Enhancing K-means Clustering Algorithm with Improved Initial Center

Madhu Yedla[#1], Srinivasa Rao Pathakota[#2], T M Srinivasa[#3]

[#]*Department of Computer Science and Engineering, National Institute of Technology Calicut*
Calicut, Kerala, India -673601

[1]yedlamadhu@gmail.com, [2]sri.nevaehnitc@gmail.com, [3]srini_2007@nitc.ac.in

***Abstract*— Cluster analysis is one of the primary data analysis methods and k-means is one of the most well known popular clustering algorithms. The k-means algorithm is one of the frequently used clustering method in data mining, due to its performance in clustering massive data sets. The final clustering result of the k-means clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In this paper a new method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. According to our experimental results, the proposed algorithm has the more accuracy with less computational time comparatively original k-means clustering algorithm.**

***Keywords*—** Clustering, Data Mining, Data partitioning, Initial cluster centers, K-means clustering algorithm. Cluster analysis.

I. INTRODUCTION

Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised means that clustering does not depends on predefined classes and training examples while classifying the data objects. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [6], [9]. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Clustering is an crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc.

Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step [9].

Numerous methods have been proposed to solve clustering problem. One of the most popular clustering method is k-means clustering algorithm developed by Mac Queen in 1967. The easiness of k-means clustering algorithm made this algorithm used in several fields. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups [1], [2], [4], [5], [7], [9]. The k-means clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently. However, k-means algorithm is highly precarious in initial cluster centers. Because of the initial cluster centers produced arbitrarily, k-means algorithm does not promise to produce the peculiar clustering results. Efficiency of the original k-means algorithm heavily rely on the initial centroids [2], [5]. Initial centroids also have an influence on the number of iterations required while running the original k-means algorithm. The computational complexity of the original k-means algorithm is very high, specifically for massive data sets [2]. Various methods have been proposed in the literature to enhance the accuracy and efficiency of the k-means clustering algorithm. This paper presents an enhanced method for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity.

This paper is organised as follows. Section 2 presents an overview of k-means algorithm and a short analysis of the existing clustering methods. Section 3 introduces proposed method. Section 4 describes about the time complexity of the proposed method. Section 5 experimentally demonstrates the performance of the proposed method. And the final Section 6 describes the conclusion and future work.

II. THE K-MEANS ALGORITHM

One of the most popular clustering method is k-means clustering algorithm. It generates k points as initial centroids arbitrarily, where k is a user specified parameter. Each point is then assigned to the cluster with the closest centroid [3], [4], [10]. Then the centroid of each cluster is updated by taking the

mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again we calculate new centroids and assign the data points to the suitable clusters. We repeat the assignment and update the centroids, untill convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids [2]. Pseudocode for the k-means clustering algorithm is described in Algorithm 1. The Euclidean distance between two multi-dimensional data points $X = (x_1, x_2, x_3... x_m)$ and $Y = (y_1, y_2, y_3... y_m)$ is described as follows:

$$d(X,Y)=\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + .. + (x_m - y_m)^2} \quad (1)$$

Algorithm 1: The k-means clustering algorithm [2]

Require: $D = \{d1, d2, d3, ..., d_i, ..., d_n\}$ // Set of n data points.
k // Number of desired clusters
Ensure: A set of *k* clusters.
Steps:
1. Arbitrarily choose *k* data points from *D* as initial centroids;
2. **Repeat**
Assign each point d_i to the cluster which has the closest centroid;
Calculate the new mean for each cluster;
Until convergence criteria is met.

Although k-means has the great advantage of being easy to implement, it has some drawbacks. The quality of the final clustering results of the k-means algorithm highly depends on the arbitrary selection of the initial centroids. In the original k-means algorithm, the initial centroids are chosen randomly and hence we get different clusters for different runs for the same input data [10]. Moreover, the k-means algorithm is computationally very expensive also. The computational time complexity of the k-means algorithm is *O(nkl)*, where *n* is the total number of data points in the dataset, *k* is the required number of clusters and *l* is the number of iterations [2]. So, the computational complexity of the k-means algorithm is rely on the number of data elements, number of clusters and number of iterations.

