ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Efficient Algorithm to fix Initial Centroid for Clustering High Dimensional data

M.Saranya*
Research scholar
Sri Ramakrishna College of Arts and Science for women,
Coimbatore, India.
saranya.sara@gmail.com

P. Krishnakumari
Associate Professor,
Sri Ramakrishna College of Arts and Science for Women,
Coimbatore, India.
kkjagadeesh@yahoo.co.in

Abstract: Clustering is one of the primary data analysis methods and K-Means clustering is one of the most well known popular clustering algorithm. The K-Means algorithm is one of the most frequently used clustering method in data mining, due to its performance in clustering massive data sets. The final clustering result of the K-Means clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. Because of the initial cluster centers produced arbitrarily, K-Means algorithm does not promise to produce the consistent clustering results. Efficiency of the original K-Means algorithm heavily rely on the initial centroids. The number of distance calculations increases exponentially with the increase of the dimensionality of the data. An algorithm is proposed which uses the Principal Component Analysis (PCA) in the first phase that simplifies the analysis and visualization of multi dimensional data set and in the second phase, a method is proposed to find the initial centroids to make the algorithm more effective and efficient. The proposed algorithm is implemented using MATLAB 7.10. This method provides more accurate results with less computational time compared to the existing algorithm.

Keywords: Clustering, K-Means, PCA.

I. INTRODUCTION

Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [6]. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters One of the most popular clustering method is K-Means clustering algorithm developed by Mac Queen in 1967 [7]. The easiness of K-means clustering algorithm made this algorithm used in several fields. The K-Means clustering algorithm is a partitioning clustering method that separates data into k groups [5]. The K-Means clustering algorithm is more prominent because it clusters massive data rapidly and efficiently. Because of the initial cluster centers produced arbitrarily, K-Means algorithm does not promise to produce the consistent clustering results. Efficiency of the original K-Means algorithm heavily rely on the initial centroids [2], [12]. Initial centroids also have an influence on the number of iterations required while running the original K-Means algorithm. The computational complexity of the original K-Means algorithm is very high, specifically for massive data sets [13].

The quality of the clustering algorithm greatly depends on the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. The computational complexity of original K-Means algorithm is very high, especially for large data sets. In addition the number of distance calculations increases exponentially with the increase of the dimensionality of the data. When the dimensionality increases usually, only a small number of dimensions are relevant to certain clusters, but data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. Moreover when dimensionality increases, data usually become increasingly sparse, due to which data points located at different dimensions can be considered as all equally distanced and the distance measure, which, essentially for cluster analysis, becomes meaningless. Hence, attribute reduction or dimensionality reduction is an essential data-preprocessing task for cluster analysis of datasets having a large number of attributes.

Traditional K-Means algorithm for cluster analysis developed for low dimensional data, often do not work well for high dimensional data. Also the computational complexity increases rapidly as the dimension increases. Hence, to improve the efficiency, PCA is applied on original data set, so that the correlated variables exist in the original dataset would be transformed to possibly uncorrelated variables, which are reduced in size. Before applying PCA the dataset needs to be normalized, so that any attribute with larger domain will not dominate attributes with smaller domain. The resulting reduced data set obtained from the application of PCA will be applied to a K-Means clustering algorithm [13].

A. M. Fahim, A.M.Salem, F.A.Torkey and M.A.Ramadan [1] proposed an enhanced method for assigning data points to the suitable clusters. In the original K-Means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive. In Fahim's approach the required computational time is reduced when assigning the data elements to the appropriate clusters. But in this method the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. K. A. Abdul Nazeer and M. P. Sebastian, [2] proposed an enhanced algorithm to improve the accuracy and efficiency of the K-Means clustering algorithm. In this

algorithm two methods are used, one method for finding the better initial centroids and another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time.

