Quality Improvement of Partition based Technique

Rinkal Dabra

Dept. of Computer Science & Engineering , RNCET, KUK University, Panipat ,Haryana, INDIA

Kirti

Dept. of Computer Science & Engineering, RNCET, KUK University, Panipat,Haryana, INDIA

Abstract— When looking at data for the purpose of classification, various systems are used now a days for collecting data and managing it in a large databases in the organizations. The raw data is collected and high level of information is extracted from it: information which is useful for decision making, for exploration, and for better understanding of the phenomena generating the data. Traditionally this task of extracting information was done with the help of analysis where one or more analysts with the help of statistical techniques provide summaries and generate reports. Such an approach fails when there is a large volume of data. Hence various tools are required to aid the automation of analysis. Thus, knowledge discovery of databases was evolved which is "automatic": extraction of patterns of information from data. Combining, the group of the objects of a database into meaningful subclasses such that similarity of objects in the same group is maximized and similarity of objects in different groups is minimized, which is known as cluster formation. But my focus is on partitioning methods only, and proposing a new algorithm for automatic discovery of data clusters.

Index Terms —Data, Cluster Formation, Data set, K-mean, K-medoids.

I. INTRODUCTION

Importance of Collecting the data that reflect business or scientific activities to achieve competitive advantage is widely recognized now. Powerful systems for collecting data and managing it in large databases are in place in all large and mid-range organizations. The value of raw data(collected over a long time) is on the ability to extract high-level information: information useful for decision support, for exploration, and for better understanding of the phenomena generating the data. Traditionally this task of extracting information was done with the help of analysis where one or more analysts with the help of statistical techniques provide summaries and generate reports. Such an approach fails as the volume and dimensionality of the data increase.

Hence tools to aid the automation of analysis tasks are becoming a necessity. Thus, data mining was evolved which is "automatic": extraction of patterns of information from data.

Data mining, is the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. OR Data Mining also refer to extracting or mining from large amount of data. This is also sometimes referred to as Knowledge Discovery from Data (KDD) . Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge- driven decisions. Data mining tools can answer business questions that traditionally were time consuming to resolve.

Clustering, the grouping of the objects of a database into meaningful subclasses such that similarity of objects in the same group is maximized and similarity of objects in different groups is minimized, is called clustering . Thus the objects are clustered or grouped based on the principle of "maximizing the intraclass similarity and minimizing the interclass similarity". Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. Clustering methods can be divided into various types: Partitioning methods, Hierarchical methods, Density based methods, Grid-based methods, Model based methods, Probabilistic techniques, Graph theoretic and Fuzzy methods. But focus of thesis is on partitioning methods only , and proposing a new clustering algorithm for automatic discovery of data clusters.

A. Key Management and LAN:

Many international organizations produce a huge quantity of information in a week than many other persons could read in a lifetime. The situation is even more alarming in worldwide networks like Internet. Everyday hundreds of megabytes of data are distributed around the world, but it is no longer possible to monitor this increasingly rapid development – the growth is nearly exponential. So the basic problems with the management of the data can be summarized below:

- €Analysis required unearthing the hidden relationships within the data i.e. for decision support.
- €DBMS gave access to the data stored but no analysis of data.
- €Size of databases has increased and needs automated techniques for analysis as they have grown beyond manual extraction.

II. METHODOLOGY

As already stated, dissertation is based on methodology of Iterative Relocating Technique of Partitional Clustering i.e. k-means and k-medoids. All the algorithms are implemented in C++ language and results are studied in same environment.

But, before going through the study of these algorithms, their implementation view and results there are following terms used in the implementations of these algorithms DISTANCE:

Both of these algorithms are based on Distance among objects, hence it is necessary to introduce with the distance which we have taken in this dissertation. Let X={xi | i=1, 2, ...n} be a data set with n objects, k is the number of clusters, mj is the centroid of cluster cj where j=1, 2,...k. Then the algorithm finds the distance between a data object and a centroid by using the following Euclidean distance formula[1,9,10] Euclidean distance formula= √| xi – mj |2 Starting from an initial distribution of cluster centers in data space, each object is assigned to the cluster with closest center, after which each center itself is updated as the center of mass of all objects belonging to that particular cluster.

