ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

K-Means Clustering With Initial Centroids Based On Difference Operator

Satish Chaurasiya[1], Dr.Ratish Agrawal[2]

M.Tech Student, School of Information and Technology, R.G.P.V, Bhopal, India

Assistant Professor, School of Information and Technology, R.G.P.V, Bhopal, India

ABSTRACT: Clustering is a well known technique in data mining which is used to group together data items based on similarity property. k-means clustering algorithm is a one of the commonly used cluster analysis method for extracting useful information in terms of grouping data. However one of its drawbacks is that initial centroid points are selected randomly because of which algorithm has to re-iterate number of times, and also accuracy of the K Means algorithm depends much on the chosen central values. This paper first reviews existing methods for selecting selecting initial centroid points, followed by a proposed method for selecting the initial centroid points .

KEYWORDS: Data mining, Clustering, K-means Clustering, Initial cluster centroids, Difference operator.

I. Introduction

Clustering is the process by which a set of data objects is partitioned into subsets such that the data elements of a same cluster are similar to one another and different from the elements of other clusters [1].The set of clusters as outcome from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clusters on the same data set. The partitioning is performed by the clustering algorithms. Cluster analysis now days has wide range of applications in machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

There are different methods of clustering such as Partitioning Method, Hierarchical Method, Density Based Method and Grid Based Method. In partitioning method given k is the number of partitions to construct, a partitioning method creates an initial partitions then an iterative relocation technique is used to improve the partitioning by interchanging objects from one cluster to another. A hierarchical method tries to create a hierarchical decomposition of the given set of data objects. In Density Based Method of clustering a given cluster is partitioned as long as the density in the neighborhood reaches threshold. In Grid-based methods the object space are quantized into a finite number of cells that form a grid structure and then all clustering operations are performed on these grid structures

A forward difference operator, denoted by D, is defined by the equation $Df(x) = f(x+h) - f(x)$, where h is a constant denoting the difference between successive points of interpolation or calculation.

II. K-means Algorithm

K-means is one of the least complex unsupervised learning algorithm based on decomposition that tackle the notable clustering problem. In traditional K-means algorithm K is used as a parameter, Divide n object into K clusters, to create relatively high similarity between elements in the same cluster, relatively low similarity between elements in other clusters and minimize the total distance between the values in each cluster to the cluster center. The mean value of each cluster is the cluster center. The calculation of similarity is done by mean value of the cluster objects. The distance between the objects is measured by using Euclidean distance. The closer is the distance, bigger is the similarity of two objects, and vice versa.

Algorithm: k-means. The k-means algorithm for clustering, where each cluster's center is represented by the

mean value of the objects in the cluster.
Input: k: number of clusters,
D: a data set with n objects.
Output: k clusters.
Method:
1. Randomly select k objects as the initial centre clusters.
2. repeat
3. (re)assign each object to the cluster to which it is most similar, based on the mean value of the
objects in the cluster.
4. Update the cluster means.
5. Until no change.

III. Performance Analysis

A. *ADVANTAGES:*

- K-means is a classical algorithm to resolve clustering problems simply and quickly and it is easy to implement and understand.
- Better efficiency in clustering high dimensional data..
- Complexity of k-mean algorithm is O(ntk) where n represent number of objects, t is number of iteration and k is number of cluster.
- Often terminate at local optimum.

B. *DISADVANTAGES*:

- In K-means algorithm user need to supply number of cluster that is k.
- It is sensitive to the initial centroids [2] and change in initial centroids leads to different clustering results
- Unable to handle noisy data and outliers
- Does not apply directly to categorical data

IV. Existing Methods

The earliest method to initialize K-means was suggested by Forgy in 1965. Forgy's method involves selecting initial centroids randomly from the dataset. The Forgy approach takes advantage of the fact that if we selects points randomly we are more likely to choose a point near a cluster centre by virtue of the fact that this is where the highest density of points is located [3].

McQueen proposed an approach know in literature as McQueen Approach [4] in 1967. He proposed a method very similar to the Forgy method. He suggested that, as with the FA above, K instances are selected at random from the database as seeds. The next step is where MA differs from FA. Instead of assigning all remaining instances to one of the K nearest seed locations, and iterating the K-means algorithm until it converges, we assign one instance at a time, in the order they occur in the database, to the nearest cluster centre. After each instance is assigned, the K-means algorithm is run before the next instance is assigned to a cluster. A drawback of this technique is the computational complexity,several iterations of the K-means algorithm are needed after each instance is assigned, which in a large database is extremely burdensome.

