IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March, 2013
**ISSN: 2320 - 8791**
**www.ijreat.org**

# New algorithm to get the initial centroids of clusters on multidimensional data

**D Mariammal[1], M Gowthami[2], N Sindhujaa[3]**

**[1,2,3] Department of Computer Science and Engineering**
**Agni College of Technology, Anna University**
**Chennai,Tamilnadu, India**

**Abstract**

Cluster analysis is one of the efficient techniques in the field of data mining and k-means is one of the most well known familiars and partitioned based clustering algorithms. K-means clustering algorithm is widely used in clustering. The performance of k-means algorithm will affect when clustering the multidimensional data. In this paper, a heuristic approach for performing k-means clustering on multidimensional data based on attribute (column) with maximum range is proposed. It also proposes an improvement on the classic k-means algorithm to produce more accurate clusters. The proposed algorithm comprises of a *O(n(kt+ logn))* heuristic method, based on sorting and partitioning the input data, for finding the initial centroids in accordance with the data distribution.

***Keywords-*** *Data Mining; Clustering; Enhanced k-means Algorithm and multidimensional data; Improved Initial Centroids; Sorting and Partitioning*

## 1.Introduction

In order to survive and succeed in today's competitive global environment, decisions need to be made quickly and correctly. Data warehouse is an architectural construct of information system that provides users with current and historical decision support information. But the information in data warehouse is very hard to access. New and efficient algorithm is needed to extract the information from the data warehouse.

Clustering algorithm [1] is one of the efficient data analysis methods to group the set of data as cluster. A cluster is the collection of data objects that are similar to one another within the same cluster and are dissimilar in other cluster.

The k-means algorithm [1, 4, 5, 6, 7] helps to produce clusters for many practical applications in emerging areas like bioinformatics [2, 3]. But the computational complexity of the standard k-means algorithm is very high. Algorithm results are based on choice of initial centroids. This paper gives a heuristic method to sort and partitioning the data for find initial centroids, thereby improving the efficiency of the k-means algorithm.

## 2. K-Means Clustering Algorithm

There are two goals for clustering algorithms:

i. Determining good clusters and doing so efficiently.
ii. Clustering is used in application domains such as in data mining and knowledge discovery, statistical data analysis, data classification and compression, medical image processing and bioinformatics.

The original k-means clustering algorithm consists of two separate phases: the first phase is to define *k* centroids, one for each cluster. The next phase is to take each point belonging to these given data set and assigning it to the nearest centroid. The first phase is completed when all the points are included in some clusters, and an early grouping is done. We need to recalculate the new centroids at this point, as the inclusion of new points may owe a change in the cluster centroids. Once we find *k* new centroids, a new cluster is to be created between the same data points and the nearest new centroids, generating a loop. As a result of this loop, the *k* centroids may change their position in each iteration. Eventually, a situation will be reached where the centroids do not move anymore. It indicates the convergence of clustering. Pseudo code for the k-means clustering algorithms listed as Algorithm1 [4].

---

**Algorithm 1**: k-means clustering algorithm

---

Input:
D = {d1, d2,.....,d} //set of *n* data items.
*k* // Number of desired clusters
Output:
A set of *k* clusters.

Steps:

1. Arbitrarily choose *k* data-items from D as initial centroids;
2. **Repeat**
   2.1 Assign each data item *di* to the cluster which has the closest centroid;
   2.2 Calculate the new mean of each cluster;

**Until** convergence criterion is met.

## 3. Related Work

Many researches have been undergone by the researchers to improve the k-means algorithm [8, 9, 10]. A deviation of the k-means clustering algorithm is the k-modes [9, 11] method. The mean of cluster is replaced with modes in this algorithm. This method produce the cluster based on modes instead of taking mean. Mr. Huang [9], produces a k-prototype algorithm by combining the k-means and k-modes algorithm to cluster the data. This method defines the dissimilarity measure by taking into account both numeric and categorical attributes.

The systematic method to find the initial centroids has been proposed by Fang Yuan et al. [10]. The obtained centroids by this method are consistent with the data distribution. A cluster produced by this approach has better accuracy, over the original k-means algorithm. Even though, it does not provide any efficiency improvement in the k-mean algorithm.

