A new k-Means Clustering Technique with Modified Initial Cluster Centers and Centroids

E. U. Oti[1]; M. O. Olusola[2]; W. K. Alvan[3]; G. Oberhiri-Orumah[4]

[1]Department of Statistics,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria.

[2]Department of Statistics,
Nnamdi Azikiwe University, Awka, Nigeria.

[3]Department of Physics with Electronics,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria.

[4]Library Department,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria.
E-mail: eluchcollections@gmail.com[1]

***Abstract* — *In this paper, a new k-means clustering method is proposed which addresses the initial cluster center problem in the k-means algorithm based on binary search techniques, and it also updates cluster centers (centroid). In the initialization phase, the initial cluster centers are generated using the modified binary search property approach, while in the updating phase, cluster centroids are updated using an algorithm depending on if a point is added to a cluster or a point is removed from a cluster. Various results showed that the proposed method performed favorably with four of the existing methods: Lloyd, MacQueen, Faber, and Astrahan methods considered in terms of accuracy, efficiency, and minimization of the within-cluster sum of squares for k clusters both in the simulation and in the real-life data situations*.**

Keywords: *K-Means Clustering, Binary Search, Centroids, Euclidean Distance, Data-points.*

I. Introduction

Clustering is a branch of statistical multivariate analysis and an unsupervised classification mechanism in pattern recognition (Kaufman and Rousseeuw, 1990; Jain et al., 2000). It is a method for classifying like groups of a data set into the same cluster and unlike groups into different clusters (MacQueen, 1967; Anderberg, 1973; Hartigan, 1975; Everitt et al., 2011; Mirkin, 2013; Yuan and Yang, 2019). Cluster analysis is a powerful data exploratory and descriptive approach to forming data groups and it reveals the features and structure information of a given data set and is carried out purely based on similarities or dissimilarities (Johnson and Wichern, 2002). Cluster analysis could be divided into the following categories: hierarchical clustering (Hartigan, 1975; Kaufman and Rousseeuw, 1990), mixture-model clustering (McLachlan and Basford, 1988; McLachlan and Krishnan, 1997), objective-function-based clustering, and partition clustering (Bezdek, 1981; Yang, 1993).

The aim of classification is the difficult task of accurately grouping n points (objects) into K homogeneous clusters. There are many algorithms developed for data clustering; of these, perhaps the best-known of it is the k-Means clustering (Ball and Hall, 1967; Chan et al., 2004). The K-Means partitions a data set into k non-overlapping clusters so that a point $x_i \in X = \{x_1, x_2, \dots, x_i\}$ is assigned to a single cluster $s_k \in S = \{s_1, s_2, \dots, s_K\}$ through the iterative minimization of the criterion in Equation (1).

$$W(S, C) = \sum_{k=1}^{K} \sum_{i \in s_k} d(x_i, c_k) \qquad (1)$$

where $d(x_i, c_k)$ is the dissimilarity between x_i and its respective centroid $c_k \in C = \{c_1, c_2, \dots, c_K\}$, the center of gravity of the cluster s_k. The k-Means criterion allows the use of any distance function.

The purpose of this paper is to propose a k-means clustering method that generates initial cluster centers using a binary search approach and also updates cluster centroids depending on if a point is added to a cluster or a point is removed from a cluster and also compares it with existing ones like Lloyd's and MacQueen's methods.

The rest of this paper is organized as follows: section 2 discusses the materials and methods which consist of methods like Lloyd, MacQueen, Faber, Astrahan, and the proposed k-Means clustering method. Furthermore, section 3 is centered on experimental results and discussion, while section 4 is the conclusion of the paper.

II. Materials and Methods

Several k-means clustering methods were aimed to classify points or objects to be analyzed into well-separated groups (clusters). Four k-means clustering methods and the proposed method will be discussed in this paper. The rationale behind this developed method is based on the assumption that an optimal clustering solution with k clusters can be obtained through local search.

To be able to use any of the methods, the number of clusters present in the data need to be known; multiple runs or trials will be necessary to find the best number of clusters. There is no best method, as the tendency of generating global optimum depends on the characteristics of the data set, size, and the number of variables in the cases. The k-means clustering methods have two phases of iteration namely: the assignment or initialization phase which involves an iterative process where each data point is assigned to its nearest centroid using any metric of choice; the next is the centroid update phase, where clusters centroids are updated given the partition obtained by the previous phase. The iterative process stops when no data point change clusters or some maximum number of iterations is reached.

