Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

CENTROID INITIALIZATION IN K-MEANS CLUSTERING USING GATCAM

William Rupert Waboke[1], Mustapha Aminu Bagiwa[2] , Ayodele Afoloyan Obiniyi[2], Adekunle Isiaka Obasa[1]

[1]Department of Computer Science, Air Force Institute of Technology, Kaduna
[2]Department of Computer Science, Ahmadu Bello University, Zaria

*Corresponding Author Email Address: rupert.william@afit.edu.ng Phone: +2347065531350

ABSTRACT
Clustering is one of the most widely used machine learning techniques in data processing. Clustering has a wide range of applications, including market research, pattern recognition, data analysis, and image processing, among others. The k-means algorithm is one of the most extensively used clustering algorithms, although it does not guarantee convergence to the global minimum solution because it uses randomization as its initialization of the centroid. Several studies have offered various techniques for solving the problem, including heuristic and meta-heuristic search optimization algorithms. The implementations of k-means continue to rely on random initialization; these solutions have not been successful in addressing the issue of convergence in k-means. This paper proposes GATCAM, an enhanced genetic algorithm with a two-point crossover and adaptive mutation, for centroid initialization in k-means clustering. GATCAM, the proposed approach, improved k-means accuracy by 4.2% for the wine dataset and 1.23% for the Irish dataset while also increasing the likelihood that k-means will converge to the global minimum. According to the experimental results, GATCAM k-means obtain higher accuracy with fewer iterations than standard genetic algorithm-initialized k-means (SGA).

Keywords: Clustering, Machine Learning, Genetic Algorithm, Centroid Initialization

INTRODUCTION

Clustering is one of the most important ML algorithms in data analysis (Jain, 2010; Oyelade *et al.*, 2019). It is used in the pre-processing stage for other data analysis tools. Clustering is the process of partitioning a large dataset into many similar units without having any prior knowledge of the labels. Usually, data points belonging to the same group are as identical as possible, while data points in different groups are as dissimilar as possible. Clustering, according to (Omran et al., 2007; Kaushik, 2016) is broadly classified as hierarchical and partition based as described in figure 1. Hierarchical clustering is achieved either through the use of agglomerative or divisive techniques, while partition-based clustering is achieved through the use of either a hard (K-means) or soft (fuzzy) clustering technique. Soft clustering is also known as fuzzy clustering because data points may belong to more than one cluster with varying degrees of probability. In hard clustering, data points belong to distinct clusters (Bora & Gupta, 2014). Clustering is applicable in various scientific fields, such as biology, genetics, and market segmentation, among many others (Pauletic et al., 2019).

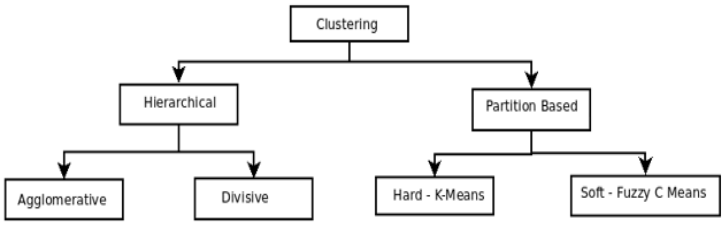

Figure 1: Chart showing different classifications of clustering (Kaushik, 2016)

The partition-based iterative K-means technique partitions a dataset into K separate, non-overlapping subgroups, with each data point belonging to just one cluster (MacQueen, 1967). It keeps intra-cluster data points close and inter-cluster data points far apart (Cao et al., 2009) . According to (Peña et al., 1999; Fränti & Sieranoja, 2019) K-Means is the most used clustering algorithm due to its simplicity and interpretability. However, due to its random selection of the initial centers, k-means does not produce good clusters (Barakbah & Helen, 2005; Chang et al., 2018). Depending on the initial centers, empty clusters can result in poor clustering outcomes (Liu et al., 2018; Muhima et al., 2020) and reduced system accuracy. Figure 2 illustrates the random initial centroid, which luckily gave the accurate clusters in Figure 3. In another instance with the same data points, k-means were unlucky, and random initialization in figure 4 produced the very poor cluster in figure 5. Many studies have been done to improve k-means' accuracy (Ikotun *et al.*, 2023) , but no single method has successfully produced an acceptable initialization of k-means. As such, k-means continued to use random initialization in real-life applications.

