clustering is the process of grouping physical or abstract objects into classes of similar objects. These groups of similar objects are called clusters. Objects in one cluster are very similar to other objects in that particular cluster but very dissimilar when compared to objects in other clusters. Portraying data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It portrays many data objects by few clusters, and hence, it models data by its clusters. Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. In this paper we give a broad description of different clustering algorithms and methods that exist in data mining.

Index Terms— Clustering, data mining, clusters, clustering techniques

I. INTRODUCTION

Clustering is the process of grouping physical or abstract objects into classes of similar objects. These groups of similar objects are called clusters. Objects in one cluster are very similar to other objects in that particular cluster but very dissimilar when compared to objects in other clusters. Clustering analysis follows the principles of minimizing distance between similar objects and maximizing distance between dissimilar objects. Clustering Analysis does not depend on predefined class names as nothing will be know beforehand. It's follows unsupervised learning of machine learning concept. Cluster analysis as such is an iterative process of knowledge discovery or reciprocal multi-goal optimization. It will often be necessary to modify criterion until get accomplish the desired properties.

Basic example for clusters is shown in Fig 1. Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for. This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict [1]. These models are often referred to as unsupervised learning models, since there is no external standard by which to judge the model's classification performance. There are no right or wrong answers for these models. Their value is determined by their ability to capture interesting groupings in the data and provide useful descriptions [1].

Cluster 1
Cluster 2
Cluster 3

Fig 1. Example for clusters

II. REQUIREMENTS AND APPLICATIONS OF CLUSTERING

Emblematic requirements [2] [3] of clustering are as below:

1. Scaling: While dealing with large databases, we would need highly scalable clustering algorithms.

2. Capability to deal with different types of parameters and attributes: Clustering algorithms should be able to used on any kind of data such as numerical, binary data, characters etc.

3. Diagnosis of clusters with random shape: The algorithm should be able to detect any cluster with

arbitrary shape without depending only on particular shape and size.

4. High Dimensionality: Algorithms should not only deal with low dimensional data but also deal with high dimensional data.

5. Ability to deal with noisy data: Database might contain noisy, missing or inconsistent data. Algorithm must be capable of dealing with those kinds of data.

6. Interpretability: The results after applying clustering algorithms should be interpretable, coherent and usable.

Clustering can be applied to many fields in different ways [4] [5]. Few are listed below:

1. Marketing: Aid marketing team to discover different groups of people in their customer base and develop programs specific to a group.

2. WWW: Clustering can be used in online document classification:

3. Clustering analysis can be used for Pattern Recognition

4. Image Processing: Clustering is used in segmentation phase of DIP which is very critical to get correct processing results.

5. Climate: Cluster analysis is used to find patterns in atmospheric pressure of polar regions.

III. DIFFERENT DATA CLUSTERING TECHNIQUES

Main categories of clustering methods are [6]

A. Partitioning Clustering
B. Hierarchical Method
C. Density-based Clustering
D. Grid-Based Clustering
E. Model-Based Clustering

A. Partitioning Clustering Method

In Partitioning method, n objects are divided into k clusters such that k is less than or equal to n and each of the clusters have at least one element. It's an iterative process which improves cluster analysis by reassigning the clusters till no more changes are possible or convergence is reached. This method is suitable for small to medium sized data sets.

In this method, we will choose a arbitrary characteristic or a principle according to which we divide the data and we modify it iteratively till it becomes more reasonable.

There are 2 partitioning algorithms: K-means Algorithm and K-medoids Algorithm [6].

i) K-means Algorithm:

In this algorithm, the data set is divided into k clusters. It is divided such that each of the k clusters contains atleast 1 data element in it. The goal of the K-means algorithm is to find the clusters that minimize the distance between data points and the clusters [7].

The algorithm works as follows:

1. Given data set elements are divided into k clusters with each cluster containing more or less the same number of elements.

2. For each data element do the following till convergence is reached and no more reassignment of data elements are possible

- Assign each object to the cluster to which the data object is most similar. This is assignment is based on the mean value of data objects in the cluster.

- Update the mean of the cluster

This algorithm stops when the assignments do not change from one iteration to next.

