Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 15 Issue 7 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Modified Version of the K-Means Clustering Algorithm

By Juhi Katara & Naveen Choudhary

Maharana Pratap University of Agriculture and Technology, India

Abstract- Clustering is a technique in data mining which divides given data set into small clusters based on their similarity. K-means clustering algorithm is a popular, unsupervised and iterative clustering algorithm which divides given dataset into k clusters. But there are some drawbacks of traditional k-means clustering algorithm such as it takes more time to run as it has to calculate distance between each data object and all centroids in each iteration. Accuracy of final clustering result is mainly depends on correctness of the initial centroids, which are selected randomly. This paper proposes a methodology which finds better initial centroids further this method is combined with existing improved method for assigning data objects to clusters which requires two simple data structures to store information about each iteration, which is to be used in the next iteration. Proposed algorithm is compared in terms of time and accuracy with traditional k-means clustering algorithm as well as with a popular improved k-means clustering algorithm.

Keywords: *clustering, data mining, initial centroids, k-means clustering.*

GJCST-C Classification : *B.2.4 B.7.1*

A Modified Version of the K-Means Clustering Algorithm

Juhi Katara [α] & Naveen Choudhary [σ]

Abstract- Clustering is a technique in data mining which divides given data set into small clusters based on their similarity. K-means clustering algorithm is a popular, unsupervised and iterative clustering algorithm which divides given dataset into k clusters. But there are some drawbacks of traditional k-means clustering algorithm such as it takes more time to run as it has to calculate distance between each data object and all centroids in each iteration. Accuracy of final clustering result is mainly depends on correctness of the initial centroids, which are selected randomly. This paper proposes a methodology which finds better initial centroids further this method is combined with existing improved method for assigning data objects to clusters which requires two simple data structures to store information about each iteration, which is to be used in the next iteration. Proposed algorithm is compared in terms of time and accuracy with traditional k-means clustering algorithm as well as with a popular improved k-means clustering algorithm.

Keywords: clustering, data mining, initial centroids, K-means clustering.

I. Introduction

Data mining refers to using a variety of data analysis techniques and tools to discover previously unknown, valid patterns and relationship in large dataset[5]. Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions. Clustering may be defined as preprocessing step in all data mining algorithms in which the data objects are divided into clusters which contains high intra-cluster similarity and low inter-cluster similarity [3], [10].

Clustering can be applied to a wide range of fields like pattern recognition, marketing, image processing etc[3]. Clustering algorithms are mainly divided into partitioning, hierarchical, density based, grid based, model based clustering algorithms.

Partitioning clustering algorithm first creates an initial set of k partition, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that tries to improve the clustering by moving objects from one class to another. Hierarchical clustering algorithm creates a hierarchical decomposition of the dataset using some criterion. The method can be categorized as being either agglomerative or divisive, based on how the hierarchical decomposition is designed. Density based clustering algorithm uses notion of density for clustering data objects. It either grows clusters according to the density of neighborhood objects or according to some density function. Grid based clustering algorithm first quantizes the object space into a finite number of cells that form a grid structure, and then performs clustering on the grid structure. Model based clustering algorithm attempts to optimize the fit between the given data and some mathematical model.

K-means clustering is a partitioning clustering technique in which clusters are formed with the help of centroids. It follows unsupervised, non deterministic and iterative approach towards clustering. K-means clustering is processed by the minimization of the average squared Euclidean distance between the data objects and the cluster centroids. The result of the k-means clustering algorithm is affected by the choice of initial centroid. Distinct initial centroid might result in distinct final clusters. Centroid of the cluster may be defined as the mean of the objects in a cluster. It may not necessarily be a member of the dataset.

II. Traditional K-means Clustering Algorithm

K-means clustering is the most popular clustering algorithm [9]. In the traditional k-means clustering given dataset is classified into k numbers of disjoint clusters, where the value of k is given as input to the algorithm. The algorithm is implemented in two phases.In the first phase k centroids are selected randomly. In the second phase assignment of each data object to the closest centroid cluster is done. Distance between data objects and centroids is generally calculated by Euclidean distance. When all data objects are assigned to any of the k clusters, first iteration is completed and an early grouping is done. After completion of first iteration recalculation of centroids are done by taking mean of data objects of each cluster. As k new centroids are calculated, a new assignment is to be done between the same data objects and new centroids, generating loops which results in number of iterations. As a result of this loop k centroids and data objects may change their position in a step by step manner. Ultimately the situation will occur where the centroids do not update anymore. This means the