2.1 Lloyd's Method

Lloyd (1982) proposed a method that is widely known as the standard k-means algorithm; it is a batch algorithm that is based on the minimization of the average squared Euclidean distance between the data items and the cluster centers and it treats the data set as a discrete distribution. The error function for a discrete distribution is defined as

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} f(x) d(c_i, x_i) \qquad (2)$$

In Equation (2) above, $d(c_i, x_i)$ is the distance function of the data point x_i and cluster center c_i. The first step of the algorithm begins with choosing the number of clusters k and its initial centroids or cluster centers. It could be done by either using k random observations or from the k observations that are the farthest from one another in the data space. Initialization of the centroids occurs only once, and once the initial centroids have been chosen, iterations are done in the following two steps. First, the data set is assigned to cluster centroids (centers), using any of the distance metrics. All cases assigned to a centroid are said to be part of the centroid subspace c (R^d) (Morissette and Chartier, 2013). Second, update the value of the centroid by using the mean of the data points (cases) assigned to the centroid.

Algorithm 1: *The Lloyd's (Standard) Algorithm.*

1. Choose k data objects representing the cluster centroids.
2. Assign each data object of the entire data set to the cluster having the closest centroid.
3. Compute a new centroid for each cluster by averaging the data observations belonging to the cluster.
4. If at least one of the centroids has changed, go to step 2, otherwise go to step 5
5. Output the clusters.

2.2 MacQueen's Method

MacQueen (1967) proposed MacQueen's algorithm, which is often referred as the basic k-means algorithm which is an online or incremental algorithm. MacQueen's method is similar to the Lloyd's Method, but the main difference is that the centroids are updated by re-calculating the points (cases) any time it is moved. Once the initial centroids have been chosen in the same way as Lloyd's algorithm, the iterations follow:

For each case (x_i) in turn, after arbitrarily partitioning points into clusters, we compute the coordinates ($\bar{x}_i^{\prime s}$) of the cluster centroid (mean), likewise, the Euclidean distance is computed for each point from the group centroids and reassigned each point to the nearest group. If a point is moved from its initial position, the cluster centroid must be recalculated or updated before computing the squared Euclidean distance.

If the centroid of a case belongs to the nearest subspace, no change is made. If another centroid is closest to the subspace, the case is re-assigned to the other centroid and the centroids for both the old and new subspaces (centers) are recalculated as the mean of the cases. When we see that each point is currently assigned to the clusters with the nearest centroid, the process stops.

Algorithm 2: *The MacQueen's (Basic) Algorithm.*

1. Choose k points as initial cluster centroids.
2. Assign each object to the cluster that has the closest centroid.

3. When all objects have been assigned, re-compile the positions of the k centroids.

4. If at least there is a change in one of the centroids, repeat step 2 and 3, otherwise go to step 5.

5. Output the clusters.

2.3 Faber's Method

Faber (1994) proposed the Faber's method which is popularly known as the continuous k-means algorithm. The continuous k-means algorithm is faster than the standard k-means algorithm and it is also different from the standard k-means algorithm in two ways.

First, the reference points in the continuous k-means algorithm are chosen as a random sample from the whole population of data points, while in the standard k-means algorithm the initial reference points are chosen more or less arbitrarily.

Secondly, the way the data points are treated during the update process; that is during the iteration, the standard k-means algorithm examines all of the data points in sequence while the continuous k-means algorithm examines only a random sample of data points. If the data set is very large and the sample is a representative of the data set, the continuous k-means algorithm should converge much faster than the algorithm that examines every point in the sequence. To be precise, the continuous k-means algorithm adopts MacQueen's method of updating the centroids during the initial partitioning, when the data points are first assigned to clusters (Faber, 1994).

Theoretically, random sampling represents a return to Macqueen's original concept of the algorithm as a method of clustering data over a continuous space. In Macqueen's formulation, the error measure E_i for each region R_i is given by

$$E_i = \int_{x \in R_i} f(x) \parallel x - z_i \parallel^2 dx \qquad (3)$$

where $f(x)$ is the probability distribution function, which is a continuous function defined over the space, x is the data point and z_i is the centroid of the region R_i; while E_i is the total error measure. Hence, a large set of discrete data point can be seen as a large sample as well as a good estimate of the continuous probability density $f(x)$. Then it suffices that a random sample of the data set can also be a good estimate of $f(x)$. Such a sample yields a representative set of cluster centroids and a reasonable estimate of the error measure without using all the points in the original data set.