Simple Cluster Seeking method was suggested by Tou and Gonzales in 1974. In this approach, the first seed Initialize with the first object in the database and then it calculate the distance between this seed and the next object in the database, if calculated distance is greater than some threshold then it is selected as the second seed, otherwise move to the next object in the database and repeat the process. Once the second seed is chosen move to the next object in the database and calculate the distance between it and the two seeds already selected, if both these distances are greater than the threshold then select as the third seed. This process is repeated until K seeds are chosen [5].

Katsavounidis et al. (1994) proposed a method termed as the KKZ algorithm [6]. This algorithm starts the first seed by choosing a point x, preferably one on the 'edge' of the data. The point furthest from x is chosen as the second seed. The distance of all points to the nearest of the first and second seeds is calculated. The point which is the furthest from its nearest seed is chosen as the third seed. We repeat the process of choosing the furthest point from its nearest seed until K seeds are chosen. This method has major drawback that any noise in the data, in the form of outlying data points, will pose difficulties. Any such outlying points will be preferred by the method but will not necessarily be near a cluster centre.

Bradley and Fayyad present a technique for initializing the K-means algorithm [7]. The process begin by randomly breaking the data into 10 small sub-subsets. Then perform a K-means clustering on each of the 10 subsets, all starting at the same initial seeds which are chosen using Forgy's method with provision that empty clusters at termination will have their initial centers re-assigned and the sub-sample will be re-calculated. The result of the 10 runs is 10K centre points. These 10K points are then themselves input to the K-means algorithm and the algorithm run 10 times, each of the 10 runs are initialized using the K final centroid points among one of the 10 subset runs. The result thus obtained are initial cluster centroids for the K-means algorithm. The main advantage of the method is that it increases the efficiency of the result by the fact that initial centroids are obtained by multiple runs of the K-means algorithm. The major disadvantage of this initialization method is that it requires more computational effort.

Koheri Arai et al. proposed an algorithm for initialization of centroids for K-means algorithm. In this algorithm both k-means and hierarchical algorithms are utilized. This method uses all the clustering results of k-means in certain times. Then, the result is transformed by combining with Hierarchical clustering algorithm to find the better initial cluster centers for k-means clustering algorithm [8].

Yunming Ye et al. proposed a new method for selecting initial cluster centers in k-means clustering. This method first identifies the high density neighborhoods from the data set and then the central points of the neighborhoods are selected as initial centers [9].

K. A. Abdul Nazeer et al. suggested an enhanced algorithm for finding initial clusters. This method starts by calculating the distances between each data point and all other data points in the dataset. Next it find out a pair of data points which are closest to each other and it forms a set S1 consisting of these data points. These two data points are then deleted from the data point set D. Then it find the data point which is closest to the data points in the set S1. And this point is added to S1 and is deleted from dataset D. This process is repeated until the number of elements in the set S1 reaches a threshold. At that point repeat the second step and form another data-point set S2. This process is repeated till k such data point sets are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids [10].

Neha Aggarwal and Kirti Aggarwal represented a Mid-point based k-mean clustering algorithm in which auto-generate initial partition rather than randomly selection [11]. If dataset include the negative value attributes, then all the attributes are changed to positive space by subtracting each data point attribute with the minimum attribute value in the data set. This transformation is required because the distance from origin to each data point is calculated in the algoritthm. So if there are both positive and negative values in database, then for different data point’s similar Euclidean distance will be obtained and which will result in incorrect selection of initial centroids. After that distance of each data object from origin is calculated. Based on these calculated distance data object is sorted. In next step, dataset is divided into k partition. For each partition, mid-point is calculated which is used as initial center for that partition.

Raed T. Aldahdooh and Wesam Ashour proposed a new distance based method of initial centroid selection for k-means clustering algorithm [12]. This algorithm starts by selecting random point as initial centroid and then performs some calculation to verify whether the selected point is noise or not. To verify noisy point, the distance between selected centroid and each point in data set is computed and then sort the data points based on the resulted distances, then divide data set into k partition with N points and the average distance between each pair of N points is computed. If this distance is greater than predefined threshold valve the selected random point is neglected and another point is selected for first center, otherwise second initial center is selected randomly, this process continue until k-partition is generated. After that, arithmetic mean is calculated for each partition which is taken as cluster center.

V. Initial Centroids Based on Difference Operator.

K-Means algorithm select K numbers of cluster centroid points from the given dataset randomly. But this random selection leads to more number of iterations also selecting different centroid points every time gives different clusters thus leading to almost different output each time. To overcome these problems centroid determination method has been proposed which uses forward difference operator to determine the best possible cluster centroid points which will at least reduce the number of times K-Means algorithm need to re-iterate and thus enhance efficiency of K-Means algorithm.