DATA SET:

Data set is the group of n data points on which the partitioning approach is to be applied which is of m dimension. Multidimensional data can include any number of measurable attributes. Each independent characteristic, or measurement, is one dimension. The consolidation of large, multidimensional data sets is the main purpose of the field of cluster analysis.[7] In this dissertation, data set (Dn) is to be used of two dimensional and having twenty data points. Each data points is represented by a pixel (x,y) where x is represented data point's first dimension and and y is represented by data point's second dimension respectively. Visualization of data set are also produced in the form of set of twenty pixels.

K-MEANS STEP:

The basic step of k-means clustering is to give the number of clusters k and consider first k objects from data set D as clusters & their centroid. Then the K-means algorithm will do the three steps below until convergence Iterate until stable (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroid
3. Group the object based on minimum distance.

K-MEDOIDS OR PARTITION AROUND MEDOID(PAM) STEP:

The basic step of K-Medoid or PAM clustering is to give the number of clusters k and consider first k objects from data set Dn as cluster & their medoid. Then the k-medoid or PAM algorithm will do the following five steps below until convergence Iterative until stable(=no objects change group)

1. Design clusters and Compute the Distance over all the clusters

2. Randomly select non-medoid objects as Compute the New Distance w.r.t. new.

III. OBJECTIVES

The Value of particular clustering method will depend on how closely the reference points represent the data as well as how fast the program runs. The speed of any algorithm can be represented by computation complexity and whereas the concern of the term "closely the reference points represent the data" it can be measured by several methods like sum of squared error, effect of outliers etc. Hence, the clustering technique can be modified in following areas:

COMPLEXITY

An algorithm is thus a sequence of computational steps that transform the input into the output. An algorithm is said to be correct if, for every input instance, it halts with the correct output. Algorithms devised to solve the same problem often differ dramatically in their efficiency. Complexity is measurement of an algorithm that shows the efficiency of algorithms mainly in two following terms:

Space Complexity : Space is the measurement of memory size that is used by the algorithm and by its different types of variables to implement. However, it is no more considerable now-a-days, because modern systems have a lot of memory to use and execute.

Computational Complexity : Computational is the measurement of time that an algorithm takes to execute for different inputs. Time is most critical and important because the algorithm using less time are more attractive than other algorithm of same domain. The time complexity on aforesaid algorithms depends on two things: *Size of Data Set, Dimensions of Data, Iterations and Number of Clusters*.

SUM OF SQUARE-ERROR

Each object in each cluster, the distance from object to its cluster center is squred and the distances are summed. In other words, square error is the sum of distance for all objects in the data set with respect to the centroid or medoid of belonging to that object's cluster. It can also be defined as:

$$E = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} |p_{ij} - m_j|^2 \right)$$

where E is the sum of square error for all objects in the data set; p_{ij} is the point in i^{th} object in j^{th} space/cluster representing a given object, m_j is the mean of j^{th} cluster (both p_{ij} and m_j are multidimensional). This criterion tries to make the resulting k clusters as compact and as separate as possible.. It is also refered to as "global optima" which is mostly ignored in the k-means clustering in order to find the local optima i.e. distance of data points of cluster from the centroid of that cluster.

OUTLIER

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. Aforesaid k-means algorithm is very much sensitive for outliers . A cluster is defined as a set of density-connected objects which is maximal w.r.t. density rechability and the *noise* is the set of objects not contained in any cluster.

PRIOR SPECIFICATION OF NUMBER OF CLUSTERS

It is always an issue of discussion about the value of k because different values for k will given different result and it is also necessary to determine the required numbers of k in advance because it take k as input parameter for deciding the number of clusters to be made. It has the disadvantage of being a gradient descent like search such that it is easily trapped in local optima. But it is generally not possible to estimate the numbers of clusters beforehand.