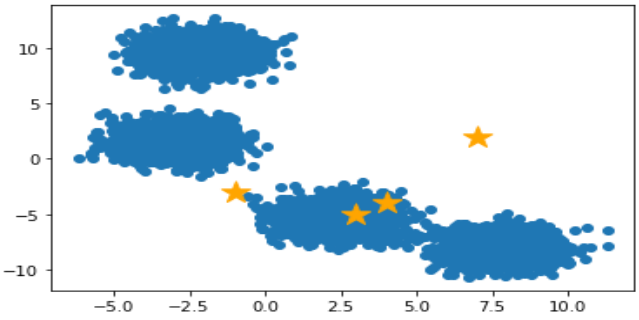

Figure 2: Random Initialization

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

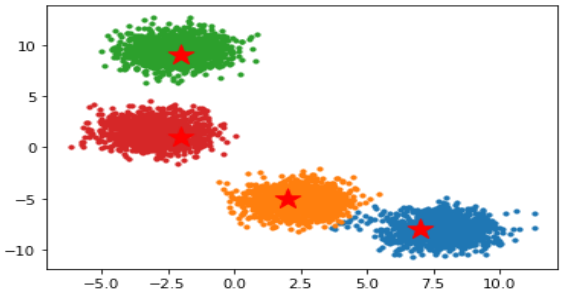

Figure 3: Correctly Clustered Data Points

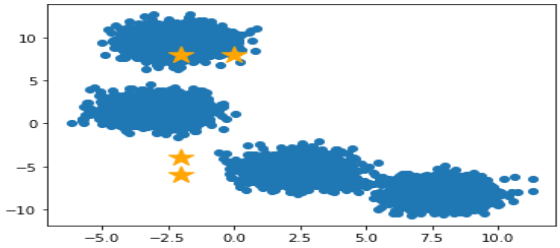

Figure 4: Random Initialization

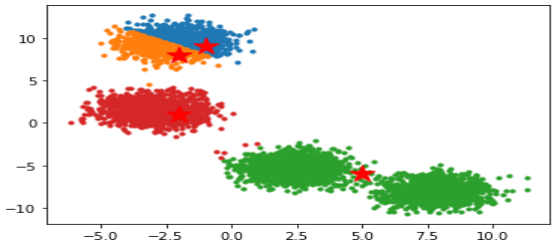

Figure 5: Incorrectly Clustered Data Points

Over the last decade, nature-inspired meta-heuristic algorithms for solving optimization problems have been developed (Fong *et al.*, 2014) this includes genetic algorithms, ant colony optimization, particle swarm optimization, cuckoo search, the firefly algorithm, the bat algorithm, and the bee algorithm, among many others. These algorithms have been demonstrated to be highly effective in solving a wide range of problems. They are thought to be capable of overcoming the limitations of K-means clustering algorithms in terms of getting trapped in the local optima due to their susceptibility to initialization and outliers (Muhima et al.,2020). This paper proposed a novel approach using a genetic algorithm with two-point crossover and adaptive mutation (GATCAM) for the k-means initialization procedure. This approach enhances k-means accuracy and guarantees that k-means do not become trapped at local minima. The rest of the paper is organised as follows: Section Ii investigates related literature, while Section III discusses materials and procedures. Section IV reported the experimental data, which were then discussed in Section V, and ultimately concluded in Section VI.

Related Literatures

K-means clustering approaches and initialization procedures have been extensively studied. Many scholars have offered various strategies to overcome the constraint of becoming caught in local minima. Heuristic and meta-heuristic approaches are among these strategies. (Ikotun *et al.*, 2023). For instance, (Barakbah & Helen, 2005) proposed an optimized k-means algorithm, the algorithm distributes the initial centroids according to data points' proximity to the middle. The algorithm was straight forward: It first measures the grand mean, denoted by m, then the nearest data point to m, denoted by c_1, distance between m and c_1, denoted by d_1, and c_1 becomes the first cluster automatically as shown in Figure 6 To get the second cluster, it obtain the second nearest data point to the center m, denoted by C_2 and a distance d_2, such that $d_2 > d_1$ and $d(c_2, c_1) \geq d_i$. The algorithm worked well for a dataset with less than four clusters but failed for datasets with a higher number of clusters.

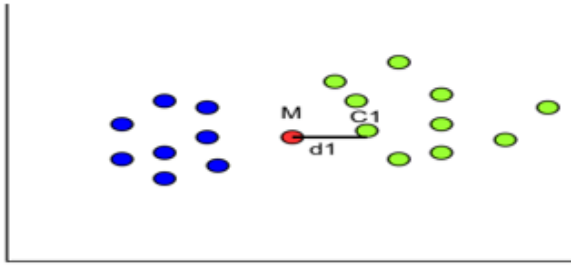

Figure 6: Determining The First Centroid For K-Means (Barakbah & Helen, 2005)

(Barakbah, 2006) calculates the initial centroids based on the highest distance weights of the centroids to stabilize the positioning of the initialized centroids in data distribution. Different from (Barakbah & Helen, 2005), who scattered the initial centroids based on the closeness of each data point to the grand mean of the data. In this technique, the grand mean of the data is calculated as the center of the data distribution. The distance measure is calculated by comparing each data point to the grand mean. As shown in Figure 7, the first centroid is chosen from the data point with the largest distance metric. To establish the second initial centroid, the distance metric from the first centroid is computed again and added to the distance metric from the first centroid. This is done to prevent data points that are too close to the previously picked centroids from being selected before the iteration process continues.

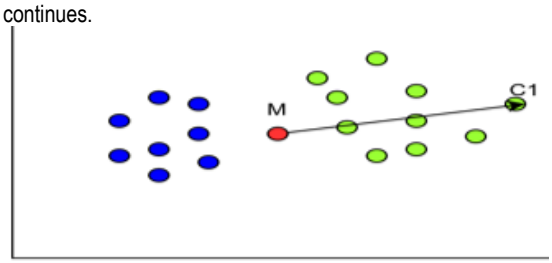

Figure 7: Determining the first centroid for k-means (Barakbah, 2006)

(Barakbah & Kiyoki, 2009), inspired by the thought process of placing a set of "pillars" in a house or building to make it sturdy, proposed the "pillar algorithm." This algorithm simulates how the pillars of a building should be placed so that the building might be able to stand up to the pressure of the roof. This is comparable to the positioning of the centroids in the k-means algorithm, which improves the performance of the K-means clustering algorithm (Barakbah, 2006). From the data points, the one with the highest cumulative distance metric is chosen as the first initial centroid. The

next initial centroids are chosen by changing the cumulative distance metric between each data point and all prior initial centroids, and then selecting the data point with the greatest distance as the new initial centroid. This repeated approach is required to designate all of the first centroids. This method also has a way to keep outlying data from being chosen as the initial centers. The results show that the suggested method is useful for improving the results of K-means clustering. However, when clustering huge quantities of data with multiple outliers, the approach takes a longer time to calculate the cumulative distance since its complexity is O((k+h+1) n), where k is the number of clusters, h is the number of outliers, and n is the number of data items to set the initial centroids.

Singh & Kaur, (2013) suggested a procedure for initializing the centroid that was extremely simple. In this technique, the arithmetic mean of the entire set of data points is calculated, and the mean point becomes the first centroid. The dataset is split into two parts, after which, the mean of these two pieces is determined. These will be the cluster's second and third centroid, respectively. This procedure is replicated until a total of k cluster centres is discovered. It asserts that the algorithm is more efficient than the standard k-means algorithm, which randomly selects the initial centroids, however, it does not guarantee convergence to the optimal in every run of the algorithm.

