Volume 4, Issue 12, December 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Several Technical Methods for Selecting Initial Cluster Centers in K-Means Clustering Algorithm

Uday Kumar S
PG Scholar, Dept. of CSE,
NMAMIT, Nitte, India

Naveen D Chandavarkar
Asst. Professor, Dept. of CSE
NMAMIT, Nitte, India

Abstract— Drastic growth of digital data is an emerging area of concern which has led to concentration of Data Mining technique. The actual data mining task is the programmatic or semi-programmatic analysis of large quantities of data to extract hidden interesting patterns such as groups of data records, which is usually referred as Cluster Analysis or Clustering. Clustering is the classification of data objects into different groups, or more incisively the partitioning of a data set into subsets (clusters), so that the data objects of each cluster share common characteristics.

Several clustering algorithms have been proposed among which k-means is one of the simplest unsupervised learning algorithm that will solve the well-known clustering problem. K-means clustering generates a specific number of disjoint, flat clusters. The k-means algorithm requires k initial cluster centers that must be specified beforehand and are randomly selected. However, k-means algorithm must be executed several times to obtain good clustering result which directly depends on the initial random selection of k. This paper illustrates certain techniques for selecting the k initial cluster centers thereby restricting the random selection of initial clusters. The techniques specified in this paper significantly improve the clustering effect when compared to traditional k-means algorithm.

Keywords— Clustering, Data Mining, Density Area, Distance Measure, K-Means Algorithm, Initial Clusters Centroids, Mapreduce.

I. INTRODUCTION

With the fast growth of information technology, various organizations are accumulating huge amount of data every day. However, these accumulated data is highly unstructured and it may consist of important and useful information hidden in it. The process of extracting hidden pattern and knowledge from huge data set and transforming it into human understandable structure for further processing is called as data mining which can also be called as knowledge discovery from data. The knowledge extraction can be achieved by using a sequence of iterative steps which process the data at each step by using different algorithm. Data mining consists of following six common classes of tasks:

Anomaly detection (Outlier/deviation/change detection) – The identification of peculiar data records, that might be interesting or data errors that require further investigation.

Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a shopping mall might collect data on customer purchasing behavior. Using association rule learning, the shopping mall can find which items are frequently bought together and use this information for improving the marketing strategies.

Clustering – is the process of discovering classes and structures in the data that are in some way similar, without using known structures in the data.

Classification – is the process of generalizing known structure to apply to new data.

Regression – attempts to find a function which models the data with the least error.

Summarization – providing a more compact representation of the data set, including visualization and report generation.

Most commonly used data mining techniques for extracting hidden patterns in the data are clustering and classification analysis. Classification is a data mining technique which assigns data items in a collection to a target classes which must be previously defined. Clustering is also used to segment the data but it segments the data into groups (classes) that were not defined in prior. Clustering is especially used for exploring the data. Clustering in data mining is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. Clustering analysis is an unsupervised machine learning method and it is also an active research field in data mining which has very extensive applications [1]. There are two types of clustering algorithms: 1) Hierarchical algorithm and 2) Partition algorithm. Hierarchical clustering algorithm finds successive clusters using previously generated clusters (example: Agglomerative algorithm and divisive algorithm.) Partition clustering algorithm constructs various partitions and then evaluates them by some criterion (example: k-means and fuzzy c-means). The main goal behind the clustering process is that the objects within a group must be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, will results in better or more distinct the clusters.

