ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
Research Article
www.ijptonline.com

IMPROVEMENT ON INITIAL SEED SELECTION FOR K-MEANS ALGORITHM

Aman Shrivastav, Bharat Nalwaya, Sudish Kumar, Rajaprabha M N
SITE, VIT University, Vellore, Tamilnadu, India.
Email: aman.shrivastav2015@vit.ac.in

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract:

Clustering is the technique used to group set of objects in a way such that each object or data point in the group (called a cluster) is similar to each other in the same cluster is highly dissimilar to other objects of another cluster. One of the method of clustering is k-means algorithm which groups the data point or objects using the mean values. One major problem of the k-means algorithm is that it starts with random points selection as the centre (also called Centroids or Seeds). It causes an extremely high negative effect on the performance of the algorithm. In this paper, we are going to propose a method for selection of initial centres which will improve the performance of the algorithm.

Keywords: Centroids, Clustering, k-means, Seeds.

1. **Introduction:** Clustering is the activity of ordering a given set of data items into a set of groups which are highly dissimilar to other group called clusters. It makes the data items within a group are more related to each other and unrelated to the points of other groups. Clustering is one of the active area of research, which is applicable for many fields such as bioinformatics, fraud detection, marketing, pattern recognition, data mining, image processing, economics, etc. Cluster analysis is an important tool in data analysis. Mainly clustering algorithms are divided in two categories: Partitional algorithms and Hierarchical algorithms. In a hierarchical clustering technique, whole dataset is divided into smaller datasets in hierarchical manner. In partitional clustering technique, whole dataset is divided into desired number of smaller dataset in a single step. K-means is a clustering technique used in partitioning manner. In K-means algorithm clustering is done based on Euclidean distance of each data points from the cluster centres. In these algorithm, centres to start the algorithm are chosen randomly which is the major disadvantage of this algorithm. Due to random selection of initial centres (or initial seeds), algorithms take more number of iteration to complete

which causes the reduction in the performance. This paper has proposed the method for selection of initial centres, which improves the performance of algorithm.

2. Literature Survey

Md. Sohrab Mahmud, Md. Mostafizer and Md. Nasim Akhtar [1] proposed an improved k-means algorithm according to which we first calculate the weighted average score of data by multiplying weight with each attribute then adding these values and dividing the sum by total number of data objects. After this we sort the entire average score using merge sort. After sorting is done, we arrange the data object and calculate the mean then we take the nearest possible data point of the mean as initial centroid. This algorithm completes the k-means algorithm in less number of iteration because the initial centres are calculated in a strategic way not in a random way. Anand M. Baswade and Prakash S. Nalwade [2] proposed a method for selection of initial centroids that has overcome the traditional method of k-means algorithm. According to this method first the average of given set of objects is calculated and chosen as the first initial centroid. Now from the given set of objects we choose the next centroid in such a way that its Euclidean distance is maximum from other centroids and this process is repeated till we find all desired initial centres. K. KarteekaPavan, A.V. Dattatreya Rao, G.P. Sridhar and AllamAppa Rao [3] proposed the algorithm which is also related to improvement of k-means algorithm. It is an outlier insensitive algorithm. It also performs well on synthetic and real data sets. According to this algorithm, a set C is chosen with k initial centres from given n data objects. After that a distance matrix of order *nXn* is obtained by calculating the distance of each point from all other points. After that a sum matrix is obtained with entries as the sum of the distance of a point from all other points. A high-density point is chosen from sum matrix and set as the first initial centre. Then the distance of all m/k points is calculated from the nearest point in set C and their distance is added up and the minimum distance point is chosen as the next centroid. We repeat this step till all the k initial centres are found. Then the conventional method of k-means algorithm is followed. So, SPSS is also known as the Single Pass algorithm with unique solution.

3. Proposed Algorithm

In this enhanced method of improved k-means algorithm, we calculate the distance of each data point from the origin and multiply the distances by its weightage. Weightage of the data point is the number of attributes of that data point. So, we get a unique weighted distance for each data point. In the subsequent step, weighted distances are sorted and according to this sorted distances, we sort the original data points. Now we select the required k centres in a systematic way such that all centres are chosen from n/k difference positions of the sorted order. After that we follow the traditional k-means algorithm with this chosen cluster centre.

Input: Dataset = {d_1, d_2, d_3..., d_i…., d_n} set of n data points.

d_i = {a_1, a_2, a_3…., a_i…., a_m} set of attributes of one data point.

k number of required clusters.

Output: set of k clusters.

4. Algorithm

Step 1: Calculate the distance of each data point from the origin.

Step 2: Find the weighted distance W_{di} for each data item d_i as following:

$$W_{di} = \sum_{i=1}^{m}(ai * dist(di))$$

Step 3: Sort obtained weighted distances and according to this sorted distances, sort the original data points.