Few drawbacks of this algorithm is we need to specify number of clusters, if this specification is not appropriate the grouping will not be proper. K-means is unable to handle noisy data and is sensitive to outliers [7].

ii) K-medoids Algorithm:

Unlike in K-means algorithms where mean is calculated for each object in a cluster, in K-medoids a representative object called medoid is chosen for each cluster at each iteration. The partitioning is done based on minimizing the following [8]

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

where Ci is the cluster which contains the object i and d(i,k) is the distance between the objects i and k.

Algorithm works as follows [9]:

1. Select k objects as medoids initially.

2. Repeat until no change,

a. Remaining objects are allocated nearest clusters.

b. A non-medoid object is chosen

c. Swapping cost for medoid object and non-medoid object are selected.

d. If total cost is negative swap operation is performed to from new set of k-medoids.

Drawback of K-medoids algorithm is that we need to specify number of clusters in advance and result depends on the initial partition.

A main advantage of this algorithm is that it is not sensitive to outliers and noisy data. Comparison between the above two algorithms [10] is given in Table1.

k-means	k-medoids
Complexity is O(ikn)	Complexity is O(i k(n-k)^2)
More efficient	Comparatively less efficient
Sensitive to outliers	Not Sensitive to outliers
Convex shape is required	Convex shape is not must
Number of clusters need to be specified in advance	Number of clusters need to be specified in advance
Efficient for separated clusters	Efficient for separated clusters and small data sets

Table 1: Comparison of k-means and k-medoid algorithms

B. Hierarchical Method

The hierarchical clustering method decomposes hierarchically the given data set .It usually results in a tree structure, where each cluster node consists of one or more child nodes. In the first step of hierarchical clustering, the algorithm will look for the two similar data objects and merge them to create a new data object, which is the average of the two merged data objects.

Each iterative step takes the next two most similar objects and merges them. This process is continued until there is one large cluster containing all the original data objects.

In hierarchical clustering, we assign N clusters for N data items. We then merge most similar pair of clusters into a single cluster. We repeat this procedure till all the clusters are merged into a single large cluster.

There are two types of hierarchical clusters Agglomerative clusters and Divisive clusters as shown in Fig 2.

i) Divisive Clustering:

Divisive Clustering is also called as top-down method. Here all the data elements are grouped into a single cluster and during successive iteration dissimilar objects are divided into two clusters. Finally we continue this process till the end recursively until there is one cluster for each data item.

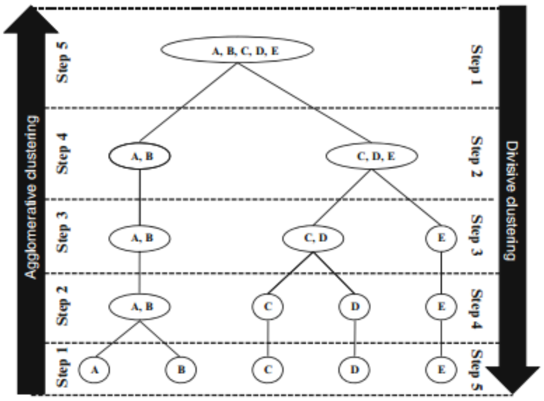
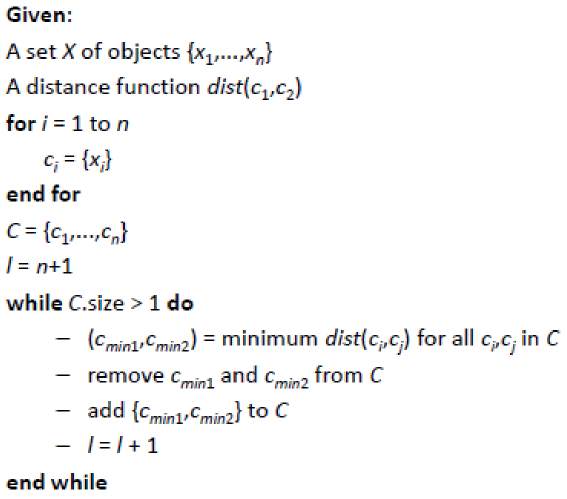


Fig.2 Types of Hierarchical Clustering

ii) Agglomerative Clustering:

This is reverse procedure of divisive clustering. Agglomerative method is also called as bottom-up approach. Here, each data item is assigned a cluster and in successive steps, two most similar clusters are merged to get a single cluster. This is repeated until all the clusters are merged into a single large cluster.

