New Method for Finding Initial Cluster Centroids in K-means Algorithm

Harmanpreet Singh
M-Tech Research Scholar
Shri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab, India

Kamaljit Kaur
Assistant Professor
Shri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab, India

ABSTRACT

Data Mining is special field of computer science concerned with the automated extraction of patterns of knowledge implicitly stored in large databases, data warehouses and other large data repositories. Clustering is one of the Data Mining tasks which is used to cluster objects on the basis of their nearness to the central value. It is a method of grouping objects automatically. In clustering elements within same cluster are more similar than the elements in other clusters. K-Means is one the method of clustering which is used widely because it is simple and efficient. The output of the K Means depends upon the chosen central values for clustering. So accuracy of the K Means algorithm depends much on the chosen central values. The original K Means method chooses the initial cluster centroids randomly which affects its performance. This paper presents a new method for finding initial cluster centroids for K Means.

General Terms

Centroids, Complexity, Dataset, Modified K-Means, K-Means

Keywords

Arithmetic Mean, Clustering, Cluster Distance, Efficiency Partitions

1. INTRODUCTION

Clustering is the process of partitioning a set of data objects into subsets such that the data elements in a cluster are similar to one another and different from the elements of other clusters [1]. The set of clusters resulting from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clusterings on the same data set. The partitioning is not performed by humans, but by the clustering algorithm. Cluster analysis has wide range of applications in business intelligence, image pattern recognition, Web search, biology, and security [2].

There are many methods of clustering which include: Partitioning Method, Hierarchical Method, Density Based Method and Grid Based Method. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique to improve the partitioning by moving objects from one cluster to another. A hierarchical method creates a hierarchical decomposition of the given set of data objects. In Density Based Methods a given cluster as long as the density (number of objects or data points) in the neighbourhood exceeds some threshold. Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure [1].

2. K-means Algorithm

The k-means clustering algorithm was developed by Mac Queen in 1967. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups [2]. Despite being used in a wide array of applications, the K-Means algorithm is not exempt of drawbacks. Some of these drawbacks have been extensively reported in the literature. The most important is that the K-Means algorithm is especially sensitive to initial starting conditions (initial clusters and instance order) [3]. Various methods have been devised to solve this problem but there is always an efficiency and accuracy trade off. This paper reviews various algorithms for choosing initial centroids in K-means.

```
Algorithm: K-means algorithm for clustering
Input: number of clusters k and a dataset of n objects.
Output: a set of k clusters
1. Arbitrarily choose k objects as the initial centre clusters.
2. repeat
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster;
5. Until no change;
```

Fig 1: K-means Algorithm

3. EXISTING METHODS

Various methods for the calculation of initial clusters in K-means algorithm are given below:

The earliest method to initialize K-means was proposed by Forgy in 1965. Forgy’s method involves choosing initial centroids randomly from the database. This approach takes advantage of the fact that if we choose 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 [4].

Simple Cluster-Seeking (SCS) method was suggested by Tou and Gonzales. This method initializes the first seed with the first value in the database. It then calculates the distance between the chosen seed and the next point in the database, if this distance is greater than some threshold then this point is chosen as the second seed, otherwise it will move to the next instance in the database and repeat the process. Once the second seed is chosen it will move to the next instance in the database and calculate the distance between this instance and the two seeds already chosen, if both these distances are greater than the threshold then select the instance as the third seed. This process is repeated until K seeds are chosen [5].

KKZ method is named after the first alphabet of last name of each of the persons who had proposed the method. In the first step a point x is chosen as the first seed, this point is preferably at the edge of the data. Then the method finds a point furthest from x and this point will be the second seed. Then the method calculates the distance of all points in the dataset to the nearest of first and second seed. The third seed is the point which is the furthest from its nearest seed. The process of choosing the furthest point from its nearest seed is repeated until K seeds are chosen [6].

Bradley and Fayyad suggested a new technique for finding initial cluster centroids in K-means algorithm. In the first step the data is broken down randomly into 10 subsets. In the second step K-means algorithm is applied on each of the 10 subsets, the initial centroids for these are chosen using Forgy's method. The result of the 10 runs of the K-means algorithm is 10K centre points. These 10K points are then given as input to the K-means algorithm and the algorithm run 10 times, each of the 10 runs initialized using the K final centroid locations from one of the 10 subset runs. The result thus obtained is initial cluster centroids for the K-means algorithm [7].

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

Samarjeet Borah and Mrinal Kanti Ghose proposed Automatic Initialization of Means (AIM) algorithm. In this method the original dataset D is first copied to a temporary dataset T. The algorithm is required to run n times i.e. equal to the number of objects in the dataset. The algorithm selects the first mean of the initial mean set randomly from the dataset. Then this object (which is selected as mean) is removed from the temporary dataset. Then the distance threshold is calculated by employing a certain procedure. This method calculates the average distance with existing means of a new object which is considered as the candidate for a cluster mean. If the candidate satisfies the distance threshold then it is considered as a new mean and is deleted from the temporary dataset. The algorithm detects the total number of clusters automatically. This algorithm also has made the selection process of the initial set of means automatic. AIM applies a simple statistical process which selects the set of initial means automatically based on the dataset [9].