2.4 Astrahan Method

Astrahan (1970) proposed a method that begins with the selection of a large number of cluster centers scattered throughout the measurement space. These are components of the natural clusters being sought and cluster boundaries are determined by assigning each point to its nearest center. The center of each cluster is then recalculated. This assignment and center recalculation process is repeated until it converges, as indicated by some measure of cluster compactness. The natural clusters are then approached by a process of combining the closest clusters interspersed with reassignment and center recalculation.

***Algorithm 3**: The Astrahan's Algorithm.*

1. Set $d_1 = \frac{1}{n(n-1)} \sum_{1=1}^{n-1} \sum_{j=i+1}^{n} \|y_i - y_j\|$.
2. Pick the $y_i \in Y$ entity with the largest density within the d_1 radius; there should be at least a distance of d_1 between y_i and all the other centroids.
3. If the number of centroids is smaller than k, return to step 2.

2.5 The Proposed Method

For the initialization phase of the proposed method, the initial cluster points are generated by using the unique property of the binary search algorithm to find the value of the middle point given as

$$A[Mid] = \frac{A[beg] + A[end]}{2} \qquad (4)$$

The above property of the binary search algorithm is modified to generate the initial cluster point for k-means where

- A[beg] is replaced by A[max]
- A[end] is replaced by A[min]
- 2 is replaced by k number of clusters
- A[mid] is replaced by any variable such as M
- Plus symbol is replaced by minus symbol

Now, Equation (4) is formulated in another form as

$$M = \frac{A[max] - A[min]}{k} \qquad (5)$$

The generalization of Equation (5) is written as

$$M_i = \frac{max(A_i) - min(A_i)}{k} \qquad (6)$$

Equation (6) is used to calculate the value of the variable M that specifies the range of the initial cluster center (Kumar and Sahoo, 2014). The initial clusters for the proposed method are generated using Equation (7)

$$C_K = \min(A_i) + (k-1)M \qquad (7)$$

For the updating phase of the proposed method, the cluster centroids will be updated or recalculated before computing the squared Euclidean distance. The ith coordinate, where $i = 1,2,\dots,k$ of the centroid is updated using Equations (8) and (9) below:

$$C_i, new = \frac{N_k C_i + C_{ij}}{N_k + 1} \qquad (8)$$

If the jth point is added to the cluster

$$C_i, new = \frac{N_k C_i - C_{ij}}{N_k - 1} \qquad (9)$$

If the jth point is removed from the cluster

Here N_k is said to be the number of points (cases) in the old cluster with centroid $c^T = (C_1, C_2, \dots, C_K)$ or perhaps the cluster size and centroid C_K is a multidimensional vector that minimizes the sum of square's distance to clusters elements. If a point or case is closest to the centroid of a particular subspace where the case is not moved to another cluster implies that the case will not be reassigned but if a case is closest to the centroid of a particular subspace where the case is moved to another cluster implies that the case will be reassigned and updated (Oti, et al. 2019). The stopping rule is to end when there is no further change of cluster membership observed.

The Proposed K-Means Algorithm [Data set (n) and k]

1. Set the number of clusters (k) where $k = 1,2,\dots,p$.
2. Generate the range of the initial cluster centers using Equation (5).
3. Obtain the initial cluster centers using Equation (6).
4. Calculate the squared Euclidean distance $d^2(x,y) = \sum_{i=1}^{d}(x_i, y_i)^2$ and apply minimum distance rule to determine what cluster list a data point, I should be assigned to.
5. Update within cluster centroid, C_k, using Equating (8) or (9) depending on if a point is added to a cluster or a point is removed from a cluster.
6. The stopping rule is to end the iteration when there is no further change of cluster membership observed.
7. Output results.

III. Results and Discussion

This section shows the performance comparison of the modified k-means method and the existing six k-means clustering methods using R statistical software (R version 3.2.2) that supports window 64 bit operating system. We conducted experiments using one simulated data set and two real-life data sets to ensure the efficiency of the proposed k-means method. The number of clusters k used are two and three, since research has proven that the optimal number of clusters k will either be two, three, or four using methods like elbow, the silhouette, and the gap statistic methods (Kaufman and Rousseeuw, 1990).

The performance of the proposed method was evaluated using total intra-cluster variance and accuracy parameters, after which was ranked.

Total intra-cluster variance: The total intra-cluster variance is defined as the sum of the squared distance between points and the corresponding centroid. That is; $W(C_K) = \sum_{x_i \epsilon c_k}(x_i - \mu_k)^2$ where

- x_i is the data point belonging to the cluster c_k.
- μ_k is the mean value of the points assigned to the cluster c_k.