S. Deelers and S. Auwatanamongkol [6] proposed an algorithm to compute initial cluster centers for K-Means clustering. Data in a cell is partitioned using a cutting plane that divides cell in two smaller cells. The plane is perpendicular to the data axis with the highest variance and is designed to reduce the sum squared errors of the two cells as much as possible, while at the same time keep the two cells far apart as possible. Cells are partitioned one at a time until the number of cells equals to the predefined number of clusters, k. The centers of the K cells become the initial cluster centers for K-Means. The experimental results suggest that the proposed algorithm is effective, converge to better clustering results than those of the random initialization method. The research also indicated the proposed algorithm would greatly improve the likelihood of every cluster containing some data in it.

Kohei Arai and Ali Ridho Barakbah [5] proposed a new approach to optimize the initial centroids for K-Means. It utilizes all the clustering results of K-Means in certain times, even though some of them reach the local optima. Then, the result is transformed by combining with Hierarchical algorithm in order to determine the initial centroids for K-Means. The experimental results provide accurate results and improved clustering results as compared to some clustering methods.

Moth'd Belal and Al-Daoud [10] proposed a new algorithm to initialize the clusters. The proposed algorithm is based on finding a set of medians extracted from a dimension with maximum variance. The proposed algorithm avoids many unnecessary distance calculations by applying an efficient partial distance strategy. Experimental results show that the proposed algorithm gave better results than both the exhaustive and the conventional PD(partial distance) algorithms when applied to real data sets. Rcently an algorithm is proposed to find the better initial point [12]. In this algorithm the dataset is partitioned to k equal sets and the midpoint is fixed as initial centroid [12]. To improve the accuracy, a new method is proposed to find the initial centroids. The algorithm takes the datapoints that have maximum distance among the various data points. The main advantage of this approach is to obtain better clustering results with reduced complexity and also provides more accurate results with less computational time.

II. METHODOLOGY

A. Existing System

In the existing algorithm [12] , the data points are sorted in ascending order and then partitioned into k equal sets at random and the midpoint is fixed as initial centroid for each dataset. The method used for finding the initial centroids is computationally expensive.

B. Proposed Algorithm

As original K-Means clustering algorithm often does not work well for high dimension, the algorithm applies PCA on original data set in the first phase to obtain a reduced dataset and then a new algorithm is proposed to find the better initial centroid in the second phase. The Euclidean distance is calculated for all the data points. The data point that has the maximum distance is chosen as initial centroid. For more than two clusters the algorithm chooses the data point that have the next maximal distance. The determined data points are fixed as initial centroids and the remaining data points are assigned to appropriate clusters. As the quality of the final clusters heavily depends on the selection of the initial centroids, the new method proposed chooses data objects as initial centroids whose squared Euclidian distance is maximum among all the data objects, to make the algorithm more effective and efficient.

Algorithm: The proposed algorithm is as follows:

Input: $X = \{d1, d2, \ldots\ldots, dn\}$ // set of n data items.

K // Number of desired clusters.

Output: A set of k clusters

// **Phase-1:** Apply PCA to reduce the dimension of the data set

i. Organize the dataset in a matrix X.

ii. Normalize the data set using Z-score.

iii. Calculate the singular value decomposition of the data matrix.

iv. Calculate the variance using the diagonal elements of D.

v. Sort variances in decreasing order.

vi. Choose the p principal components from V with largest variances.

viii. Form the transformation matrix W consisting of those p PCs.

viii. Find the reduced projected dataset Y in a new coordinate axis by applying W to X.

//**Phase-2:** Find the initial centroids

ix. Compute the Euclidean distance between each data points in the set Y.

x. Choose the two data points x_o, y_o such that the distance (x_o,y_o) is maximum.

xi. set m=1.

xii. Centroid [m]=x_o, centroid[m+1]=y_o, m=m+1;

xiii. while (m<=k) // for k more than 2

a. Choose the next maximal distance between the data points (x_i,y_i) such that the data point y_o gets repeated.
b. centroid[m]=y_o.
c. m=m+1.

xiv. For each data point in the dataset Y, find the nearest cluster center from the list centroid that is closest and assign that datapoint to the corresponding cluster.

xv. Update the cluster centers in each cluster using the mean of the data points, which are assigned to that cluster.

xvi. Repeat steps 14 and 15 until there are no more changes in the value of the centroids.

