Efficient Cluster Initialization Method Using Principal Component Analysis

Yousef Kh. Majdalawi

Computer Information Systems Department, the University of Jordan, Amman 11942, Jordan
E-mail: ymajdal@ju.edu.jo
Tel: +962-77-7487105

Abstract: Clustering is a very well known technique in data mining, pattern recognition and image processing, used to group data according to shared characteristics or a degree of convergences. One of the most widely used clustering techniques is the k-means algorithm. Solutions obtained from this technique are dependent on the initialization of cluster centers (centroids). Whenever the initial centroids are closed to the representative one in each cluster, k-means algorithm gives better results. In this article I proposed a new method to initialize the clusters. The proposed method is based on the Principal Component Analysis (PCA). A comparison made between the conventional (random) and proposed method is performed. The new (proposed) method when applied to different data sets showed good results.

[Yousef Kh. Majdalawi. **Efficient Cluster Initialization Method Using Principal Component Analysis.** *J Am Sci* 2013;9(12): 333-336]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 45

Keywords: Clustering, K-means, Principal Component Analysis, Data Mining, Pattern Recognition, Image Processing.

1. Introduction

Clustering techniques have received attention in many areas including engineering, medicine, biology and data mining. The purpose of clustering is to group together data points, which are close to one another. The k-means algorithm [1] is one of the most widely used techniques for clustering.

The k-means algorithm starts by initializing the K cluster centers. The input vectors (data points) are then allocated (assigned) to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest data points. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers.

The steps of the k-means algorithm are written below:-

1. Initialization: choose K input vectors (data points) to initialize the clusters,
2. Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster,
3. Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster,
4. Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

However, it has been reported that solutions obtained from the k-means are dependent on the initialization of cluster centers [2]–[4].

There are two simple approaches to cluster center initialization: 1) Selecting the initial values randomly, 2) Choosing the first K samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters [5]. Therefore, different methods have been proposed in literature [6]–[8].

In the following sections, in section 2 a new algorithm is proposed for cluster initialization. The proposed algorithm finds a set of medians extracted from a dimension with maximum variance to initialize clusters of the k-means. The method can give better results when applied to k-means.

The rest of this paper presents the experimental results in section 3, then finally the conclusions in section 4.

2. Proposed Algorithm

The idea of the proposed method is based on Principal Component Analysis (PCA).

Principal Component Analysis (PCA) is a powerful tool that has been applied in many application areas such as data mining, intrusion detection and image processing [9]. It is a widely used technique for dimension reduction.

It is based on transforming a large number of variables (dimensions) into a smaller number of uncorrelated variables. This is done by finding a few orthogonal linear combinations of the original variables with the largest variance.

Given a data set with D variables, it is possible to construct a new set of p variables, p < D which are a linear transformation of the original dimensions [10]. The flow chart of original PCA is presented in Figure 1 [11].

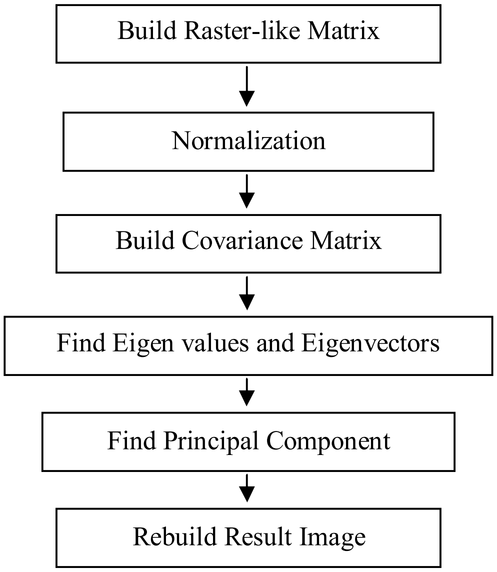

Figure 1: Flow chart of original PCA.

The first principal component of the transform is the linear combination of the original variables with maximum variance can keep the most characteristics of the sample points [12].

The proposed method works as follows:

1. Perform PCA,
2. Select the first Principal Component (PC),
3. Divide the dataset into k groups,
4. Find the average of each group,
5. Reconstruct the data for each group,
6. Use the resulting k data points as initial centroids,
7. Run the k-means algorithm with the initial centroids obtained from step 6.

