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 \( cv_{max} \) and sort it in any order.
3. Divide the data points of \( cv_{max} \) 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.
The tables above show that the results obtained from the new algorithm are better in all cases. This is also true when using different numbers 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.

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.

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