International Journal of Pure and Applied Mathematics
Volume 118 No. 18 2018, 1565-1570
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

INITIALIZING CENTROIDS FOR K-MEANS ALGORITHM –AN ALTERNATIVE APPROACH

V.R.Geetha,
M.TECH (IT-2nd YEAR),
Hindustan Institute of Technology & Science,
Padur.

DR.K.RameshKumar (HOD)
Department of Information Technology
Hindustan Institute of Technology & Science,
Padur.

ABSTRACT: Clustering is the method of grouping the Data points based on their similarity. Clustering does not require any class labels. Many algorithms are existing to perform clustering that includes centroid based clustering, Hierarchical clustering, and Agglomerative clustering. The classic K-means algorithm is a centroid based clustering algorithm. Even though the algorithm is efficient, there is no specific method for choosing the initial centroids. This paper proposes an alternative method for choosing the initial centroids. Clustering finds its application in many real world domain that includes Wireless Sensor Networks, Mobile Computing, Data Analytics, Customer Segmentation, Portfolio Management, etc.

KEY WORDS: Data mining, Clustering, K-means, Centroids Initialization.

INTRODUCTION :

Data Mining is the process of digging out any useful information or patterns from the existing collection of Big Data. Big Data is the data that is collected from various devices and internet. The data stored could be processed to get useful insights. Such insights could be extctracted with the help of Data mining algorithms. Out of which clustering is an important method Clustering helps in obtaining the groups from the data points. The datasets that do not have any clear labels can be grouped with clustering process. To perform clustering some standard algorithms are being used. K-means is one of those algorithms. But the use of K-means have one drawback. That is, its centroids are not initialized in a procedural way. This paper focus on the initialization of the centroids for the existing K-means algorithm. Measures are available to determine the goodness of the clustering that has been performed. One of them is the Reduced Intra-cluster Distance. Intra-cluster distance can be defined as the distance between the data points within the same cluster. It should be low so that the data points are meant to be packed closely hence improving the cluster quality. The intra-cluster distance is calculated as the total sum of squares of the data points within the same cluster.

EXISTING ISSUES:

The notable issue of the existing K-means algorithm is the random selection of the initial centroids. The classic K-means select its centroids randomly. No proper method is suggested for the process. Problem arises when the algorithm undergoes multiple runs .For each run of the same algorithm and the same dataset, different clustering results will be obtained. This is not a reliable method when working with many practical applications like wireless sensor networks .For example, while implementing clustering with random centroids for clustering the wireless sensor nodes the results will be random which could not be accepted in the practical world.

EXISTING ALGORITHM:

1. Input the Dataset.
2. Choose the number of clusters.
3. Select the centroids randomly based on the number of clusters.
4. Find distance between the available data points and the picked up random centroids. The distance is calculated using the Euclidean Distance or Manhattan Distance.
5. Assign the particular data point to the cluster which has minimum distance.
6. Re-compute the mean of the particular cluster to select the new cluster head.
7. Repeat the above process for the remaining data points.

PROPOSED SOLUTION:

This paper proposes an alternate method for initializing the centroids for the K-means algorithm .The idea is to pick up the initial cluster heads based on some methodology that could improve the quality of the resulting cluster and the results are expected to be same for different number of runs and the dataset.

PROPOSED ALGORITHM:

1. Input the Dataset.
2. Choose the number of clusters.
3. Find the Distance between every data point and the origin. The distance is calculated using the Euclidean Distance or Manhattan Distance.
4. Sort the data points based on the above calculated Distance.
5. Divide the sorted data points into K number of clusters chosen.
6. Pick the middle data point from each resulting cluster and make it the initial cluster heads.
7. Find the distance between each data points to the above chosen centroids.
8. Assign data points to the cluster whose distance is minimum.
9. Re-compute the Cluster mean once allocated.
10. Repeat the above process for all other data points in the set.

MATHEMATICAL EXPLANATION:

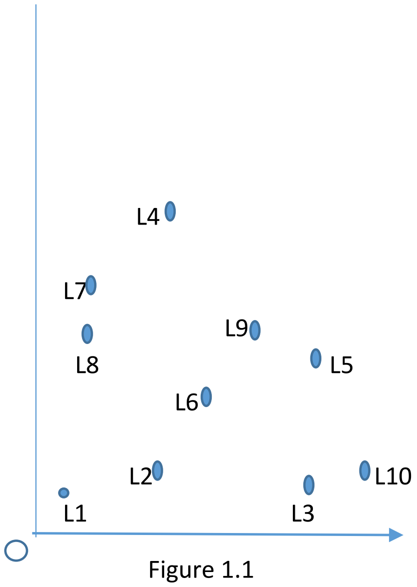

Figure 1.1

Consider the sample data points on the space be length1, length2, length3, length4 till length10.They are represented on the figure as L1, L2 L3 …L10.Their distance from the origin should be calculated with the above formula. Hence data points are given below in sorted order of their distances.

