Modify Initialization k-means Clustering Algorithm to Generate Initial Centroids

Lamia AbedNoor Muhammed

College of Computer and Mathematics

University of Al-Qadissiyah

bon1491988@yahoo.com
Modify Initialization k-means Clustering Algorithm to Generate Initial Centroids

Abstract

K-means is one of the most common clustering techniques used with numeric data. Different issues are conducted in k-means algorithm in order to reach the optimum solutions with best situations, weather producing good results or the ways used to produce the results efficiently. Initial centroids of this algorithm play important role, so the generation initial centroids attracting more work. However, this paper aims to discuss a new proposed step to improve the generation of initial centroids i.e. modification the first iteration of k-means algorithm. The experiment work of this paper would be applied with one of the famous data that is "iris", this data is suited with k-means algorithm. The experiments were tested with the origin k-means algorithm in two parameters: "execution time" and "cost function" that is represented by sum square error SSE. The results are promise work with this modification

Keywords: centroids, clustering, initial centroids k-means clustering

I. Introduction

Clustering algorithm has a broad attraction and usefulness in exploratory data analysis[1]. Thousands of clustering algorithms have been proposed in the literature in many different scientific disciplines[2]. The k- Means clustering algorithm is an old algorithm that has been intensely researched owing to its ease and simplicity of implementation[1]. The research works in this field have been varied in different issues. Initialize the k-means algorithm with center objects that are called centroid is a critical problem because this algorithm is sensitive to these values and may reach to local minima. So there are some research work to deal with this problem. One technique that is commonly used to address the problem of choosing initial centroids is to perform multiple runs, each with a different set of randomly chosen initial centroids, and then select the set of clusters with the minimum sum of the squared error (SSE)[3]. The work in [4], suggested the creation of initial centroids through generate iteratively these centroids from n objects. At each time select the centroid that has maximum distances from the previous centroids. Other
work in [5], the creation of initial centroids is performed through selection portion of data and then generate the best ones from. In [6], k-means algorithm was modified i.e. partitions the whole space into different segments and calculates the frequency of data point in each segment. The segment which shows maximum frequency of data point will have the maximum probability to contain the centroid of cluster. In[7] , compute the distance between each data point and all other data points. Then the put the closest data points in set data points in subsets that represent the cluster and compute the centroids for each one by averaging the points data in subset. In[8], the proposed algorithm search about the points data that has maximum distance between them in order to present as centroids.

II. Clustering

Clustering analysis is an important technique in the rapidly growing field known as exploratory data analysis and is being in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, and remote sensing[9]. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression[10]. Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments. Cluster analysis is a tool for exploring the structure of the data that does not require the assumptions common to most statistical methods. It is called "unsupervised learning" in the literature of pattern recognition and artificial intelligence[9].

III. K-Means Clustering

The most popular partitional algorithm among various clustering algorithms is K-mean clustering. K-means is an exclusive clustering algorithm, which is simple, easy to use and very efficient in dealing with large amount of data with linear time complexity[11]. In 1967 MacQueen
first proposed k-Means clustering algorithm. k-Means algorithm is one of the popular partitioning algorithm. The idea is to classify the data into k clusters where k is the input parameter specified in advance through iterative relocation technique which converges to local minimum[1]. K-means algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the of the cluster. The cluster’s mean is then recomputed and the process begins again[11] as shown in Algorithm(1) [12].

Algorithm(1): Traditional k-means Algorithm

a) Select k points as initial centroids
b) Repeat
c) From k clusters by assigning each point to its closest centroid
d) Recompute the centroid of each cluster until centroid does not change

The objective function of K-mean algorithm is to minimize the intra-cluster distance and maximize the inter-cluster distance based on the Euclidean distance[11]. So to measure the quality of a clustering , sum square error (SSE) is used i.e. the error for each point is computed. The SSE is formally defined as follows[3]:

\[ SSE = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i) \]  

(1)

In the above equation, \( m_i \) is the center of cluster \( C_i \), while \( d(x, m_i) \) is the Euclidean distance between a point \( x \) and \( m_i \). Thus, the criterion function \( E \) attempts to minimize the distance of each point from the center of the cluster to which the point belongs [13].

The K-means algorithm requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. Cluster initialize with one object that is center for each cluster. A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid, the average of all the points in the cluster[12].