Author α σ: Department of Computer Science & Engineering, College of Technology and Engineering, Maharana Pratap University of Agriculture and Technology, Udaipur, Rajasthan, India.
e-mail: katarajuhi@gmail.com

convergence criterion for clustering is achieved. In this algorithm generally Euclidean distance is used to find distance between data objects and centroids [3]. Between one data object $X = (x_{1}, x_{2}, \ldots x_{n})$ and another data object $Y = (y_{1}, y_{2}, \ldots y_{n})$ the Euclidean distance $d(X, Y)$be calculated as follows:

$$d(X, Y) = \{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2\}^{1/2}$$

```
Algorithm 1 : The Traditional K-Means Clustering Algorithm [3]

Input:  D = {d_1, d_2,......,d_n}  //set of n data objects.
        k    // number of required clusters.
Output: k clusters
Steps:
1. Randomly select K data objects as initial
   centroids from D.
2. Calculate the distance between each data object
   d_i (1<=i<=n) and all k cluster centroids c_j
   (1<=j<=k) , then allocate data object d_i to the
   cluster which has closest centroid.
3. Calculate new mean for each cluster.
   //new mean is the updated centroid of cluster.
4. Repeat step 3 and 4 until no change in the
   centroid of cluster.
```

III. Drawbacks of Traditional K-means Clustering Algorithm

Traditional K-means clustering algorithm has several drawbacks. The major drawback of traditional K-means clustering algorithm is its performance is mainly depends on the initial centroids, which are selected randomly and resulting clusters are different for different runs for the same input dataset. Another drawback includes distance calculation process of traditional k-means algorithm which takes long time to converge clustering result, as it calculates the distance from each data object to every cluster centroids in each iteration while there is no need to calculate that distance each time. As in the resulting clusters some data objects still remains in the same cluster after several iteration. It affects the performance of the algorithms. One more drawback of k-means clustering is the requirement to give number of clusters formed as input by the user.

IV. Related Work

Xiuyun Li et al. [1] proposed enhanced k-means clustering algorithm based on fuzzy feature selection. This algorithm generates weight of feature important factor to describe the contribution of each feature to the clustering and makes use of FIF to improve the similarity measure and then achieve the improved clustering result.

Wang Shunye et al. [5] proposed an improved k-means clustering algorithm in the optimal initial centroids based on dissimilarity. This algorithm achieves the dissimilarity to reflect the degree of correlation between data objects then uses a Huffman tree to find the initial centroids. It takes less amount of time because the iteration diminishes through the Huffman algorithm.

Shi Na et al. [3] proposed an improved k-means clustering algorithm to increase efficiency of k-means clustering algorithm. This algorithm requires two simple data structures to store information in every iteration which is to be used in the next iteration. The improved algorithm does less calculation, which saves run time.

Mohammed El Agha et al. [4] proposed improved k-means clustering algorithm which has ElAgha initialization that uses a guided random technique as k-means clustering algorithm suffers from initial centroids problem. ElAgha initialization outperformed the random initialization and enhanced the quality of clustering with a big margin in complex datasets.

K.A Abdul Nazeer et al. [2] proposed an algorithm to enhance accuracy and efficiency of the k-means clustering algorithm. This algorithm consist of two phases. First phase is used to determine initial centroids systematically so as to produce clusters with better accuracy. Second phase is used for allocating data objects to the appropriate clusters in less amount of time. This algorithm outputs good clusters in less amount of time to run.

V. Proposed Algorithm

In this section a modified algorithm is proposed for improving the performance of k-means clustering algorithm. In the paper [3], authors proposed an improved k-means clustering algorithm to improve the efficiency of the k-means clustering algorithm but in this algorithm the initial centroids are selected randomly so this method is very sensitive to the initial centroids as random selection of initial centroids does not guarantee to output unique clustering result. In the paper [5], authors proposed an improved k-means clustering algorithm in the optimal initial centroids based on dissimilarity. However this algorithm is computationally complex and requires more time to run.

In this paper we proposed a new approach for selecting better initial centroids which outputs the unique clustering result and increases the accuracy of basic k-means clustering algorithm and proposed approach is combined with the algorithm of paper [3] for allocating the data objects to the suitable cluster. The algorithm of paper [3] is referred as shina improved k-means clustering algorithm in this paper. We compared the traditional k-means clustering algorithm, shina improved k-means clustering algorithm [3] and proposed algorithm in terms of time and accuracy parameters.