SEEDING

Seeding or Initialization means the way to initialize the value to the k-clusters with data objects which plays a very important role in determining the clusters in aforesaid algorithms. Choosing these cenres implizitly defines a clustering – for each center, we set one cluster to be the set of data points that are closer to that center than to any other. A probabilistic approach is also used to choose the initial centers as cluster which gives better results e.g. k-means ++. The results in partitional clustering strongly depends on the initial guess of centroids (or assignment) . Careful seeding can give the better result than simple mean and there are different algorithms which can give these kind of results e.g. k-means ++ .

VALIDATION

Validation is the measurement of similarity between the actual classes or groups and clusters results of any clustering technique. It can be measured by way of different measures like Purity, Entropy, F-measures and Coefficient of Variation. In which the Coefficient of Variation is the best measurement of Validation . Coefficient of Variation is a measure of the data dispersion. The CV is defined as the ratio of the Standard deviation to the mean. CV is a dimensionless number that allows comparison of the variation of populations that have significantly different mean values. In general the large the CV value is , the greater the variability is in the data. Coefficient of Variation can be calculated as follows:

1. Find the mean of all Classes individually in the data set:

$$\bar{x} = {}_{i=1}\sum^{n} x_i /n$$

where $\bar{x}$ is the mean of each class having n number of objects.

2. Next calculate the Global Mean of All classes by using their means :

$$\bar{x_g} = {}_{j=1}\sum^{N} \bar{x_j} /N$$

where $\bar{x_g}$ is the Global Mean of all N number of classes and x_j is the mean of that classes.

3. Find the Standard Deviation of all the Classes :

$$s=\sqrt{}_{j=1}\sum^{N} (\bar{x_j} - \bar{x_g})^2/N$$

4. Find the Coefficient of Variation (CV) :

$$CV = s/ \bar{x_g}$$

Where CV is the Coefficient of Variation, s is the Standard Deviation and $\bar{x_g}$ is the Global Mean.

In the aforesaid way we have to calculate the CV of Clusters identified by the Clustering and then find the DCV i.e. Difference of Coefficient of Variation between the CV of Classes and CV of Clusters. As less the DCV, as much the clusters are close to the actual classes.

The objectives of this dissertation is to modify the existing partitional algorithms i.e. k-means and k-medoids in respect of the followings aforesaid terms: Va*lidation, Computational Complexity, Prior specification of number of clusters and Sum of Square Error.*

IV. DISCUSSION ABOUT RESULTS OF K-MEANS AND K-MEDOIDS

A critical and careful look on the aforesaid results and from various literature indicates the following shortcomings are existing in the k-means and k- medoids clustering algorithms.

LIMITATIONS OF K-MEANS.CLUSTERING

1. In partition based k-Means clustering algorithms, the number of clusters (*K*) needs to be determined beforehand.

2. The algorithm is sensitive to an initial seed selection (Starting cluster centroids).Due to selection of initial centroid points; it is susceptible to a local optimum and may miss the global optimum. It may converge to suboptimal solutions. This means suboptimal classifications may be found, requiring multiple runs with different initial conditions. The selection of spurious data points as a center may lead to no data points in the class, with the outcome that the center cannot be updated.

3. It can model only a spherical shape of clusters. Thus the non convex shape of clusters cannot be modeled in center based clustering.

4. It is sensitive to outliers since a small amount of outliers can substantially influence the mean value.

5. Due to the nature of iteration scheme in producing the clustering result, it begins at starting cluster centroids and iteratively updates these centroids to decrease the square error. But it is not confirmed how many time it iterates which is not relevant for incremental data sets. It may take a high number of iterations to converge. Such number of iterations cannot be determined beforehand and may change from run to run. Result may be bad with high dimensional data.

6. It cannot be used for clustering problems whose results cannot fit in main memory, which is the case when data set has very high dimensionality or desired number of clusters is too big.

