


A New Approach in Metaheuristic Clustering: Coot Clustering

*¹Gökhan KAYHAN, ²İsmail İŞERİ

*¹ Corresponding Author, Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Türkiye, gkayhan@omu.edu.tr 

² Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Türkiye, ismail.iseri@omu.edu.tr 

Abstract

As a result of technological advancements, the increase in vast amounts of data in today's world has made artificial intelligence and data mining significantly crucial. In this context, the clustering process, which aims to explore hidden patterns and meaningful relationships within complex datasets by grouping similar features to conduct more effective analyses, holds vital importance. As an alternative to classical clustering methods that face challenges such as large volumes of data and computational complexities, a metaheuristic clustering method utilizing Coot Optimization (COOT), a swarm intelligence-based algorithm, has been proposed. COOT, inspired by the hunting stages of eagles and recently introduced into the literature, is a metaheuristic method. Through the proposed COOT metaheuristic clustering method, the aim is to contribute to the literature by leveraging COOT's robust exploration and exploitation processes, utilizing its dynamic and flexible structure. Comprehensive experimental clustering studies were conducted to evaluate the consistency and effectiveness of the COOT-based algorithm using randomly generated synthetic data and the widely used Iris dataset in the literature. The same datasets underwent analysis using the traditional clustering algorithm K-Means, renowned for its simplicity and computational speed, for comparative purposes. The performance of the algorithms was assessed using cluster validity measures such as Silhouette Global, Davies-Bouldin, Krznowski-Lai, and Calinski-Harabasz indices, along with the Total Squared Error (SSE) objective function. Experimental results indicate that the proposed algorithm performs clustering at a competitive level with K-Means and shows potential, especially in multidimensional datasets and real-world problems. Despite not being previously used for clustering purposes, the impressive performance of COOT in some tests compared to the K-Means algorithm showcases its success and potential to pioneer different studies aimed at expanding its usage in the clustering domain.

Keywords: Clustering, Metaheuristic, Coot Optimization, K-Means

1. INTRODUCTION

With advancements in science and technology, various data mining methods are employed to transform increasingly complex and irregular large-scale data into meaningful insights through computer programs. Clustering, which involves identifying hidden patterns and meaningful relationships within data, poses one of the challenging problems in the field of data mining. It entails grouping data based on shared features and is commonly preferred as an unsupervised learning technique [1]. Numerous classical and heuristic algorithms exist for solving clustering problems. Among classical clustering algorithms, the K-means (KM) algorithm remains widely used due to its speed and simplicity in operations [2]. However, challenges such as getting trapped in local solutions due to erroneous initial parameter selections and slow convergence for large datasets persist in traditional methods, prompting the development of new techniques. In recent years, metaheuristic optimization

algorithms, which excel in global searches and avoid getting stuck in local solutions, have been frequently employed to overcome the difficulties encountered by traditional algorithms [3].

Many clustering studies in the literature utilize KM and metaheuristic algorithms. For example, in the clustering of multidimensional data, a new approach for preventing local solution traps by considering the farthest points for initial cluster center selection has been proposed for KM and metaheuristic Particle Swarm Optimization (PSO)-based clustering methods [4]. In another study, KM clustering was used to group five different countries, including Turkey, based on economic and financial indicators such as inflation rates and stock indices [5]. Another application involved clustering samples from 45 different crude oil sources based on their physicochemical properties using the KM algorithm [6]. To mitigate local optima issues in the traditional KM algorithm, some studies have incorporated Levy flight

equations. In the medical field, an image segmentation application for brain tumor detection utilized Otsu thresholding and KM clustering algorithms together [7]. A hybrid clustering method was proposed for segmenting brain MR images by combining the KM algorithm with the metaheuristic Gray Wolf Optimization Algorithm (GWO), demonstrating its success [8]. Comparing the metaheuristic Sine Cosine Algorithm (SCA) with classical methods yielded satisfactory results in segmenting multiple reference images [9]. An evolutionary metaheuristic Genetic Algorithm (GA)-based clustering method was suggested and observed to be successful when compared to KM [1]. A hybrid method created by combining Whale Optimization Algorithm (WOA) with a classical clustering technique was proposed and its success examined [10]. A Gray Wolf Optimization (GWO)-based clustering method was suggested and proven to be superior in many datasets compared to six known metaheuristic-based clustering algorithms [2].