Fahim A M et al. [8] proposed an efficient method for assigning data-points to clusters. The computational complexity of the k-means algorithm is very high due to the calculation of distances between data points and all the centroids for each iteration. To reduce the number of distance calculations, heuristics method has been used. But this method does not provide a guarantee for the accuracy of the final clusters.

Abdul Nazeer and Sebastian proposed an algorithm [12], to produce a cluster with improved efficiency and accuracy. Data warehouse is a multi dimensional data model due to incredible growth of multi dimensional dataset, conventional data base querying methods are inadequate to extract useful information, so researchers nowadays is forced to develop new techniques to meet the raised requirements. Such large expression data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of features/attributes. Hence, to improve the efficiency and accuracy of mining task on multidimensional data, the data must be preprocessed by an efficient method. But multidimensional data values are not deal by any existing algorithm.

## 4. Proposed Algorithm

The basic idea of this algorithm is to determine the initial centroids of the clusters in a heuristic manner, so as to ensure that the centroids are chosen in accordance with the distribution of data. The method involves sorting the input data set and partition the sorted data set into ‘k’ number of sets where ‘k’ is the number of clusters to be formed. Mean values of each of these sets aretaken as the initial centroids.

This paper proposes a heuristic method to deal with multidimensional data items. Each multidimensional data contain attribute like *di1, di2,…..dim*, where *m* is the number of attributes or columns in each data item. The attribute with maximum range is determining first in this case [6], where range is the difference between the maximum and the minimum element in the column. Initially, we determine the maximum and minimum element of each attribute and compute the range of values for each column. Then we select the attribute having maximum range. By using the heap sort algorithm [12] technique the data set are sorted in increasing order, based on attribute with maximum range. The list of sorted data points are partitioned into ‘k’ equal sets. Finally, the arithmetic means of each of ‘k’ sets are computed. These means are taken as the initial centroids for the clusters formation.

After identifying the initial centroids as described above, the data points are assigned to various clusters by using the same method used in the second phase of the original k-means algorithm. Each data point *di* is assigned to the cluster which is having the closest centroid. To determine the distance between the data points and the centroid the Euclidean distance are used. The proposed algorithm is outlined below as Algorithm 2.

---

**Algorithm 2**: Proposed Algorithm for Clustering

---

**Input:**
D = {d1, d2,.....,dn} // set of *n* data items.
*k*// Number of clusters.
**Output:**
A set of *k* clusters.
**Steps:**

1. For each column of the data set, determine the *range* as the variation between the maximum andthe minimum element;
2. Identify the column with maximum *range*;
3. Sort the entire data set increasing order based on the column having the maximum *range*;
4. The sorted data set are partitioned into 'k' equal parts;
5. Determine the arithmetic mean of each part obtained in Step 4 as *a1, c2,….ak*; Take these mean values as the initial centroids.
6. **Repeat**
   6.2 Assign each data item *di* to the cluster which has the closest centroid;
   6.3 Calculate new centroid of each cluster;

   **Until** convergence criterion is met.

---

The selection of initial centroids is varied from original k-mean algorithm. Instead of choosing randomly, the proposed algorithm determines the initial centroids in a more meaningful way, in accordance with the distribution of data. Consequently, the algorithm converges much faster than the original k-means algorithm. Moreover, since the method for determining the initial centroids is based on the technique of sorting, this phase requires less time compared to other similar approaches available in the literature [10].

## 5. Time Complexity Analysis

In the original k-means algorithm, the initial centroids are selected randomly. As a result, the centroids are recalculated again and again before the algorithm converges and the data points are assigned to their nearest centroids. Since complete redistribution of data points takes place based on the new centroids, this procedure takes time *O(nkt)* where *n* is the number of data-points, *k* is the number of clusters and *t* is the number of iterations.

For the Enhanced algorithms discussed in [10],the first phase of identifying the initial centroids takes *O(n2)* time even though it produces better results compared to the original k-means algorithm.