Yunming Ye et al. proposed a new method for effectively selecting initial cluster centers in *k*-means clustering. This method identifies the high density neighborhoods from the data first and then selects the central points of the neighborhoods as initial centers [10].

K. A. Abdul Nazeer et al. proposed 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. Then it find out a pair of data points which are closest to each other and it forms a set A1 consisting of these data points. These two data points are then deleted from the data point set D. It then find the data point which is closest to the data points in the set A1. Then this point is added to A1 and is deleted from dataset D. This process is repeated until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data-point set A2. This is repeated till k such sets of data points 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 [11].

Madhu Yedla et al. proposed a simpler algorithm for choosing the initial clusters. The proposed algorithm first checks whether the given data set contain the negative value attributes or not. If the data set contains the negative value attributes then all the data points are transformed to the positive space by subtracting the each data point attribute with the minimum attribute value in the given data set. In the next step, for each data point the distance from the origin is calculated. Then, the original data points are sorted accordance with the sorted distances. After sorting partition the sorted data points into k equal sets. In each set take the middle points as the initial centroids. These initial centroids lead to the better unique clustering results [12].

4. MODIFIED K-MEANS

The efficiency of the K-Means algorithm depends heavily on the chosen initial clusters. So much research has been done to find initial clusters in the K-Means algorithm. But the problem of finding initial clusters still persists. No one method is able to completely solve the problem. There is trade off between these two. When we try to maximise efficiency then accuracy suffers and on the other hand if one tries to maximise accuracy then efficiency suffers. In this section an algorithm is proposed that tries to balance both the accuracy and efficiency of the K-Means algorithm. The method for finding initial centroids is given below:

Algorithm: Finding the Initial Cluster Centroids

Input: set of n data items and k Number of desired clusters

Output: A set of k initial centroids.

Steps:

1. User supplies the value of k i.e. number of clusters.
2. Arithmetic mean of the whole data is calculated, this will be the first cluster centre.
3. Data is then divided into two parts.
4. Mean of these two parts is then calculated these will be second and third cluster centres respectively.
5. This process is repeated until k cluster centres are found.

Fig 2: Finding Initial Cluster Centroids

The efficiency of the algorithm depends upon the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. A dataset is chosen on which both the original k-means algorithm and the modified k-means algorithm are applied. The results of both the algorithms are given below:

Table 1: Comparison of K-Means & Modified K-Means

K	Sum of Cluster Distance	
--------	**K-Means**	**Modified K-Means**
2	40303.2	40303.2
3	31834.5	31834.5
4	12293	12293
5	8458.19	10289.3
6	6319.24	6319.24
7	5775.18	5608.38
8	3452.63	5318.96
9	2918.74	3642.32
10	4340.43	3260.17
11	3374.08	2495.98
12	1968.08	1968.08
13	2329.73	1549.78
14	2219.35	1501.35
15	1688.49	1318.63
16	1823.67	1289.75
17	1598.6	930.133
18	1576.21	920.467
Total	**132273.3**	**130843.2**

The results showed that the new algorithm minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances with the increasing value of k. This shows that the new algorithm is more efficient than the original k-means algorithm which chooses the initial centroids randomly. The chart foe the results obtain is given below:

Dataset on which the algorithms are applied consist of two rows is given below:

Row 1 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 4, 60, 58, 57, 59, 62, 35, 48.

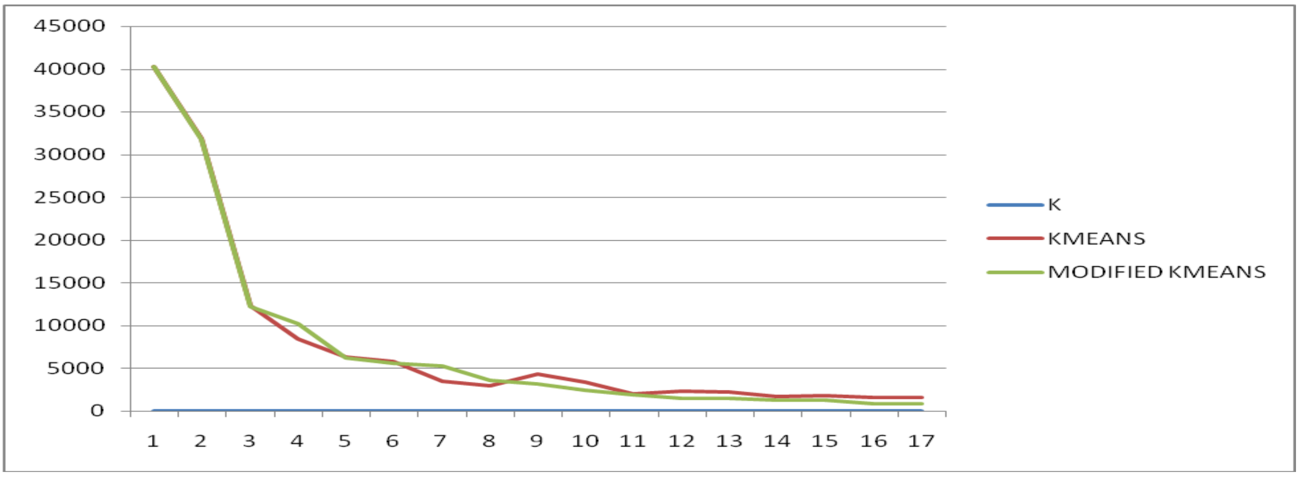
Row 2 = 9, 1, 2, 5, 78, 69, 48, 56, 47, 35, 59, 78, 33, 35, 66, 55, 44, 211, 4, 59, 26, 85, 47, 15, 32, 65, 89, 4, 12, 45, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 1, 9, 78, 6, 5, 4, 3, 2, 1.

