AN AVERAGE-BASED APPROACH FOR INITIAL CENTROID SELECTION IN K-MEANS ALGORITHM

A. Shafiq, M. Rehman and M. Anjum

Department of Computer Science, Lahore College for Women University, Lahore, Pakistan
Corresponding author's email: dr.mrehman13@gmail.com

ABSTRACT: The underlying research work was focused on one of the standard k-means issue of initial centroid selection. An average based approach was used for avoiding random cluster initialization. The experiments of this study showed that the results obtained with proposed method were better and consistent. It was concluded that the proposed method had less classification error, reduced total number of iterations and took less execution time than random initialization method.

Key words: Data clustering, Partitioned-based clustering algorithms, K-means, Initial centroids.

(Received 28-06-2016 Accepted 28-11-2016)

INTRODUCTION

Development in sensing and storage technology and significant progress in programs like internet search, video surveillance and digital imaging have increased data sets not only in volume but also in dimensions. In addition to progress in the volume of data, the diversity of data has also increased. This growth in both the size and diversity of data entails development in methods to automatically recognize procedure and precise the data (Jain, 2010).

Data clustering is a data mining and data analysis method, that produces refined views to the in-built structure of a data set by separating it into a number of disjoints or overlapping classes. Many researchers have addressed the clustering problem in many disciplines which reflects its extensive demand and effectiveness as one of the steps in exploratory data analysis (Velusamy *et al.*, 2014).

The two broad classes of clustering algorithms include hierarchical and partitioned clustering algorithms (Vijayarani *et al.,* 2014). K-means is one of the most extensively used partitioning based clustering algorithms. It is one of the most effective data mining algorithms for being simple, having ability to be easily scaled up and modified to a variety of environments and program fields (Tommikarkkainen, 2006). The result of k-means algorithm depends highly upon the choice of initial centroids (Erisoglu *et al*., 2011).

Clustering is a technique of data mining that is used to group the data on the basis of similarities and dissimilarities among data points (Siraj and Abdoulha, 2007). It is an unsupervised organization of patterns into sets called clusters; where a cluster is collection of items which are similar to items in same clusters and are different to the items belonging to other clusters (Velmurugan, 2012). An ideal cluster is basically a set of points that is isolated and compacted (Naldi and Campello, 2015 and Verma *et al.*, 2012). The chief goal of clustering is descriptive *i.e.* to discover a new set of categories to group a set of points, patterns or items (Jain *et al.*, 1999).

Cluster analysis is another notation used for data clustering. It is a process of putting similar data into groups. It can find different types of similarity measures to categorize classes depending on the applications and their associated data. The precise definition of 'cluster' gave rise to many clustering algorithms with each using a distinct induction principle. These algorithms simultaneously discover all clusters as a partition of data without imposing a hierarchy. The basic aim of these algorithms is to decompose the set of items into a set of pre-set number of disjoint groups (Fayyad *et al*., 1996).

In these methods, the whole data is separated into clusters, such that each data point has one cluster and each cluster has at least one data point. Such algorithms want that user should pre-set the number of clusters. An in-depth enumeration process of all partitions is required for achieving global optimality. Due to finite number of data points and distant partitions, exhaustive search methods can be used to avoid local minima problem. However, this is only true in theory because it is a NP-hard problem to find global optimal partition and exhaustive methods are not useful in practice (Tommikarkkainen, 2006).

The underlying research was conducted to propose an average based method for initial centroid selection instead of random selection. The performance comparison of standard k-means and proposed method is provided based on classification error, execution time and number of iterations taken by each algorithm.

MATERIALS AND METHODS

The algorithms were implemented in visual studio 2015 using C# language and Iris data set whose

specifications are shown in Table-1. The data set was taken from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html). The data set contained 3 classes named as Iris setosa, Iris versicolor and Iris virginica (Erisoglu *et al., 2011*). Fifty samples were placed in each class. Both algorithms were tested setting 'k' to different values. The data was used after applying Gaussian normalization. The purpose of normalizing data was to equalize the scale and the variability of the attributes (James, 2014). The results obtained in each case were compared.

Table: 1. Dataset.

Dataset	Cardinality	Classes	Attributes
Iris	150	3	4

The error percentage criterion was used to compare clustering results of both algorithms. As the data set held 3 classes, the error percentage was calculated using k=3. The error percentage was obtained from total number of items in the set of data and total misclassified objects using following formula:

$$Error = \frac{\varepsilon}{n} \times 100$$

Where ε was total number of misclassified objects and *n* was the total number of objects in the entire data set (Bangoria, 2014).

Further, the clustering results were also compared against total number of iterations and execution time taken by each algorithm to reach final clustering. The comparison was made for different values of k.

