***I.J. Modern Education and Computer Science*, 2015, 6, 37-42**
Published Online June 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.06.06

A Proposed Modification of K-Means Algorithm

Sharfuddin Mahmood
American International University- Bangladesh, Dhaka, 1213, Bangladesh
Email: smahmood@aiub.edu

Mohammad Saiedur Rahaman, Dr. Dip Nandi, Mashiour Rahman
American International University- Bangladesh, Dhaka, 1213, Bangladesh
Email: msr.aiub@gmail.com
Email: {dip.nandi, mashiour}@aiub.edu

Abstract—K-means algorithm is one of the most popular algorithms for data clustering. With this algorithm, data of similar types are tried to be clustered together from a large data set with brute force strategy which is done by repeated calculations. As a result, the computational complexity of this algorithm is very high. Several researches have been carried out to minimize this complexity. This paper presents the result of our research, which proposes a modified version of k-means algorithm with an improved technique to divide the data set into specific numbers of clusters with the help of several check point values. It requires less computation and has enhanced accuracy than the traditional k-means algorithm as well as some modified variant of the traditional k-Means.

Index Terms—Clusters, clustering algorithms, Euclidian distance, Data Mining

I. Introduction

Nowadays data have become as an asset for human beings. Almost in every sector data is stored for future utilization. As the data grows in size, the technique to utilize these data becomes more challenging.

This large amount of data is then used for discovering new knowledge. Essentially, data mining (DM) utilizes these large data warehouses to extract the information [1]. Data mining is a technique by which users try to gain knowledge from the stored data. It is frequently used in knowledge discovery system. Additionally data mining is used to design artificially intelligent systems, machine learning process and statistical systems [2] [3].

Clustering is an application of Data mining. Clustering is concerned with grouping together objects that are similar to each other but different from the object that belongs to other clusters [4]. It is a way to classify raw data reasonably so that one can find out the hidden pattern that may exist in the dataset [5]. Clustering is one of the key technologies used in data mining field and also in machine learning sector. It is also utilized in many areas such as knowledge discovery, pattern recognition and classification, data compression and vector quantization. It also plays an important role in the field of biology, geology, geography and marketing [6].

To assign an object into a specific cluster requires extensive calculation. One of the most popular algorithms for doing so is the K-Means algorithm. In this algorithm, K-random data points are chosen as cluster centers. Then the Euclidian distance of every member is calculated from the k-cluster centers. Members are then assigned to any cluster by this distance.

This process is repeated several times unless any object moves to another cluster [7]. This algorithm is very popular for distributing data objects into desired clusters. However the main problem of this algorithm is that it requires a high number of computations and hence takes an extensive time.

There are many improved version of k-means algorithm proposed by the researchers. In the paper [8] [9], researchers proposed to maintain two data structures to save the current minimum distance and current assigned cluster name. This process reduces the computation in a large scale and provided a better performance than the traditional K-means algorithm. But the main problem of this algorithm was that the "current minimum distance" is not always the correct minimum distance.

As mentioned above, the main objective of this paper is to propose modification of the k-means algorithm which will have the same functionality as the traditional k-means algorithm. But it will overcome the problems discussed in the previous section. A more efficient and less computationally expensive algorithm will be proposed in this paper.

In different sections of this paper, the total procedure and the way of advancement of the work along with the related background studies are explained elaborately. Here is the glimpse of the paper-

In Background Studies, related previous works are presented in a summarized way. The work related to the improved k-means clustering algorithm, improved initial center theory and better initial cluster concepts are focused here.

In Proposed Algorithm section, the algorithm is proposed. The algorithm and some graphical descriptions are provided here.

The Result Analysis section is focused on result analysis. The outcome of the proposed algorithm is discussed here along with the comparison with other modification of traditional K-means algorithm.

And finally the draw backs and future work is described in the conclusion section.

II. Related Works

Before starting the proposed algorithm, it is essential to discuss about the related background studies. In this section the related ideas and researches are discussed in brief. A lot of work has been done with the k-means algorithm. Here it will be discussed from the very beginning for a better understanding.