III. RESULTS AND DISCUSSION

The proposed algorithm is implemented in MATLAB 7.10 with Intel(R) Core 2 duo CPU with a RAM capacity of 2 GB. The algorithm is tested with three datasets Iris [11], height-weight [11] and New-Thyroid [11] dataset. The datasets are taken from the UCI machine learning repository. The result is found to be more accurate. The original K-Means algorithm requires the initial centroids to be chosen at random. The proposed algorithm fixes the initial centroids systematically to assign the data points to the appropriate clusters. The Figure 1 shows the Sample output screen for proposed algorithm for Height Weight Dataset.

.

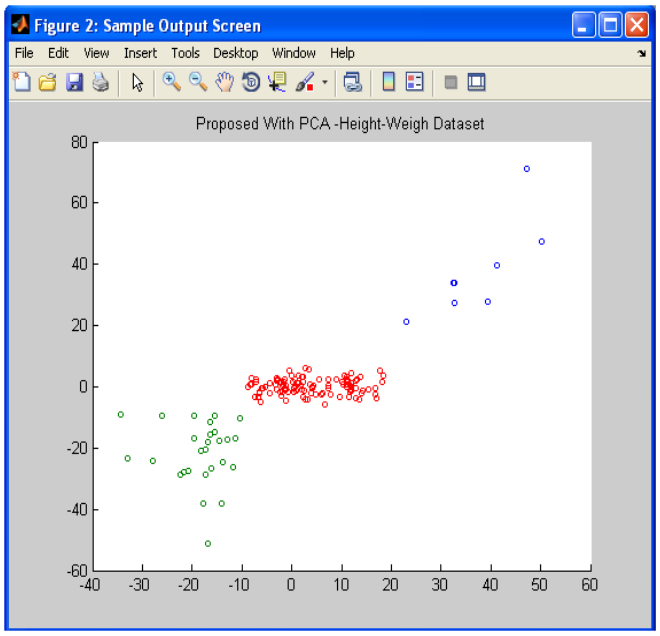

Figure 1: Sample Output screen for the proposed algorithm with PCA for Height-Weight dataset.

Figure 2 represents the time taken by the proposed method for the three datasets namely Iris, New-Thyroid and Echocardiogram. Figure 3 represents the accuracy of the proposed algorithms for the three datasets compared to the existing algorithm [12]. Rand Index is used to calculate the accuracy.

The proposed algorithm is implemented for the dataset Iris of data size 150 with 4 dimensions. By applying PCA the dimension is reduced to 3. The proposed method produces 96% accuracy. The algorithm is also implemented for the dataset Height-weight of data size 150 with 2 dimensions. By applying PCA the dimension is reduced to 1dimension.The proposed method produces 94% accuracy. The algorithm is also implemented for the dataset New-Thyroid of data size 215 with 6 dimensions. By applying PCA the dimension is reduced to 4 dimensions. The proposed method produces 96% accuracy.