Following algorithm illustrated agglomerative clustering procedure [11]:



```
Given:
A set $X$ of objects $\{x_1,...,x_n\}$
A distance function $dist(c_1,c_2)$
for $i$ = 1 to $n$
    $c_i = \{x_i\}$
end for
$C = \{c_1,...,c_n\}$
$l = n+1$
while $C$.size > 1 do
    – $(c_{min1},c_{min2})$ = minimum $dist(c_i,c_j)$ for all $c_i,c_j$ in $C$
    – remove $c_{min1}$ and $c_{min2}$ from $C$
    – add $\{c_{min1},c_{min2}\}$ to $C$
    – $l = l + 1$
end while
```

C. Density-Based Clustering

In this method, we use density to discover randomly shaped clusters. A density based cluster can be defined as maximal set of density-connected points.

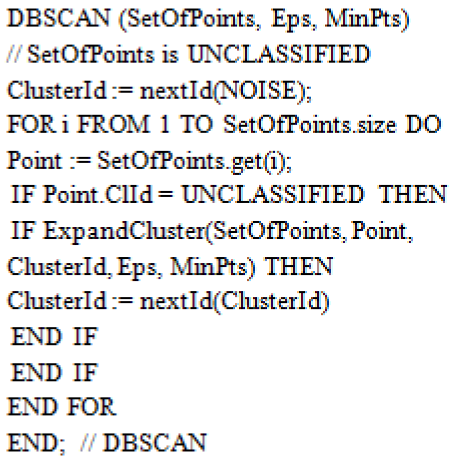
Main density based clustering methods are DBSCAN and OPTICS.

i) DBSCAN (Density Based Spatial Clustering Applications with Noise):

DBSCAN is a density based clustering method which is used to find random shaped clusters. It is based on the concept of 'density reachability' and 'density connectivity' and grows clusters with respect to density of the neighborhood objects.

Density Reachability - A point "i" is said to be density reachable from a point "j" if point "i" is within ε distance from point "j" and "j" has sufficient number of points in its neighbors which are within distance ε [12]. Density Connectivity - A point "i" and "j" are said to be density connected if there exist a point "k" which has sufficient number of points in its neighbors and both the points "i" and "j" are within the ε distance. This is chaining process [12].

Basic DBSCAN algorithm [13]:



```
DBSCAN (SetOfPoints, Eps, MinPts)
// SetOfPoints is UNCLASSIFIED
ClusterId := nextId(NOISE);
FOR i FROM 1 TO SetOfPoints.size DO
Point := SetOfPoints.get(i);
 IF Point.ClId = UNCLASSIFIED  THEN
 IF ExpandCluster(SetOfPoints, Point,
ClusterId, Eps, MinPts) THEN
ClusterId := nextId(ClusterId)
 END IF
 END IF
END FOR
END;  // DBSCAN
```

where Set Of Points can be whole of database or the data set obtained in previous iteration. Eps(ε) and MinPts are the global density parameters which represents neighborhood and minimum number of points required to form a cluster respectively.

Advantages of this algorithm [12] are i) no need to specify number of clusters in advance ii) can discover random sized and random shaped clusters iii) also has capability to find noise in the data.

Disadvantages of this algorithm [12] are i) Fails if clusters of varying densities are used ii) does not work well with high dimension objects.

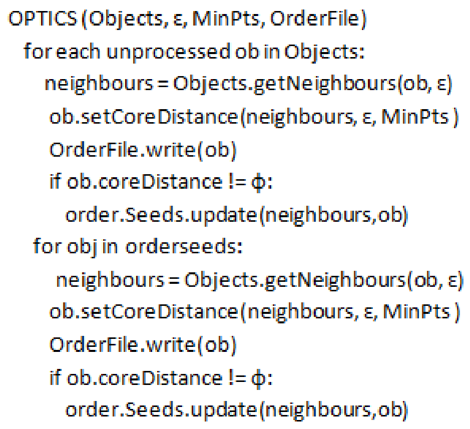
ii) OPTICS (Ordering Points to Identify Clustering Structure):

OPTICS is a variation of DBSCAN which generates an intensified ordering of the structure of data item's cluster. OPTICS is actually generalization of DBSCAN in which Eps(ε) is set to maximum. This algorithm addresses one of the major limitations of DBSCAN of failing to identify clusters when variable densities are used by finding clusters in spatial data. This algorithm adds two more attributes core distance and reachability distance [14].