The Minimum-Average-Maximum (MAM) initialization technique was proposed by (Dhanabal & Chandramathi, 2013). During initialization, the MAM algorithm took into account outliers by starting with the first data point instead of picking one at random. It converged faster and with greater precision. The algorithm picks the first data point as the initial seed P_1. Then the distance from P_1 to all other data points is calculated using the Euclidean distance. The distance with the maximum value is marked P_2. Again, from P_2 find the distance to all other data points and choose the data point with a minimum distance called P_3. From P_3 find the distance to all other data points and choose the maximum distance which is marked as P_4. Then calculate the mean P_5 by adding P_3 and P_4 and dividing them by the number of rows in the dataset. If $K = 2$, choose P_3 and P_4 as the initial centroids. If $K = 3$ then P_3 , P_4 and P_5 becomes the initial centroids. However, if $K > 3$, then the centroids would be calculated using $\frac{(P_{i-1}+p_{i+1})}{n}$ where n is the number of rows in the dataset. This proposed algorithm worked effectively for a small dataset.

K-means is computationally expensive for large datasets. It requires $O(Knt)$ which is proportional to the number of data items, clusters, and iterations. Singh and Yadaf, (2013) reduced the processing time by $k-1$ iterations. This they achieved by calculating the distance for each data point to the nearest cluster and comparing the data points to the current and previous centroid on the second iteration; if the new distance is less than the previous distance, further computation is not necessary, otherwise it necessitates computing the interval between clusters. This concept was extremely successful, particularly for the data sets that have a large number of clusters, but was less effective on datasets with few clusters.

Bhusare & Bansode, (2014) enhanced the work of (Barakbah and Kiyoki, 2009) by calculating the accumulated distance between each data point and the mean. The data points with the greatest distance between them would be called the initial centroids. This approach uniformly distributes the original centroid in the function space to minimize the distance between them.

This reduces the difficulty of (Barakbah and Kiyoki, 2009) by eliminating the procedure of selecting the initial centroids, thereby reducing the complexity as the number of iterations increases.

Durga et al., (2020) compared out a study by combining genetic and nature-inspired optimization algorithms with the K-Means for clustering. (Mustafi & Sahoo, 2019) suggested a heuristic-based algorithm for optimizing the k-means clustering algorithm's initialization. The objective was to improve k-means efficiency while maintaining the necessary number of clusters, k. To find the best solution and the necessary number of clusters, Mustafi and Sahoo, (2019) used a mixture of GA and DE which essentially focuses on enhancing the initial selection of the centroids, which is also utilized by the k-means algorithm in order to ensure that the number of clusters is always returned in each run of the algorithm. However, because of the GA approach, the algorithm was slow. This paper proposed GATCAM for k-means initialization. The GATCAM algorithm is a genetic based algorithm in which the crossover and mutation phases use two-point crossover and mutation suggested by Bouhmala et al, (2015) and Jiang et al., (2020)

MATERIALS AND METHODS

Data Source

This paper used two benchmark datasets, Wine and Iris, which were downloaded from the UCI machine learning dataset repository. The key descriptions of the dataset are discussed in the following subsection:

Iris Dataset

In the fields of machine learning and statistics, the Iris dataset is well-known. The dataset consists of four characteristics (length and breadth of sepals and petals) for 50 samples of three Iris species (Iris setosa, Iris virginica, and Iris versicolor) (Kotu & Deshpande, 2019). The Iris dataset is often used as a dataset for machine learning and statistics beginners since it is small, straightforward, and simple to visualize. The attributes and scatter plot of the Iris dataset are shown in Table I and Figure 6, respectively.

Table I: Description of Iris Dataset

Attributes	Description of attributes
sepal length	Sepal length in cm
sepal width	Sepal width in cm
petal length	Petal length in cm
petal width	Petal width in cm

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

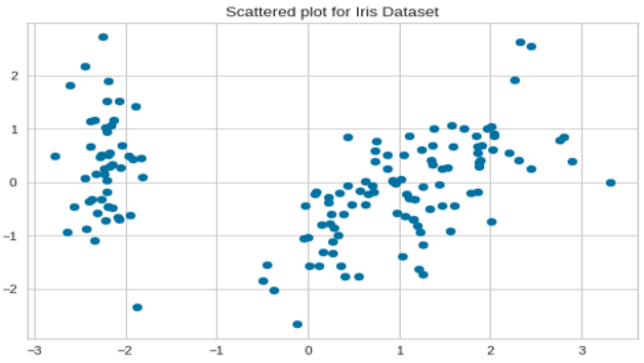

Figure 6: The Iris Data Set Scattered Plot

Wine Dataset

A wine dataset is a collection of information about different types of wines that can be used for analysis and research. It typically includes information such as the wine's origin, grape variety, alcohol content, price, ratings, and tasting notes (Kumar et al., 2020). Wine datasets can be used for a variety of purposes, such as analyzing consumer preferences, identifying trends in the wine industry, and training machine learning models to predict wine quality or classify different types of wines. There are several publicly available wine datasets that can be used for research and analysis. This paper used the dataset of a chemical study of wines cultivated in a specific region of Italy. The data set contains 178 samples representing three varieties of wine, with the results of 13 chemical analyses recorded for each sample. The variable's type has been changed to categorical. There are no missing values in the data, and it is entirely numerical, with a three-class target variable for classification (Kumar et al., 2020). Table II gives a detailed description and attributes of the wine dataset and the scattered plot (see figure 7 below).

