A Comparative Study on Partition-based Clustering Methods

[1]Mridumurchana Kashyap, [2]Surakhita Gogoi,[3] Rabinder Kumar Prasad
[1]Student, [2]Student, [3]Assistant Professor
[123]Computer Science and Engineering,
[123]DUIET, Dibrugarh, Assam,India

Abstract: **Clustering analysis is one of the essential data analysis tools that separate a group of data objects into similar sets called clusters. In Partition-based clustering method, identifying the initial centroid is challenging task. This paper presents a review of some partition-based clustering method that improves the selection of initial centroid value and enhances the quality of clustering to some extent.**

***Keywords*- Data mining, spatial data Clustering, Partition-based method, k-means**

I. INTRODUCTION

Data mining is known as the way toward investigating information to separate interesting examples and learning. Data mining is utilized for investigation reason to dissect diverse sort of information by utilizing accessible information mining devices. Data mining is the way towards finding designs in vast informational indexes including strategies at the convergence of machine learning, measurements, and database frameworks. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in two categories- descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions. In [1], the following definition is given:

Data mining is the process of exploration and analysis, by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns and rules. Data mining is an inter disciplinary sub field of computer science which involves computational process of large datasets' patterns discovery. The goal of this advanced analysis process is to extract information from a dataset and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, data base systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases[2]. Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. The superiority of a clustering result depends on equally the similarity measure used by the method and its implementation [3].

II. CLASSIFICATION OF CLUSTERING:

Clustering algorithms can be categorized into partition-based algorithms hierarchical-based algorithms, density-based algorithms, grid-based algorithms and model-based algorithms.

A. ***Partitioning Algorithms*:** Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \sum \sum (||p - m_i||)^2 \quad (1)$$

Where p is the point in a cluster and m_i is the mean of the cluster. The cluster should exhibit two properties, they are

I. Each group must contain at least one object.
II. Each object must belong to exactly one group. The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor result due to overlapping of data points [3]. There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning around Medoids).

The advantages and disadvantages of partitioning method:

Advantages:

I. Relatively scalable and simple.

II. Suitable for datasets with compact spherical clusters that are well-separated.

Disadvantages:

I. Poor cluster descriptors.
II. High sensitivity to initialization phase, noise and outliers.

B. ***Hierarchical Clustering****:* Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to 'n' clusters each containing a single object. Hierarchical Clustering is classified as:

a. *Agglomerative Nesting:* It is bottom-up approach[4]. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance.
b. *Divisive Analysis*. It is top-down approach[4]. It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own.

The advantages and disadvantages in hierarchical clustering methods are:

Advantages

I. Embedded flexibility regarding the level of granularity.
II. Well suited for problems involving point linkages, e.g. taxonomy trees.

Disadvantages

I. Inability to make corrections once the splitting/merging decision is made.
II. Lack of interpretability regarding the cluster descriptors.

C. ***Density-based Clustering:*** Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. It is able to find the arbitrary shaped clusters and handle noise. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The density based algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) is commonly known. The Eps and the Minpts are the two parameters of the DBSCAN [5]. The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius (ε) must contain at least minimum number of points (MinPts) [5].

The advantages and disadvantages of density based clustering:

Advantages

I. Discovery of arbitrary shaped clusters with varying size.
II. Resistance to noise and outliers.

Disadvantages

I. Poor clusters descriptors.
II. High sensitivity to the setting of input parameters.

D. ***Grid-based Clustering:*** Grid Density based clustering is concerned with the value space that surrounds the data points not with the data points. This algorithm uses the multi resolution grid data structure and use dense grids to form clusters. It first quantized the original data space into finite number of cells which form the grid structure and then perform all the operations on the quantized space. Grid based clustering maps the infinite amount of data records in data streams to finite numbers of grids. Its main distinctiveness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm self-governing of the number of data points in the original data set[6]. Grid Density based algorithms require the users to specify a grid size or the density threshold, the problem here arise is that how to choose the grid size or density thresholds.

E. ***Model-based Clustering:*** Model based clustering methods are that a model is hypothesized for each clusters to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods [7].

III. RELATED WORK

As discussed earlier, in context to improve the quality of clustering several attempts are been taken to choose the initial centroids.