Algorithm: selecting the Initial Cluster Centroids
Input: dataset containing n items and k Number of clusters
Output: A set of k initial centroids.
Steps:

1. Sort the given data set
2. Apply forward difference operator on sorted data set obtained from step 1
3. Select k maximum forward difference distances from step 2
4. Partition the data set into k maximum forward difference distances
5. In each partition, take the mean as the initial centroid.

VI. Simulation and Results

We have implemented the k-Means with initial centroids based on difference operator in R with different datasets. The data values and the value of k are the only inputs required by this proposed algorithm. Here we used the value of k=4 and then at the beginning we use traditional k-Means algorithm and then we used proposed method for selection of initial centroids instead of selecting initial centroids randomly. By using k-Means with initial centroids based on difference operator we obtained good clustering results. The experiment is conducted 5 times for different sets of values. In each experiment, the iteration and execution time was computed and taken the average iteration and execution time of all experiments. The k-Means with initial centroids based on difference operator is better than selecting the initial centroids randomly.

No of cluster	Algorithm	Iterations
k=4	K-means	4
k=4	K-Means with initial centroids based on difference operator	1

Table 1: Resulted iteration

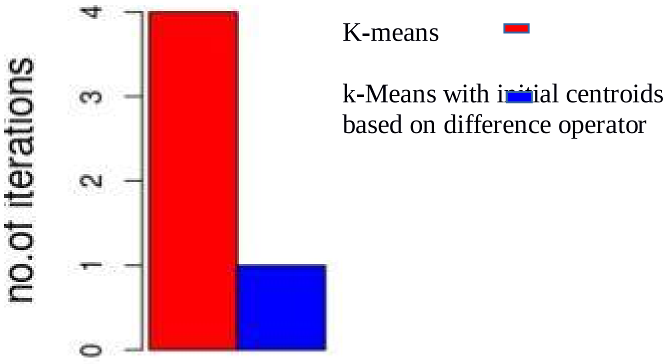

Figure 1: Resulted iteration

VII. Conclusion

As the results of the k-mean clustering method are highly dependent on the selection of initial centroids, so there should be a systematic method to determine the initial centroids which makes the k-mean algorithm to converge in global optima and unique clustering results in less number of iteration. An overview of the existing methods along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper to overcome the deficiency of the traditional K-means clustering algorithm. The new method suits for both uniformly and non-uniformly distributed data. The proposed method uses a systematic and efficient way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional K-means algorithm.

REFERENCES

1. Jiawei Han, Data mining: concepts and techniques (Morgan Kaufman Publishers, 2006).
2. Pena, J.M., Lozano, J.A., Larranaga, P, An empirical comparison of four initialization methods for the K-Means algorithm, Pattern Recognition Letters 20 (1999) pp. 1027-1040.
3. Anderberg, M, Cluster analysis for applications (Academic Press, New York 1973).
4. MacQueen, J. B., 1967. Some methods for classification and analysis of multi-variate observation. In: In Le Cam, L.M and Neyman, J., editor, 5 Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.
5. Tou, J., Gonzales, Pattern Recognition Principles (Addison-Wesley, Reading, MA, 1974).
6. Katsavounidis, I., Kuo, C., Zhang, Z., 1994. A new initialization technique for generalized lloyd iteration. IEEE Signal Processing Letters 1 (10), 144-146.
7. Bradley, P. S., Fayyad, Refining initial points for K-Means clustering: Proc. 15th International Conf. on Machine Learning, San Francisco, CA, 1998, pp. 91-99.
8. Koheri Arai and Ali Ridho Barakbah, Hierarchical k-means: an algorithm for centroids initialization for k-means, Reports of The Faculty of Science and Engineering Saga University, vol. 36, No.1, 2007.
9. Ye Yunming, Advances in knowledge discovery and data mining (Springer, 2006).
10. K. A. Abdul Nazeer and M. P. Sebastian, Improving the accuracy and efficiency of the k-means clustering algorithm, Proceedings of the World Congress on Engineering, London, UK, vol. 1, 2009.
11. Neha Aggarwal and Kirti Aggarwal (2012a), A mid-point based k-mean clustering algorithm for data mining, International Journal on Computer Science and Engineering, Vol. 4, No. 06.
12. FRaed T. Aldahdooh and Wesam Ashour (2013), DIMK-means ―Distance-based Initialization Method for K-means Clustering Algorithm, I.J. Intelligent Systems and Applications, 02, pp. 41-51.