L1, L7, L8, L2, L4, L6, L9, L3, L5, L10

Let the number of clusters be 3.So, divide the data points into 3 clusters in their sorted order. Pick the middle point as the initial cluster head. Here, the points are L7, L4, L3 the initial centroids chosen.

Initial clusters are,

C1--->L1, L7, L8

C2--->L2, L4, L6

C3--->L9, L3, L5, L10

Find the distance between each data point and the cluster heads to find the nearest one. Here the nearest cluster to point L1 is C2. Assign L1 to C2. Recalculate the mean of C2 as it is assigned a new point. Now the recalculated mean of the cluster C2 will be ……… Which will be the next centroid point of C2. Continue the above process for all the available data points say L10.The final cluster formed will be in the following group.

C1--->L8, L7, L4

C2--->L1, L2, L6, L9, L5

C3--->L3, L10

So, for every run of the algorithm the procedure for picking up the initial centroids remain the same and the resulting cluster will also be the same. This process is more regularized than the existing K-means algorithm with initial centroids.

Following the above procedural way improves the quality of the resulting cluster by reducing the Intra-cluster distance. The Intra-cluster Distance is one of the method for evaluating the "Goodness" of the cluster [1].

EXPERIMENTAL RESULTS:

The proposed algorithm has been implemented using the **R** programming language with the sample Datasets and the Results has been visualized.

Table 1.1

Dataset-1	DIABETES
Dataset-2	CARS
Source	R - datasets

Both the above datasets have been implemented in R and compared with the Classical K-means algorithm.

The Bar Cart for the implementation is shown below.

For DIABETES Dataset:

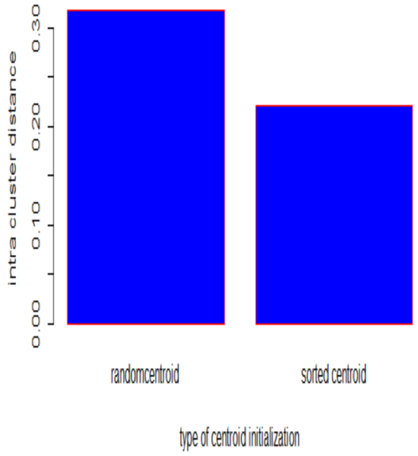

Figure 1.2

The sample Dataset has 7 number of attributes and 280 rows that makes a small dataset. The proposed algorithm has been run on R programming language which is especially for Statistical Data Analysis. The result obtained is compared with the classic K-means and is visualized as a Bar chart.

For CARS Dataset:

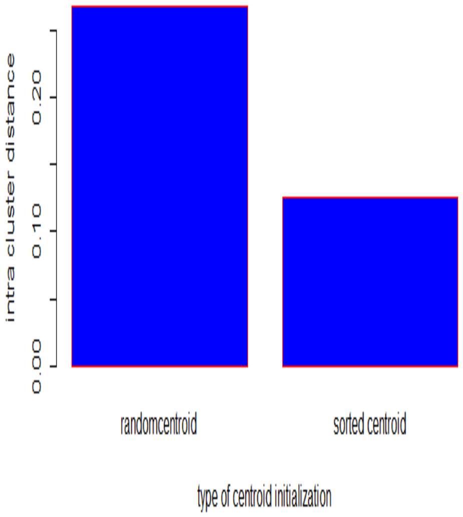

Figure 1.3

In both the cases, we can see the reduced Intra-cluster distance compared with the Classic K-means with randomly initialized centroids. Hence this paper focuses only on obtaining quality clusters, the performance based on the run-time has not been evaluated. With the current super-fast machines, the Quality of the clustering overrides the total run time performance. We can say, the algorithm holds good for Data sets of smaller size. The results are tabulated in the next part of the paper.

CONCLUSION:

The experimental results has been tabulated below. From the table we can come to the conclusion that the proposed algorithm reduces the intra-cluster distance and the results are consistent for multiple runs of the algorithm. Since the algorithm still depends on the user's choice for selecting the number of clusters needed, it may be considered as the future work for finding the optimal number of clusters for the given dataset.

Some standard method should be advised to decide upon the cluster numbers needed. Also, this paper focussed on the Intra-cluster Distance alone and can be enhanced to verify more advanced quality measures in the future. The algorithm is verified for clustering but the verification for the same over classification can be made as a future work.

REFERENCES:

1. Text book of “Monte Carlo Study on thirty internal measures for cluster analysis” by Glenn Milligan.

2. Applying Multivariate Statistical Analysis-STAT505.

3.“K-Means Clustering with Initial Centroids Based on Difference Operator”,Dr.Ratish Agarwal, Satish Chaurashiya-International Journal of Innovative Research in Computer and Communication Engineering, Vol 5, Issue 1, January 2017,ISSN 2320-9801.

4. “A novel hybrid clustering algorithm for microblog topic detection”, Yanmei Zhang. International conference on material science (MRSEE2017).

5. “Modifying initialization K-means clustering algorithm to generate initial centroids”, Lamia Abed Noor Muhammad-Journal of AlQadissi of Computer Science- v. 6, n. 2, p. 176-185, Aug. 2017. ISSN 2521-3504.