A modified version of the COOT algorithm is introduced in a study to address potential drawbacks, such as the possibility of becoming stuck in local minima. Two novel techniques, Opposition-Based Learning and Orthogonal Learning are incorporated into this new version. Named mCOOT, this algorithm has been tested on the dimensionality reduction problem and has been demonstrated to be superior to similar algorithms in terms of classification accuracy and the number of selected features. These findings highlight the effectiveness and practical potential of the proposed algorithm [11]. In another study, a novel hybrid COOT-ANN model is proposed, where the COOT algorithm, previously unutilized in training ANNs, is employed for classification tasks. The weight and bias values of a single hidden layer ANN model are optimized using the COOT algorithm instead of traditional gradient descent algorithms. The performance of the proposed ANN model is assessed in classification tasks using four distinct datasets (wine, breast cancer, iris, glass) through experimentation [12]. Additionally, a new approach proposed in a paper combines deep convolutional neural networks (HDCNN) with the COOT algorithm to predict disease risks. Initially, an improved crossover-based Levy flight optimization algorithm (ICLFDO) is utilized to process unstructured textual data. Subsequently, the HDCNN-COOT approach is implemented for more accurate disease predictions. Furthermore, the classifier determines the future disease risks for patients. The effectiveness of the proposed model is evaluated using data obtained from the University Hospital of Ludwig Maximilian University of Munich, Germany, comprising 29,477,035 data items from 36,082 patients. The model demonstrates superior performance in classification accuracy and classifier performance across five different datasets in experimental results [13].

In one study, a modified version of the COOT optimization algorithm, called MCOOT, was introduced to solve the community detection problem. MCOOT enhances the exploration and exploitation capabilities by introducing some modifications to the basic COOT method, thereby providing more effective performance in community detection problems. The results of the study demonstrate that MCOOT exhibits superior or comparable performance

compared to other optimization methods. Therefore, MCOOT is suggested as a competitive solution for community detection problems [14]. In another study, a meta-model-based approach has been developed for multi-objective optimization in real building designs. This method starts with building performance simulation using EnergyPlus™ and then combines it with the Modified Coot Optimization Algorithm (MCOA) and artificial neural network meta-models (ANN-MM). The aim of this approach is to minimize the sample generation used for training and validation to achieve accurate optimization results. The obtained results are compared with the Pareto front obtained through simulation-based optimization, resulting in a 75% reduction in computational power [15]. In another study, a research is presented where six different meta-heuristic algorithms are employed to address the community detection problem. These algorithms have been adapted to be effective in solving CD problems. Additionally, a fast approach has been proposed to reduce the time cost when solving the problem. Experimental results indicate that the COOT algorithm is more effective than others, and the CommunityID-based approach enables faster solutions. Therefore, it is concluded that COOT can be an effective alternative method for community detection problems, and the CommunityID-based approach can provide significant solutions in larger networks [16]. A study utilizing the COOT algorithm focuses on gene selection strategy. The aim of the study is to utilize microarray analysis of gene expression for disease and cancer diagnosis and prognosis. However, identifying gene biomarkers is challenging in microarray cancer classification due to the complexity of different cancer types and the high dimensionality of the data. Therefore, the study proposes a gene selection strategy using the binary version of the COOT optimization algorithm, called BCOOT, to identify genes targeted for cancer and disease classification. Three different binary COOT variants are proposed: BCOOT, BCOOT-C, and BCOOT-CSA. These algorithms are tested in conjunction with a pre-filtering technique such as minimum redundancy maximum relevance (mRMR). The experiments demonstrate that the BCOOT-CSA approach outperforms other techniques in terms of prediction accuracy and the number of selected genes [17]. Another study introduces a hybrid approach combining machine learning algorithms with expert medical knowledge for precise classification of brain MRIs. In the proposed classification system, a comprehensive feature set is extracted using GLCM. Additionally, the feature extraction process is enhanced using COOT optimization, resulting in improved features. Finally, a model trained with CNNs achieves increased accuracy in classifying new images [18]. COOT optimization has been utilized to predict disease risk using patients' medical data, and a COOT-based hybrid deep convolutional neural network (HDCNN) is proposed. Within the scope of the study, unstructured textual data was processed using an improved crossover-based levy flight optimization algorithm (ICLFDO). Subsequently, disease prediction was performed using the HDCNN-COOT approach. The effectiveness of the proposed model was evaluated on a large dataset obtained from the University Hospital of Ludwig Maximilian University of Munich, Germany. Experimental results demonstrate that the

proposed model achieves higher classification accuracy and improved performance of classifiers [19].

In this study, the aim is to achieve accurate and effective clustering using the swarm intelligence-based metaheuristic method called COOT optimization algorithm, known for its flexible structure in handling multidimensional data and its adaptability to complex data structures, to overcome the issues of getting trapped in local optima with classical clustering algorithms.

