Review of Existing Methods for Finding Initial Clusters 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
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. It has found many applications in the field of business, image processing, medical etc. 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. This paper presents the various methods evolved by researchers for finding initial clusters for K Means.

General Terms
Accuracy, Centroids, Complexity, Dataset, Initial Clusters, K-Means

Keywords
Automatic Initialisation of Means (AIM), Cluster Centre Initialisation (CCIA), Simple Cluster-Seeking (SCS).

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:

3.1 Forgy's Method
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]. In their research paper M.E. Celebi et al. revealed that cluster centroid initialization methods such as Forgy, Macqueen, and max-min often perform poorly and there are other methods with same computational requirements which can give better results [5].

3.2 Simple Cluster Seeking Method
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 [6]. The schematic diagram of SCS method is given below:

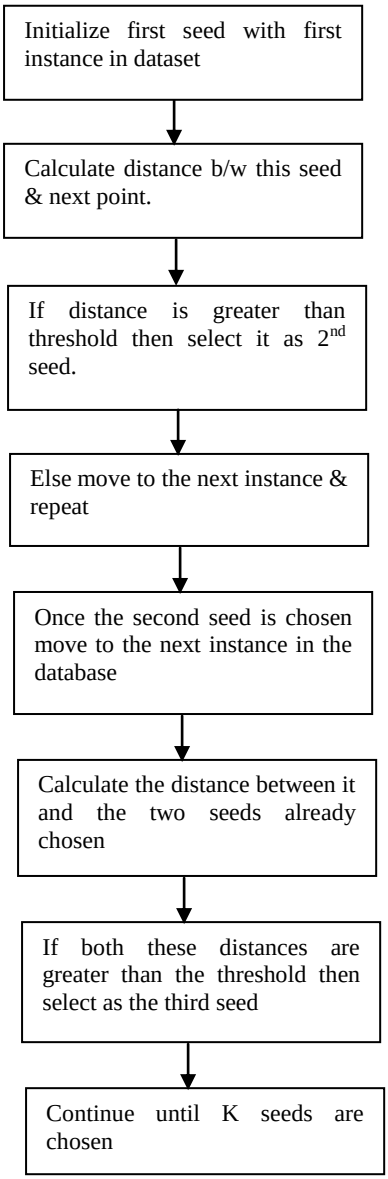

Fig 2: Simple Cluster Seeking Method

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.

3.3 KKZ Method

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 [7]. The schematic diagram of KKZ method is given below:

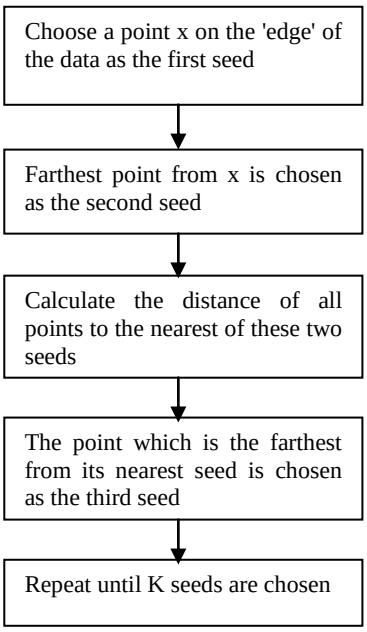

Fig 3: KKZ Method

The KKZ method is attractive in practice because it is simple for decision of unique initial centers. However, the KKZ method sometimes finds bad clusters because unfortunately it depends on outlier data points [8]. 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 [9].

3.4 Bradley and Fayyad's Method

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 [10].

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 [11]. In this method we have to run the K-means algorithm multiple numbers of times which increases the time taken by the method to produce the desired result. Also, this method requires more memory to store intermediate results of multiple runs of K-means. This makes the use of this method limited to situations where computational time, space and speed does not matter.

3.5 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). 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 choose its seeds by examining each of the m attributes individually to extract a list of K' > K possible seed locations. Then the DBMSDC algorithm is invoked and points which are close together are merged until there are only K points remaining [12].

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 [13]. The experimental results of CCIA show the effectiveness and robustness of the method in solving several clustering problems [14]. The main limitation of the method is that it results in higher computational cost, as it involve density calculations.