Accuracy: Accuracy is defined as the ratio of the total number of correctly classified instances divided by the total number of correctly plus incorrectly classified instances denoted by Acc. in percentage.

3.1. Simulated Data

The simulated data was generated randomly from a Gaussian (Normal) distribution with a dimension of 300 rows($m = 1,2,\dots,300$), n = 60 and 2 columns (categories or attributes) that are divided into two and three clusters (that is, k = 2,3). The 60 data points are sampled as follows: The first 30 points are sampled uniformly at random from a Gaussian distribution with $\mu = -1$ and $\sigma = 1$, while the remaining 30 points are sampled uniformly at random from a Gaussian distribution with $\mu = 1$ and $\sigma = 1$. For each of the remaining $m - 1$ dimensions, all 60 points are sampled uniformly at random from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. Thus, this obtained a number of well-separated Gaussians with the true centers providing a good approximation to the optimal clustering.

Shown below is the summary table of the results of experiments and data analysis of six existing methods when the number of clusters k is two and three respectively:

Table 1: Summary results of simulated data when the number of clusters k = 2 and 3.

Methods	**When K = 2**				**When K = 3**				**Combined Rank**
	μ	σ	Acc.(%)	Rank	μ	σ	Acc.(%)	Rank	
Lloyd	1.58	0.52	75.4	3	2.25	0.78	71.4	5	8
MacQueen	1.50	0.50	78.1	2	1.92	0.71	82.0	2	4
Faber	1.72	0.55	73.7	4	2.30	0.75	79.7	4	8
Astahan	1.90	0.62	69.5	5	2.14	0.73	81.3	3	8
Proposed Method	1.45	0.42	84.0	1	1.76	0.68	84.6	1	2

From the above results of the simulation generated randomly, when the number of clusters k = 2, the proposed method performed best with a minimum standard deviation of 0.42 and accuracy rate of 84 percent; and when the number of clusters k = 3, the proposed method performed better than the existing method with a minimal standard deviation of 0.68 and accuracy rate of 84.6 percent considering the fact that the variance (the total within-cluster sum of squares) is minimized; it measures the compactness (i.e. goodness) of the clustering which is meant to be as small as possible, also, high accuracy indicates how better the method is.

3.2. Real-Life Data

To understand how efficient these methods are under more practical circumstances, we run a number of experiments on two data sets which consist of the iris data set, and the breast cancer Wisconsin (diagnostic) data set. The two data sets are from UC-Irvine Machine Learning Repository. Each experiment involves solving k-means problem on a set of points in a real dimensional space.

3.2.1. Iris Data Set

The iris flower data set is a multivariate data set with 150 rows (instances) which is divided into 3 instances each, where each class refers to a type of iris plant (iris setosa, iris versicolor, and iris virginica): the number of columns (attributes) is 4 which consist of sepal length, sepal width, petal length and petal width (Fisher, 1936). The summary table of the results of the experiments when the number of clusters k is two and three are shown in Table 2.

Table 2: Summary results of iris data when the number of clusters k = 2 and 3.

Methods	**When K = 2**				**When K = 3**				**Combined Rank**
	μ	σ	Acc.(%)	Rank	μ	σ	Acc.(%)	Rank	
Lloyd	1.35	0.40	81.3	3	1.56	0.79	82.8	2	5
MacQueen	1.30	0.35	83.5	2	1.51	0.74	85.2	1	3
Faber	1.35	0.48	76.5	5	1.53	0.84	81.5	4	9
Astrahan	1.46	0.45	80.2	4	1.87	0.92	77.4	5	9
Proposed Method	1.26	0.32	89.7	1	1.55	0.80	82.4	3	4

From the above experiments and summary table on the iris data set, it is observed that when the number of clusters k = 2, the proposed method performed better than the other existing methods with a standard deviation of 0.32 and 89.7 percent accuracy, and when the number of clusters k = 3, the MacQueen's method performed better with a standard deviation of 0.74 and accuracy of 85.2 percent, while the proposed method performed better than Faber's method and Astrahan's method with a minimal standard deviation of 0.80 and accuracy of 82.4 percent.

3.2.2. Breast Cancer Wisconsin (Diagnostic) Data Set

The breast cancer Wisconsin (diagnostic) data set is a multivariate data from UC-Irvine Machine Learning Repository which consist of 569 number of instance and 36 numbers of attributes. The summary table of the results of the experiments when the number of clusters k is two and three respectively is shown in Table 3.