7. SSE increase proportionally as the number of clusters *k* is increased in comparison of k-medoids.

LIMITATIONS OF K-MEDOIDS CLUSTERING

1. K-medoids takes more time to cluster than k-mean this is one of the big drawback of this clustering methodology.

2. Swapping of non-medoid object with medoid object create complexity and undetermined iteration because scheme in producing the clustering result, it begins at starting cluster medoid and iteratively updates these medoids to decrease the Sum Square Error. But it is not confirmed how many time it iterates which is not relevant for those data sets that are of incremental nature. It may take a high number of iterations to converge. Such number of iterations cannot be determined beforehand and may change from run to run.

3. It is not suitable for Large Data Set application because of its random selection and hence it is executed on the partition of data set separately and again combined in the same way which create requirement of large space.

THE PROPOSED WORK : D- M CLUSTERING

Data Mining is data driven method to discover hidden knowledge from massive data sets. Several data mining tasks have been identified and one of them is clustering. In clustering grouping is accomplished by finding similarities between data objects according to characteristics found in actual data objects. It is of many types like partitioning based, model based, density based, grid based etc. Proposed work of this dissertation is on partitioning based clustering. A new partitioning based algorithm is proposed named D-M (Density Means) clustering algorithm.

Density Means clustering algorithm, the proposed work, provides its own way to form the clusters and solve the aforesaid problems that arises from k-Means & k-Medoids. It clusters the objects based up on the distance between object and existing cluster centroids. Moreover, it does not require prior knowledge of number of clusters to be formed.

Density Means clustering approach is the new methodology to cluster the data objects into number of groups, which is unknown initially. Number of groups (clusters) is some positive integer. The grouping is done by measuring the distance between object and centroid. Objects are iteratively grouped into the existing clusters or a new cluster formation is done with those objects based up on the Distance Determination Factor. Thus the purpose of this clustering is to classify the data on the basis of distance dynamically. It could improve the chances of finding the global optima with careful selection of initial cluster. In this algorithm data objects are stored in secondary memory and transferred to main memory one at a time. Only the cluster representatives are stored permanently in main memory to alleviate space limitations.Therefore, a space requirement of this algorithm is very small, necessary only for the centroids of the clusters. This algorithm is non–iterative and therefore its time requirement is also small.

Distance Determination Factor allows maximum permissible distance between data object and centroid of any cluster. It can vary according to the density of database.

V. DISCUSSION ABOUT RESULT OF D-MEANS

As already mentioned in previous chapter that the focus of this work is modification in k-means and k-medoids in terms : Va*lidation, Computational Complexity, Prior specification of number of clusters and Sum of Square Error.* Hence going through these points we consider them one by one as follows:

VALIDATION

As discussed in the objective, Validation of clustering can be measured by way of Coefficient of Variation of data set. In this sample, in order to simplify the conclusion, two cases are considered as follows:

First case*:* Consider the data set as a collection of two classes and compare the results of clustering with it where the number of clusters is two and

Second case*:* Consider the data set as a collection of four classes and compare the results of clustering with it where the number of clusters is four.