Many clustering algorithms are available, in that k-means is most widely used in data mining applications because of its simplicity and fast processing ability of voluminous data sets. K-means clustering is an algorithm to classify or to group the objects based on features/attributes into K number of groups, where k is a positive integer. The clustering is

done by minimizing the sum of squares of distances between the data points and the corresponding cluster centroid. The algorithm accepts the count of clusters and the initial set of cluster centroids as input parameters. It computes the distance between each data point and the initial set of cluster centroid. The point which has least distance from a particular centroid is then added to that cluster. After the assignment of all data points to the respective clusters, the centroids of the clusters are recalculated. Euclidean distance is the most commonly used method for distance measure. Suppose there are two points defined as P=(x1(p), x2(p), x3(p)….) and Q=(x1(q), x2(q), x3(q)….). The Euclidean distance between two multi-dimensional data points is defined as:

$$D(P, Q) = \sqrt{(x1(p) - x1(q))^2 + (x2(p) - x2(q))^2 + \dots} = \sqrt{\sum_{j=1}^{p} (xj(p) - xj(q))^2} \quad (1)$$

The main part of this algorithm is to obtain a minimal squared difference between the centroid of the cluster and the data point in the dataset.

$$|X^{(j)}_i - C_j|^2 \quad (2)$$

Where X_i is the value of the data in the dataset and C_j is the value of the centroid of the cluster [2].
The primary steps of the algorithm are as follows:

- Select k points as initial centroids arbitrarily.
- Calculate the distance (Euclidean distance) between all the data points in the dataset and k centroids. Add each data points which have least distance from the centroids into the nearest cluster. (Arrange each data into the nearest cluster.)
- After processing all the points, compute the average of all the points in each cluster. The average of each cluster is used as the new centroid of that particular cluster.
- Calculate the difference between the new centroid with the original one in the same cluster. if the difference is smaller than the threshold or the number of iterations of the algorithm has been reached for the maximum, the algorithm is over. Otherwise, the new centroid is substituted for the original ones. Return to step two and continue.[1]

The k-means algorithm is relatively efficient and fast. It computes result at O (tkn**),** where n is the number of objects or points, k is the number of clusters and t is the number of iterations. The computational complexity of the k-means algorithm mainly depends on the number of clusters, number of data elements and the number of iterations.

K-means algorithm has made its footprint in many areas, ranging from unsupervised learning of neural network, Pattern recognitions, machine vision, image processing, Artificial intelligence, Classification analysis, and many others.

Although k-means is simple and faster to use, it has its own limitations. The accuracy of the final clustering highly depends on the arbitrary selection of the initial centroids. This paper discusses the different methods for selecting initial cluster center rather than selecting random initial center and provides a comparative report on the methods being illustrated in this research study.

II. TECHNIQUES FOR SELECTING INITIAL CLUSTER CENTERS

Method 1: This method illustrates the research work by Madhu et al.[3], the algorithm described here initially checks whether the data sets given as an input contains any negative value attributes or not. If it finds any negative value in the data set, then all the data points in the data set must be transformed. The transformation can be achieved by subtracting each data point attribute value from the minimum attribute value.

Transformation is required in this algorithm, because it computes the distance from the origin to each data point in the data set. During this distance calculation if the data set consists of negative attribute data point, then for different data points as shown in figure 1 we may get the same Euclidean distance from the origin. This causes the improper selection of the initial centroids. Hence, to overcome this problem all the data points are transformed to the positive point's space. Then for all the data points we will get unique distance from the origin. If the data set does not contain any negative value attributes, then the transformation is not necessary.

In the next step of the algorithm, for each data point it calculates the distance from the origin. Then sort all the distances and based on this data points are sorted. After sorting all data points, divide the sorted data points into k equal sets. In each set it will take the middle points as the initial centroids. These initial centroids yields better unique clustering results.

Next, algorithm computes the distances between each data point in the data set to all the initial centroids. The next stage is to reduce the required computational time by using an iterative process which makes use of a heuristic approach. In the iterative process initially the data points are assigned to the clusters having the closest centroids. It uses two arrays; one is ClusterId which will store the ClusterId of each data points. The ClusterId of a data point denotes the cluster to which it belongs. Second array used is NearestDist which will store the present nearest distance of each data point. The NearestDist of a data point denotes the present nearest distance from closest centroid.