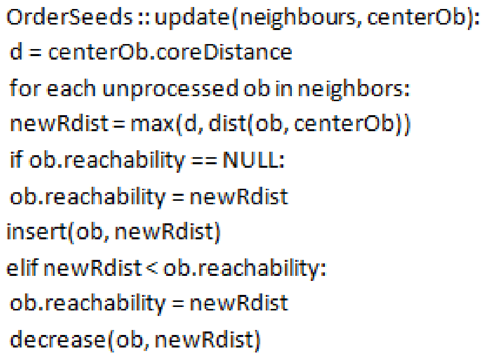
Core-distance- It is the smallest distance ε' between a cluster x and an object in its ε-neighborhood such that x would be a core object.

Reachability-distance- It of x is the smallest distance such that p is density-reachable from a core object z.

OPTIC algorithm steps are follows [14]:



```
OPTICS (Objects, ε, MinPts, OrderFile)
  for each unprocessed ob in Objects:
     neighbours = Objects.getNeighbours(ob, ε)
     ob.setCoreDistance(neighbours, ε, MinPts )
     OrderFile.write(ob)
     if ob.coreDistance != ϕ:
        order.Seeds.update(neighbours,ob)
   for obj in orderseeds:
       neighbours = Objects.getNeighbours(ob, ε)
     ob.setCoreDistance(neighbours, ε, MinPts )
     OrderFile.write(ob)
     if ob.coreDistance != ϕ:
        order.Seeds.update(neighbours,ob)
```



```
OrderSeeds :: update(neighbours, centerOb):
d = centerOb.coreDistance
for each unprocessed ob in neighbors:
newRdist = max(d, dist(ob, centerOb))
if ob.reachability == NULL:
ob.reachability = newRdist
insert(ob, newRdist)
elif newRdist < ob.reachability:
ob.reachability = newRdist
decrease(ob, newRdist)
```

D. Grid-Based Clustering Algorithm

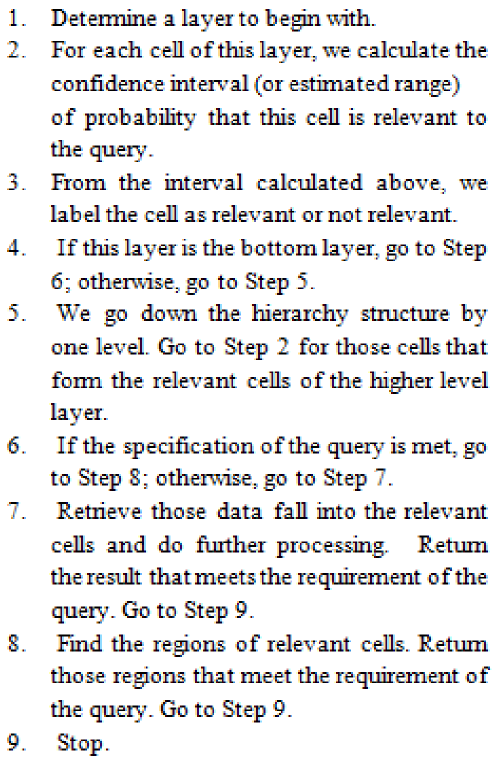
The Grid-Based clustering analysis uses multiresolution grid data structure. It divides underlying attributes into cells or grids. All the clustering operations are performed in these cells. One of the main advantages of grid-based clustering approach is that they have fast processing time and is not dependent of number of data elements but it only dependent of number of cells in the space.

Here we present the following grid based algorithms: STING, CLIQUE and WaveCluster.

i) STING (Statistical Information Grid)

STING is a grid based clustering technique in which spatial area is divided into cells and uses hierarchical structure. STING architecture is illustrated in Fig.3. Statistical information such as mean, min-max value etc are computed in advance and stored in cells.

STING algorithm works as follows [15]:



1. Determine a layer to begin with.
2. For each cell of this layer, we calculate the confidence interval (or estimated range) of probability that this cell is relevant to the query.
3. From the interval calculated above, we label the cell as relevant or not relevant.
4. If this layer is the bottom layer, go to Step 6; otherwise, go to Step 5.
5. We go down the hierarchy structure by one level. Go to Step 2 for those cells that form the relevant cells of the higher level layer.
6. If the specification of the query is met, go to Step 8; otherwise, go to Step 7.
7. Retrieve those data fall into the relevant cells and do further processing. Return the result that meets the requirement of the query. Go to Step 9.
8. Find the regions of relevant cells. Return those regions that meet the requirement of the query. Go to Step 9.
9. Stop.