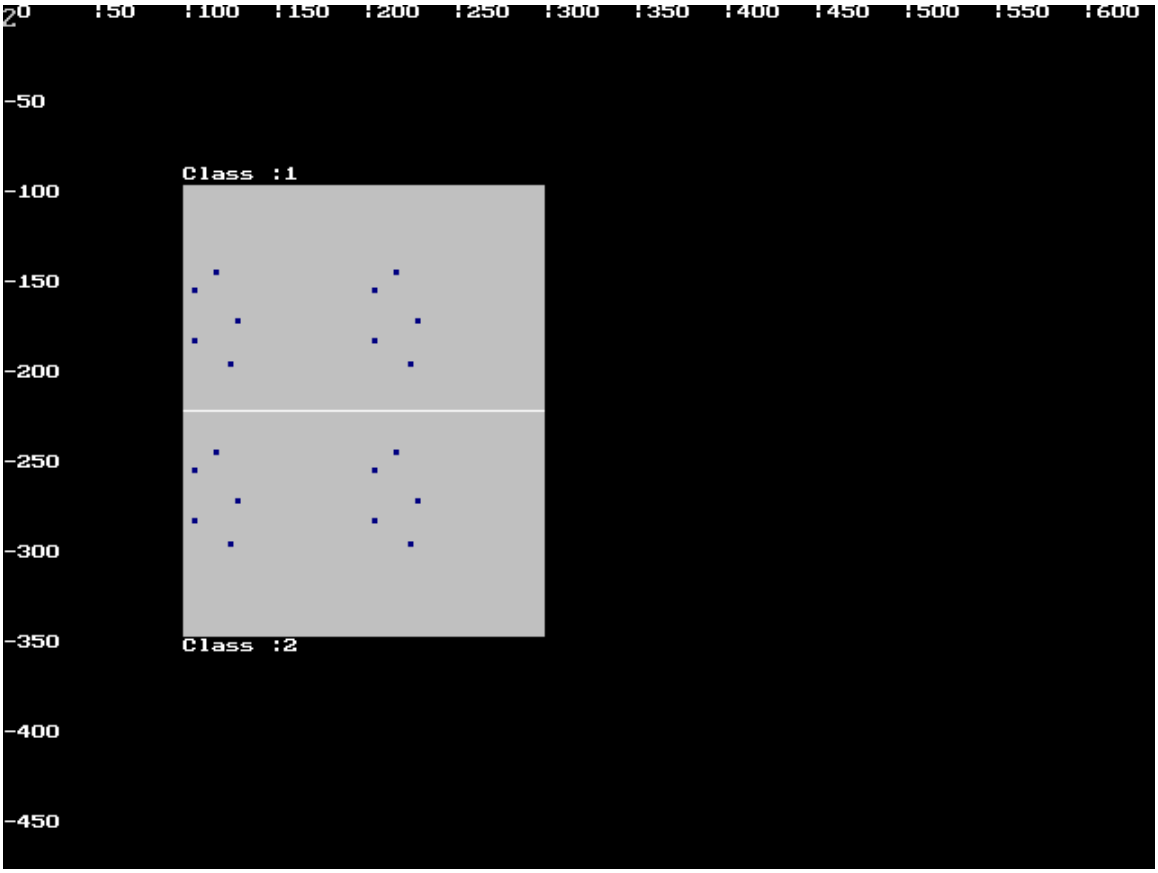

Fig. 4.15 Partition of Data Set in Two Classes

Results of First Case: In this case, data set is divided in two set of classes as shown in figure 4.15 and then compute the Coefficient of Variation for these classes by aforesaid way. The CV for this class set is 0.1181.

With respect to two classes, compute the CV of those algorithm's value having two clusters as output and then, compare the CV of them with actual CV of class partition. The comparison shown in Table 4.1 clearly shows that the Difference between the actual classes and clusters is same as in k-means and D-mean but vary in k-medoids. In other words the clusters given by k-medoids are not much valid as others because Difference of CV of others clusters are zero whereas the DCV of PAM is 0.0668.

COMPARISON TABLE:

Name of algorithm	**Coefficient of Variation**
k-means	*0.1181*
k-medoids	*0.1849*
D-M	*0.1181*

Table 4.1 Comparison of CV of Algorithm's.

Results of Second Case: In this case, data set is divided in four set of classes as shown in figure 4.16 and then compute the Coefficient of Variation for these classes by aforesaid way. The CV for this class set is 0.2990.

With respect to four classes, compute the CV of those algorithm's value having four clusters as output and then, compare the CV of them with actual CV of class partition. The comparison shown in Table 4.2 clearly shows that the Difference between the actual classes and clusters is same as in k-medoisd and D-mean but vary in k-means. In other words, the clusters given by k-mean are not much valid as others because the Difference of CV of others clusters are zero whereas the DCV of k-means is 0.1630.