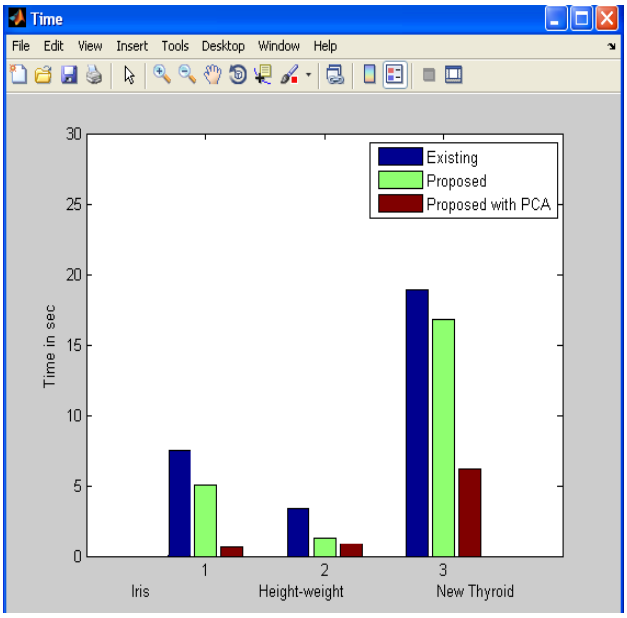

Figure 2: Chart representing the time of the datasets

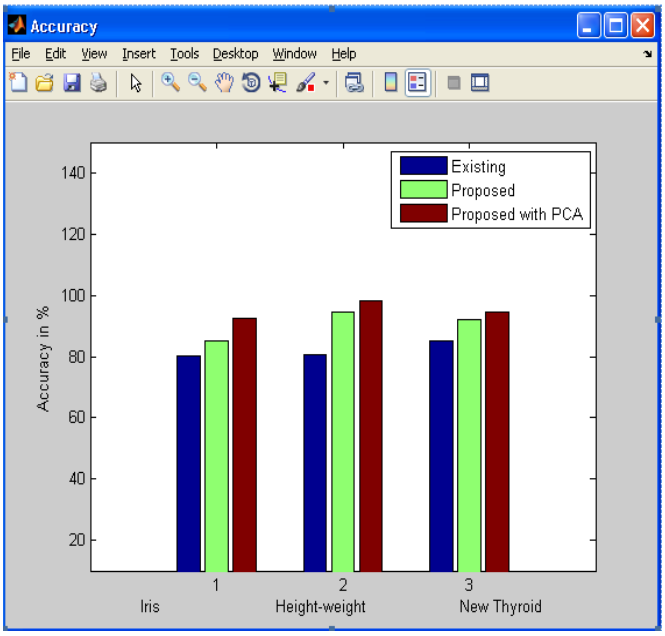

Figure 3: Chart representing the accuracy of the datasets

Table I shows the performance of the algorithms for the three datasets. The computational time of the proposed algorithm is less when compared to the existing algorithm. The accuracy of the proposed algorithm is found to be higher than the existing algorithm.

Table I Performance comparison of the algorithms

Dataset	Number of clusters	Algorithm	Accuracy %	Time (seconds)
Iris	3	Existing Algorithm	77	5.63
		Proposed Algorithm	94	5.05
		Proposed Algorithm +PCA	96	0.66
Height-Weight	3	Existing Algorithm	82	169.94
		Proposed Algorithm	93	0.27
		Proposed Algorithm +PCA	94	0.08
New-Thyroid	3	Existing Algorithm	82	1.40
		Proposed Algorithm	93	2.57
		Proposed Algorithm +PCA	96	0.11

IV. CONCLUSION

In this paper a new algorithm is proposed which finds the better initial centroids and the dimension is reduced using PCA. Using the proposed algorithm the given data set is partitioned in to k clusters in such a way that the sum of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are

maintained to be as large as possible. The experimental results show that the proposed algorithm provides better accuracy with less computational time compared to the existing algorithm [12]. Though the proposed method gives better quality results in all cases, over random initialization methods, still there is a limitation associated with this, i.e. the number of clusters (k) is required to be given as input. Evolving some statistical methods to compute the value of k automatically is suggested for future research.

V. REFERENCES

[1] 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.

[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] Chen Zhang and Shixiong Xia, "K-Means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[5] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[6] S. Deelers and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4.

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

[8] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," bioinformatics, Vol. 24, p. 1359-1366, 2008.

[9] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation" IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002.

[10] Moth'd Belal and Al-Daoud "A New Algorithm for Cluster Initialization" World Academy of Science,, Engineering and Technology, Vol.4, 2005.

[11] R.Indhumathi and Dr.S.Sathiyabama "Reducing and Clustering high Dimensional Data through Principal Component Analysis" International Journal of Computer Applications, December 2010.

[12] Madhu Yedla, Srinivasa Rao Pathakota and T M Srinivasa "Enhanced K-Means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, Vol.1(2), 2010,121-125.

[13] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath and Milu Acharaya "A hybridized K-Means clustering approach for high dimensional dataset", International journal of Engineering, Science and Technology, volume 2, No.2, 2010, pp. 59-66.