A method was proposed in Rauf et al. [8], to calculate the initial centroids which reduces the number of iterations and improves the elapsed time. The algorithm works in two phases. In the first phase, the cluster size is fixed and the output of the first phase is the initial centers. In the second phase, the cluster sizes vary and the output of this phase are the finalized clusters.

A systematic method for finding the initial centroids was presented in Abdul et al.[9]. The centroid obtained by this method is consistent with the distribution of data. Hence, it produces clusters with better accuracy as compared to original k-mean algorithm.

Initial starting centroids algorithm based on k-means was proposed by Agha et al.[10]. The algorithm uses guided random technique .The experimental results shows that the algorithm outperformed the traditional random initialization and improved the quality of clustering with a big margin especially in the complex datasets.

A comparative analysis of various clustering methods, with an emphasis on their computational efficiency, was presented in Celebi et al.[11]. Eight commonly used linear time complexity initialization methods were compared on a large and diverse collection of datasets using various performance criteria. Experimental results were presented using non-parametric statistical tests. It was concluded that popular initialization methods often perform poorly.

A method for finding the initial-centroids was also proposed by Yedla[12]. The method works in the cases where the input-data is uniformly distributed. But in the cases where the input data is non-uniform that is,where most of the data items in each group lie towards the boundary of the group, this method does not produce good clustering result.

A selection method for initial centroid in k-means clustering was proposed in Aldahdooh et al. [13]. It provides a detailed performance assessment of the proposed initialization method over many datasets with different dimensions. The experimental results shows that the proposed initialization method is more effective and converges to more accurate clustering results than those of the random initialization method.

Cluster centre initialization algorithm for iterative clustering algorithm was presented in Khan et al.[14]. The algorithm was based on the fact that very similar objects form the core clusters and the cluster membership remains the same. However , the outliers are more susceptible to a change in cluster membership. Hence, these similar patterns aid in finding initial centers.

IV. ALGORITHMS

k-*means Algorithm(random selection of centers[2]*

k-means 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 set through a certain number of clusters. The main idea is to define k centers , one for each cluster. These centers should be placed in cunning way because of different location causes different results. So, the better choice is to place the as much as possible far away from each other. The next step is to take each point belonging to a given dataset and associate it to the nearest center. When no point is pending,the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k centroids , a new binding has to be done between the same dataset points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function known as Squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (|\|x_i - v_j\||)^2 \qquad (2)$$

Where, $(\|x_i - v_j\|)^2$is the Euclidean distance between x_i and v_j,

c_i, is the number of data points in ith cluster,

c is the number of cluster centers.

Algorithmic steps of k-mean clustering:

Let X={$x_1, x_2, x_3 \ldots . x_n$} be the set of data points and V={ $v_1, v_2, \ldots . v_c$,} be the set of centers:

1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data points and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster centers using:

$$Vi = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i$$

where c_i represents the number of data points in the ith cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop , otherwise repeat from step 3.

k-means Algorithm(initial centroid selection):[15]

An attempt to systematically select the initial centroid is explained below. This is proposed so that the multiple runs of the same algorithm for the same dataset do not produces different results, and so as to get a good quality of the results. The first step of the proposed algorithm includes that the given points are plotted in a two dimensional space. All the data points should have positive valued attributes. If not the negative valued attributes should be transformed to positive by subtracting each data point attribute with the minimum attribute value in the given dataset. This transformation is required because in the proposed algorithm the distance of each data point from the origin has to be calculated. If the data points are not transformed then there is a chance that for different data point, the same Euclidean distance from the origin is obtained, which will result in incorrect selection of initial centroids.

Algorithmic steps of improved k-mean:

The algorithm is :

Input: A dataset D containing n data points.

- D=$\{d_1, d_2, \ldots d_n\}$
- Number of desired clusters k.

Output: k number of initial centroids

Steps:

1. For each data point in D calculate the distance from the origin.
2. Sort the distances obtained in the previous step. In accordance with these distances sort the original data point.
3. Divide the sorted data point into k number of equal partitions.
4. In each partition, calculate the mean of the data points. These mean values will be taken as the initial centroids to be used in the k-mean algorithm.