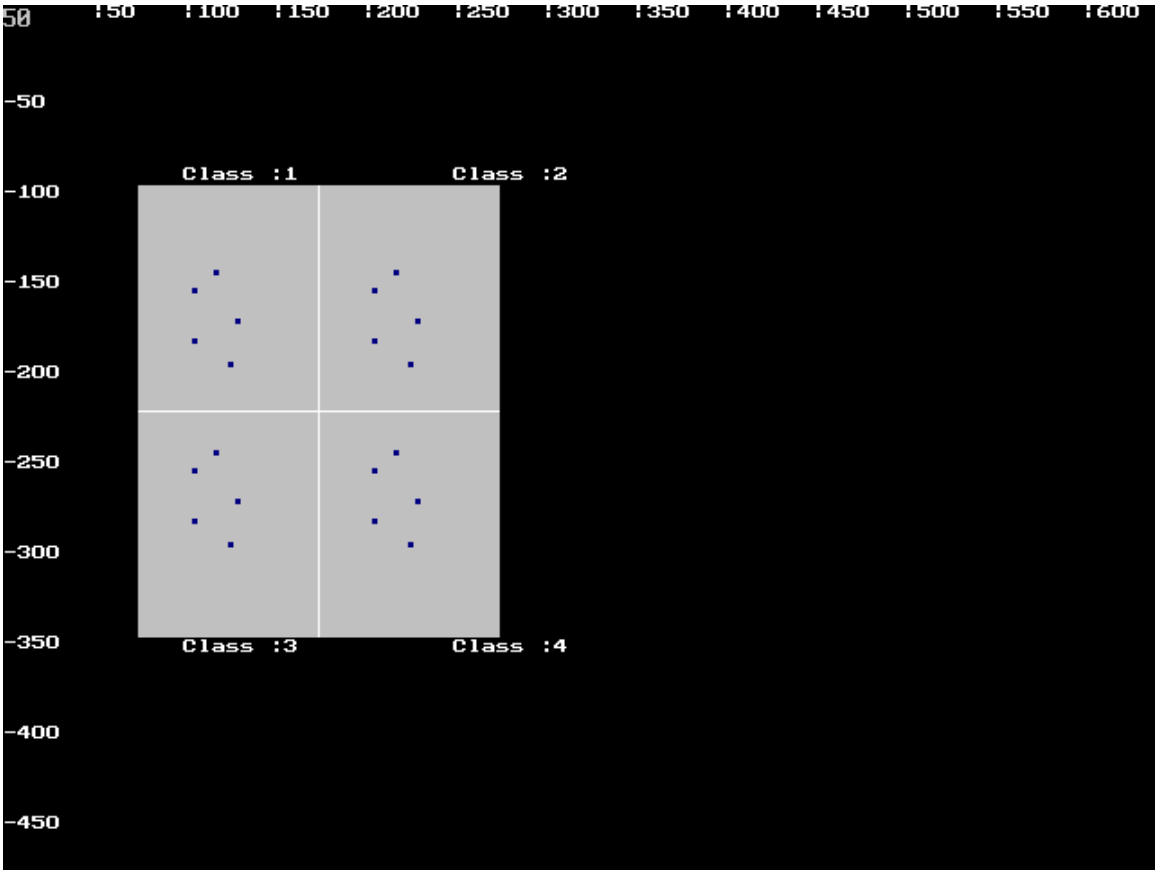

Fig. 4.16 Partition of Data Set in Four Classes

COMPARISON TABLE:

Name of algorithm	**Coefficient of Variation**
k-means	*0.1360*
k-medoids	*0.2990*
D-M	*0.2990*

Table 4.2 Comparison of CV of Algorithm's.

Hence, by looking at the tables 4.1 and 4.2 it is crystal clear that the probability of validity or change in coefficient of variation is less in density-means Clustering in comparison with k-means and k-medoids.