3.6 Hierarchical K-means Algorithm

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 [14]. S.A. Majeed et al. used the Hierarchical k-means algorithm as a clustering technique for training and testing the isolated Malay digits feature vectors. Both the speed of k-means algorithm and the precision of hierarchical algorithm are given preference in Hierarchical K-means. The system showed promising results in recognition accuracy [15].

3.7 Automatic Initialization of Means

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 [16].

The strength of the Automatic Initialization of Means for finding of initial clusters lies in its automation of the description of value of K and detection of initial clusters. But the method is also not free from discrepancies. Firstly, it requires the calculation of distance threshold and various other calculations which makes it computationally expensive. Second limitation relates with the need of storing the data in two data structure. Storing same data at two places makes the computations slow and also requires double the space as compared to other methods. This increases the space and time complexity of method.

3.8 K. A. Abdul Nazeer's Method

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 [17].

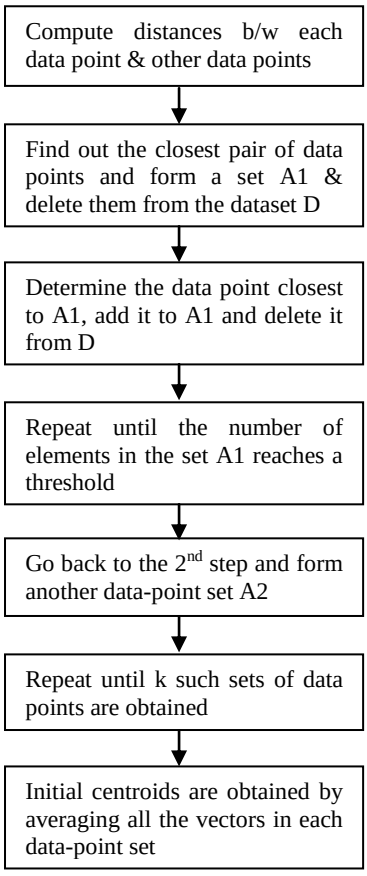

Fig 4: Enhanced algorithm for finding initial clusters

This algorithm uses two methods; one method is used for finding the better initial centroids and another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time [18]. Though the algorithm produces good initial centroids but it also suffers from the same computational limitation, that is, it is computationally very expensive.

3.9 Mid-point based K-means Clustering

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. The transformation is done because in the proposed algorithm we calculate the distance from origin to each data point in the data set. So, for data points, which have same values but differ only in sign we will get the same Euclidean distance from the origin. This will result in incorrect selection of the initial centroids. To overcome this problem all the data points are transformed to the positive space. If data set contains the all positive value attributes then the transformation is not required. 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 [18]. The Schematic diagram of the method is given below:

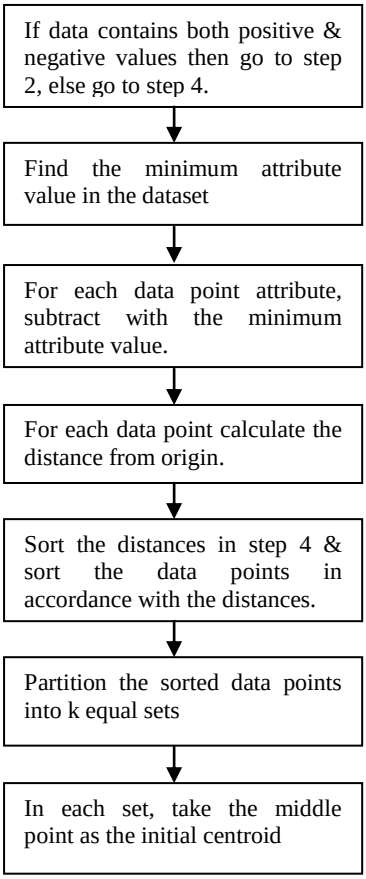

Fig 5: Mid-point based k-means clustering

4. COMPARISON AMONG EXISTING METHODS

The comparison of various methods used for calculating initial clusters in K-means algorithm is given below:

Various methods for choosing initial clusters in K-means algorithm are presented above along with their merits and demerits. Tabular comparison of the methods is given below:

Table 1: Comparison among existing methods

Method Name	Strengths	Weaknesses
Forgy's Method	• Simplest method. • Give quick results. • User do not have to supply any threshold value	• Randomness in choosing initial clusters gives extreme results
Simple Cluster Seeking Method	• Allow the user to control the distance b/w cluster centers.	• Dependency on the order of data points in the database. • The user has to supply a threshold value.
KKZ Method	• Simple method for choosing unique initial clusters. • Does not depend on any threshold value.	• Outliers pose a challenge to this method.
Bradley & Fayyad's Method	• Increases the efficacy of the result by running the K-means algorithm many times.	• It requires a lot of computational effort.
Cluster Centre Initialization Method	• Improved & consistent clustering results.	• It requires higher computational cost, as it involve density calculations.
Hierarchical K-means	• It combines speed of K-means algorithm with precision of Hierarchical clustering.	• Computational cost is more as it combines & run two algorithms.
Automatic Initialization of Means	• Its strength lies in automation of the description of value of K and detection of initial clusters.	• It needs two data structures for storing the data.
K.A. Abdul Nazeer's Method	• This algorithm produces good clusters in comparatively less amount of computational time.	• Still computational time is high in this method.
Mid-point based K-means Clustering	• This method is simplest of all the methods which do not find initial clusters randomly.	• Comparatively its accuracy is less than other methods.

The Forgy's method is simplest of all the methods and can be used if there is limitation of computational resources and quick results are required. Simple Cluster Seeking method increases the accuracy of the K-means algorithm but in this method the user needs to specify the threshold value. The

KKZ method is useful in practice because it is simple for decision of unique initial centres. But it produces bad initial clusters if outliers are present in the data. The method proposed by Bradley and Fayyad is computationally very expensive as it requires multiple runs of the K-means algorithm. Cluster Centre Initialisation Method also suffers from high computational costs. Similarly other methods proposed by Koheri Arai et al., K. A. Abdul Nazeer et al. and Samarjeet Borah et al. are computationally very expensive. The method proposed by Madhu Yedla et al. is simple and require less computational time and space as compared to all other methods except Forgy's method.

It can be seen that no single method is able to make the choosing of initial clusters both efficient and accurate. Nearly all of the methods presented in the paper have high computational costs. Moreover the use of these methods depends upon the priority of the user i.e. whether he wants highly accurate results irrespective of the high computational cost involved or he can compromise the accuracy of the results and requires methods with low computational costs. If the user wants to minimize interaction with the system then the Automatic Initialisation of Means method is a good choice.

5. CONCLUSION

In this paper various methods for choosing initial clusters in K-means algorithm are presented along with their merits and demerits. It can be seen that no single method is able to make the choosing of initial clusters both efficient and accurate. Nearly all of the methods presented in the paper have high computational costs. Moreover the use of these methods depends upon the priority of the user. So clearly there is a need to develop a new method which combines the merits of various methods to produce both accurate and efficient results.

6. 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] M. E. Celebi, H. Kingravi, P. A. Vela, A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm, Expert Systems with Applications, 40(1), 2013, pp. 200-210.

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

[7] Katsavounidis, I., Kuo, C., Zhang, Z., A new initialization technique for generalized lloyd iteration, IEEE Signal Processing Letters 1 (10), 1994, pp. 144-146.

[8] Takashi Onoda, Miho Sakai, Seiji Yamada, Careful Seeding Method based on Independent Components Analysis for k-means Clustering, Journal Of Emerging Technologies In Web Intelligence, vol. 4, No. 1, February 2012.

[9] Stephen J. Redmond, Conor Heneghan, A method for initialising the K-means clustering algorithm using kd-trees, Pattern Recognition Letters 28(8), 2007, pp. 965-973.

[10] 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.

[11] Fernando Bacao, Victor Lobo, Marco Painho, Self-organizing maps as substitutes for K-means clustering, Computers and Geosciences, vol. 31, Elsevier, 2005, pp. 155-163.

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

[13] Shehroz S. Khan, Shri Kant, Computation of initial modes for k-modes clustering algorithm using evidence accumulation, 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 2007, pp. 2784-2789.

[14] 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.

[15] S. A. Majeed, H. Husain, S. A. Samad, A. Hussain, Hierarchical k-means algorithm applied on isolated Malay digit speech recognition, International Conference on System Engineering and Modeling, vol. 34, Singapore, 2012.

[16] 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.

[17] 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.

[18] 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.