For each data point, the distance from the origin is calculated using the Euclidean distance measure as given below.

Origin O (0,0)

Data point: A(x, y)

Euclidean distance between O-A will be: $\sqrt{(x-0)^2+(y-0)^2}$

Then these distances are sorted in ascending or descending order. According to these sorted distances the corresponding original data points is divided into k equal partitions. Then, for each partition mean of the data point is calculated as:

Let a partition contains three data points (x_1, y_1), (x_1, y_2) and (x_3, y_3). Their mean will be $((x_1+x_2+x_3)/3, (y_1+y_2+y_3)/3)$. The mean values for each partition are taken as initial centroids. The centroids thus taken are well suited for every type of dataset. The data points with uniformly and evenly distributed values and also for those in which values are not uniformly distributed over the partition but most of them are concentrated towards any of the boundaries of the partition.

Advantages of original k-means:

1. Fast ,robust and easier to understand.
2. Relatively efficient O(tknd), where n is objects , k is clusters , d is dimension of each object and t is iterations. Normally, k,t,d<<n.
3. Gives tight and well-separated clusters.

Advantages of improved k-means:

1. The main advantage over the original k-means algorithm is that once the initial centroids are systematically determined, the number of iterations required to reach the convergence criteria are reduced to great extent.

Disadvantages:

1. The learning algorithm requires prior specification of the number of cluster centers.
2. The uses of Exclusive Assignment- If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
3. It is unable to handle noisy data and outliers.
4. Applicable only when mean is defined .i.e. fails for categorical data.
5. Algorithm fails for non-linear dataset.
6. Randomly choosing of cluster centers cannot lead us to fruitful results.

V. EXPERIMENTAL RESULTS:

A comparative study on various partition-based algorithms discussed. In this paper, we have chosen *k*-mean and *improved k*-mean method and implemented them in a Windows environment using R. Several self-generated datasets(i.e. synthetic datasets) were taken to evaluate the performance of those methods.

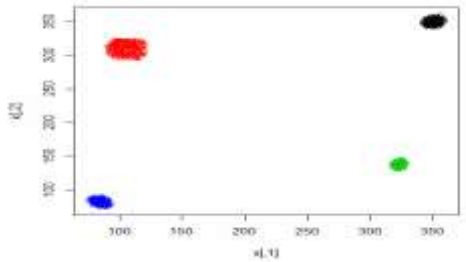

Fig1(a):*k*-means (N=439 and *k*= 4)

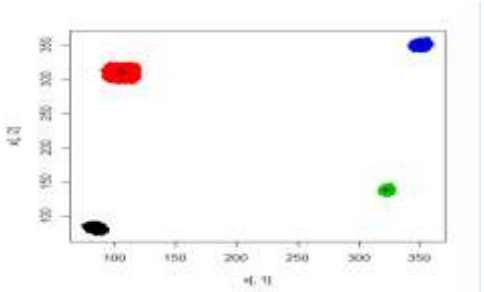

Fig1(b): Improved *k*-mean(N=439and*k*=4)

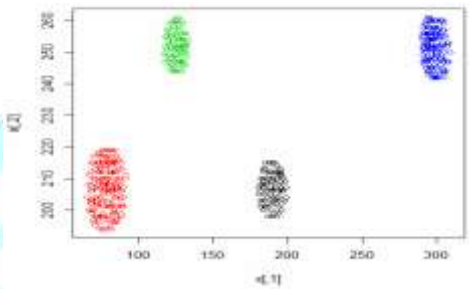

Fig 2(a): k-means(N=514 and *k*=4)

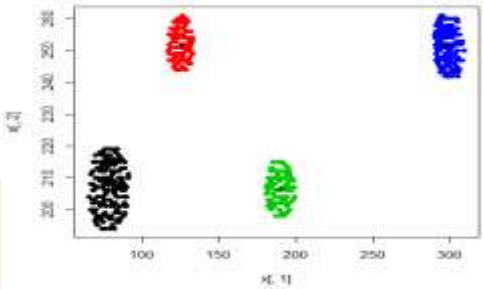

Fig 2(b): Improved k-means (N=514 and *k*= 4)