The proposed method choose initial centroids in two steps. At step 1, it calculated the average of attributed values for each object and then created range based on these average values for each cluster depending on total number of clusters. It then assigned each object to the cluster whose range matched to the particular object's average value. In this way, k initial partitions were created. A check was also performed at this step to ensure that each cluster contained at least one object. At step 2, mean of each cluster was calculated and these mean points were marked as initial centroids. At step 3, sum of squared difference was calculated for each object from all cluster centers and the object was assigned to the cluster having closest cluster center. These steps for all objects resulted in new partitions. At step 4, new cluster centers were obtained taking mean of each cluster. The algorithm stopped when no change was seen at step 3 and 4. The partition obtained at this point was marked as final clustering produced by the algorithm.

RESULTS AND DISCUSSION

The results for both algorithms were found for eight unlike values of 'k'. Each algorithm was executed 5 times for each value of k. The mean value of the 5 trials taken for measuring execution time of each algorithm was tabulated as shown in Table-2.

The performance of both algorithms with respect to total execution time taken and total number of iterations is graphically drawn as shown in Fig-1 and Fig-2.

The clustering results obtained with the k-means algorithm using initial centroids calculated by the proposed method were improved. The proposed method produced consistent clusters as compared to random initialization. The results gained by suggested method were better in terms of total number of iterations exchanged, elapsed CPU time and classification error percentage. The error percentage was calculated to check the accuracy of both algorithms. The comparison of error percentage was calculated is shown in the form of bar chart in Fig-3.

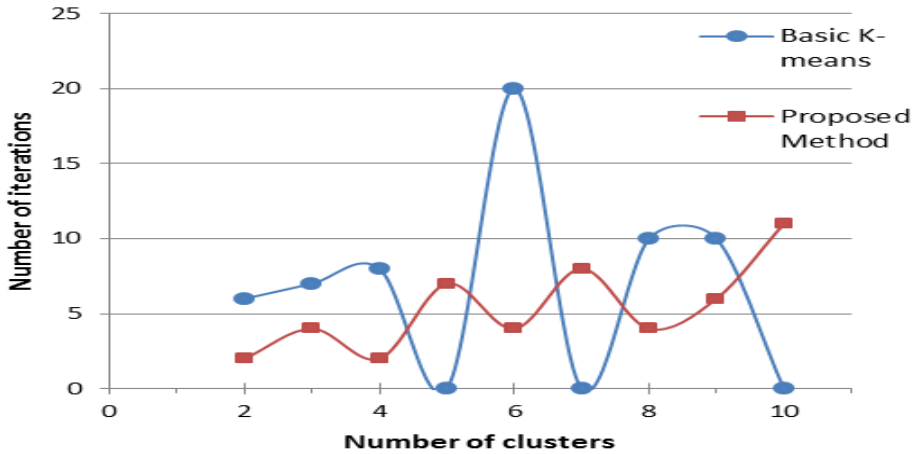

Fig-1: Number of iterations for k-means and proposed method with better initial centroids

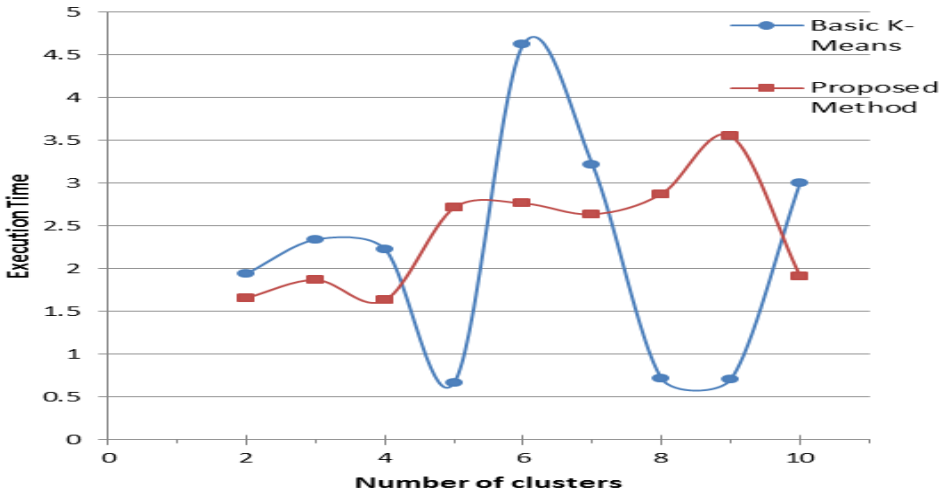

Fig-2: Execution time in seconds for k-means and proposed method with better initial centroids