3. Experimental Results

As discussed in [6] and [12] there is no general proof of convergence for the k-means clustering method. However, there exist some techniques for measuring clustering quality. One of these techniques is the use of the Sum of Square Error (SSE), representing distances between data points and their cluster centers. This technique has been suggested in [6] and [14].

The technique allows two solutions be compared for a given data set, the smaller value of SSE, the better solution.

The proposed method has been applied to two set. The first data set is the well known Ruspini [14] data set, while the second set, containing data points in 2, 4, and 8-dimensional formats, representing the well known Baboon image. Since no good method for initialization exists [15] and [16], we compare against the standard method for initialization: randomly choosing an initial starting points. Table 1 shows the initial results (initial SSE values) of the proposed method when applied to the first data set. The table shows that the proposed method works better than random initial centroids. These results are shown in Figure 2.

Table 1**:** The SSE values for Ruspini data set.

K	Random	Proposed	Deviation (Proposed-Random)
2	141771	89337	-52434
4	37547	20698	-16849
8	21392	20350	-1042
12	18497	12012	-6485

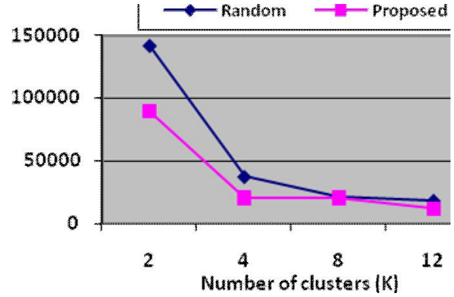

Figure 2: Graph that present SSE values in Table 1.

Tables 2, 3, 4 and Figures 3,4, 5 are presenting initial results (initial SSE values) when applied on the second data sets with different dimensions, for both random and proposed methods. The tables and Figures show that the results obtained from the new algorithm are better in different dimensions (2D, 4D, 8D).

As shown in Tables 1, 2, 3, 4 and Figures 2, 3, 4, 5 there is a remarkable reduction in SSE values that leads to better solutions when k-mean algorithm is applied.

Table 2: The SSE values for baboon data set (2D).

K	Random	Proposed	Deviation (Proposed-Random)
2	1.36384e+007	9.39992e+006	-4.24E+06
4	4.9146e+006	4.72002e+006	-1.95E+05
6	5.14676e+006	3.50381e+006	-1.64E+06
8	4.02572e+006	3.00218e+006	-1.02E+06
12	2.881e+006	2.53566e+006	-3.45E+05

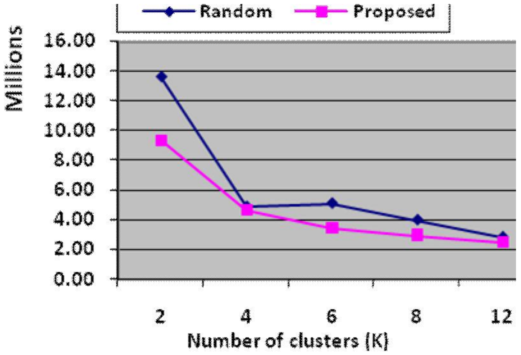

Figure 3: Graph that present SSE values in table 2.

Table 3: The SSE values for baboon data set (4D).

K	Random	Proposed	Deviation (Proposed-Random)
2	1.07713e+007	9.62684e+006	-1.14E+06
4	5.81384e+006	5.02862e+006	-7.85E+05
6	4.33441e+006	3.89765e+006	-4.37E+05
8	4.02572e+006	3.41302e+006	-6.13E+05
12	5.27143e+006	2.97799e+006	-2.29E+06

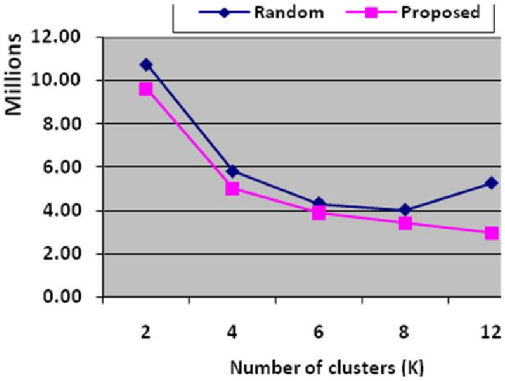

Figure 4: Graph that present SSE values in table 3.

