International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

A New Algorithm for Cluster Initialization

Moth'd Belal. Al-Daoud

Abstract**—**Clustering is a very well known technique in data mining. 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. In this article we propose 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 algorithm has been applied to different data sets and good results are obtained.

Keywords**—** clustering, k-means, data mining.

I. 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. 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].

Two simple approaches to cluster center initialization are either to select the initial values randomly, or to choose 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, a new algorithm is proposed for cluster initialization. The proposed algorithm finds a set of medians extracted from the dimension with maximum variance to initialize clusters of the k-means. The method can give better results when applied to k-means.

II. THE NEW PROPOSED ALGORITHM

The idea of the algorithm is to find the dimension with maximum variance, sorting it, dividing it into a set of groups of data points then finding the median for each group, using the corresponding data points (vectors) to initialize the k-means.

The method works as follows.

1. For a data set with dimensionality, *d*, compute the variance of data in each dimension (column).
2. Find the column with maximum variance; call it *cvmax* and sort it in any order.
3. Divide the data points of *cvmax* into *K* subsets, where *K* is the desired number of clusters.
4. Find the median of each subset.
5. Use the corresponding data points (vectors) for each median to initialize the cluster centers.

III. EXPERIMENTAL RESULTS

As discussed in [6], [9], 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], [10].

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

The proposed method has been applied to two sets of random and real data points to compute different sets of clusters. The first data set (which contains different data points and different dimensional formats) was generated randomly, 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 [11], we compare against the standard method for initialization: randomly choosing an initial starting points. In this paper the average of 8 initial runs was chosen for the random method.

Tables 1, 2 and 3 are presenting initial results (initial SSE values) when applied on the first data sets, for both random and new methods.

M. B. Al-Daoud is with the Computer Information Systems Department, University of Jordan, Amman-Jordan, (e-mail: mba@ju.edu.jo).

TABLE I
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 1, WITH 2D

No. Clusters	Rand Init SSE	New init SSE
32	782,945	578,956
64	689,380	459,747
128	476,167	309,289
256	327,995	202,818

TABLE 2
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 1, WITH 4D

No. Clusters	Rand Init SSE	New init SSE
32	761,785	521,753
64	673,119	531,798
128	561,222	229,303
256	489,554	149,564

TABLE 3
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 1, WITH 8D

No. Clusters	Rand Init SSE	New init SSE
32	891,037	351,444
64	803,092	237,378
128	596,800	158,005
256	378,848	113,067

The tables above show that the results obtained from the new algorithm are better in all cases. This is also true when using different number of clusters.

Tables 4, 5 and 6 are presenting final results (after applying the k-means algorithm) on the first data sets, for both random and the proposed methods using the same stopping criteria.

TABLE 4
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 1, WITH 2D

No. Clusters	Rand Final SSE	New Final SSE
32	456,115	455,982
64	333,064	324,498
128	225,118	208,102
256	155,353	142,985

TABLE 5
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 1, WITH 4D

No. Clusters	Rand Final SSE	New Final SSE
32	372,017	358,861
64	262,124	230,065
128	186,715	152,785
256	141,300	104,550

TABLE 6
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 1, WITH 8D

No. Clusters	Rand Final SSE	New Final SSE
32	302,342	260,689
64	225,931	170,064
128	154,368	115,960
256	101,304	82,846

The tables above show that the final results obtained from the new algorithm are better in all cases. The results also show that final results are much better when applying the proposed method on higher dimensions.

Tables 7, 8 and 9 are presenting initial results (initial SSE values) when applied on the second data sets (the baboon data with different dimensions), for both random and new methods. The tables show that the results obtained from the new algorithm are better in all cases. This is also true when different numbers of clusters are used.

TABLE 7
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 2, WITH 2D

No. Clusters	Rand Init SSE	New init SSE
32	666,405	101,020
64	256,306	72,340
128	135,263	7,475
256	86,628	46,085

TABLE 8
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 2, WITH 4D

No. Clusters	Rand Init SSE	New init SSE
32	349,746	108,447
64	189,961	86,761
128	166,977	74,174
256	89,871	61,662

TABLE 9
RANDOMLY AND NEW INITIAL VALUES FOR DATA SET 2, WITH 8D

No. Clusters	Rand Init SSE	New init SSE
32	180,202	125,958
64	166,986	113,990
128	147,316	94,618
256	105,565	82,086

Tables 10, 11 and 12 are presenting final results (after applying the k-means algorithm) on the second data sets, for both random and the proposed methods using the same stopping criteria.

TABLE 10
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 2, WITH 2D

No. Clusters	Rand Final SSE	New Final SSE
32	133,702	82,276
64	81,645	62,852
128	55,251	43,376
256	37,007	32,182

TABLE 11
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 2, WITH 4D

No. Clusters	Rand Final SSE	New Final SSE
32	108,289	87,310
64	84,011	73,160
128	69,174	60,666
256	61,662	50,217

TABLE 12
RANDOMLY AND NEW FINAL VALUES FOR DATA SET 2, WITH 8D

No. Clusters	Rand Final SSE	New Final SSE
32	97,686	96,975
64	85,715	84,496
128	75,940	75,503
256	67,053	65,681

The results above show that final results (after running the k-means) obtained from our proposed algorithm are always better when applied on the second data set.

IV. SUMMARY

In this paper we propose a new algorithm to initialize the clusters of the k-means algorithm. The proposed algorithm finds a set of medians extracted from the dimension with maximum variance to initialize clusters. Two data sets were used, with different number of clusters and different dimensions. In all experiments, the proposed algorithm gave best results in all cases, over randomly initialization methods, getting better quality results when applied to k-means algorithm. This is also true when different sets of cluster centers are used.

REFERENCES

[1] J. MacQeen, 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] N. Venkateswarlu and P. Raju. Fast isodata clustering algorithms. pattern recognition Vol. 25 (3), 1992, pp. 335-342.
[10] A. Gersho and R. Gray, Vector quantization and signal compression, CAP, 1992.
[11] M. Meila and D. Heckerman, An experimental comparison of several clustering methods, Microsoft Research Technical Report MSR-TR-98-06, Redmond, WA, 1998.