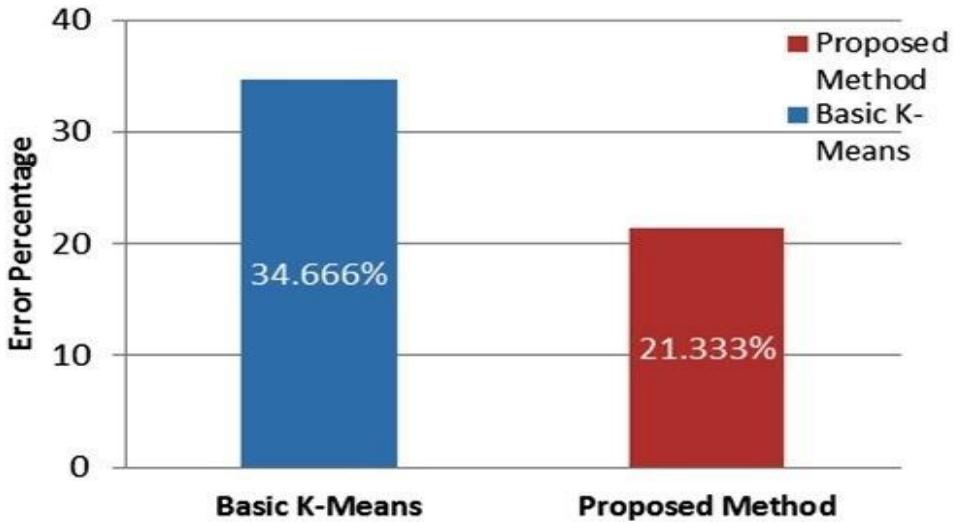

Fig-3: Error percentage for k-means and proposed method with better initial centroids

A comparison was made with the results produced in this study and other related work that used different data set, technique and comparison criteria but dealt the same issue of initialization in k-means algorithm. Deelers and Auwatanamongkol (2007) proposed an algorithm to compute better initial cluster centers. The results obtained by this algorithm showed reduced clustering error and converged to better clustering results reducing running time of k-means for large data sets. The results were found to be consistent with the current study. Chen and Shixiong (2009) also proposed a method that didn't require user to give number of clusters in advance and avoid random assignment of centers using 'sub-merger' strategy. The results of this study also showed similar results to current study as the proposed method recovered all inherent clusters. However, random selection, in this study, missed some clusters due to bad initial centroids selection and the current study considered this issue. Erisoglu *et al*. (2011) proposed a method to compute initial cluster centers for k-means with a focus on improving the performance of k-means algorithm by making better initial centroids selection. The results in their study showed reduced classification error, number of iterations and random index compared to random selection method which was improved in the current study. The current approach was adopted to improve results of existing studies and produced superior results compared to existing studies.

Conclusion: The proposed method was very effective for small value of 'k'. The experimental results showed consistent, better and improved clustering results as compared to randomly chosen initial centroids. Also, the proposed algorithm was easier to implement than earlier suggested methods for calculation of initial centroids.

REFERENCES

Bangoria, B.M. (2014). Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Values. International Journal of Computer Science and Information Technologies. 5(1): 876–879

Chen, Z., and X.Shixiong (2009). K-means clustering algorithm with improved initial center. In Knowledge Discovery and Data Mining (KWDD). Second International Workshop. IEEE. 790-7.

Deelers, S., and S. Auwatanamongkol (2007). Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering. 2(4): 247-252

Erisoglu, M., N. Calis, and S. Sakallioglu (2011). A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Letters. 32(14): 1701-1705

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). From data mining to knowledge discovery in databases. AI magazine. 17(3): 37

Jain, A. K., M.N. Murty, and P.J. Flynn (1999). Data clustering: a review. ACM Computational Surveys. 31(3): 264-323

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 31: 651-666

James, M. (2014). "k-Means Clustering," in Machine Learning Using C# Succinctly. Ed. Syncfusion; Morrisville (New York). 11-35

Naldi, M. C. and R.J. Campello (2015). Comparison of distributed evolutionary k-means clustering algorithms. Neurocomputing. 163: 78-93

Siraj, F., and M.A. Abdoulha (2007). Mining enrolment data using predictive and descriptive approaches. Knowledge-Oriented Applications in Data Mining. 53-72

TommiKarkkainen, S. (2006). Introduction to partitioning-based clustering methods with a robust example. Report, Dept. Mathematical IT., University of Jyvaskyla, (Finland).

Velusamy, K., A. Sakthivel, and M. Jayakeerithi (2014). Studies on clustering and fuzzy clustering. International Journal of Computer Science and Information Technologies. 5(4): 5231-5232

Velmurugan, T. (2012). Performance Comparison between K-Means and Fuzzy C-Means Algorithms Using Arbitrary Data Points. Wulfenia Journal. 19(8): 234-241

Verma, M., M. Srivastava, N. Chack, A.K. Diswar, and N. Gupta (2012). A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications. 2(3): 1379-1384.