Analysis of Initial Centers for k-Means Clustering Algorithm

M.P.S Bhatia,Ph.D
Department of Computer Engineering,
Netaji Subhash Institute of Technology,
University of Delhi, New Delhi, India

Deepika Khurana
Department of Computer Engineering,
Netaji Subhash Institute of Technology,
University of Delhi, New Delhi, India

ABSTRACT
Data Analysis plays an important role for understanding different events. Cluster Analysis is widely used data mining technique for knowledge discovery. Clustering has wide applications in the field of Artificial Intelligence, Pattern Matching, Image Segmentation, Compression, etc. Clustering is the process of finding the group of objects such that objects in one group will be similar to one another and different from the objects in the other group. k-Means clustering algorithm is one of the popular algorithm which has gained a lot of attraction because of its simplicity and ease of implementation. k-Means algorithm's efficiency is limited because of random selection of k initial centers. Therefore, we have surveyed different approaches for initial centers selection for k-Means algorithm. We have also shown comparative analysis of Original K-Means and Data Clustering with Modified k-Means Algorithm using MATLAB R2009b. We chose Euclidean distance as the similarity measure for our implementation and results are evaluated.

General Terms
Data Mining, Clustering Algorithm, Objects.

Keywords
k-Means, Clustering, Initial Centers, Similarity measures.

1. INTRODUCTION
Data Mining is defined as mining of knowledge from huge amount of data. Using Data mining we can predict the nature and behavior of any kind of data. The past two decades has seen a dramatic increase in the amount of information being stored in the electronic format. This accumulation of data has taken place at an explosive rate. It was recognized that information is at the heart of the business operations and that decision makers could make the use of data stored to gain the valuable insight into the business. DBMS gave access to the data stored but this was only small part of what could be gained from the data. Analyzing data can further provide the knowledge about the business by going beyond the data explicitly stored to derive knowledge about the business.

Learning valuable information from the data made clustering techniques widely applied to the areas of artificial intelligence, customer – relationship management, data compression, data mining, image processing, machine learning, pattern recognition, market analysis, and fraud – detection and so on. Cluster Analysis of a data is an important task in Knowledge Discovery and Data Mining. Clustering groups the data on the basis of similarities or dissimilarities among the data elements. Clustering is the process of finding the group of objects such that object in one group will be similar to one another and different from the objects in the other group. A good clustering method will produce high quality clusters with high intra cluster distance similarity and low inter cluster distance similarity. Similarity measure used is standard Euclidean distance but there can also be other distance measures such as Manhattan distance, Minkowski distance and many others. The quality of clustering depends on both the similarity measure used by the method and also by its ability to discover some or all of the hidden patterns. The popular clustering approach can be partition based or hierarchy based, but both approaches have their own merits and demerits in terms of number of clusters, cluster size, separation between clusters, shape of clusters, etc… Some other approaches are also based on hybridization of different clustering techniques. Many Clustering algorithms use the center based cluster criterion. The center of a cluster is often a centroid, the average of all the points in a cluster. In sections 2- 4 we present various factors that influence efficiency of k-Means clustering algorithms. Section 2 explains various similarity measures. 3 section explains about how to select value of k. Section 4 explains about how to select initial centroids. Section 5 shows experiment analysis of original k means and modified k means in [1] on the basis of its implementation in MATLAB. The paper is concluded in last section.

2. SELECTION OF SIMILARITY MEASURES
The Similarity between objects is computed based on the distance between each pair of objects. The distance measures are as follows:

1. The Minkowski distance is a generalized metric that includes others as special cases of the generalized form. It is given by:

$$d(x_i, y_i) = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p} \quad \text{Eq. 1}$$

In the equation 1 $d(x_i, y_i)$ is the Minkowski distance between the data vector x =(x_1, x_2, x_3-------x_n) *and y* = (y_1, y_2, y_3-------y_n), *n* the total number of data points and p is the order of the Minkowski metric. Different names for the Minkowski distance arises from the order p:

- p = 1 is the Manhattan distance. It is known as L_1-norm or City-block distance. For two vectors of

ranked ordinal variables it is also called as called Foot-ruler distance.