III. Related Work

The original k-means algorithm is very impressionable to the initial starting points. So, it is quite crucial for k-means to have refine initial cluster centers. Several methods have been proposed in the literature for finding the better initial centroids. And some methods were proposed to improve both the accuracy and efficiency of the k-means clustering algorithm. In this paper, some of the more recent proposals are reviewed [1-5], [8].

A. M. Fahim et al. [1] proposed an enhanced method for assigning data points to the suitable clusters. In the original k-means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive. In Fahim approach the required computational time is reduced when assigning the data elements to the appropriate clusters. But in this method the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results.

K. A. Abdul Nazeer et al. [2] proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used, one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time.

Zhang Chen et al. [3] proposed the initial centroids algorithm based on k-means that have avoided alternative randomness of initial center.

Fang Yuan [4] proposed the initial centroids algorithm. The standard k-means algorithm selects k-objects randomly from the given data set as the initial centroids. If different initial values are given for the centroids, the accuracy output by the standard k-means algorithm can be affected. In Yuan's method the initial centroids are calculated systematically.

Koheri Arai et al. [5] proposed an algorithm for centroids initialisation for k-means. In this algorithm both k-means and hierarchical algorithms are used. This method utilizes all the clustering results of k-means in certain times. Then, the result transformed by combining with Hierarchical algorithm in order to find the better initial cluster centers for k-means clustering algorithm.

A. Bhattacharya et al. [8] proposed a novel clustering algorithm, called Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes. DCCA is able to produce clusters, without taking the initial centroids and the value of k, the number of desired clusters as an input. The time complexity of the algorithm is high and the cost for repairing from any misplacement is also high.

IV. Proposed Algorithm

In this section, we proposed an enhanced method for enhancing the performance of k-means clustering algorithm. In the paper [1] authors proposed an enhanced method to improve the efficiency of the k-means clustering algorithm. But in this method the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. In the paper [2] authors proposed an enhanced

algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used, one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters. In the paper [2] the method used for finding the initial centroids computationally expensive. In this paper we proposed a new approach for finding the better initial centroids with reduced time complexity. For assigning the data points we follows the paper [1], [2]. The pseudocode for the proposed algorithm is outlined as Algorithm 2.

In the proposed algorithm first we are checking, the given data set contain the negative value attributes or not. If the data set contains the negative value attributes then we are transforming the all data points in the data set to the positive space by subtracting the each data point attribute with the minimum attribute value in the given data set. Here, the transformation is required, because in the proposed algorithm we calculate the distance from origin to each data point in the data set. So, for the different data points as showed in Fig. 1, we will get the same Euclidean distance from the origin. This will result in incorrect selection of the initial centroids. To overcome this problem all the data points are transformed to the positive space. Then for all the data points as showed in Fig. 1, we will get the unique distances from origin. If data set contains the all positive value attributes then the transformation is not required.

In the next step, for each data point we calculate the distance from origin. Then, the original data points are sorted accordance with the sorted distances. After sorting partition the sorted data points into k equal sets. In each set take the middle points as the initial centroids. These initial centroids lead to the better unique clustering results. Next, for each data point the distance calculated from all the initial centroids. The next stage is an iterative process which makes use of a heuristic approach to reduce the required computational time. The data points are assigned to the clusters having the closest centroids in the next step. ClusterId of a data point denotes the cluster to which it belongs. NearestDist of a data point denotes the present nearest distance from closest centroid.