Table II: Description of the Wine Dataset

Attributes	Description of attributes
Type	The type of wine, into one of three classes, 1(59), 2(71), and 3(48).
Alcohol	Alchol
Malic	Malic acid
Ash	Ash
Alcalinity	Alcalinity of Ash
Magnesium	magnesium
Phenols	phenols
Nonflavanoids	Nonflavanoids
Proanthocyanins	Proanthocyanins
Color	Color Intensity
Hue	Hue
Delution	delution
Propline	Proline

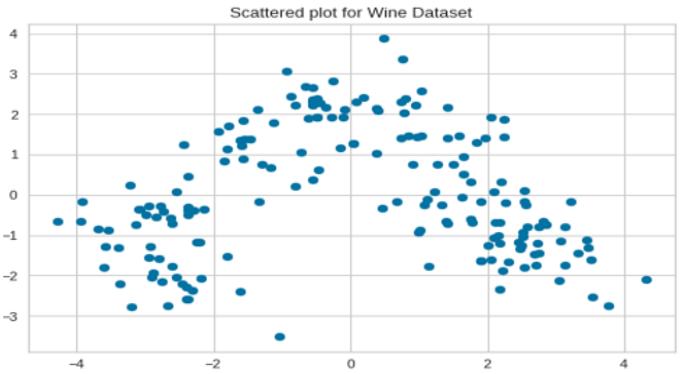

Figure 8: The Wine Dataset Scattered Plot

System Architecture

A genetic algorithm (GA) is a metaheuristic inspired by natural selection that is part of the wider family of evolutionary algorithms (EA). GAs are often employed to develop high-quality solutions for optimization and search problems by leveraging biologically inspired operators like mutation, crossover, and selection (Gerges et al., 2018). To obtain the initial centroid for the k-means clustering algorithm, this work used GA with two-point crossover and adaptive mutation (GATCAM) for centroid initialization in the k-means clustering algorithm. This improved on the performance of K-Means initialized with standard GA. Figure 9 depicts the flowchart of the system design, which begins with raw data that was acquired from a source in either comma-separated values (.csv) or text format (.txt). The dataset was visualized with plots and charts in a Pandas data frame. The dataset was then examined to find the best cleaning process for transforming categorical data to numerical values. In addition, the dataset features are reduced to a tolerable size during the data processing step. This was accomplished through the use of principal component analysis (PCA), and the dataset was then standardized or normalized. This was done because data may contain large numeric values that differ in large proportions and may be provided in several measurement units. The dataset was then used as the initial population for the genetic algorithm.

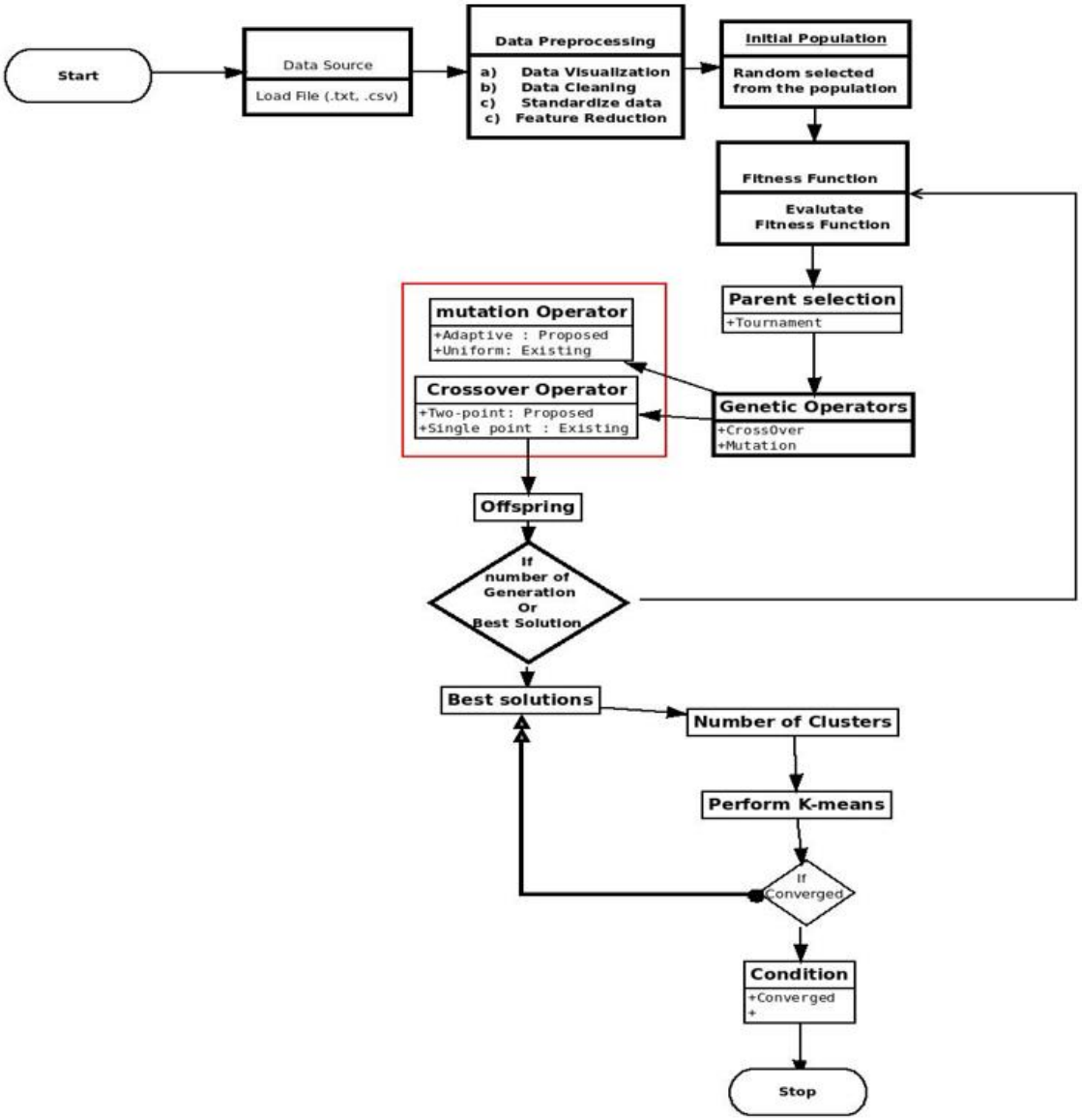

Figure 9: Flow Chart Showing the System Architecture

Data Pre-processing and Visualization

Data pre-processing is a data mining method that involves the conversion of raw data into a usable format. Data is often insufficient, inconsistent, and prone to mistakes. Data preparation addresses these problems by cleaning, converting, eliminating features, and standardizing data.