- p= 2 is the Euclidean distance. It is also known as L_2-Norm or Ruler distance. For two vectors of ranked ordinal variables the Euclidean distance is sometimes called Spearman distance.
- p = ∞ is the Chebyshev distance. Also known as L_{max}-Norm or Chessboard distance.

2. Mahalanobis distance is based on correlations between variables by which different patterns can be identified and analyzed. It gauges similarity of an unknown sample set to a known one. It is better adapted than usual Euclidean distance for non spherical symmetric distribution. It is used to measure a distance of a point to the center of a distribution. It is given by:

$$D^2 = (x - \mu)' S^{-1} (x - \mu) \qquad \text{Eq. 2}$$

Where S is the covariance matrix, D is called the Mahalanobis distance of point x from the mean of a distribution.

3. Pearson's correlation is a measure of the correlation (linear dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables. It is commonly used for analyzing gene expression data. It is given by:

$$D_{ij} = (1 - r_{ij})/2 \qquad \text{Eq. 3}$$

Where r_{ij} is the pearson's coefficient given by:

$$r_{ij} = \frac{\sum_{l=1}^{d} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\left[\sqrt{\sum_{l=1}^{d} (x_{il} - \overline{x_i} \sum_{l=1}^{d} (x_{jl} - \overline{x_j})}\right]} \qquad \text{Eq. 4}$$

4. Cosine similarity, the measure commonly used for document clustering, is given by:

$$S_{ij} = Cos \propto = \frac{x_i^T x_j}{\|x_i\| \|x_j\|} \qquad \text{Eq. 5}$$

3. SELECTION OF k

1. Value of k equals to number of classes in a dataset

We have run original k –Means algorithm and other two algorithms of [1] and [2] on Wine and Fisher-iris dataset. It is found that dataset has three defined classes and on changing the number of clusters from k=3 to k=4,5and 6 the performance degrades as many points get wrongly clustered.

2. Value of k to be determined by the algorithm

For this user needs to input initial value of k and algorithm will ran for that number of k until a threshold specified by the cluster validity is achieved. This threshold for validity can be set up by using different cluster validity index like Silhouette validity index, Davies- Bouldin, Dunn's Index etc.

3. Value of determined in a preprocessing step

By applying preprocessors, such as hierarchical clustering as preprocessor to k-Means clustering that will determine the right number of clusters which then will be used as input k to k-Means clustering algorithm. There could also be other methods based on some hypothesis, taking input from user, some statistical measures to determine initial value of k.

The selection of appropriate value of k is of great importance. If the fixed number of cluster is very small then there is a chance of putting dissimilar objects into same group and suppose the number of fixed cluster is large then the more similar objects will be put into different groups.

4. SELECTION OF INITIAL CLUSTER CENTERS

This Section reviews existing method for selection of initial clusters. As in the original k-Means, random initial cluster centers are chosen that leads to converge to local minima each time algorithm is run. The research shows that there could be better method to choose initial cluster center so that results do not vary on different runs of the algorithm on the same dataset.

The first approach discussed in [1] optimizes the original k – Means algorithm by proposing a method on how to choose initial clusters. The author proposed a method that partitions the given input data space into k * k segments, where k is desired number of clusters. After portioning the data space, frequency of each segment is calculated and highest k frequency segments are chosen to represent initial clusters. If some parts are having same frequency, the adjacent segments with the same least frequency are merged until we get the k number of segments. Then initial centers are calculated by taking the mean of the data points in respective segments to get the initial k centers. By this process we will get the initial which are always same as compared to the original k – means algorithm which always select initial random centers for a given dataset.

This paper [2] proposes a systematic approach to determine the initial centroids so as to produce clusters with better accuracy. To determine initial centroids, for this compute the distance between each data point and all other data points in the set D. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. Then again form another data-point set A2. Repeat this 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 points to the cluster centroids.

The author in literature [3] uses Principal Component Analysis (PCA) for dimension reduction and to find initial cluster centers. The variable with the highest Eigen value calculated using PCA is taken as first principal component along which partitioning is done, on the basis of which k subsets are formed and k median values are taken as initial k centers.

In [4] first data set is pre-processed by transforming all data values to positive space. The distance of each point is calculated from origin and then data is sorted and divided into k equal sets and then middle value of each set is taken as initial center point.

