Survey on Exiting Method for Selecting Initial Centroids in K-means Clustering

[1]Trupti M. Kodinariya, [2]Dr. Prashant R. Makwana
[1] Research Scholar in JJT University, [2]Director – Research center
[1]Department of Computer Engineering,
[1]Jhunjhunu, Rajasthan - India

***Abstract* - Clustering is one of the Data Mining tasks that can be used to cluster or group objects on the basis of their nearness to the central value. K-means clustering algorithm is a one of the major cluster analysis method that is commonly used in practical applications for extracting useful information in terms of grouping data. But the standard K-means algorithm is computationally expensive by getting centroids that provide the quality of the clusters in results. This paper presents the various methods evolved by researchers for finding initial clusters for K Means.**

***Index Terms* - Binary Splitting, Clustering, Cluster Centre Initialization Method, Forgy's Approach, Kaufman Approach, K-means Clustering, Kernel Principle Component Analysis based Method Macqueen Method, Simple Cluster Seeking method**

I. INTRODUCTION

Clustering is one of the most popular fields in machine learning and has various applications. Its aim is to partition a dataset into such subgroups that samples in the same group share more similarities than those from different groups. Although there are various kinds of clustering methods, K-means type clustering is more widely used as it converges fast. Its popularity can be attributed to several reasons. First, it is conceptually simple and easy to implement. Virtually every data mining software includes an implementation of it. Second, it is versatile, i.e., almost every aspect of the algorithm (initialization, distance function, termination criterion, etc.) can be modified. This is evidenced by hundreds of publications over the last fifty years that extend k-means in various ways. Third, it has a time complexity that is linear.

On the other hand, k-means has several significant disadvantages. First, it requires the number of clusters, K, to be specified a priori. Second, it can only detect compact, hyperspherical clusters that are well separated. Third, due its utilization of the squared Euclidean distance, it is sensitive to noise and outlier points since even a few such points can significantly influence the means of their respective clusters. This can address by outlier pruning or using a more robust distance function such as City-block distance. Fourth, it is highly sensitive to the selection of the initial centers. Adverse effects of improper initialization include empty clusters, slower convergence, and a higher chance of getting stuck in bad local minima. Fortunately, all of these drawbacks except the first one can be remedied by using an adaptive initialization method .In this paper, we focus initialization of center problem.

The paper is organized as follows. A generic version of K-Means is described in Section 2. Section 3 contains a review of different initialization methods for selection of initial centroid in the published literature. Section 4 concludes the paper.

II. GENERIC VERSION OF K-MEANS ALGORITHM

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.
The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is needed to re-calculate k new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop it may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.
Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$W(S,C) = \sum_{k=1}^{K} \sum_{i \in S_k} \|y_i - c_k\|^2$$

Where S is a K-cluster partition of the entity set represented by vectors y_i ($i \in I$) in the M-dimensional feature space, consisting of non-empty non-overlapping clusters S_k, each with a centroid c_k ($k=1,2,\ldots K$).
The algorithm is composed of the following steps:

1. Place k points in the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids.

4. Repeat Step 2 and 3 until the centroids no longer move.

III. DIFFERENT APPROACHES TO SELECTION OF INITIAL CENTROIDS IN K-MEANS CLUSTERING

Researchers are always been conducted to improve the accuracy and efficiency of the K-means algorithm. Some of these innovative approaches to K-Means clustering are discussed in this survey. Though the time complexity is not improved, these proposals could fix the initial centroids and the inconsistency in clustering got reduced.

A. Forgy's Approach

The earliest method to initialize K-means was proposed by Forgy in 1965. In Forgy's approach (FA), initial centroids are selected by randomly. 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 [1]. This method has no theoretical basis, as such random clusters have no internal homogeneity.

Strength:

- Simplest method.
- Give quick results.
- User do not have to supply any threshold value

Weakness

- Randomness in choosing initial clusters gives extreme results

B. MacQueen Method

In 1967, McQueen proposed an approach know in literature as McQueen Approach [2]. He proposed a method very similar to the FA method. He suggested that, as with the FA above, K instances are chosen 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 convergence, 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. This is, in essence, an on-line version of the K-means algorithm. 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.

C. Simple Cluster Seeking method

Simple Cluster Seeking method was proposed by Tou and Gonzales in 1974. In this method, the first seed Initialize with the first object in the database; then it calculate the distance between this seed and the next object in the database, if it is greater than some threshold then select it as the second seed, otherwise move to the next object in the database and repeat. 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 chosen, if both these distances are greater than the threshold then select as the third seed. We continue until K seeds are chosen. This process is repeated until K seeds are chosen [3].