Advantages of STING are:
- Query-independent
- Parallel processing
- Efficient
- Incremental updating

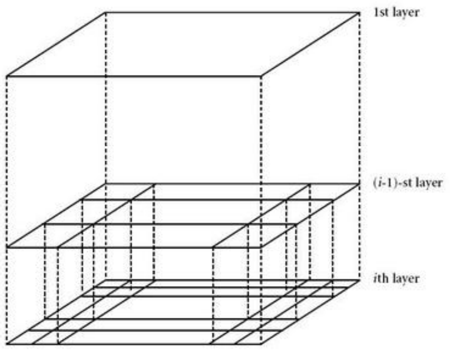
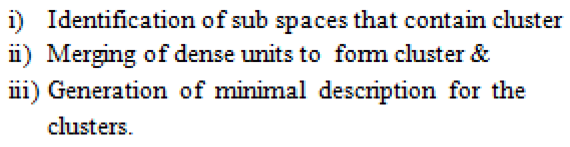


Fig.3 Basic structure of STING grid clustering method

ii) CLIQUE (Clustering in Quest)

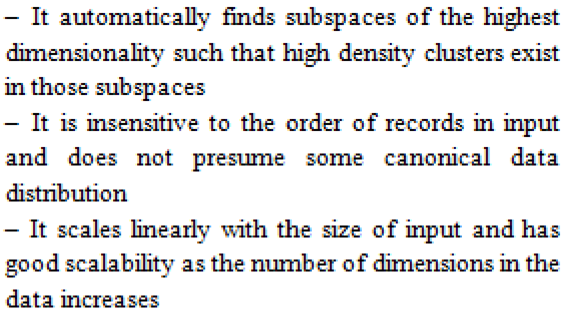
This Algorithm automatically identifies subspaces of a high dimensional spatial which allows efficient clustering. CLIQUE makes use of both grid-based and density-based concepts. It partitions the data space into equal length and non-overlapping rectangular grids.

Major steps in CLIQUE approach are:



i) Identification of sub spaces that contain cluster
ii) Merging of dense units to form cluster &
iii) Generation of minimal description for the clusters.

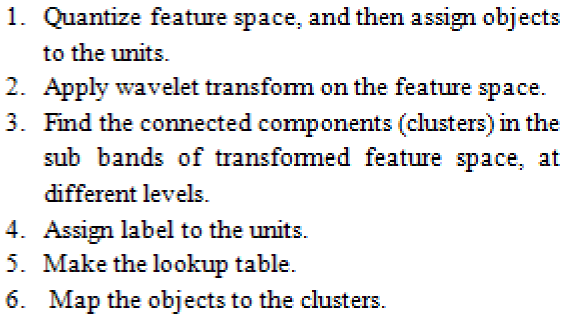
Advantages of CLIQUE are:



– It automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
– It is insensitive to the order of records in input and does not presume some canonical data distribution
– It scales linearly with the size of input and has good scalability as the number of dimensions in the data increases

iii) WaveCluster Method

Major steps in WaveCluster method are [16]:



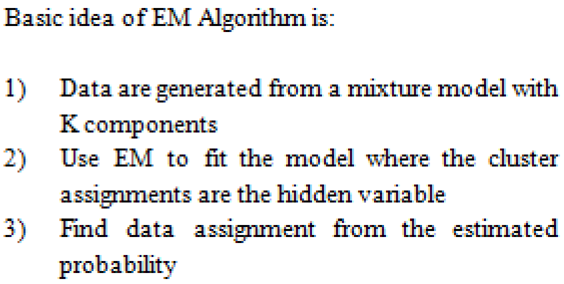
1. Quantize feature space, and then assign objects to the units.
2. Apply wavelet transform on the feature space.
3. Find the connected components (clusters) in the sub bands of transformed feature space, at different levels.
4. Assign label to the units.
5. Make the lookup table.
6. Map the objects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

E. Model-Based Clustering

Model based clustering methods tries to optimize fit the given data set and a mathematical model. This model is based on the assumption that data is generated by mixture of underlying probability distribution.

One of the Model-based Clustering methods is EM Algorithm.