PRIOR SPECIFICATION OF CLUSTERS

In D-M Clustering, the input is Data Set and Distance Determination Factor *(DDF)* and on the basis of *DDF* it is decided that any particular data objects will lie in which cluster. Firstly, measures the distance between the data object and centroids and do the followings with the centroid having minimum distance with that data object:

If (Distance > DDF) then

Design the New Cluster taking this objects as centroid

Else

Combine the object with cluster,

having minimum centroid distance from the data object

Thus, in this way we divide the data points in different set of clusters and there is no prior specification of any k to decide the cluster beforehand.

ADVANTAGES OF D-MEANS

Having looked at the available results, tables and figures indicates the following advantages can be found in D-M clustering over k-Means clustering and k-medoids clustering algorithms.

ADVANTAGES OVER K-MEANS

1. In k-means clustering algorithms, the number of clusters (*K*) needs to be determined beforehand but in D-M clustering algorithm it is not required. It generates number of clusters automatically.

2. k-means Depends upon initial selection of cluster points; it is susceptible to a local optimum and may miss the global optimum. D-M clustering algorithm is employed to improve the chances of finding the global optimum.

3. k-means is sensitive to outliers since a small amount of outliers can substantially influence the mean value. In dynamic clustering algorithm outliers can't influence the mean value. They can be easily identified and removed (if desired) because they will lie outside the given Distance Determination Factor.

4. In k-means it is not confirmed that how many times it iterates but in D-M clustering it is known.

ADVANTAGES OVER K-MEDOIDS

1. In k-medoids clustering algorithms, the number of clusters (*K*) needs to be determined beforehand but in D-M clustering algorithm it is not required. It generates number of clusters automatically.

2. D-M Clustering takes less time than K-medoids.

3. In D-M there is one iteration for all objects in data set hence the complexity is largely less than k-medoids. Besides above, some key features like Sum of Squared Error, Validity and Complexity etc. are more favourable in D-means clustering than k-means and k-medoids.

VI. CONCLUSION

In partitioning based clustering algorithms, the number of final clusters (k) needs to be determined beforehand. Also, algorithms have problems like susceptible to local optima, sensitive to outliers, memory space, validity and unknown number of iteration steps required to cluster. So an algorithm which explores number of clusters automatically considering all these problems is an active research area and most of the research has focused on effectiveness and scalability of the algorithm. In this paper, a partitioning based algorithm, D-M (Density Means), clustering has been considered which generate clusters automatically.

VII. ACKNOWLEDGEMENT

The authors are quite thankful to the Computer science department at RN College of Engineering for supporting our work.

REFERENCES

[1] J. Han and M. Kamber K. Data Mining: Concepts and Techniques. Morgan Kaufman, 2000.
[2] P.SBradley,UsamaM.Fayyad.,"InitialPoints for K-Means Clustering.", Advances in Khowledge Discovery and Data Mining. MIT Press.
[3] "Introduction to Data Mining and Knowledge Discovery" 3rd Edition by Two Crows Corporation.
[4] Teknomo, kardi. "K-Means clustering tutorial."
[5] Hand & Others: Principal of Data Mining
[6] Pavel Brakhin: "Survey of Clustering Data Mining Techniques" .
[7] Vance Faber, "Clustering and the Continuous k-Means Algorithm".
[8] Introduction of Clustering:Computer Science Department,Rutgers University
[9] Shang ,Yi &Li, Guo-Jie(1991) New Crossover Operators In Genetic Algorithms, , P. R. China: National Research Center for Intelligent Computing Systems (NCIC)
[10] Singh ,Vijendra & Choudhary ,Simran (2009) Genetic Algorithm for Traveling Salesman Problem: Using Modified Partially-Mapped CrossoverOperator,sikar,Rajasthan,India: Department of Computer Science & Engineering, Faculty of Engineering & Technology, Mody Institute of Technology & Science, Lakshmangarh
[11] Su, Fanchen et al(2009) New Crossover Operator of Genetic Algorithms for the TSP, P.R. China: Computer School of Wuhan University Wuhan.