In modern world data is becoming more and more important day by day. Large organizations tend to keep their data safe and store them in such a way that it can be used in the future. As the volume of the data grows, the need to preserve it in a structured way is becomes challenging. In data mining technology, this challenge is handled in a structural way. In this technology data are mined in such a way that useful information can be obtained from this data if needed. In present world data is a continuous by product of every business that can be used for good.

Clustering is a process that organizes a set of objects into disjoint classes. These classes are referred as clusters [9]. It is a version of unsupervised classification and hence does not depend on predefined classes. It tries to partition a data set based on specific features so that within a cluster the objects are more similar than the object in different one [4] [10-11].

K-means algorithm is a partition-based clustering method [12-15].In the traditional k-means algorithm k random points are chosen from the whole dataset as the primary cluster center. Then the distance between each object and each cluster center is calculated. Finally each object is assigned to the nearest cluster. To find the distance Euclidian Formula is used.

So the whole algorithm can be described as below [16]:

Input: N-objects and number of cluster K

Output: K- Custer each having 0<n<N data members.

Arbitrarily select k objects as initial luster centers.

Calculate distance between each object xi and each cluster center.

Assign each object to the nearest cluster. For calculating the distance Euclidian formula is used. The formula is:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_i)^2}$$

Where i=1…..N; j=1……..k;

d(xi, mi) is the distance between data i and cluster j.

Calculate the average of every object in each cluster as new cluster center, using the following formula.

$m_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$Where i=1,2……..k; Ni is the number of object in current cluster i.

Repeat until convergence criterion is met.

In traditional K-Means algorithm a lot of calculations are needed as this a brute force algorithm. As a result we have to compute the distance of each object from its cluster center whether or not it moves in or out from the cluster. Even for the closest object from one cluster center needs to be make sure that it is not close to any other cluster center. That results in a large number of calculations.

Reference [8] [9] shows that the researchers proposed an enhanced algorithm that have an improved initial center. Here they sorted the whole data set according to the distance calculated from the origin. Then the sorted dataset is divided into equal k-portions. This value of k depends on the dataset as well as the need of the users. After the center point of each dataset is taken as the initial cluster center, all other attribute distance is calculated with respect to these centers using Euclidian distance formula. After that the data points are assigned to the nearest cluster center also known as centroid. Two data structures are used in this method: One for storing the label of current cluster center and other for saving the distance from current cluster center. After the first iteration the cluster center is recalculated and the distance is measured again. If the distance is less than the distance calculated in the previous stage then the cluster center and the distance is updated otherwise the point remains in the same cluster. This process continues until convergence criterion is satisfied. These two data structures turn this modification into a better algorithm. As the performance raises and the algorithm provides a better accuracy than the traditional k-means algorithm. But the problem of this algorithm is that the current minimum distance of the previous iteration may not be the best minimum distance for next iteration. As every time the cluster centers are recalculated, the distance of every object changes. But using the minimum distance from the previous iteration may lead to some miscalculation.

Reference [17] shows that the researchers proposed an algorithm for finding a better initial center as well as the number of initial center that is the value of K. Here the algorithm is based on k-means with the ability to avoid the alternative randomness. With the help of sub merger strategy the categories are combined and resultant algorithm gave a better performance than the traditional k-means algorithm.

III. Proposed Algorithm

In the previous section it was mentioned that a new and better variant of K-Means algorithm will be proposed in this paper. In this section the algorithm and its definition will be provided. Some figures will also be there for a better understanding.

In the traditional K-Means algorithm, the distance between a cluster center and each data point is measured in every iteration. This makes the algorithm more complex and increases the number of computations. To reduce that, a check point value will be used in the proposed algorithm. This will reduce the computation in a large scale. To start with:

Consider a data set where N-data members/objects are available. Each object have Di (i=1…..n) attributes. The output will be K clusters. K will be defined by the user as required.

The Proposed algorithm:

- ➔ Step 1: Find the Euclidian distance of each data object from the origin (0i,0j……0n).

Here we randomly select N data objects as initial origin. Then we find out the Euclidian distance between each data object with respect to the origin.

- ➔ Step 2: Sort the N-data object in ascending order according to the distance found in the previous step.
- ➔ Step 3: Divide the data set into K equal clusters. K will be determined according to the user requirement or on the type of the data set. This will act as the primary cluster.