The advantage of this method is that it allows the user to control the distance between different cluster centers. But the method also suffers from some limitations which include, the dependency of the method on the order of the points in the database, and, more critically, the user must decide on the threshold value.

D. Binary Splitting Method (BS)

Binary Splitting (BS) method proposed by Linde, Buzo and gray in 1980 [4] which was intended for use in the design of Vector Quantiser codebooks. It calls upon a hierarchical clustering (divisive-top down clustering) to initialize the centroid within k-means clustering. In the method, individual cluster centroids and set of cluster centroids are called as codewords and codebook respectively. The K-means algorithm is first run for K = 1 (like in divisive clustering, initially all objects in single cluster). Then it calculates center c as per k-mean clustering. The cluster centre found, c, is split into two clusters, $c+\in$ and $c-\in$, where $\in$ is some small random vector. The K-means algorithm is run again with these two points as seeds. When convergence is reached, the two cluster centres are again split to make four seeds and the K-means algorithm is run again. The cycle of split and converge is repeated until a fixed number of clusters is reached, or until each cluster contains only one point. This method clearly carries increased computational complexity, since after each split the K-means algorithm must be run. In addition, the quality of the clustering after each split depends upon the choice of $\in$, since this encourages the direct in which each cluster will be split.

E. Kaufman Approach

Kaufman and Rousseeuw proposed a method in which the first seed is selected as the most centrally located instance [5]. Next they examine which of the points in the database, which when chosen as the next seed, will produce the greater distance with previously selected centroids. The procedure is as follow:

For every non-selected point wi

For every non-selected point wj, calculate Cji=max(Dj-dji, 0) where dji is the distance between wi and wj, Dj is the distance between wj and its nearest centroid.

Calculate $\sum j$ Cji

Select next seed wi which maximize $\sum j$ Cji

Once the second seed is chosen then choose the third seed in the same fashion and continue until K seeds are chosen. Again the first obvious drawback of this algorithm is the considerable amount of computation involved in choosing each seed. If this algorithm is to be considered useful for large databases a subsample of the instances must be used instead when finding the seeds.

F. KKZ Method

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

G. Bradley and Fayyad Method

Bradley and Fayyad present a technique for initializing the K-means algorithm [7]. They begin by randomly breaking the data into J randomly small sub-subsets. They then perform a K-means clustering on each of the J subsets, all starting at the same set of 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 J runs is JK centre points. These JK points are then themselves input to the K-means algorithm and the algorithm run J times, each of the J runs initialized using the K final centroid locations from one of the J subset runs. The resulting K centre locations from this run are used to initialize the K-means algorithm for the entire dataset. The main advantage of the method is that it increases the efficiency of the result by the obvious fact that initial centroids are obtained by multiple runs of the K-means algorithm. The major drawback of this initialization method is that it requires a lot of computational effort. This makes the use of this method limited to situations where computational time, space and speed does not matter.

H. Likas Dynamically Deterministic Global Search Method

Likas et al. present a global K-means algorithm which is in incremental approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N (N being size of data set) executions of the k-means algorithm from its initial positions [8]. This method is required larger computation cost. In addition to these, they employ a variation of the kd-tree to initialize each of run of this comparison method. They simply use the kd-tree to create K buckets and use the centroids of each bucket as seeds.

I. Cluster Centre Initialization Method

Khan and Ahmad proposed a method for finding initial cluster centroids in K-means algorithm and named it Cluster Centre Initialization Method (CCIA) [9]. CCIA method is based on the use of Density-based Multi Scale Data Condensation (DBMSDC). DBMSDC method is used for estimating the density of the data at a point, based on their density it then sort the points. A point is chosen from the top of the sorted list and all points within a radius inversely proportional to the density of that point are pruned. It then moves on to the next point in the list which has not been pruned and repeat. This process is repeated until a desired number of points remain. This method chooses its seeds by examining each of the m locations. Then the DBMSDC algorithm is invoked and points which are close together are merged until there are only K points remaining. The strength of the method is that the initial cluster centers computed by using this are found to be very close to the desired cluster centers with improved and consistent clustering results. The main limitation of the method is that it results in higher computational cost, as it involve density calculations.