In literature [5] a dynamic solution to k –Means is proposed that algorithm is designed with preprocessor using silhouette validity index that automatically determines the appropriate number of clusters, that increase the efficiency for clustering to a great extent.

In [6] a method is proposed to make algorithm independent of number of iterations that avoids computing distance of each data point to cluster centers repeatedly, saving running time and reducing computational complexity.

In the literature [7] dynamic means algorithm is proposed to improve the cluster quality and optimizing the number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. In the former case it works same as K-Means algorithm. In the latter case the algorithm computes the new cluster centers by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality.

5. RESULTS

We have implemented Original k -Means and algorithm in [1] using MATLAB R2009b. Using Original k- Means results of different values of k is shown for Fisher iris dataset in table 1, that shows that using value of k equals to specified number of classes in a dataset results in better clustering. The dataset contains 150 instances and 4 attributes.

TABLE I – Results of Fisher-iris dataset using Original k-Means

No. of clusters	No. of iterations	No. of points misclassified	Execution time
k=3	8	28	3.118
k=4	8	111	5.72
k=5	5	119	2.2

Comparison between two approaches analysing merits and demerits of both approaches is shown in Table 2.

TABLE II – Analyzing merits and demerits of two approaches

Basis	Original k-Means	Modified k-Means
1. Convergence to Local optima	Yes, because of random selection of centers.	Achieve global optima because selection of centers is consistent with data distribution.
2. Dead unit problem	May be there , because when centers are randomly chosen it is possible to select a point which is far away from other points, so it will never be updated	Never, because center is selected from that segment that has highest number of points.
3. Execution time	More	Less
4. Space complexity	Less	More because , phase 1 to calculate initial centroids requires additional space to store the frequency of data points in each segment.

Figure 1 to 3 compares the result of Original k means algorithm and algorithm in [1] on random dataset of 178 two-dimensional points. Results shown using column chart are for algorithm 1 Original k means and 2 for modified approach discussed in [1].Experiments are done on different values of k and performance is evaluated on the basis of no. of iterations, no of points misclassified and execution time for both algorithms.

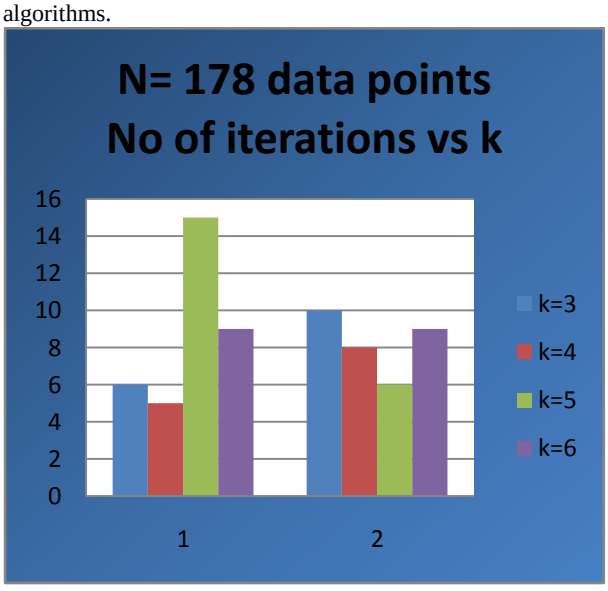

Fig 1: Comparison of no. of iterations vs no. of clusters

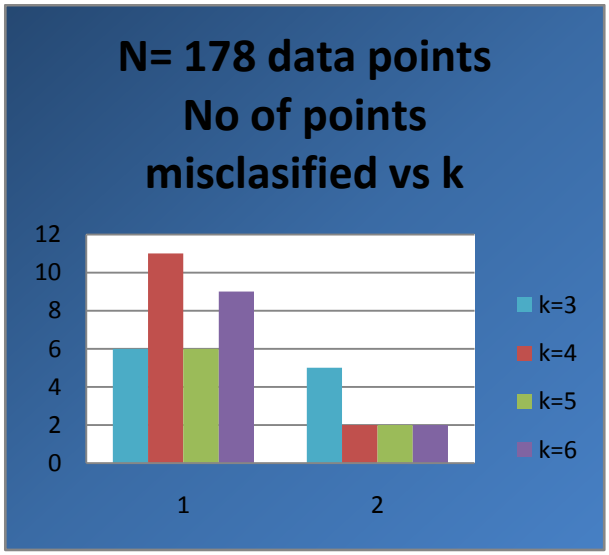

Fig 2: Comparison of no. of points misclassified vs no of clusters.