For setting up the initial cluster this step is necessary. Depending on the number of cluster needed we now divide the whole data set into equal portion. For every situation this may not be the case as there may not be equal numbers of object in every data set. As example, if we have 1000 data objects and we have to divide them into 3 clusters then the cluster may have 333,333,334 numbers of objects in each of it.

- ➔ Step 4: For each cluster, consider the middle point as the primary cluster center. That is, if there is N data members and K clusters, the primary cluster center will be Γ ((n/k)/2)˥th object.

As this data set is obtained from the distance from the initial origins, so the center points will be the most significant points in each clusters from which all the data objects will mostly have a unified distance.

- ➔ Step 5: Find the distance between the cluster centers. If there are K clusters, there will be K-distances. Divide the distance by 2 and store the value in Dij (i, j=0,1,…k). Here Dij denotes the middle point of the distance from cluster center i to cluster center j. This Dij will be used as a check point value.

For example, if cluster A and B have cluster centers A_i and B_i, and suppose the center point of the distance between A_i and B_i will denote a point where the distance is equal from both cluster centers. As a result it can be utilized to determine the new cluster for any data objects if needed.

- ➔ Step 6: Find the Euclidian distance of each data object di (i=1. k) from the cluster center it is assigned to.
- ➔ Step 7: Compare di with the distance stored in Dij.

If the distance is less than or equal to Dij, then the object stays in the previous cluster.

That is, the distance from the current cluster center is less than the distance from the center points of two cluster centers. As a result we can conclude that this object is closer to its current cluster. Hence we do not need to calculate the distance from other cluster center. This check point value will ensure that we need less computation.

Else calculate the Euclidian distance of the data object with respect to the center with which the distance crossed the check point value. That is, if Dij is exceeded and the object was previously in the cluster with center i, then compute the distance with respect to cluster center j.

Means the object may be closer from the other cluster center. To be sure about this, we have to calculate the distance with respect to other cluster center.

Now compare the distances. Assign the data object to the cluster from whose center it has a shorter distance.

Recalculate the cluster centers by taking the mean of every objects currently present in one cluster. This point can be an imaginary point which has no existence in our current data set or can be any current object of our dataset. This will not affect the outcome of our algorithm.

Go back to step 4 and repeat until the convergence criteria is met. That is no data object is moving from one cluster to another cluster after the cluster center is changed. That results in the object of the cluster remains same, hence the center also remains unchanged. Now we can draw the conclusion that we have achieved the final clusters. That is we grouped together similar objects in each, which may be different from other clusters.

Fig. 1 shows two cluster centers 1 and 2 and one object Ni. The figure also shows the center point D_{12} between the cluster centers. We assume that the object Ni was previously assigned to cluster center 1. Now we find the distance of Nij from cluster 1. If the distance is less than or equal to the mean distance of two centers (D_{12}), then the object stays in the previous cluster.

Fig. 2 shows the other possibility. If the current distance of any object with respect to the cluster center is greater than the mean distance of two clusters then we calculate the Euclidian distance of the object with respect to the other cluster center. If the new distance is less than the previous one, then the object moves to the new cluster.