Table 3: Summary results of breast cancer Wisconsin (diagnostic) data when the number of clusters k = 2 and 3.

Methods	When K = 2				When K = 3				Combined Rank
	μ	σ	Acc.(%)	Rank	μ	σ	Acc.(%)	Rank	
Lloyd	1.08	0.269	91.4	2.5	1.59	0.78	67.8	5	7.5
MacQueen	1.08	0.269	91.4	2.5	1.17	0.60	86.3	1	3.5
Faber	1.28	0.305	90.8	4	1.53	0.73	68.1	4	8
Astrahan	1.36	0.346	87.2	5	1.49	0.66	73.5	3	8
Proposed Method	1.01	0.266	92.1	1	1.44	0.63	76.2	2	3

It was observed that when the number of clusters k = 2, the proposed method performed better than the other methods with a minimal standard deviation of 0.266 and an accuracy of 92.1 percent. When the number of clusters k = 3, the MacQueen's method outperformed every other method with a standard deviation of 0.60 and an accuracy of 86.3 percent. The performance of the proposed method was relatively more efficient than Lloyd's method, Faber's method and Astrahan's method with a standard deviation of 0.63 and an accuracy of 76.2 percent. From the combined ranking of the breast cancer Wisconsin (diagnostic) data set, our proposed method is the best in minimization of the total intra-cluster variance.

IV. Conclusion

In this paper, we have presented a new k-means clustering method that uses the modified property of the binary search algorithm to generate the initial cluster center at its initialization (assignment) phase, while the updating phase of the proposed method updates the within-cluster centroid, C_k, using Equation (8) or (9) depending if a point is added to a cluster or if a point is removed from a cluster. The proposed method performed favorably in comparison with existing methods in terms of minimizing the total intra-cluster variance. From the experimental summary results considering the combined ranks, the new k-means method was more effective than most existing methods both in simulation and in real-life data sets used when the number of clusters k = 2 and 3.

Acknowledgement

The authors thank the reviewers and editors for their worthwhile comments and suggestions.

References

Anderberg, M. R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Astrahan , M. M. (1970). Speech analysis by clustering or hyperphonene method: Issue 124 of Memo (Stanford Artificial Intelligent Project).

Ball, G. H. and Hall, D. J. (1967). A clustering technique for summarizing multivariate data: Bahavioral Science 12(2), pp. 153-155.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press New York.

Chan, E. Y., Ching, W. K., Ng, M. K. and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measure. Pattern recognition; 37(5), pp. 943-952.

Everitt, B., Landau, S., Leese, M., Stajl, D. (2011). Cluster Analysis, 5thedition, John Wiley and Sons.

Faber, V. (1994). Clustering and the continuous k-means algorithm: Los Alamos Science, 22, 138-144.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems, "Annals of Eugenics, 3, 179-188.

Hartigan, J. A. 1975. Clustering Algorithms. New York: John Wiley and Sons.

Jain, A. K., Duin, R. P. W. and Mao, J (2000). Statistical Pattern Recognition: A review. IEEE Transaction on Pattern Analysis and Machine Intelligence 22,4-37.

Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis: 5th Edition, Eaglewood Cliffs, NJ: Prentice-Hall.

Kaufman, L., and Rousseeuw, P. J. (1990). Finding Groups in Data, An Introduction to Cluster Analysis. Wiley Series, New York: John Wiley and Sons.

Kumar, Y and Sahoo, G. (2014). "A New initialization method to originate initial cluster centers for k-means algorithm". In: International Journal of Advanced Science and Technology, Vol.62. pp. 43-54.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transaction on Information Theory, 28 (2), 129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, (1), 281-297. Berkeley, CA: University of California Press.

McLachlan, C. J. and Basford, K. E. (1988). Mixture Models: Inference and Application to Clustering. Marcel Dekker, New York.

McLachlan, C. J. and Krishnan, T. (1997).The EM Algorithm and Extensions. Wiley, New York.

Mirkin, B. (2013). Clustering: A Data Recovery Approach, Second Edition (Chapman and Hall/CRC Computer Science and Data Analysis).

Morissette, L. and Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology, 9 (1), 15-24.

Oti, E. U., Onyeagu, S. I. and Slink, R. A. (2019). A modified k-means clustering method for effective updating of cluster centroid: Journal of Basic Physical Research, 9(2), 123-137.

Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. Multidisciplinary Scientific Journal, 2(2), 226-235.

Yang, M. S. (1993). A Survey of Fuzzy Clustering, Mathematical and Computer Modeling 18, 1-16..