Data visualization is the use of visual components such as graphs, charts, and maps to create a graphical representation of data. It aids in the comprehension of anomalies, patterns, and trends in data. by putting the information in a visual context. Data visualization helps us understand what the data is all about. This study employed Pandas, Matplotlib, and Seaborn technologies to perform data preprocessing and visualize the data.

Initial Population

The initial population was drawn at random from the search space. The entire dataset forms the search space from which the initial population was randomly drawn through a selection method. This work used tournament selection because fitness values with higher values have better chances of being selected for reproduction in the next generation. The selection procedure was carried out after the evaluation of the fitness function of the population. The selection of the parent was based on performance. Lower-ranked chromosomes are discarded in this situation, and the remaining population is employed for reproduction.

GATCAM k-Means Clustering

Figure 10 provides a flow chart of the system architecture that explains the GATCAM algorithm. The algorithm concentrated on selecting the initial clusters that are both well separated from one

another and evenly distributed throughout the search space. The motivation for doing this is to ensure that the search covers a maximal amount of the search space before the clustering phase begins. The GATCAM algorithm employed a two-point crossover and an adaptive mutation that solved the problem of k-means convergence to the local minima. The GATCAM utilized the fitness function that maximizes the centroids. This was to ensure that the centroids are adequately dispersed in the search space.

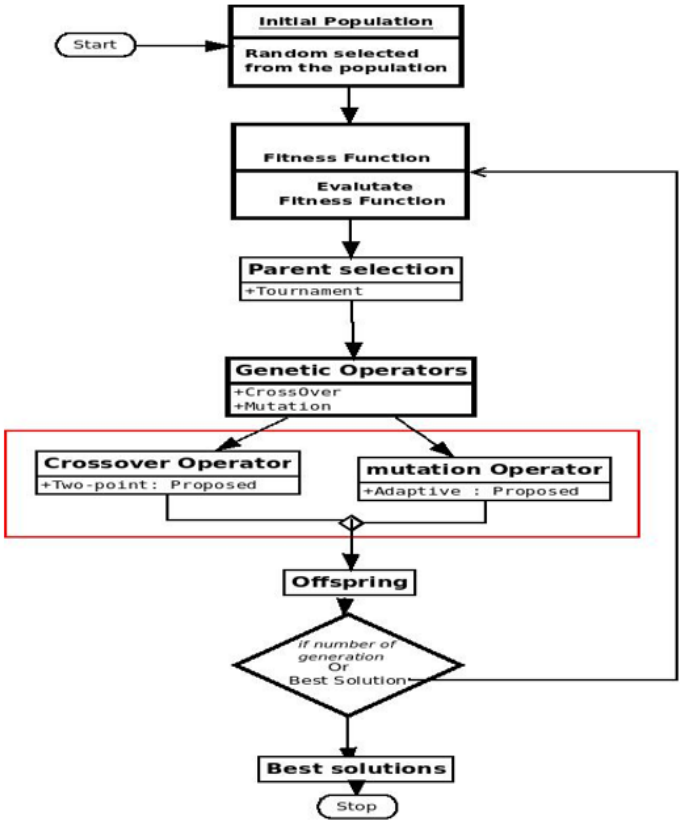

Figure 10: Flow Chart showing GATCAM Algorithm

Algorithm I GATCAM Algorithm

1. *Randomly initialize populations p.*
2. *Determine fitness of population*
3. *Until convergence repeat*
4. *Select parents from population*
5. *Crossover and mutation*
6. *mutation on new population*
7. *Calculate fitness for new population*

Programming Language

The GATCAM algorithm for the centroid initialization in the K-means clustering algorithm was implemented in the Python programming language. pandas, numpy, sklearn, matplotlib, and a variety of other open-source libraries and an application programming interface (API) were utilized in the processing and visualizing of data, as well as the implementation of the solution. The study employed the Jupyter notebook, which is a text editor.

Evaluation of Fitness Function

The fitness of an individual in a genetic algorithm is the value of an objective function for its phenotype. In calculating fitness, the chromosome has to be first decoded and the objective function has to be evaluated. The fitness does not only indicates how good the solution is, but also corresponds to how close the chromosome is to the optimal one. The fitness function for the centroid initialization of the k-means is presented in equation 1

$$maximise = \sum dist\,(C_k, m) + \sum\sum dist\,(c_i, c_j) \quad (1)$$

(Mustafi & Sahoo, 2019)

where m is the center of the entire solution, c_i, c_j are the centroid of the i^th and the j^{th} Clusters respectively, $dist(\ )$ is the distance metrics used in the calculation, and k is the number of clusters.

This research work used a floating point encoding format , the chromosome used was an array of size $K * nooffeatures$, where $nooffeatures$ is the dimensionality of the data space.

The first term of the objective function was used to locate the centers that were of far distant from the center of the total data set, while the second term guarantee that the centers were widely separated from each other.

RESULTS

In this section, presented results from the experiment and compared with findings from previous studies. The study used tables and graphs to illustrate the findings in order to improve the quality of comprehension.

Table III. Experimental result showing the performance of GATCAM and SGA K-Means

Initialization Methods	Accuracy		Inertia		Silhouette Score		DBI		Iteration	
	Wine	Iris	Wine	Iris	Wine	Iris	Wine	Iris	Wine	Iris
GATCAM k-Means	88.6	66.12	259.8	117.7	0.57	0.5	0.5	0.71	6	5
SGA K-Means	84.4	64.89	260	166.2	0.56	0.51	0.6	0.72	6	8

Comparative Analysis

GATCAM k-means performance was assessed using accuracy, Inertia, the silhouette coefficient, DB-index and number of iteration. Figure 9 and 10 plots the actual dataset before clustering.