Now for each cluster the new centroids are calculated by taking the mean of its data points. Then for each data points, algorithm computes the distance from the new centroid of the present nearest cluster. If this distance is less than or equal

to previous nearest distance, then the data point stays in the same cluster, otherwise for each data point it calculates the distance from all centroids. After calculating the distances, the data points are assigned to the appropriate clusters and the new ClusterID's and new NearetsDist's values are updated. This process of reassigning is repeated until the convergence criterion is met [3].

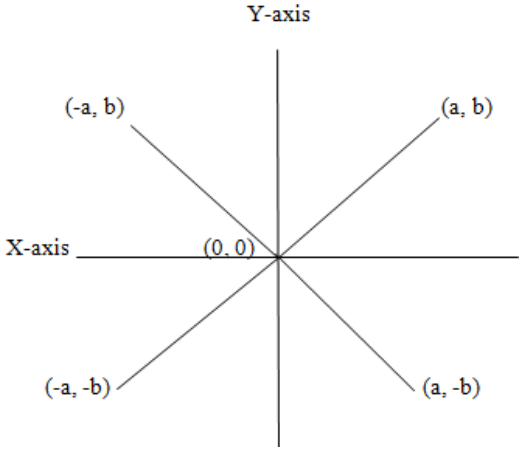

Fig 1.Data points in Two Dimensional Space

Method 2: This method illustrates the research work by Jiangang Qiao et al. [4], the algorithm is used for choosing initial cluster centers which consists of three parts;

a). Reducing the number of dimensions:

In this step in order to speed up the selection of initial cluster centers, a two dimensional subspace is selected from the feature space, i.e. two main variables which are most representative for the original data are selected for initializing the cluster centers. The first variable must be the one with maximum absolute value of the coefficient of variation (CV), where the coefficient of variation can be determined using,

$$CV_j = |S(x_j)/\ddot{x}_j|,\ j=1, 2, ..., p \qquad (3)$$

Where $S(x_j)$ is the standard deviation, $\ddot{x}j$ is the mean of the jth attribute variable, and p is the number of features.

Next for selecting the second variable this method makes use of the correlation coefficient (CC) of the variables. The correlation coefficient is defined as,

$$CC_{jj'} = \sum_{i=1}^{n}(xij - \ddot{x}j)(x - \ddot{x}j) \,/\, \sum_{i=1}^{n}(xij - \ddot{x}j)^2 \sum_{i=1}^{n}(xij - \ddot{x}j)^2 \qquad (4)$$

The lowest absolute value of correlation coefficient $CC_{jj'}$ means the j^{th} and j'^{th} attribute columns are most independent. The variable which is most independent from the first main variable is selected as the second main variable. Based on these two variables this method describes the selection of cluster center.

Consider a data set X, let v1 and v2 be the two main variables which is determined in the previous step. Now construct a new data set X' consisting of only variables v1 and v2. A radius parameter R is computed using the data in X' and this parameter is used to choose the cluster centers.

b). Determination of a radius R:

- Choose 100 data points arbitrarily from X'.
- Compute the distance between each of the data point and its nearest neighbour point in X'.
- Find the maximum of 100 distances, the value of R is four times of this maximum distance.

c).Choosing the candidate cluster centers:

Choose a data point from the set X' which is marked as the first candidate center. Now compute the mean of the data points that exist in the circle cantered at the candidate center having the radius R, a data point which is very nearest from the mean is selected as the new position of the cluster center. Repeat this procedure until the center attains a stable position. The selection of remaining new candidate center uses a probability measure i.e. farther a point is from the selected centers, higher is the probability for choosing that point as a new center. For each data points calculate distance from the cluster centers found previously, the shortest distance from each of the points to the centers is divided by the maximum distance from all the data points to the centers. This ratio of distances of each point is used as the probability for the data point to be chosen as the next new candidate cluster center.