Algorithm 2: Modified k-means clustering algorithm

```
Input :   D = {$d_1, d_2, d_3, \ldots\ldots\ldots\ldots d_n$}
                //Dataset of n data  objects
            k   // Number of required clusters
Output :  A set of k clusters.
Steps :
1.  Calculate distance from origin of each data
    object $d_n$ in the dataset D.
2.  Apply sorting on the distances obtained in
    step 1. Sort the data objects according to
    distance.
3.  Now divide the sorted data objects into k equal
    sets.
4.  Select the middle data object as the initial
    centroid from each set.
5.  Calculate the distance between each data
    object $d_i(1<=i<=n)$ and all k cluster centroids
    $c_j$ $(1<=j<=k)$ as Euclideandistance $d(d_i, c_j)$.
6.  For each data object $d_i$ find the closest
    centroid $c_j$ and assign $d_i$ to that cluster j.
7.  Repeat step 8 to 11 until no change in the
    centroid of clusters.
8.  Store the cluster number in array cluster[ ].
    Set cluster[i]=j.
9.  Store the distance of data object from the
    closest centroid in the array dist[ ].
     Set dist[j] = $d(d_i, c_j)$.
10.  For each cluster $j(1<=j<=k)$ recalculate the
     cluster centroid.
11.  For each data object $d_i$
    11.1  Compute its distance from the new
          computed centroid of the present cluster.
    11.2  If this distance is less than or equal to
          the present closest centroid, the data
          remains in the same cluster.
  Else
      11.2.1 For every centroid $c_j$ $(1<=j<=k)$
             Calculate the distance $d(d_i, c_j)$ ,
             then assign $d_i$ to the cluster which
             closest centroids.
     End For
```

In the proposed algorithm distance of each data object from origin is calculated. Then the original data objects are sorted accordance with the sorted distance. Insertion sort is used for sorting in this paper. Now divide the sorted data objects into k equal sets. Take middle data object as the initial centroid from each set. This process of selecting centroid outputs better unique clustering result. Now for every data object in the dataset calculate distance from every initial centroid. The next step is an iterative process which reduces the required time to run. The data objects are assigned to the cluster which has the closest centroid. Two data structures *cluster [ ]* and *dist[ ]*are required to store information about the completed iteration of the algorithm. Array *cluster [ ]* stores the cluster number of data object from which it belongs to and array *dist [ ]* stores the distance of every data object from closest centroid. Next, for each cluster obtained in completed iteration the new centroid is calculated by taking the mean of its data objects.

Then for each data object the distance is calculated from the new calculated centroid of its present cluster. If this distance is less than or equal to the previous closest distance, the data object remains in the same cluster otherwise for every remaining data object,calculate the distance from all new calculated centroids. Next, the data objects are assigned to the cluster which has the closest centroids. Now array *cluster* and *dist* are updated storing new values obtained in this step. This reassigning process is repeated until no change in the centroids of cluster.

VI. Experimental Results and Discussion

All the experiments are carried out on core i3 Intel based PC machine with 4 GB RAM, running on WINDOWS 7 64 bits operating environment and Programming Platform is MATLAB version R2013a.

In this paper two different datasets are taken from the UCI repository of machine learning databases [6] to test the performance of the proposed k-means clustering algorithm and for comparing the traditional k-means clustering algorithm, shina improved k-means clustering algorithm [3] and proposed algorithm of this paper. IRIS and WINE datasets are selected as the test datasets [6]. The values of attributes are numeric.

A brief introduction of the datasets used in experimental evaluation is given in the table below:

Table 1 : Characteristics of datasets

Dataset	Number of attributes	Number of instances
Iris	4	150
Wine	13	178

a) Iris dataset

Iris dataset contains the three classes of iris flower: setosa, versicolour and virginica. This dataset contains 150 instances and three classes. In iris dataset, each class contains 50 instances with four attributes: sepal length, sepal width, petal length, petal width.

b) Wine dataset

This dataset contains the chemical analysis of wine in the same region of Italy but three different cultivators. The dataset contains 178 instances and three classes with 13 attributes. First class contains 59, second class contains 71 and third class contains 48 instances. The attributes of dataset are alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavonoids, nonflavanoids phenols, proanthocyanins, Color intensity, hue, OD280/OD315 of diluted wines and proline.

The same datasets are given as input to all the algorithms. Number of k is given three for both the datasets. Experiment compares proposed k-means clustering algorithm with the traditional k-means clustering algorithm and with the shina improved k-means [3] in terms of time and accuracy.