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

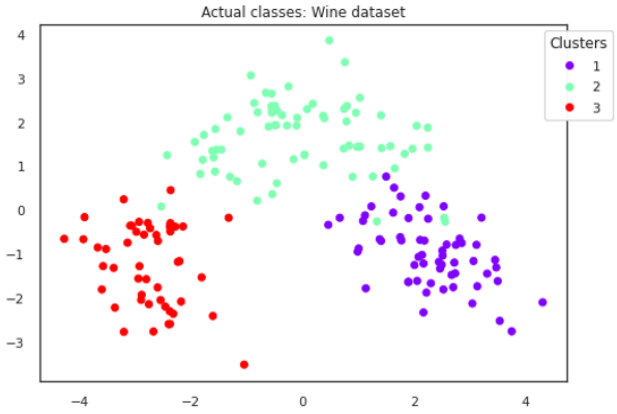

Figure 11: Actual Classes for Wine Dataset

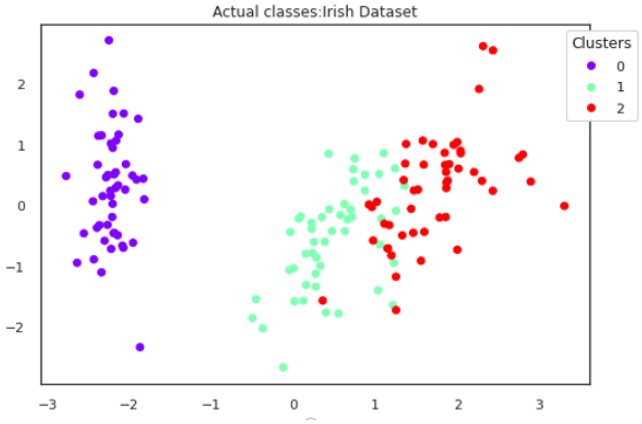

Figure 12: Actual Classes for Iris Dataset

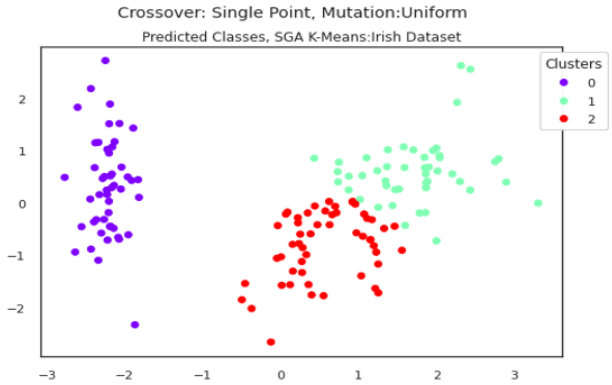

Figure 13: k-means Clustering, GA centroid Initialization with Single point crossover and uniform mutation for Iris Dataset

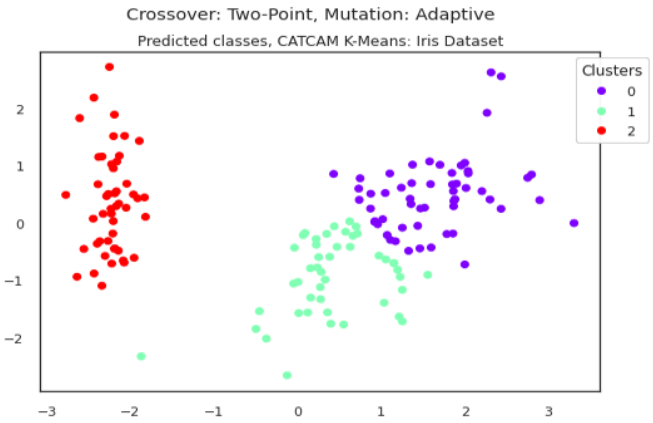

Figure 14: k-means Clustering, GA centroid Initialization with Two Point Crossover and Adaptive Mutation for Iris Dataset

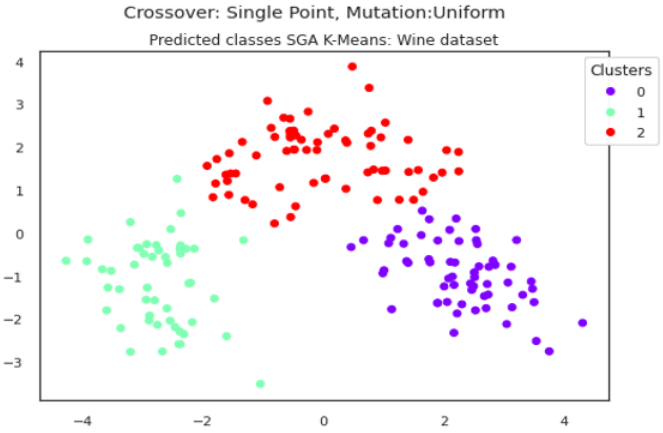

Figure 15: k-means Clustering, GA centroid Initialization with Single Point Crossover and Uniform Mutation for Wine Dataset

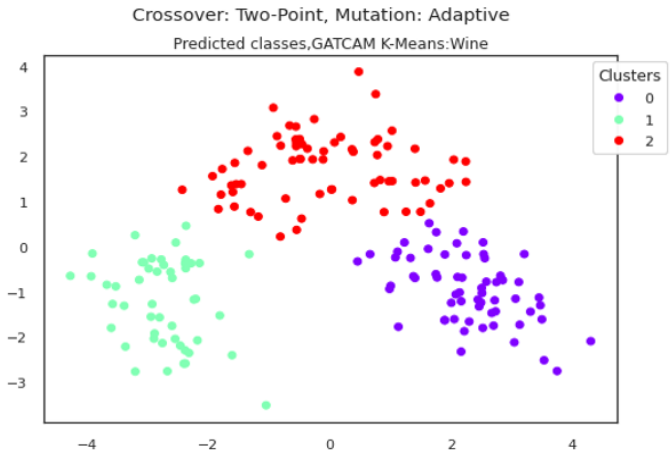

Figure 16: k-means Clustering, GA centroid Initialization with Two Point Crossover and Adaptive Mutation for Wine Dataset