Different initializations can lead to different final clustering because K-means only converges to local minima. One way to overcome the local minima is to run the K-means algorithm, for a given K, with multiple different initial partitions and choose the partition with the smallest squared error[2].
IV. Work Algorithm

The work algorithm contains proposed steps that would be applied in the generation initial centroids. The first iteration in k-means algorithm is modified in order to generate the initial centroid. The random selection of points data are candidate as centroids. Then execute the first iteration with the proposed modification as shown in Algorithm(2) i.e. the initial random points are selected to be initial centroid and then compute the distance between each point and the centroid. The minimum distance will make the point would be assigned to that cluster with minimum distance. After that the specific fraction from the difference (between the point and this cluster centroid) will be added to this centroid as shown bellow.

\[ c(i,j) = c(i,j) + r \times (x(p,j) - c(i,j)) \]  \hspace{1cm} (2)

where \( c \) represents centroid for cluster \( i \), \( x(p,j) \) is the point \( p \) in data space that was assigned to cluster \( i \), \( r \) is the fraction value.

**Algorithm(2): proposed steps for first iteration of k-means algorithm**

1- Input \( k \) random points as centroids for \( k \) clusters
2- Assign parameter \( r \) with value range between 0.1 and 1.0
3- Repeat
4- Present data point \( x \)
5- Assign point \( p \) to the cluster \( i \) with minimum distance
6- Recompute centroid \( c \) of cluster \( i \) as :
   \[ c(i,j) = c(i,j) + r \times (x(p,j) - c(i,j)) \]
   \[ j : \text{features of data point} \]

V. Practical Work

A. Experiment work

The practical work was applied with experiment data that is known "iris" and the code was performed in MATLAB9000a. The traditional algorithm as shown in Algorithm(1) first was executed then proposed algorithm as shown in Algorithm(2) would be executed with different fraction value \( r \) parameter that
range(0.1, 0.2, …., 1.0). The aim of the practical work is extracting the SSE and execution time for different runs of traditional and proposed algorithms. The runs are subjects to the following issues:

1- Execute the traditional and proposed algorithms for different no. of cluster (k) i.e. k=2, k=3, k=4, k=5.

2- Assign initial centroid with random points.

3- For each cluster no. (k), there are 5 different runs with different initial centroids for each algorithm.

4- The initial centroid that was applied in traditional algorithm would be used in proposed algorithm with different runs for different (R) parameter values.

B- Results and Discussion

The results were produced from the experiment are in two sides, SSE as shown in Table(1) and execution time in Table(2). The values of SSE/execution time were computed for traditional k-means algorithm and proposed algorithm with different r range(0.1, …., 1.0). Each cell in these tables represent SSE/execution time for five runs with five different random initial centroids according to the cluster no. that use. In order to compare between the two algorithms, the initial centroids data itself were used in different algorithm with different clusters.

The results in Table(1) show the superior of the proposed algorithm for different parameters in contrast to traditional algorithm and this more clear in Figure(1) case(a, b, c, d), where the two lines represent the two algorithms; traditional algorithm and proposed algorithm with r=0.1. Other issue is concerned with the parameter r, the result reveal the better to use the less value for r parameter, however the good results.

In other side, the execution time is another measure that was used in comparing the two algorithms. The execution time results are shown in Table(2) that organized in the same form of table(1). The result consolidate the superior of the proposed algorithm. The execution time of the proposed algorithm in general about half of traditional algorithm execution time. Figure(2) in for case(a, b, c, d) shows graphically this comparing for traditional algorithm and proposed algorithm with parameter (r=0.1).
Figure (1) SSE values produced for (4 cases), each one represents different no. of clusters k for traditional and proposed algorithm with r=0.1
Figure (2) execution time for (4 cases), each one represents different no. of clusters k for traditional and proposed algorithm with r=0.1
VI- conclusion

The comparison of the two algorithms assessed that the proposed algorithm results are more prompt to use it. However achieving good results through two measurements; SSE and execution time can be exploited in developing the traditional k-means algorithm in various cases such as:

1- Generating the clusters with different trails using the proposed algorithm, then choose the best ones
2- Using the these initial centroids in traditional algorithm in order to access the better final step with less time.

Reference