In the proposed algorithm discussed in this paper, the step to find the maximum and minimum element of each attribute of the data set requires *O(n)* time where *n* is the number of data items. The time required to find the maximum and minimum element of all the attributes of the data set is *O(nm)* where *m* is the number of column in each data item. The range of each column (difference between maximum and minimum elements) can be calculated in constant time and the time required to find the attribute with maximum range is *O(m)* where *m* is the number of column in the data set. In the next step the data items are sorted based on the attribute with the maximum range can be computed in *O(n logn)* time using Heap Sort. Time complexity for partitioning the *n* data items into *k* equal sets and finding the mean of each set is *O(n)*. Thus the overall time complexity for determining the initial centroids of a data set containing *n* elements is *O(ns logn)*, as *m* is much less than *n.*

The second phase is allocating data points to clusters is the same as that of the original k-means algorithm. The loop consisting of the allocation of data-points to the nearest clusters and the subsequent recalculation of centroids is executed repetitively until reach the convergence criteria. This procedure takes time *O(nkt)* where *n* is the number of data-points, *k* is the number of clusters and *t* is the number of iterations. Nevertheless, the algorithm converges in less number of iterations as the initial centroids are computed in a strategic manner in tune with the data distribution. Thus the overall time complexity of the proposed algorithm is the maximum of *O(n logn)* and *O(nkt)*, i.e. *O(n (kt + logn)).*

## 6. Experimental Results

To test the accuracy and efficiency of the proposed algorithm, multivariate data sets with known clustering available at the UCI repository of machine learning databases [13] were used. The input data sets

are E-Coli[14], Breast Cancer-Wisconsin [15] and Thyroid [16]. The same sets of data are given as input to the original k-means algorithm, the enhanced algorithm [8] and the proposed algorithm.

The standard k-means and the enhanced k-means algorithms require the values of the initial centroids also as input, and the value of *k*. The experiment is conducted for different sets of values of the initial centroids, which are chosen randomly. The data values and the value of *k* are the only inputs required by the proposed algorithm. The accuracy of clustering is determined by comparing the clusters obtained by the experiments with the already available pre-determined clusters in the UCI data set. The percentage accuracy and the time taken for each experiment are calculated.

## 7. Conclusion

The k-means algorithm is one of the widely used for clustering large sets of data. But the original algorithm does not always guarantee the accuracy of the final clusters based on the selection of initial centroids. Moreover, the computational complexity of the standard k-means algorithm very high which leads to the need to reassign the data points a number of times, during every iteration of the loop. This paper presents an improved k-means algorithm using a *O(n logn) novel* heuristic method for determining the initial centroids. This method ensures that the initial centroids are generated depend on the distribution of the data. These results in clusters shows better accuracy compared to the original k-means algorithm. Experimental results have shown that the proposed algorithm produces better clusters in less computation time compared to the standard k-means algorithm.

A limitation of the proposed algorithm is still required to give the k, number of desired clusters as an input. To compute the values of k, depending on the data distribution using some statistical methods is suggested for future research.

## References

[1] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[2] Amir Ben-Dor, Ron Shamir and Zohar Yakini, "Clustering Gene Expression Patterns," Journal of Computational Biology, 6(3/4): 281-297, 99.

[3] Daxin Jiang, Chum Tong and Aidong Zhang, "Cluster Analysis for Gene Expression Data," IEEE Transactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

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

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

[6] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[7] Stuart P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory, 28(2): 129-136.

[8] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.

[9] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, (2):283–304, 1998.

[10] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithmto Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[11] Chaturvedi J. C. A, Green P, "K-modes clustering," J. Classification, (18):35–55, 2001.

[12] T H Cormen, C E Leiserson, R L Rivest and C Stein, Introduction to Algorithms, Second Edition, MIT Press, 2001.

[13] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-arningdatabases

[14] Paul Horton and Kenta Nakai, "A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins," Intelligent Systems in Molecular Biology, 109-115.St. Louis, USA, 1996.

[15] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *SIAM News*, Volume 23, Number 5, pp 1 & 18, 90.

[16] Coomans, D., Broeckaert, M. Jonckheer M. and Massart D.L.,"Comparison of Multivariate Discriminant Techniques forClinical Data - Application to the Thyroid Functional State," Meth. Inform. Med. 22 , pp. 93-10, 1983.