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

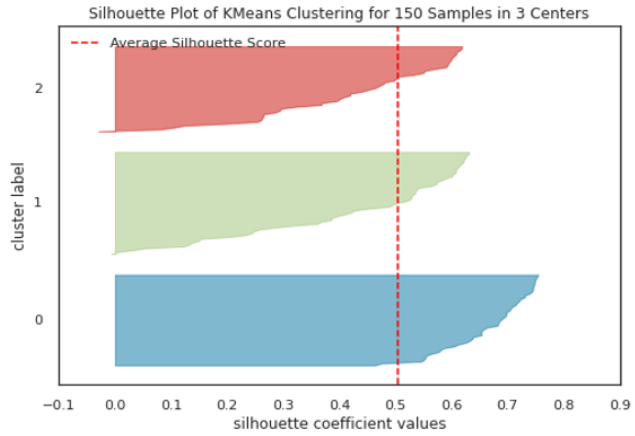

Figure 17: Silhouette Score for Wine Dataset

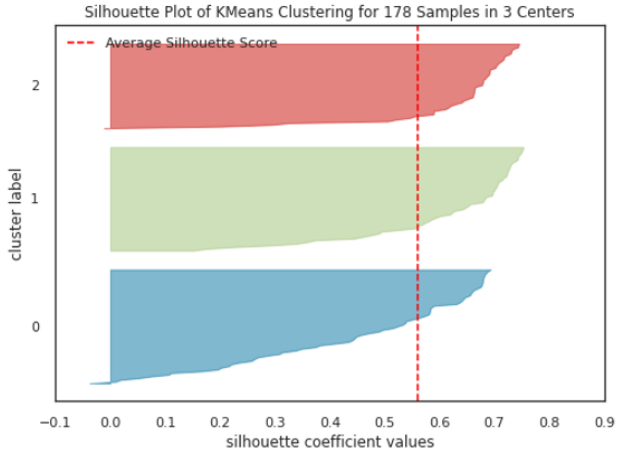

Figure 18: Silhouette Score for Iris Dataset

DISCUSSION

Due to its simplicity and ease of interpretation, the K-means is one of the most commonly utilized clustering algorithms in real-world applications. However, because of the random initialization, it has the issue of convergent convergence to the local minimum. As such, numerous studies have sought to remedy this problem, but the problem still tarries. The GATCAM technique for the K-means initialization was proposed in this study. The GATCAM is a genetic algorithm-based two-point crossover and adaptive mutation algorithm. The GATCAM's performance was evaluated in the experiment using two benchmark datasets (wines and arise).

The random and the SGA initialization were compared to GATCAM. Tables III provide the experimental findings for the performance measures of GATCAM and SGA centroid initialization for k-means. The GATCAM outperforms the SGA in terms of accuracy and initiation. There was no significant difference in the silhouette score and the DBI for k-means in Figs 14 and 15. The result indicates the performance of the k-means rather than the initialization. as evident in the accuracy and the inertia. The GATCAM initialization enhanced the performance of k-means compared to the other SGA initializations in k-means clustering, and the result shows that the GATCAM produces high-quality clusters. The adaptive mutation in GATCAM reduced the randomness of genetic algorithms, which is one of the major drawbacks of SGAs, in line with the conclusion of Marsili et al., (2000). This finding also agrees with that of Al-shboul & Myaeng, (2017) and (Kumara et al., 2019) in affirming that the initialization of the centroid affects the performance of k-means.

The silhouette value and the DBI did not vary significantly; this is because the metric tested the quality of the clusters that are produced by the clustering techniques and not the initialization method. The GATCAM initialization generated higher-quality clusters than SGA because it offered well-separated initial data point values that took outliers into account. This validates the findings and makes the results valid, which agree with those of Fränti and Sieranoja (2019).

Conclusion

K-means is one of the most widely used clustering algorithms in data mining. It is fast, simple to implement, and easy to interpret. However, it does not guarantee convergences to the global minimum due to its random initialization of the centroid during implementation. Results are usually subjected to several subjective interpretations, and this has led researchers to suggest heuristics and meta-heuristic techniques to solve the problem. As such, there has not been an acceptable and satisfactory solution as the problem still continues. This paper proposes the GATCAM k-means clustering algorithm. The performance of GATCAM K-means was compared with that of SGA K-means. The study used two benchmark datasets (wine and iris) to evaluate the performance, using the silhouette score, inertia, Bouldin index, completeness score, and number of iterations metrics. The experimental results show that GATCAM k-means outperform SGA k-means in accuracy and inertia. The findings in this paper agree with those of other known research. However, there was no significant difference between the silhouette score and the Bouldin index. This is because the result showed the quality of the clustering and not the initialization. The GATCAM k-means generates more accurate clusters with fewer iterations, as shown in Table III. The GATCAM initialized the centroid in such a way that the centroids are well spread out and outliers are considered based on the fitness function used.

This study suggests that future research should look into different types of crossover and mutation methods to improve the initialization of the k-means clustering algorithm and its performance.

REFERENCES

Al-shboul, B. and Myaeng, S. (2017) 'Initializing K-Means using Genetic Algorithms', (November 2009).

Barakbah, A.R. (2006) 'A New Algorithm for Optimization of K-Means Clustering with Determining Maximum Distance Between Centroids', *Industrial Electronics Seminar (IES) 2006*, (October 2006), pp. 240–244.

Barakbah, A.R. and Helen, A. (2005) 'Optimized K-means : an algorithm of initial centroids optimization for K-means', in *Seminar on Soft Computing, Intelligent System, and Information Technology (SIIT) 2005*.

Barakbah, A.R. and Kiyoki, Y. (2009) 'A pillar algorithm for k-means optimization by distance maximization for initial centroid designation', *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009-Proceedings*, pp. 61–68. Available at: https://doi.org/10.1109/CIDM.2009.4938630.

Bhusare, B.B. and Bansode, S.M. (2014) 'Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm', 3(4), pp. 1317–1322.