Table 4: The SSE values for baboon data set (8D).

K	Random	Proposed	Deviation (Proposed-Random)
2	1.61556e+007	1.4124e+007	-2.03E+06
4	1.12798e+007	1.045e+007	-8.30E+05
6	240723	19211	-221512
8	9.2757e+006	9.21158e+006	-6.41E+04
12	5.27143e+006	2.91441e+006	-2.36E+06

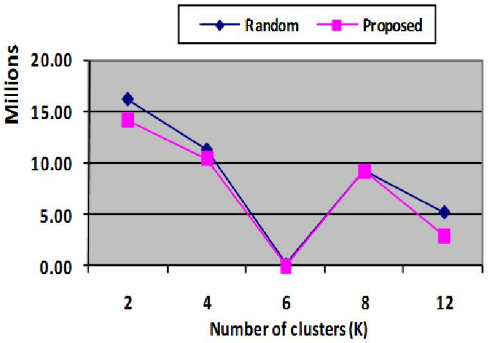

Figure 5: Graph that present SSE values in table 4.

4. Conclusions

In this paper we propose a new algorithm to initialize the clusters of the k-means algorithm. Two data sets were used, with different number of clusters and different dimensions. In all experiments, the proposed method gave best results in all cases, over randomly initialization methods, getting better quality results when applied to k-means algorithm.

References

1. J. MacQueen, Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. and prob, 1967, pp. 281-97.
2. P. Bradley, U. Fayyad, Refining initial points for k-means clustering, Proceedings 15th International Conf, on Machine Learning, San Francisco, CA, 1998, pp. 91-99.
3. N. Nasrabadi and R. King, Image coding using vector quantization: a review. IEEE trans. Comm. Vol. 36 (8), 1988, pp. 957-970.
4. J. Pena, J. Lozano and P. Larranaga, An Empirical comparison of four initialization methods for the k-means algorithm, Pattern Recognition Letters Vol. 20, 1999, pp. 1027-1040.
5. M. Ismail and M. Kamel, Multidimensional data clustering utilization hybrid search strategies. Pattern Recognition Vol. 22 (1), 1989, pp. 75-89.
6. G. Babu and M. Murty, A near optimal initial seed value selection in kmeans algorithm using a genetic algorithm. Pattern Recognition Letters Vol. 14, 1993, pp. 763-769.
7. C. Huang and R. Harris, A Comparison of several vector quantization codebook generation approaches. IEEE trans. Image Proc. Vol 2 (1), 1993, pp. 108-112.
8. Y. Linde, A. Buzo and R. Gray, An algorithm for vector quantizer design. IEEE trans. Comm. Vol. 28 (1), 1980, pp. 84-95.

9. Wei Wang, Xiaohong Guan and Xiangliang Zhang, "Processing of Massive Audit Data Streams for Real-Time Anomaly Intrusion Detection ". Computer Communications Journal, vol. 31, no. 1, pp. 58-72, 2008.
10. Tajunisha, N. and V. Saravanan, An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means, Internationa Journal of Database Management Systems (IJDMS), Vol.3, No.1, pp. 196-205, 2011.
11. Yi Zhou, 2010, "Principal Component Analysis Based Image Fusion Routine with Application to Stamping Split Detection", Ph.D. Thesis, Clemson University.
12. Chin-Chen CHANG, Chi-Shiang CHAN, A Watermarking Scheme Based on Principal Component Analysis Technique, INFORMATICA, Vol. 14, No. 4, 2003, pp.431–444.
13. N. Venkateswarlu and P. Raju. Fast isodata clustering algorithms. pattern recognition Vol. 25 (3), 1992, pp. 335-342.
14. E. H. Ruspini (1970)," Numerical methods for fuzzy clustering". *Inform. Sci.* 2, 319–350.
15. A. Gersho and R. Gray, Vector quantization and signal compression, CAP, 1992.
16. M. Meila and D. Heckerman, An experimental comparison of several clustering methods, Microsoft Research Technical Report MSR-TR-98-06, Redmond, WA, 1998.

11/12/2013