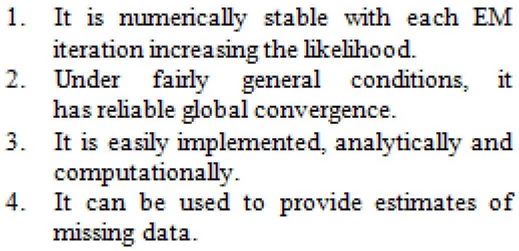
i) EM (Expectation-Maximization)



Basic idea of EM Algorithm is:

1) Data are generated from a mixture model with K components
2) Use EM to fit the model where the cluster assignments are the hidden variable
3) Find data assignment from the estimated probability

Advantages of EM algorithm are:



1. It is numerically stable with each EM iteration increasing the likelihood.
2. Under fairly general conditions, it has reliable global convergence.
3. It is easily implemented, analytically and computationally.
4. It can be used to provide estimates of missing data.

IV. CONCLUSION

Clustering is the process of grouping physical or abstract objects into classes of similar objects. These groups of similar objects are called clusters. In Partitioning method, n objects are divided into k clusters such that k is less than or equal to n and each of the clusters have at least one element. It's an iterative process. In K-Means algorithm, the data set is divided into k clusters. It is divided such that each of the k clusters contains atleast 1 data element in it. The goal of the K-means algorithm is to find the clusters that minimize the distance between data points and the clusters. In K-medoids a representative object called medoid is chosen for each cluster at each iteration. The hierarchical clustering method decomposes hierarchically the given data set .It usually results in a tree structure, where each cluster node consists of one or more child nodes. Divisive Clustering is also called as top-down method. Agglomerative method is also called as bottom-up approach .Density-based clustering method; we use density to discover randomly shaped clusters. Main density based clustering methods are DBSCAN and OPTICS. The Grid-Based clustering analysis uses multiresolution grid data structure. It divides underlying attributes into cells or grids. Grid Based methods such as STING, CLIQUE and WaveCluster are illustrated. Model based clustering methods tries to optimize fit the given data set and a mathematical model. EM Algorithm is discussed.

REFERENCES

[1] http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fnodes_clusteringmodels.htm

[2] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/sld010.htm

[3] http://churmura.com/technology/computer-science/requirements-for-clustering-in-data mining/31594/

[4] http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

[5] http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

[6] Gandhi, Gopi; Srivastava, Rohit "Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms.", International Journal of Computer Applications . Feb2014, Vol. 87, p10-13. 4p.

[7] Sharaf Ansari, Sailendra Chetlur, Srikanth Prabhu, N. Gopalakrishna Kini, Govardhan Hegde, Yusuf Hyder "An Overview of Clustering Analysis Techniques used in Data ", International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 12, December 2013

[8] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&cad=rja&uact=8&ved=0CFcQFjAF&url=http%3A%2F%2Fwww.researchgate.net%2Fpublication%2F227323522_The_Application_of_KMedoids_and_PAM_to_the_Clustering_of_Rules%2Ffile%2F79e415093a2e302452.pdf&ei=RGNGU7OaIIiVrAe09oGwDg&usg=AFQjCNERIIRauUu4CllTgNAIZfC3KtnTBw&sig2=dPmcCFsQ1gZ8GaperipyGw&bvm=bv.64507335,d.bm

[9] Jiawei Han and Micheline Kamber, "Data Mining Techniques", Morgan Kaufmann Publishers, 2000.

[10] Shalini S Singh and N C Chauhan "K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology , B.V.M. Engineering College 13-14 May 2011

[11] http://www.saedsayad.com/clustering_hierarchical.htm

[12] https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithm

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)

[14] Ankerst, M., Breunig, M., Kreigel, H.-P., and Sander, J. 1999. OPTICS Ordering points to

identify clustering structure. In Proceedings of the ACM SIGMOD Conference, 49-60, Philadelphia, PA.

[15] Wei Wang, Jiong Yang, and Richard Muntz "STING : A Statistical Information Grid Approach to Spatial Data Mining" Department of Computer Science University of California, Los Angeles February 20, 1997

[16] Gholamhosein Sheikholeslami Suro jit Chatterjee Aidong Zhang "WaveCluster: A Multi-Resolution Clustering Approach for Very Larg Spatial Databases" Proceedings of the 24th VLDB Conference New York, USA, 1998

◈◈◈