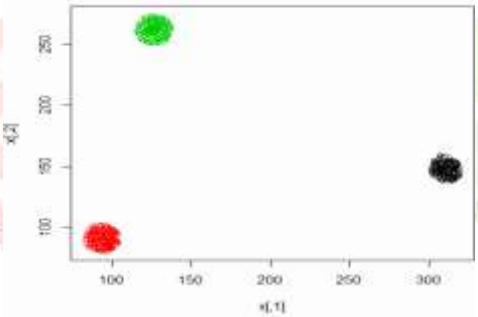

Fig 3(a): k-means(N=461 and*k*=3)

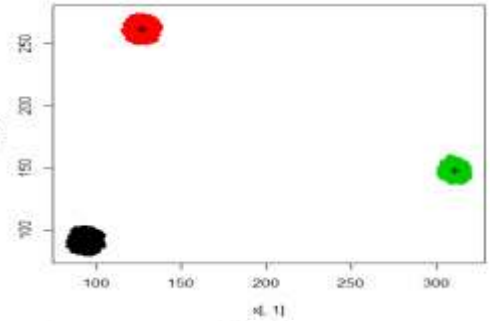

Fig 3(b): Improved k-means(N=461 and *k*=3)

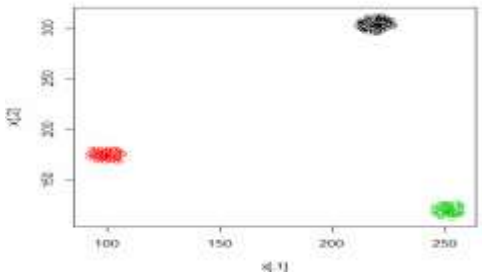

Fig 4(a):(*k*-meansN=177 and *k*=3)

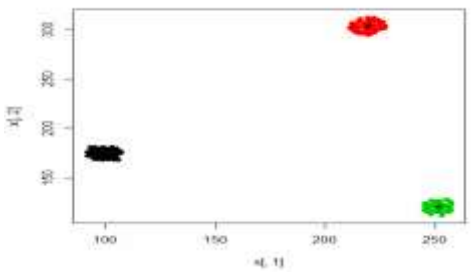

Fig 4(b): Improved *k*-means(N=177 and *k*=3)

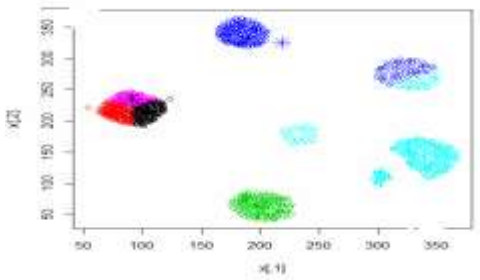

Fig 5(a): k-means(N=1111 and *k*=6)

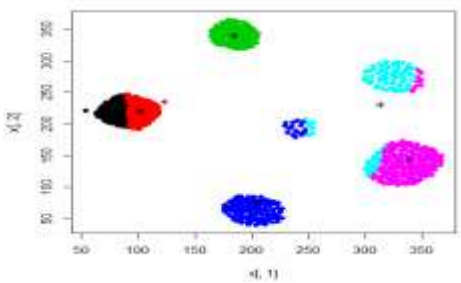

Fig 5(b): Improved *k*-means(N=1111 and *k*=6)

Figure 1(a) and Figure 1(b) represents the k-means clustering for the same dataset with Number of objects (N) = 439 and *k*= 4, where both the algorithms are well capable of showing a good quality of clustering. In Figure 2(a) and Figure 2(b) represents the k-means clustering for the same dataset with N= 514 and c=4 where both the algorithms are well capable of showing a good quality of clustering. In Figure 3(a) and Figure 3(b) represents the k-means clustering for the same dataset with N=461 and c=3 where both the algorithms are well capable of showing a good quality of clustering. Again in Figure 4(a) and Figure 4(b) represents the k-means clustering for the same dataset with N=177 and c=3 where both the algorithms are well capable of showing a good quality of clustering. But in Figure 5(a) and Figure 5(b) represent the k-means clustering for the same dataset with N=1111 where both the algorithms fails to give the good clustering results. As it can be seen that the clustering results are not as desired but the improved k-means algorithm shows little good kind of clustering result as compared to the original k-means algorithm.