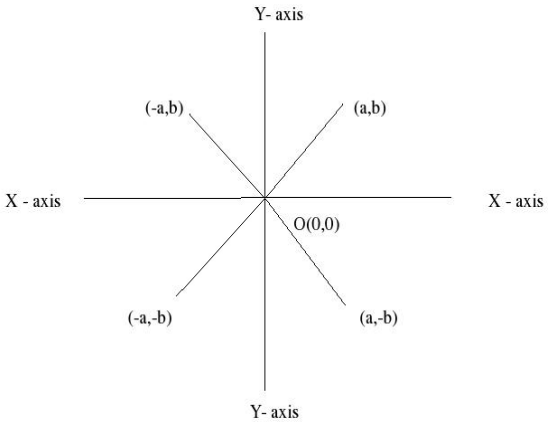

Fig. 1. Data points in Two Dimensional Space

Algorithm 2: The Enhanced Method

Require: $D = \{d_1, d_2, d_3,..., d_i,..., d_n\}$ // Set of n data points.
$d_i = \{ x_1, x_2, x_3,..., x_i,..., x_m \}$ // Set of attributes of one data point.
k // Number of desired clusters.

Ensure: A set of k clusters.

Steps:

1: In the given data set D, if the data points contains the both positive and negative attribute values then go to step 2, otherwise go to step 4.
2: Find the minimum attribute value in the given data set D.
3: For each data point attribute, subtract with the minimum attribute value.
4: For each data point calculate the distance from origin.
5: Sort the distances obtained in step 4. Sort the data points accordance with the distances.
6: Partition the sorted data points into k equal sets.
7: In each set, take the middle point as the initial centroid.
8: Compute the distance between each data point d_i $(1 <= i <= n)$ to all the initial centroids c_j $(1 <= j <= k)$.
9: **Repeat**
10: For each data point d_i, find the closest centroid c_j and assign d_i to cluster j.
11: Set ClusterId[i]=j. // j:Id of the closest cluster.
12: Set NearestDist[i]= $d(d_i, c_j)$.
13: For each cluster j $(1 <= j <= k)$, recalculate the centroids.
14: **For** each data point d_i,
14.1 Compute its distance from the centroid of the present nearest cluster.
14.2 If this distance is less than or equal to the present nearest distance, the data point stays in the same cluster.
Else
14.2.1 For every centroid c_j $(1<=j<=k)$ compute the distance $d(d_i, c_j)$.
End for;
Until the convergence criteria is met.

Next, for each cluster the new centroids are calculated by taking the mean of its data points. Then for each data point the distance calculated from the new centroid of its present nearest cluster. If this distance is less than or equal to the previous nearest distance, then the data point stays in the same cluster, otherwise for each data point we need to calculate the distance from all centroids. After calculated the distances, the data points are assigned to the appropriate clusters and the new ClusterId's are given and new NearestDist values are

updated. This reassigning process is repeated until the convergence criterion is met.

V. TIME COMPLEXITY

The required time complexity of proposed algorithm for finding the initial centroids is *O(nlogn)* in both average and worst case, where n is the number of data points. The sorting method used for sorting determines the overall time complexity for finding the initial centroids. Since the proposed enhanced method uses heap sort, its overall time complexity becomes *O(nlogn)* in both average and worst case. To get the initial clusters the required time complexity is *O(nk)*. Here, some data points stay in the cluster itself and some other data points move to other clusters based on their relative distance from old centroid and the new centroid. If the data point stays in the same cluster then the required complexity is *O(1)*, otherwise *O(k)*. In each iteration the moving of data points to other clusters is decreases. Assuming, until the convergence criteria is met, half the data points move to the other clusters from their present clusters, this requires *O(nk/2)*. Hence the total time complexity for assigning the data points is *O(nk)*, not *O(nkl)*. Therefore the total time complexity of the proposed algorithm becomes *O(nlogn)*. Hence the proposed algorithm has less time complexity compared to the original k-means clustering algorithm.

VI. EXPERIMENTAL RESULTS

We tested both the algorithms for the data sets with known clustering, Iris [11], New Thyroid [11], Echocardiogram [11], Height-Weight [12] and Diggle[13]. The same data sets are used as an input for the original k-means algorithm. Both the algorithms need number of clusters as an input. In additional, for the original k-means algorithm the set of initial centroids also required. The enhanced method finds initial centroids systematically. The enhanced method requires only the data values and number of clusters as inputs. And it does not take any additional inputs like threshold values.