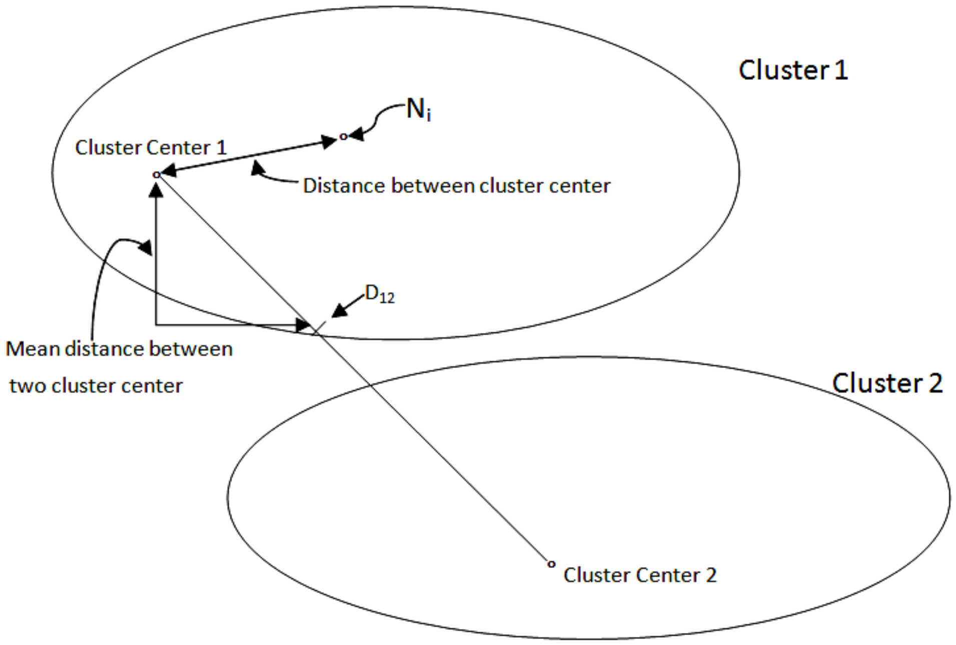

Fig: 1. Determining the cluster center depending on the distance from the center

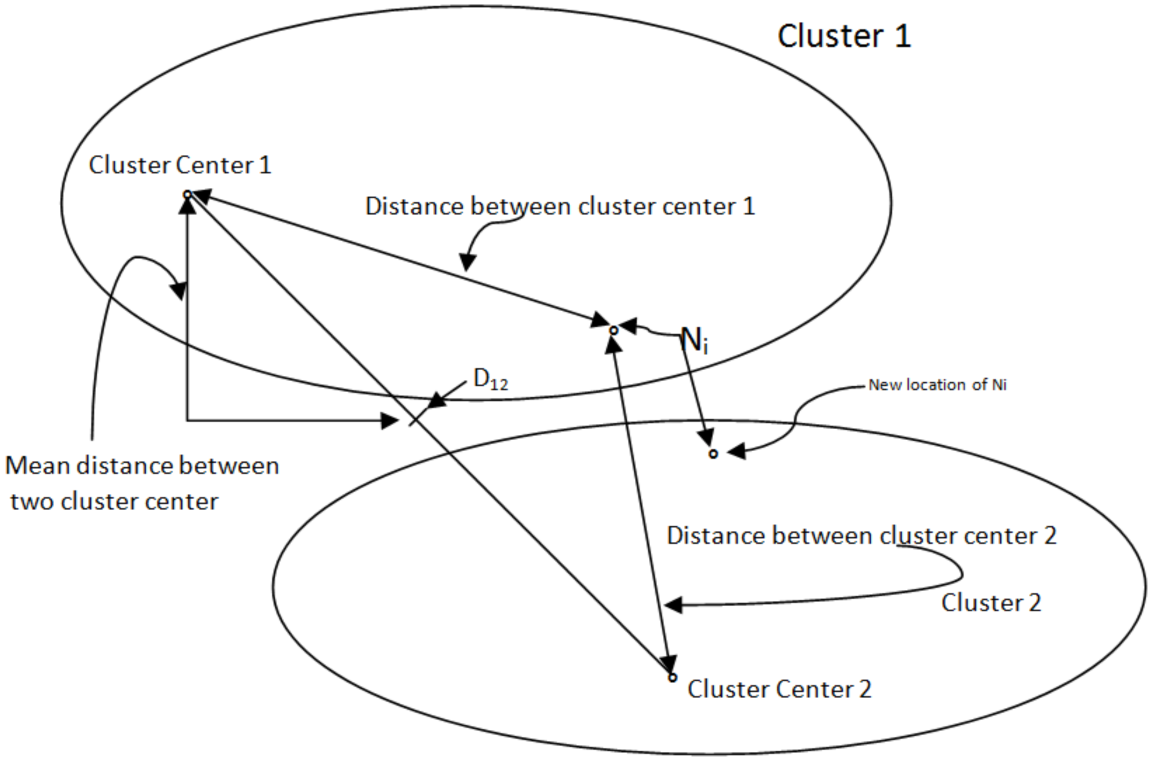

Fig: 2. Determining the cluster center depending on the distance from the center

IV. Result Analysis

In this section we analyze the effectiveness and the accuracy of our proposed algorithm. For the experiment we used Iris, New Thyroid and Echocardiogram data sets [18]. These data sets were also used in reference [9]. Therefore we can have more evident result.