After selecting the new candidate cluster center, it is relocated similarly as the first candidate center until it converges to a new position. Check whether the new position found after relocating is same as the previous cluster centers or not. If it is same, then original position of the candidate center before shifting is used. Otherwise, it is selected as the new cluster center. Repeat these above steps until all the k cluster centers are found. Finally choose k cluster centers in X which corresponds to the selected cluster centers in X' [4].

Method 3: This method illustrates the research work by Kunhui Lin et al. [5], k-means clustering algorithm with optimized initial centers based on data dimensional density is been described in this research work. The basic idea behind the algorithm is to choose the k (where k is the number of clusters) points which are farthest from each other in the high density areas as the initial center. This idea helps to exclude some isolated points that will affect the final clustering results, and also achieve final stable clustering results faster.

The necessary concepts that are used in this method are defined as follows:

- Average distance R, means the average distance between any two points of the data set.
- Density index is defined as the average count of the points in each cluster. As we know, points cannot be uniformly distributed to each cluster, so the density index must float down 20%-40% based on the experience.
- High density area, can be defined as follows:

Given a point, a circle is considered with given point as center and the average distance R as the radius; compute the number of points that exists in the circle. If the counts of points are greater than the density index, the point is in the high density area. Otherwise, the point is in the low density area.

Based on the concepts defined above, this method proposed the k-means algorithm with optimized initial centers. The steps of the algorithm are explained as follows:

1) Compute the distance of each point between all other points in the data set. The distance is calculated using Euclidean distance.
2) Calculate the average distance R of the data set using the formula

$$R = \sum_{\substack{i \le n \\ i<j \le n}} di,j \; / \; (n * (n-1)/2) \qquad (5)$$

Here the numerator is the sum of distance between all points, and the denominator is the total number of the distances between points.

3) Calculate the density index of the data set as defined above.
4) Check whether a point is in high density area. If so, add the point to high density point set. Otherwise, add the point to low density point set.
5) Choose k points from the high density point set which have the farthest distances from each other as initial cluster centers. The actual step is to sort the distances of all points in the high density point set, and select the two points which have the farthest distances assigned to the initial center set. Then compare the distances of points in the initial center set to other points, find the farthest distance to the initial center set until the number of points in the initial center is equal to the number of cluster [5].

Method 4: This method illustrates the research work by Tshibault Debatty et al. [6], the algorithm employs an iterative technique called G-means to determine the number of clusters when performing k-means clustering. The G-means algorithm uses Anderson-Darling test to verify whether a subset of data follows a Gaussian distribution. It runs K-means with increasing values of k in a hierarchical fashion until the test accepts the hypothesis that the points assigned to each center follow a Gaussian distribution.

The G-means algorithm initially starts with a small number of clusters, and increases the number of centers. In each iteration, the algorithm runs k-means to refine the current centers. If the cluster data appears does not follow a Gaussian distribution, then that cluster must be split. Otherwise retain the original center.

The steps for G-means are described as follows:

For each cluster X of center c,

1. Determine two new centers c1 and c2 in the cluster.
2. Execute K-means algorithm to refine c1 and c2
3. Let v = c1 - c2 be the vector that connects the two centers.
4. Let X' is the projection of X on v. X' is a one dimensional representation of the projection on v.
5. Normalize X', so that it has zero mean and variance equal to 1.
6. Use Anderson-Darling test to test X': If X' follows a normal distribution, keep the original center, and discard c1 and c2. Otherwise, split the cluster in two, use c1 and c2 as new centers and run the algorithm on each sub-cluster.

This algorithm takes the advantage of simplified test for Gaussian fit by the projection of data to one dimension where the test is simple to apply. It creates new centers wherever it is needed and also improves the clustering quality [6].