Accuracy: Accuracy is the ratio of correctly predicted instances divided by total number of instances.

Time: It is the amount of time that passes from the start of an algorithm to its finish.

Accuracy of clustering is determined by comparing the clustering results with the clusters already available in the UCI datasets [6]. Traditional and shina improved k-means clustering algorithm gives different accuracy and time for every run as it selects initial centroid randomly. So these algorithms are executed several time and average of accuracy and time is taken. Accuracy of proposed k-means clustering algorithm is unique at every run but time is different for each run so it is also executed several time and average of time is taken.

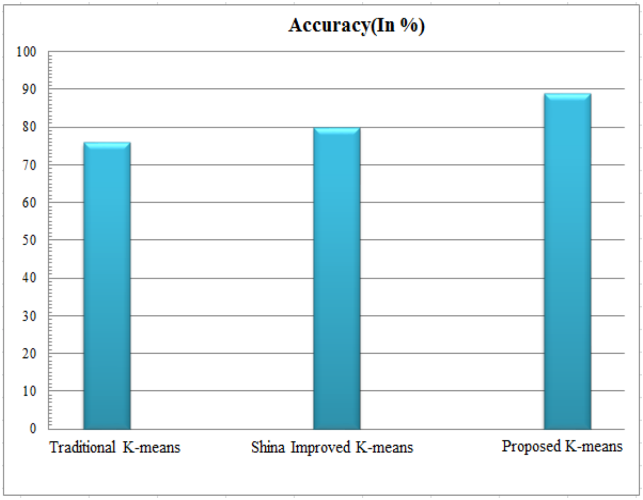

Fig. 1 : Accuracy comparison chart for Iris dataset

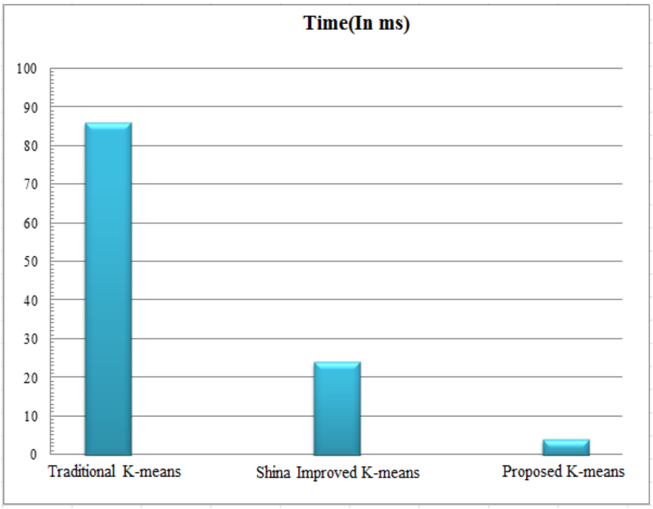

Fig. 2 : Time comparison chart for iris dataset

Table 2 : Performance comparison on Iris dataset

Parameters	Traditional K-means clustering algorithm	Shina Improved k-means clustering algorithm	Proposed K-means clustering algorithm
Accuracy (In %)	76	80	89
Time (In ms)	86	24	4