Table 1. Performance Comparison of the Algorithms

Data Set	Algorithm	Accuracy (%)
Iris	Original K-Means	63.14
	Algorithm of [9]	88.66
	Proposed Algorithm	90.0
New Thyroid	Original K-Means	67.10
	Algorithm of [9]	85.11
	Proposed Algorithm	85.11
Echocardiogram	Original K-Means	71.42
	Algorithm of [9]	80.34
	Proposed Algorithm	80.64

For IRIS dataset, the tradition K-means algorithm had an Accuracy of 63.14%. In the research proposed in [9], their accuracy was about 88.66%. With the help of our check point value, our algorithm has 90% accuracy.

For New Thyroid dataset the accuracy of traditional K-means algorithm is 67.10%. [9] and our algorithm both have the accuracy of 85.11%.

For Echocardiogram dataset the traditional algorithm had 71.41% accuracy. On the other hand [9] had 80.34 % and our algorithm have 80.64% of accuracy.

We used the center point of two clusters as our check point value. As clusters are created to group similar objects together, there is no possibility that two clusters may intersect each other. Using this theory we designed our check point value which will compare the distance of each object with respect to the cluster center. If we consider our clusters as a circle, then the middle point of two clusters is equally distant from each cluster center. We used this value to check whether the object is closed to first cluster or the second cluster. This theory reduced the problem of current minimum distance of research [8] [9].

From Table 1, it is the evident that our proposed algorithm provides better performance than the traditional k-means algorithm. It is also performing equally or better than the improved version of k-means algorithm proposed in [9].

As the second and third data set had missing values in it, we had to find out the missing values using traditional algorithm .There are lot of different ways to find out or get rid of the missing values in one data set. Each of the produces different kind of result that can be used to gain information from one data set. But in paper [9], they did not mention the method through which they found out the missing values. As a result our procedure may have produced different result than theirs. This prevented us to acquire desired result from our algorithm.

The first data set was a complete one and for that data set, our algorithm produces superior result than theirs. This is also an evident that if the remaining data sets were complete in terms of data values, we could have produced more superior result with our algorithm.

V. Conclusion

K-means algorithm is very popular algorithm in knowledge discovery systems. It helps user to divide data into similar clusters according to the similarity in nature. That is the members of same cluster shows similarity in their characteristics which may different from the characteristics of the members of other cluster. However the traditional K-means algorithm has a high time complexity that is it takes a lot of computation to finally cluster any given data set. As a result it consumes a lot of resource. This drawback is the motivation behind our research where we tried to find out a better variant of traditional K-means algorithm which will need less computation and will provide better accuracy than the traditional one.

In our proposed algorithm we have added one check point value to store the center point of the distance of two cluster centers and used it to determine the cluster any object is going to be assign to. This check point value reduced the possibility of error that we found out in some modification of traditional K-means algorithm which we discussed earlier.

We have used some datasets that our fellow researcher used to evaluate their algorithm as it will give us a better view towards the improvement of our algorithm. By using that data sets we have found out that our algorithm is producing favorable results and is working almost perfectly.

From all the discussions and evidences, we can propose that our algorithm provides better performance than the traditional k-means algorithm. Not only that, but also our algorithm is showing superior results than some improvements of the K-means algorithm.

Due to the shortage of available resources and time, we could not test our algorithm in a larger scale. In future, we are looking forward to test our algorithm with more complex and vast datasets and compare the results with traditional K-means algorithm and the better variants of it. We are also looking forward to construct our algorithm capable of clustering real time datasets. We are also planning to predict the number of cluster needed to cluster any kind of dataset and merge this algorithm with that one.

References

[1] S. AL Manaseer, A. Malibari, “Improved Teaching Method of Data Mining Course”, I.J. Modern Education and Computer Science, Second Volume, Page-15-22, 2012.