Step 4: Choose the initial centre with the n/k different positions from one centre to next centre as following: (n + k)/2k, (3n +k)/2k, …., (n (2k -1) + k)/2k

Step 5: Calculate the distance of each data point di with all initial centroids.

Step 6: Assign data point di to the cluster which has the minimum distance.

Repeat

Step 7: Calculate new cluster centres for each cluster by calculating the mean of the data points in the cluster.

Step 8: Calculate the distance of each data point d_i with all initial centroids.

Step 9: Assign data point d_i to the cluster which has the minimum distance.

Until: no data points move across the cluster.

5. Results

Testing of the proposed algorithm is done some of the well-known datasets such as: Ecoli, New Thyroid, Iris, Breast Cancer Wisconsin(Original), Echocardiogram and Height-Weight. We have performed original k-means algorithm on iris dataset 10 times, Breast Cancer Wisconsin(Original) dataset 5 times, Height-Weight dataset 7 times, Ecoli dataset 7 times, Echocardiogram dataset 7 times and New Thyroid dataset 7 times.

Now we have performed proposed algorithm once for each dataset. Accuracy and time taken by algorithms is shown in table 1.

6. Conclusion

Original k-means algorithm is sensitive to selection of initial centres. Due to this, performance of k-means algorithm is reduced. The results obtained with the use of proposed method helps the k-means algorithm to produce better clusters in less time taken as compare to traditional k-means algorithm.

Following figures show the performance of proposed algorithm in graphically manner:

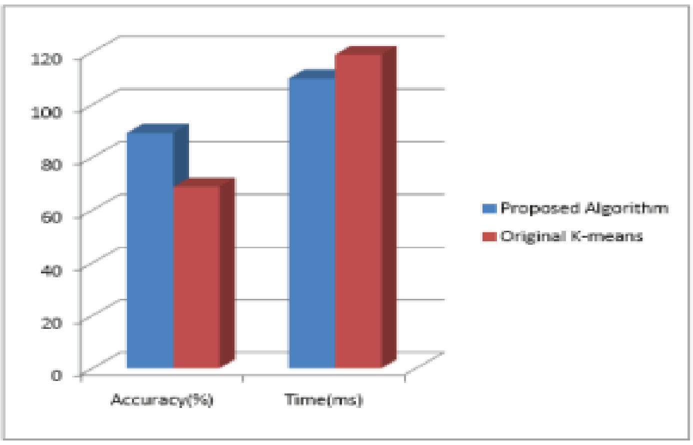

Figure 1. Iris Dataset Performance Comparison.

Table 1.Ecoli Performance Comparison.

Data Set	Number of Clusters	Algorithm	Run	Accuracy (%)	Time Taken (sec)
Ecoli	K = 3	Orinal K-means	7	77.14	0.127
		Proposed Algorithm	1	91.91	0.118
New Thyroid	K = 3	Orinal K-means	7	73.15	0.120
		Proposed Algorithm	1	86.04	0.113
Echocardiogram	K = 2	Orinal K-means	7	71.42	0.109
		Proposed Algorithm	1	82.24	0.102
Height-Weight	K = 4	Orinal K-means	7	70.28	0.102
		Proposed Algorithm	1	92	0.088
Breast Cancer Wisconsin (Original)	K = 2	Orinal K-means	5	96.19	0.142
		Proposed Algorithm	1	96.19	0.136
Iris	K = 3	Orinal K-means	10	68.93	0.119
		Proposed Algorithm	1	89.33	0.110

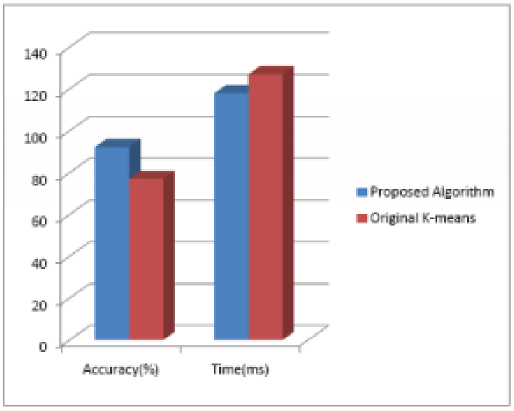

Figure 2.Ecoli Dataset Performance Comparison.

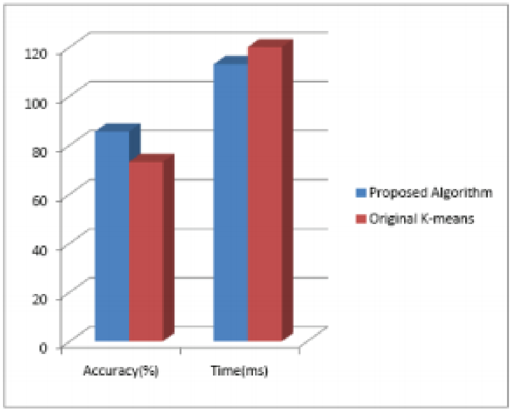

Figure 3. New Thyroid Dataset Performance Comparison.