Bora, D.J. and Gupta, A.K. (2014) 'A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm', *International Journal of Computer Trends and Technology (IJCTT)*, 10(2), pp. 108–113. Available at: https://doi.org/DOI : 10.14445/22312803/IJCTT-V10P119.

Bouhmala, N., Viken, A. and Lønnum, J.B. (2015) 'Enhanced Genetic Algorithm with K-Means for the Clustering Problem', *International Journal of Modeling and Optimization*, 5(2), pp. 150–154. Available at: https://doi.org/10.7763/ijmo.2015.v5.452.

Cao, F., Liang, J. and Jiang, G. (2009) 'An initialization method for the K-Means algorithm using neighborhood model', *Computers and Mathematics with Applications*, 58(3), pp. 474–483. Available at: https://doi.org/10.1016/j.camwa.2009.04.017.

Chang, S., Zhenzhong, X. and Xuan, G. (2018) 'Improvement of K Mean Clustering Algorithm Based on Density', *CoRR*, abs/1810.0(1). Available at: http://arxiv.org/abs/1810.04559.

Dhanabal, S. and Chandramathi, S. (2013) 'An efficient k-means initialization using minimum-average-maximum (MAM) method', *Asian Journal of Information Technology*, 12(2), pp. 77–82. Available at: https://doi.org/10.3923/ajit.2013.77.82.

Durga Bhavani, K. and Radhika, N. (2020) 'K-means clustering using nature-inspired optimization algorithms-A comparative survey', *International Journal of Advanced Science and Technology*, 29(6 Special Issue), pp. 2466–2472.

Fong, S. *et al.* (2014) 'Towards enhancement of performance of K-means clustering using nature-inspired optimization algorithms', *Scientific World Journal*, 2014. Available at: https://doi.org/10.1155/2014/564829.

Fränti, P. and Sieranoja, S. (2019) 'How much can k-means be improved by using better initialization and repeats?', *Pattern Recognition*, 93, pp. 95–112. Available at: https://doi.org/10.1016/j.patcog.2019.04.014.

Gerges, F., Zouein, G. and Azar, D. (2018) 'Genetic algorithms with local optima handling to solve sudoku puzzles', *ACM International Conference Proceeding Series*, pp. 19–22. Available at: https://doi.org/10.1145/3194452.3194463.

Ikotun, A.M. *et al.* (2023) 'K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data', *Information Sciences*, 622, pp. 178–210. Available at: https://doi.org/10.1016/j.ins.2022.11.139.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*, 31(8). Available at: https://doi.org/10.1016/j.patrec.2009.09.011.

Jiang, C. *et al.* (2020) 'An enhanced genetic algorithm for parameter estimation of sinusoidal signals', *Applied Sciences (Switzerland)*, 10(15). Available at: https://doi.org/10.3390/app10155110.

Kaushik, S. (2016) 'An Introduction to clustering and different methods of clustering', *Analytics Vidhya* [Preprint].

Kotu, V. and Deshpande, B. (2019) 'Data Exploration', *Data Science*, pp. 39–64. Available at: https://doi.org/10.1016/b978-0-12-814761-0.00003-4.

Kumar, S., Agrawal, K. and Mandan, N. (2020) 'Red wine quality prediction using machine learning techniques', in *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*. Available at: https://doi.org/10.1109/ICCCI48352.2020.9104095.

Kumara, A., Bharadwaj, H.S. and Ramaiah, N.S. (2019) 'A Survey on K-means Algorithm Centroid Initialization', *SSRN Electronic Journal*, pp. 1–3. Available at: https://doi.org/10.2139/ssrn.3372643.

Liu, Z., Bao, J. and Ding, F. (2018) 'An improved k-means clustering algorithm based on semantic model', *ACM International Conference Proceeding Series* [Preprint]. Available at: https://doi.org/10.1145/3148453.3306269.

MacQueen, J. (1967) 'Some methods for classification and analysis of multivariate observations', in *Proc. fifth Berkeley Symp. Math. Stat. Probab*. (1), p. 281<unicode>8211</unicode>297.

Marsili Libelli, S. and Alba, P. (2000) 'Adaptive mutation in genetic algorithms', *Soft Computing*, 4(2), pp. 76–80. Available at: https://doi.org/10.1007/s005000000042.

Muhima, R.R., Kurniawan, M. and Pambudi, O.T. (2020) 'A LOF K-Means Clustering on Hotspot Data', *International Journal of Artificial Intelligence & Robotics (IJAIR)*, 2(1), pp. 29–33. Available at: https://doi.org/10.25139/ijair.v2i1.2634.

Mustafi, D. and Sahoo, G. (2019) 'A hybrid approach using genetic algorithm and the differential evolution heuristic for enhanced initialization of the k-means algorithm with applications in text clustering', *Soft Computing*, 23(15), pp. 6361–6378. Available at: https://doi.org/10.1007/s00500-018-3289-4.

Omran, M.G.H., Engelbrecht, A.P. and Salman, A. (2007) 'An overview of clustering methods', *Intelligent Data Analysis*, 11(6), pp. 583–605. Available at: https://doi.org/10.3233/ida-2007-11602.

Oyelade, J. *et al.* (2019) 'Data Clustering: Algorithms and Its Applications', *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, (ii), pp. 71–81. Available at: https://doi.org/10.1109/ICCSA.2019.000-1.

Pauletic, I., Prskalo, L.N. and Bakaric, M.B. (2019) 'An overview of clustering models with an application to document clustering', in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings*, pp. 1659–1664. Available at: https://doi.org/10.23919/MIPRO.2019.8756868.

Peña, J.M., Lozano, J.A. and Larrañaga, P. (1999) 'An empirical comparison of four initialization methods for the K-Means algorithm', *Pattern Recognition Letters*, 20(10), pp. 1027–1040. Available at: https://doi.org/10.1016/S0167-8655(99)00069-0.

Singh, H. and Kaur, K. (2013) 'New Method for Finding Initial Cluster Centroids in K-means Algorithm', *International Journal of Computer Applications*, 74(6). Available at: https://doi.org/10.5120/12890-9837