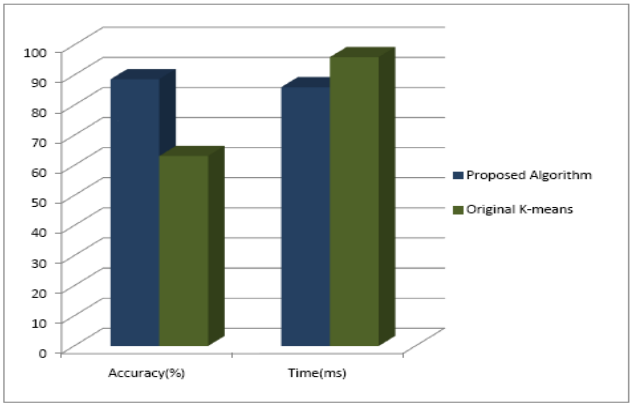

Fig. 2. Performance Comparison chart for Iris data

The original k-means algorithm is executed seven times for different sets of values of the initial centroids. In each experiment the accuracy and time was computed and taken the average accuracy and time of all experiments. Table 1 shows the performance comparison of the algorithms. The results also showed with the help of bar charts in the Fig. 2, 3, 4, 5 and 6. The results obtained show that the proposed algorithm is producing better unique clustering results compared to the k-means algorithm in less amount of computational time.