III. COMPARISON OF TECHNIQUES ADOPTED FOR INITIAL CLUSTER CENTERS SELECTION

Table 1 gives the performance wise comparison of different methods used for selecting initial cluster centers for k-means algorithm with respect to the traditional k-means algorithm. Each of these methods may be suitable for its own set of data set and has its own advantages and shortcomings with respect to data set taken for analysis. In this comparison the first method works with the relative type of data sets and it runs with lesser computational time, higher accuracy i.e. the time complexity is O (nlog n) but, the value of k is necessary to be given as input. Second method used the variable data set and it overcomes the disadvantage of previous method. This method avoids the selection of initial cluster centers in low density area. This may cause final bad clustering results by using dimension reduction, probability measure for

selecting cluster centers and relocating the cluster centers towards high density area. Hence, it improves the accuracy of the final cluster results. The drawback is the speed of this method is not appreciated.

Table-I Comparison of Different Methods With The Traditional K-Means Algorithm

Methods/Author	Best Suited Data Sets	Time complexity	Accuracy	Efficiency
Method 1/ Madhu et al.[3]	Relative Data Sets	Low	High	High
Method 2/ Jiangang Qiao et al. [4]	Variable and Distributive Data Sets	High	High	Moderate
Method 3/ Kunhui Lin et al. [5]	Numerical Data Sets	Moderate	High	Moderate
Method 4/ Tshibault Debatty et al. [6]	Large Data Sets	Low	High	High

The third method mainly considered numerical data set for analysis; it improves the stability of the k-means algorithm and achieves good performance. The main disadvantage is, by adding this optimized algorithm may cause new problem. If the size or dimension of the data set is increased, the computational time will increase substantially because of the time consuming processes are calculating the distance of all points. This problem can be solved by executing the algorithm in parallel by implementing in MapReduce framework. Fourth method is scalable to very large data set which is best suitable for this method and it overcomes the limitation mentioned in the previous method since it is implemented in Map Reduce framework with some modifications of the original algorithm. The computational cost is proportional to nk. But, this method overestimates the number of clusters.

IV. CONCLUSIONS

As clustering has been one of the significant research attentions, many clustering algorithms have been proposed in the recent years. K-means clustering algorithm is one of the most well-known clustering algorithms because of the simplicity and fast speed. However, the final clustering result mainly relies on the initial cluster centers which are selected randomly and also it is computationally very expensive. In this paper, different methods for selecting initial cluster centers so as to improve the final cluster quality has been discussed. Each method provides different solution for the problem of selecting initial cluster center and improves the stability of k-means clustering. Though each method has its own merits and demerits, MapReduce implementation of the above methods serves improved results and is highly scalable for very large data sets.

REFERENCES

[1] Qing Liao, Fan Yang, Jingming Zhao, "An Improved parallel K-means Clustering Algorithm with MapReduce", Communication Technology (ICCT), 2013 15th IEEE International Conference on. IEEE, 2013.

[2] Prajesh P Anchalia, Anjan K Koundinya, Srinath N K, "MapReduce Design of K-Means Clustering Algorithm", Information Science and Applications (ICISA), 2013 International Conference on. IEEE, 2013.

[3] Madhu Yedla, Srinivas Rao Pathakota, T M Srinivasa, "Enhancing K-means clustering Algorithm with mproved Initial Center", International journal of computer science and Information Technologies, Vol. 1(2) , 2010, 121-125.

[4] Jiangang Qiao, Yonggang Lu, "A new algorithm for choosing initial cluster centers for K-means", International Conference on Computer Science and Electronics Engineering(ICCSEE) 2013).

[5] Kunhui Lin, Xiang Li, Zhongnan Zhang, Jiahong Chen, "A K-means Clustering with Optimized Initial Center Based on Hadoop Platform", The 9th International Conference on Computer Science & Education (ICCSE 2014) August 22-24,2014. Vancouver, Canada.

[6] Thibault Debatty, Wim Mees, Pietro Michiardi, Olivier Thonnard, "Determining the k in K-means with MapReduce", 17th International Conference on Database Theory EDBT-ICDT 2014, Athens, Greece, 2014.