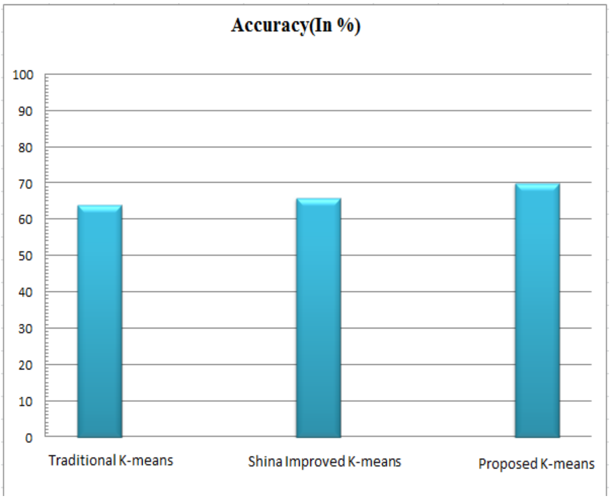

Fig. 3 : Accuracy comparison chart for Wine data set

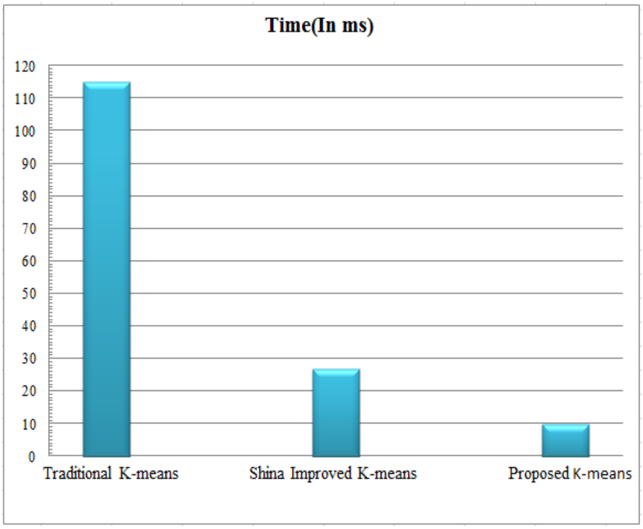

Fig. 4 : Time comparison chart for wine dataset

Table 3 : Performance comparison on Wine data set

Parameters	Traditional K-means clustering algorithm	Shina Improved k-means clustering algorithm	Proposed K-means clustering algorithm
Accuracy (In %)	64	66	70
Time (In ms)	115	27	10

The result of experiment shows that the proposed k-means clustering algorithm can output the better unique clustering result in less amount of time than traditional k-means clustering algorithm and shina improved k-means clustering algorithm [3]. As it selects better initial centroids which result in reduction of iterations. Shina improved method [3] of assigning data objects to the appropriate clusters results in less number of distance calculations. So proposed algorithm combines both this methods and results in less time to run. At the same time the proposed k-means clustering algorithm can improve the accuracy of the algorithm.

VII. Conclusion

K-means clustering algorithm is one of the most popular and an effective algorithm to cluster datasets which is used in number of fields like scientific and commercial applications. However, this algorithm has several drawbacks such as selection of initial centroid is random which does not guarantee to output unique clustering result and k-means clustering has more number of iterations and distance calculations which finally result in more amount of time to run. Various enhancements have been carried out on the Traditional k-means clustering algorithm by different researchers considering different drawbacks. The proposed algorithm combines a systematic way for selecting initial centroids and an efficient method for assigning data objects to clusters. So proposed algorithm is found to be more accurate, efficient and feasible. The value of k required number of clusters is still required to be given as an input to the proposed algorithm. Intelligent pre estimation of the value of k is suggested as a future work.

References Références Referencias

1. Xiuyun Li, Jie Yang, Qing Wang, Jinjin Fan, Peng Liu "Research and Application of Improved K-means AlgorithmBased on Fuzzy Feature Selection" in Fifth International Conference on Fuzzy Systems and Knowledge Discoveryvol 1, 2008 ieeeconference publications
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. Shi Na, Liu Xumin, Guan Yong "Research on k-means Clustering Algorithm : An Improved k-means Clustering Algorithm" in Third International Symposium on Intelligent Information Technology and Security Informatics 2010 ieee conference publications
4. Mohammed El Agha, Wesam M. Ashour "Efficient and Fast Initialization Algorithm for K-means Clustering"I.J. Intelligent Systems and Applications, 2012
5. Wang Shunye, Cui Yeqin, Jin Zuotao and Liu Xinyuan "K-means algorithm in the optimal initial centroids based on dissimilarity" in Journal of Chemical and Pharmaceutical Research, 2013

6. Merz C and Murphy P, UCIRepository of Machine LearningDatabases,Available:ftp://ftp.ics.uci.edu/pub/machine-learning-databases
7. Charles Elkan "Using the Triangle Inequality to Accelerate k-Means" in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003
8. Hong Liu and Xiaohong Yu "Application Research of k-means Clustering Algorithm in Image Retrieval System" in Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT) 2009
9. Jiawei Han, Michelinekamber and Morgan Kauffman, "Data Mining: Concepts and Techniques", 2nd edition 2006
10. Osama Abu Abbas "Comparisons between data clustering algorithms" in The International Arab Journal of Information Technology vol 5 , no. 3 , July 2008
11. OyeladeO.J, Oladipupo O.O,Obagbuwa I. C "Application of k-Means Clustering algorithm for prediction of Students Academic Performance" inInternational Journal of Computer Science and Information Security,Vol. 7, 2010
12. Chunfei Zhang, Zhiyi Fang "An Improved K-means Clustering Algorithm" in Journal of Information & Computational Science 2013