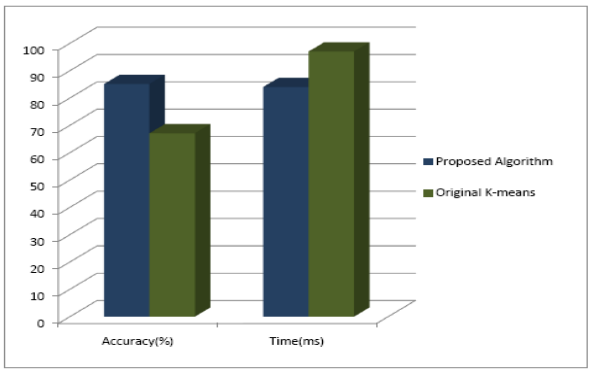

Fig. 3. Performance Comparison chart for New Thyroid data

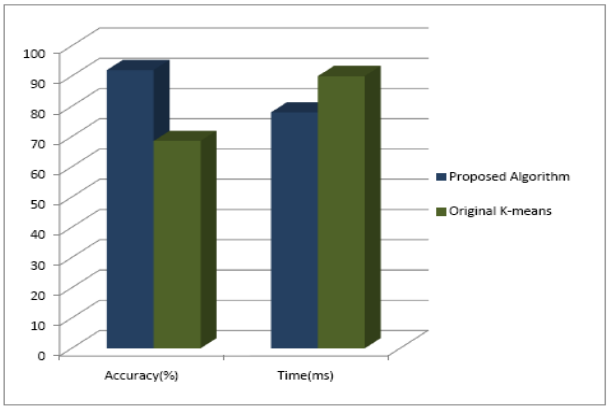

Fig. 4.Performance Comparison chart for Height-Weight data

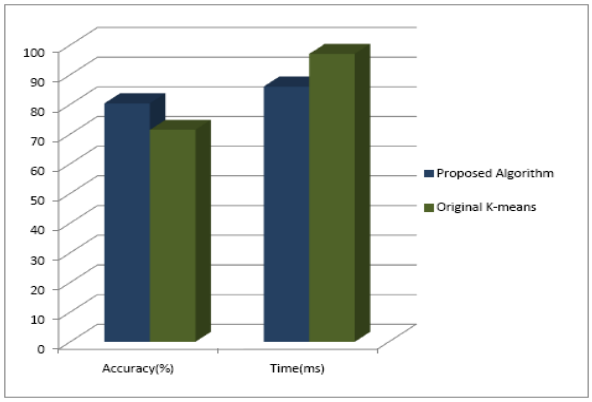

Fig. 5. Performance Comparison chart for Echocardiogram data

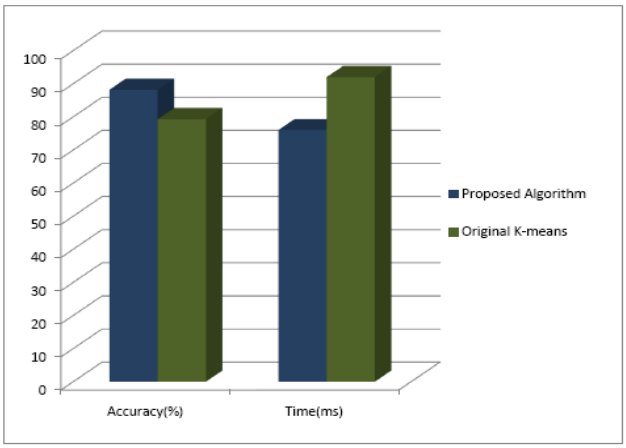

Fig. 6. Performance Comparison chart for Diggle data

TABLE I
PERFORMANCE COMPARISON OF THE ALGORITHMS

Data Set	Number of Clusters	Algorithm	Run	Accuracy (%)	Time Taken (sec)
Iris	K = 3	Original K-means	7	63.14	0.096
		Proposed Algorithm	1	88.66	0.086
New Thyroid	K = 3	Original K-means	7	67.10	0.097
		Proposed Algorithm	1	85.11	0.084
Height-Weight	K = 4	Original K-means	7	68.57	0.090
		Proposed Algorithm	1	92	0.078
Echocardiogram	K = 2	Original K-means	7	71.42	0.097
		Proposed Algorithm	1	80.34	0.086
Diggle	K = 2	Original K-means	7	79.3	0.092
		Proposed Algorithm	1	88.19	0.076

VII. CONCLUSION

One of the most popular clustering algorithm is k-means clustering algorithm, but in this method the quality of the final clusters rely heavily on the initial centroids, which are selected randomly. Moreover, the k-means algorithm is computationally very expensive also. The proposed algorithm is found to be more accurate and efficient compared to the original k-means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters. This method ensures the total mechanism of clustering in O(nlogn) time without loss the correctness of clusters. This approach does not require any additional inputs like threshold values. The proposed algorithm produces the more accurate unique clustering results. The value of k, desired number of clusters is still required to be given as an input to the proposed algorithm. Automating the determination of the value of k is suggested as a future work.

REFERENCES

[1] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," *journal of Zhejiang University*, 10(7): 16261633, 2006.

[2] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in *International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009)*, Vol 1, July 2009, London, UK.

[3] Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," *in Second International Workshop on Knowledge Discovery and Data Mining* (WKDD), pp. 790-792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," *proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 26-29, August 2004.

[5] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," *department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University*, Vol. 36, No.1, 2007.

[6] S. Deelers and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," *International Journal of Computer Science*, Vol. 2, Number 4.

[7] Mc Queen J, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, (1): 281–297, 1967.

[8] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," *bioinformatics*, Vol. 24, pp. 1359-1366, 2008.

[9] Margaret H Dunham, Data *Mining-Introductory and Advanced Concepts*, Pearson Education, 2006.

[10] Elmasri, Navathe, Somayajulu, Gupta, *Fundamentals of Database Systems*, Pearson Education, First edition, 2006.

[11] (2010) The UCI Repository website. [Online]. Available: http://archive.ics.uci.edu/

[12] Height-Weight Data. (2010). [Online]. Available: http://www.disab-ledworld.com/artman/publish/height-weight-teens.shtml

[13] Diggle Data. (2010). Available: http://lib.stat.cmu.edu/datasets/diggle