5. COMPARISON AMONG DIFFERENT METHODS

Various algorithms have been devised for improving the efficiency of k-means algorithm. These algorithms have already been discussed. It can be seen that each method has its own merits and demerits. Each method tries to increase the efficiency of k-means algorithm to achieve better cluster centroids with intra-cluster similarity and inter-cluster dissimilarity. It is observed that though these methods increase the overall efficiency of k-means and gives good initial clusters but in doing so they increase the overall complexity of the algorithm as many operations have to be applied in calculating the initial clusters. The algorithms which increase computational costs include: Cluster Centre Initialization Method, Hierarchical K-Means and Bradely and Fayyad's Method.

Some of these algorithms like Automatic Initialization of k-means require a second data structure to store the values. Some algorithms like in Simple Cluster Seeking Method and in Automatic Initialization of Means require user involvement. In these the user has to supply a threshold value on which the results depend. So overall these algorithms do sacrifice the simplicity of the k-means algorithm and increase the overall computational cost.

Various methods devised for the calculation of initial clusters for the K Means algorithm have their own merits and demerits. In choosing initial clusters two things needs to be taken care of which are accuracy and efficiency. In other words there is trade off between these two. When we try to maximise efficiency then accuracy suffers and on the other hand if one tries to maximise accuracy then efficiency suffers. Efficiency is related with the running time foe the algorithm. So we need a method that will balance both the accuracy and efficiency of the K-Means algorithm.



A new algorithm is presented in the paper to overcome the shortcomings of the previous methods of calculating initial clusters. It can be seen from the results presented in the table 1 that the algorithm is more efficient than the random initialization method used in simple k-means algorithm. Also the new method involves very few operations and its running time is similar to the simple k-means algorithm which means that it does not increase the complexity of the k-means algorithm. Another advantage of the new method is that it does not require a second data structure. User involvement is also not required in the new method. So it can be seen that the new method is better than the other methods.

6. CONCLUSION

Various researchers have devised new methods for the calculation of initial centroids for K Means algorithm. The main limitation of all the algorithms is that they are very complex and minimize the gain accrued by the simplicity of K Means algorithm. To make the process of calculation of initial centers for the clusters a new algorithm is presented in the paper. The main advantage of the proposed method is its' simplicity as it makes the process of calculation of initial centroids very simple. The results showed that the new algorithm minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances with the increasing value of k. This shows that the new algorithm is more efficient than the original k-means algorithm which chooses the initial centroids randomly.

7. REFERENCES

[1] Jiawei Han, *Data mining: concepts and techniques* (Morgan Kaufman Publishers, 2006).

[2] Margaret H Dunham, *Data mining: introductory and advanced concepts* (Pearson Education, 2006).

[3] 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.

[4] Anderberg, M, Cluster analysis for applications (Academic Press, New York 1973).

[5] Tou, J., Gonzales, Pattern Recognition Principles (Addison-Wesley, Reading, MA, 1974).

[6] Katsavounidis, I., Kuo, C., Zhang, Z., A new initialization technique for generalized lloyd iteration, IEEE Signal Processing Letters 1 (10), 1994, pp. 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] Samarjeet Borah, M.K. Ghose, Performance Analysis of AIM-K-means & K- means in Quality Cluster Generation, Journal of Computing, vol. 1, Issue 1, December 2009.

[10] Ye Yunming, *Advances in knowledge discovery and data mining* (Springer, 2006).

[11] 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.

[12] Madhu Yedla, S.R. Pathakota, T.M. Srinivasa, Enhancing K-means Clustering Algorithm with Improved Initial Centre, International Journal of Computer Science and Information Technologies, 1 (2) , 2010, pp. 121-125.