Performance Analysis:

PERFORMANCE ANALYSIS

	Time taken(in seconds)	
Data points	Original k-means	Improved k-means
177	0.35	0.42
439	0.48	0.55
461	0.50	0.59
514	0.62	0.66
1111	0.90	0.98

a. Performance analysis with respect to time

In table I, it is observed that the improved k-means algorithm takes more time in clustering, but the results obtained are somewhat tight and distinct as compared to the original *k*-means .i.e. the quality of clustering is improved.

VI. CONCLUSION

k-means is the simplest way to carry out clustering of statistical data. The methods discussed are capable of clustering in a good way for some datasets and not for all. In most of the cases the improved algorithm has shown the good quality result as compared to k-means but the time taken is found to be more. Also, it is seen that both the algorithms fail to give the accurate clustering result for some large datasets. The future scope will be somewhat related to carrying out in the direction of making the algorithm applicable for mixed data types, and also attempts will be taken to make the algorithms better for every type of datasets be it large or small. Efforts will also be given to reduce the time taken for clustering by the algorithms.

VII. REFERENCES

[1] Xingquan Zhu, Ian Davidson,"Knowledge Discovery and Data Mining: Challenges and Realities" ,ISBN978-1-59904-252,Hershey,NewYork,2007.

[2] Pooja BatraNagpal and Priyanka Ahlawat Mann , aug 2011,"Comparative Study of Clustering Algorithms" was published in International Journal of Computer Applications ,volume 27

[3] Joseph,Zernik,"DataMiningasaCivicDuty–Online Public Prisoners Registration Systems" ,International Journal on Social Media: Monitoring, Measurement,Mining,vol.-1,no.-1,pp.84-96,September2010.

[4] GabreillaSchoier and Giuseppe Borruso (2013),On model based clustering in a spatial data clustering mining context.

[5] S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems

[6] Arpit Bansal, Mayur Sharma,Shalini Goel,(2017), Improved k-mean clustering algorithm for prediction analysis using classification technique in Data Mining.

[7] Cheng-Far Tsai and Tang-Wei Huang (2012) QIDBSCAN: A Quick Density-Based Clustering Technique idea International Symposium on Computer, Consumer and Control, pp. 638-641.
[8] S. A. Rauf, S. Mahfooz , S. Khusro , H. Javed , Enhanced –means clustering algorithm to reduce number of iterations and time complexity . Middle-east J. Sci Res. 12(7), 959-963(2012).
[9] K.A.A. Nazeer, M.P. Sebestian ,Improving the accuracy and efficiency of the k-means clustering algorithm , in Proceeding of the World Congress on Engineering , vol 1, ISBN:978-988-17012-5-1,2009.
[10] Md.E. Agha, W.M. Ashour, Efficient and fast initialization algorithm for k-means clustering . Int. J. Intell. Syst. Appl.1,21-31(2012).
[11] M.E. Celebi,H.A. Kingravi , P.A. vVela , A comparative study of efficient initialization methods for the k-means clustering algorithm .J. Exp. Syst. Appl. 40(1),200-210(2013).
[12] M. Yedla, S. Rao Pathakota, T.M. Srinivasa , Enhancing k-means clustering algorithm with improved initial center, IJCSIT 1(2), 121-125(2010).
[13] R.T. Aldahdooh , W. Ashour, DIMK means :distance based initialization method for k-means clustering algorithm. Int. J. Intell. Syst. Appl. (2). 41-51(2013).
[14] S.S. Khan , A. Ahmad , Cluster center initialization algorithm for k-means clustering. Elesvior J. Pattern Recogn Lett 25, 1293 1302(2004).
[15] M. Goyal and S. Kumar,(2014) Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability.
[16] Gholam reza Esfandani, Mohsen Sayyadi and Amin Namadchian (2012) GDCLU: a new Grid-Density based Clustering algorithm, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 102-107.
[17] Amandeep kaurmann and Navneetkaur (2013) , Review paper on clustering techniques, Global journal of computer science and technology software and data engineering.