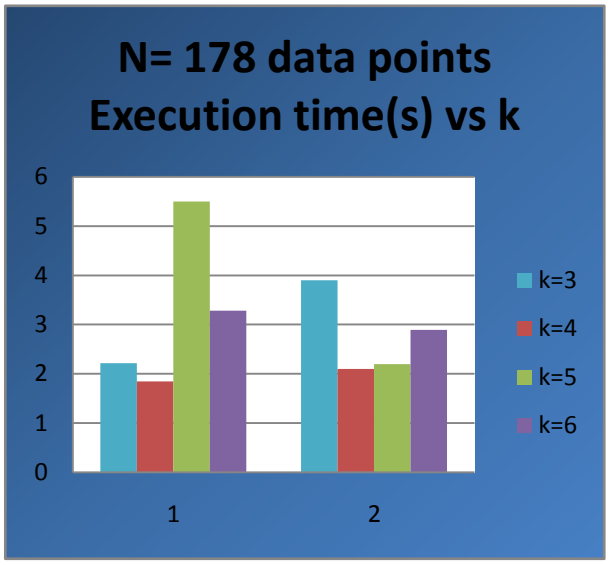

Fig 3:Comparison of execution time vs no. of clusters.

6. CONCLUSION

Existing methods for selecting the number of clusters for k-Means clustering has number of drawbacks because the initial value of k in all algorithms must be known in advance. Original k- Means also suffers from local optima because for every run it results in different centers for given value of k and input dataset. It also suffers from dead unit problem i.e. sometimes initial center that is chosen is far from rest of points in dataset , which then becomes dead as no point will be assigned to that cluster and center will never be updated. These problems are solved using enhanced approaches. Just by including one way to fix the selection of initial centers by locating the centers in segments of high frequency of data points, the modified approach solves both the problems of Original k-means. Finally the paper is concluded with comparison between two approaches. From results, we also conclude that modified k-Means takes less time to execute and also converges in less no. of iterations. Further research is going on how algorithm for clustering itself determines appropriate number of clusters.

7. REFERENCES

[1] Ran Vijay Singh and M.P.S Bhatia, "Data Clustering with Modified K-Means Algorithm", IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp 717-721.

[2] D. Napoleon and P. Ganga Lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE 2010.

[3] Tajunisha and Saravanan, "Performance Analysis of k-Means with different initialization methods for high dimensional data" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010

[4] Neha Aggarwal and Kriti Aggarwal, "A Mid- point based k –mean Clustering Algorithm for Data Mining". International Journal on Computer Science and Engineering (IJCSE) 2012.

[5] Barileé Barisi Baridam, "More work on k-Means Clustering algortithm: The Dimensionality Problem ". International Journal of Computer Applications (0975 – 8887)Volume 44– No.2, April 2012.

[6] Shi Na, Li Xumin, Guan Yong "Research on K-Means clustering algorithm". Proc of Third International symposium on Intelligent Information Technology and Security Informatics, IEEE 2010.

[7] Ahamad Shafeeq and Hareesha "Dynamic clustering of data with modified K-mean algorithm", Proc. International Conference on Information and Computer Networks (ICICN 2012) IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore 2012.

[8] Kohei Arai,Ali Ridho Barakbah, "Hierarchical K-Means: an algorithm for centroids initialization for K-Means.

[9] Data Mining Concepts and Techniques,Second edition Jiawei Han and Micheline Kamber.

[10] D.T Pham, S.S Dimov, C.D Nguyen, "Selection of k in k means clustering".

[11] Paul S. Bradley, Usama M. Fayyad, "Refining Initial Points for K-MeansClustering", 15th International Conference on Machine Learning (ICML98).

[12] K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-Means Clustering Algorithm", Proceeding of the World Congress on Engineering,vol 1,london, July 2009.

[13] T Velmurugan and T Santhanam "A survey of partition based clustering algorithms in data mining: An experimental approach". Proc. Information Technology Journal 2011.