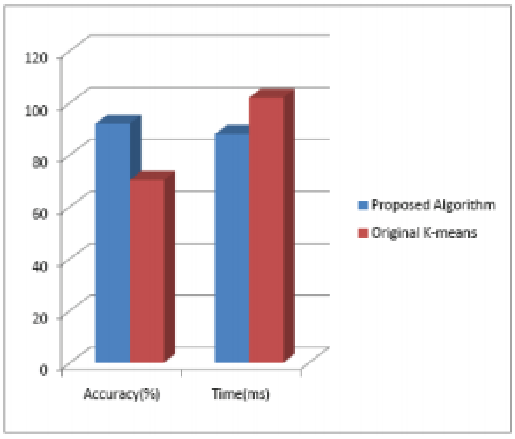

Figure 4. Height-Weight Dataset Performance Comparison.

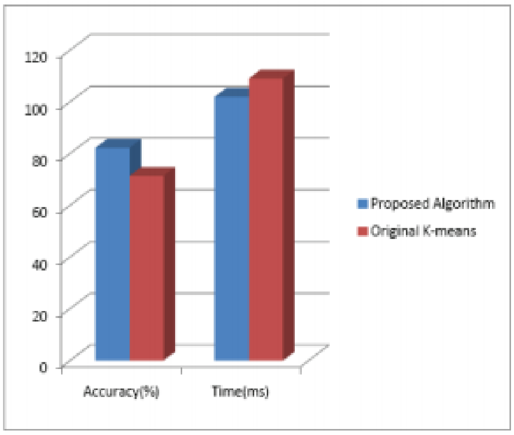

Figure 5. Echocardiogram Dataset Performance Comparison.

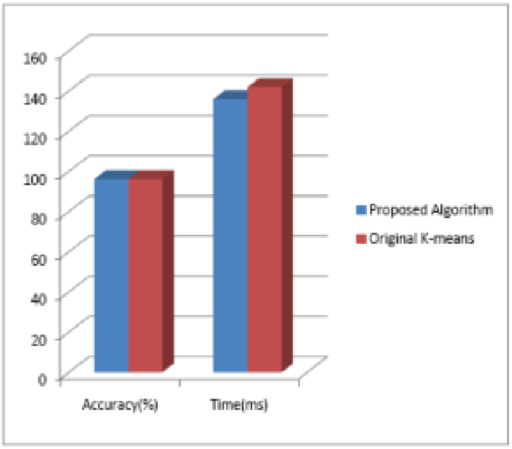

Figure 6. Breast Cancer Wisconsin (original) Dataset Performance Comparison.

7. References

1. Improvement of k-means based on weighted average by Md. Sohrav Mahmud, Md. Mostafizer and Md. Nasim Akhtar (20-22 December, 2012).
2. Selection of initial centers for k-means algorithm by Anand M. Baswade and Prakash S. Nalwade (7 July, 2015).
3. Single Pass Seed Selection Algorithm by K. KarteekaPavan, AllamAppa Rao, A.V. Dattatreya Rao and G.P. Sridhar (2011).
4. Data Mining Concepts and Techniques (Second Edition) by Jiawei Han and Micheline Kamber.
5. Improving the accuracy and efficiency of the k-means clustering algorithm by Abdul Nazeer, K A., Sebastian, M P (2009) in *International Conference on Data Mining and Knowledge Engineering (ICDMKE)*, Proceedings of the World Congresson Engineering (WCE-2009), V. 1, July 1-3, London, U.K.
6. Data Mining-Introductory and Advanced Concepts by Dunham, Margaret H (2006), Pearson Education.
7. Iris, Ecoli, New Thyroid, Echocardiogram and Breast Cancer Wisconsin(Original) data sets are available at http://archive.ics.uci.edu/ml/machine-learning-databases, (accessed on 20-10-16).
8. Height-Weight Data available at http://www.disable-world.com/artman /publish/height-weight-teens.shtml, (accessed on 20-10-16).
9. K-means Clustering Algorithm with Improved Initial center by Zhang, Chen., Xia, Shixiong. (2009), In: *Second International Workshop on Knowledge Discovery and Data Mining (WKDD),* p.790-792.
10. A New Algorithm to Get the Initial Centroids by Yuan, F Meng, Z H Zhangz, H X Dong, C R (2004), In: Proc. of the 3rd International Conference on Machine Learning and Cybernetics, p. 26-29, August.