J. Variance based method by Deelers and Auwatanamongkol

S. Deelers, and S. Auwatanamongkol proposed algorithm which performs data partitioning along the data axis with the highest variance [10]. Data in a cell is partitioned using a cutting plane that divides cell in two smaller cells. The plane is perpendicular to the data axis with the highest variance and is designed to reduce the sum squared errors of the two cells as much as possible, while at the same time keep the two cells far apart as possible. Cells are partitioned one at a time until the number of cells equals to the predefined number of clusters, K. The centers of the K cells become the initial cluster centers for K-means. It can be seen that proposed algorithm performances are comparable to the CCIA. However, the proposed algorithm is much simpler to implement than CCIA. It is heavily affected with outlier.

K. Kernel Principle Component Analysis based Method

M. Sakthi and Dr. Antony Selvadoss Thanamani proposed technique which performs data clustering with the help of Principal component obtained from KPCA [11]. In the approach, they reduce the D dimension of the N data using Kernal Principal Component Analysis (KPCA) and then they prepare another N data with d dimensions (d<D). The principle components are sorted in ascending order based on variance and then divided into k partitions. For each partition, median is founded and then they use the corresponding data points for each median to initialize the cluster centers. These initial centroids of the clusters are supplied as input to K-Means with original data set with D dimension. The KPCA has to store and diagonalizable the kernel matrix whose size is equal to square of number of data. So for large scale data set, KPCA would consume large storage space and be computationally expensive.

L. Z-score ranking based Method

Kathiresan V and Dr P Sumathi propose an algorithm to compute initial cluster centers for K-means clustering based on Z-Score ranking [12]. This technique is a statistical method of ranking numerical and nominal attributes based on distance measure.

The data are arranged based on the score values in ascending order. After that, it partitioned the ranked data into k subsets. In next step, compute the mean values of each k subsets and select the nearby value of data to the mean as the initial centroid. The time complexity of the method is in polynomial time of degree 2 without degrading the accuracy of clusters. The existing improvements of the k-means algorithm either improve accuracy or efficiency not both.

M. Mid-point based Method by Neha and Kirti Aggarwal

Neha Aggarwal and Kirti Aggarwal represented a modified k-mean clustering algorithm in which auto-generate initial partition rather than randomly selection [13]. If dataset include the negative value attributes, then all the attributes are transformed 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 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.

N. Distance based method by Raed and Wesam

Raed T. Aldahdooh and Wesam Ashour represented a new distance based method of initial centroid selection for k-means clustering algorithm [14]. The algorithm starts by choosing random point as initial centroid and then performs some calculation to verify whether the point is noise or not. To verify noisy point, Compute the distance between selected centroid and each point in data set and then sort the data points based on the resulted distances; then divide data set into k partition with N points and Compute the average distance between each pair of N points. If this distance is higher 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 treats as cluster center.

IV. CONCLUSION

In this paper various methods for choosing initial clusters in K-means algorithm are presented. It can be seen that no single method is able to make the choosing of initial clusters both efficient and accurate. There are some problems which still persist with all above methods such as K (number of cluster should be fixed beforehand); computational complexity is high of modified k-mean algorithms; some of the methods biased on ordering of data. Researches in the field of improving the performance of K - Means could not develop a widely accepted version. So, there is a need to develop a new method which combines the merits of various methods to produce both accurate and efficient results.

REFERENCES

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

[2]. 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.

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

[4]. Linde, Y., Buzo, A., Gray, R. M., 1980. An algorithm for vector quantizer design. IEEE Transactions on Communications 28, 84-95.

[5]. Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, Canada.

[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]. Likas, A., Vlassis, N., Verbeek, J. J., 2003. The global k-means clustering algorithm. In: Pattern Recognition. Vol. 36. pp. 451-461.

[9]. Khan, S. S., Ahmad, A., Cluster center initialization algorithm for k-means clustering, Pattern Recognition Letters 25 (11), 2004, pp. 1293-1302.

[10]. S. Deelers, and S. Auwatanamongkol, Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance, PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 26 DECEMBER 2007, pp. 323-328.

[11]. M.Sakthi and Dr. Antony Selvadoss Thanamani, An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA., International Journal of computer and information technologies, Vol. 2 (3). 2011, pp. 955-959.

[12]. Kathiresan V and Dr P Sumathi (2012), An Efficient Clustering Algorithm based on Z-Score Ranking method, International Conference on Computer Communication and Informatics (ICCCI -2012), 978-1-4577-1583-9/ 12 © IEEE.

[13]. 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.

[14]. Raed 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.