[2] L. Su,H. Liu, Z. Song, “A New Classification for Data Stream”, I.J. Modern education and computer science, Fourth Volume, Page: 32-39,2011.

[3] S. Jigui, Q. Keyun, “Research on Modified K-Means Data Clusters ”, Computer Engineering, Volume: 33, No: 13, page: 200-201.

[4] Mac Queen JB. “Some methods for classification and analysis of multivariate observations”. Proceeding of the Fifth Berkley Symposium Math. Stat. Prob, (1):281-297, 1967.

[5] Huang Z, “Extensions to the K-Means algorithm for clustering large data set with categorical values”, Data mining and knowledge discovery, Vol. 2, Page: 283-304, 1998.

[6] S. Deelers, S. Auwantanamongkol,”Enhancing k-mean Algorithm with Initial Cluster Center Derived from Data Partitioning along the Data Axis with the Highest Variance”, International Journal of Computer Science Vol:1, 2007.

[7] J. Han, M. Kamber, J. Pei, “Data Mining- Concepts and techniques”, Third Edition, Chapter: 7, Page: 401.

[8] S. Na, G. Yong, L. Xumin,”Research on k-means Clustering Algorithm”, Third International Symposium on Intelligent Information Technology and Security Information, 2010.

[9] M. Yedla, S. R. Pathakota, T.M. Srinivasa, “Enhancing K-means Clustering Algorithm with Improved Initial Center”, International Journal of Computer Science and Information Technologies, Vol. 1(2):121-125, 2010.

[10] A. Triantafillakis P. Kanellis, D. Martakos, “Data Warehouse Clustering on the web”, European Journal of Operational Research, 160(2):353-364, 2005.

[11] M. H. Dunham, “Data Mining- Introductory and Advanced Concepts”, Pearson Education,2006.

[12] C.C. Aggarwal, “A Human-Computer Interactive Method for Projected Clustering”, IEEE Transactions on Knowledge and Data Engineering,Vol 16(4) 448-460, 2004.

[13] 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.

[14] K. A. Abdul Nazeer, M.P. Sebastian, “Improving the Accuracy and efficiency of the K- Means Clustering Algorithm”, International Conference on Data Mining and Knowledge Engineering (ICDMKE). Proceeding of the World Congress on Engineering(WCE-2009), Volume : 1 , 2009.

[15] K. Arai, A.R. Barakbah, “Hierarchical K-Means: an algorithm for Centroids initialization for K-Means ”, Department of Information Science and Electrical Engineering Politechnique in Surabay, Faculty of Science and Engineering, Saga University, Volume 36, No: 1, 2007.

[16] J. Wang, X. Su,” An Improved K-Means Algorithm”, IEEE 3rd International Conference on Communication Software and Networks (ICCSN), 44-46, 2011,

[17] Chen Zhang, Shixiong Xia, “K-Means Clustering Algorithm With Improved Initial Center”, ISBN: 978-0-7695-3543-2, pp: 790-792.

[18] University of California, Irvine, https://archive.ics.uci.edu/ml/datasets.html.

Authors Profiles

Sharfuddin Mahmood has completed his B.Sc and M.Sc degree in Computer Science from American International University- Bangladesh. His major was Information and Database Technologies. Currently his is focusing on Data Mining technologies and algorithms. His area of research is Data mining and knowledge discovery and intelligent systems. Mr. Mahmud can be contacted at smahmood@aiub.edu

Md. Saiedur Rahaman is working as an Assistant Professor and Special Assistant in the Department of Computer Science in American International University- Bangladesh. His research interest includes Data mining, Data warehousing, Algorithms etc. He can be contacted at msr.aiub@gmail.com

Dr. Dip Nandi is working as an Assistant Professor and Head of Undergraduate Program in the Department of Computer Science in American International University- Bangladesh. His research interest includes Software Engineering, Management Information Systems, E-learning etc. He can be contacted at dip.nandi@aiub.edu

Mr. Mashiour Rahman is working as a Senior Assistant Professor and Director of Faculty of Science and IT in American International University Bangladesh. His research interest includes Algorithms, Data structure, M-learning etc. He can be contacted at mashiour@aiub.edu