The organization of the paper is as follow: In Chapter 2, an explanation is provided for the materials and methods utilized in the study. In Chapter 3, the results of the proposed method are presented comparatively. In Chapter 4, the conclusions and future directions of the study are introduced.

2. MATERIALS AND METHODS

The COOT metaheuristic clustering algorithm is applied to synthetic datasets and the Iris dataset obtained from the UCI Repository to evaluate performance [20]. The Iris dataset comprises measurements of various flower species, totaling 150 samples with attributes such as sepal and petal lengths and widths. On the other hand, synthetic dataset 1 (SV-1) comprises 400 data points, while synthetic dataset 2 (SV-2) comprises 500 data points. SV-1 dataset consists of 4 clusters, whereas SV-2 consists of 5 clusters. Each cluster contains 100 data points randomly distributed around a center. A comparison is drawn between this method and the classical KM algorithm, using criteria such as Silhouette Global (SG), Mean Davies-Bouldin (DB), Krzanowski-Lai (KL), and Calinski-Harabasz (CH) to determine cluster validity. Detailed experimental outputs are presented in the findings section. MATLAB is used to assess the effectiveness of the COOT clustering algorithm on the Iris dataset and two randomly distributed synthetic datasets. The classical KM algorithm is also implemented on the same datasets. Each algorithm yields four different performance index values, enabling a comparative analysis between the proposed COOT algorithm and KM. The steps of the KM algorithm are outlined in Figure 1. The initial selection of clusters and centroids significantly influences the algorithm's performance. Additionally, KM tends to converge toward local solution points, potentially incurring high costs when handling large datasets.

The algorithms in this study aim to minimize the Total Sum of Squared Errors (SSE) as the objective function in each iteration. SSE represents the sum of the squares of distances between each data point and its assigned cluster centroid. In each cycle, the objective is to find cluster centroids that minimize the SSE value. A smaller SSE indicates homogeneity and similarity among the data points within clusters. Here, k denotes the number of clusters, m_j represents the j -th cluster center, $|g_j|$ denotes the number of elements in the j -th cluster, x_i signifies the i -th data vector, $\|x_i - m_j\|$ represents the Euclidean distance, and $j=(1,2,\dots,k)$. SSE is computed as shown in Equation 1.

$$SSE = \sum_{j=1}^k \sum_{i=1}^{|g_j|} \|x_i - m_j\|^2 \quad (1)$$

Performance indices are utilized as quality measures to assess and compare the performance of clustering algorithms. These indices offer insights into the accuracy of the clustering process. In clustering algorithms, the aim is to have high similarity within clusters among their own elements and low similarity across different clusters.

The Mean Davies-Bouldin (DB) index is computed by taking the average of the total sums of maximum similarities between each cluster and other clusters, considering k clusters. A lower value of this index indicates successful clustering, signifying homogeneity within clusters and significant dissimilarity among clusters [21]. Here, k represents the number of clusters, x denotes cluster elements, m_i stands for the i -th cluster center, $|g_j|$ represents the number of elements in the j -th cluster, and $d(m_i, m_j)$ symbolizes the distance between i -th and j -th cluster centers. σ_i and σ_j represent the averages of the distances between data vectors and their respective centers within i -th and j -th clusters (as in Equation 2). Mean DB is calculated as shown in Equation 3.

$$\sigma_i = \frac{1}{|g_i|} \sum_{x \in g_j} \|x - m_i\|^2 \quad (2)$$

$$\text{Mean DB} = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(m_i, m_j)} \right) \quad (3)$$

- 1 **Initially**, select random $\mu_1, \mu_2, \dots, \mu_k$
- 2 For each data point x_i , use the following formula to determine the closest centroid:
$$\operatorname{argmin}_j \|x_i - \mu_j\|^2$$
- 3 Here, $\|x_i - \mu_j\|^2$ denotes the squared Euclidean distance between x_i and centroid μ_j
- 4 Assign each data point to its closest centroid.
- 5 Calculate new cluster centroids based on the assigned data points:
$$\mu_j = \frac{1}{|c_j|} \sum_{x_i \in c_j} x_i$$
- 6 Here, c_j represents the set of data points assigned to centroid j .
- 7 Stop if there is no change in centroids or upon reaching a certain number of iterations.

Figure 1. Pseudo code for Kmeans

The Calinski-Harabasz (*CH*) index used in evaluating clustering performance takes into account the similarities between clusters, yielding a higher value when clusters are well-separated. Ideally, clusters should exhibit high intra-cluster homogeneity and low inter-cluster similarity for an effective clustering. A high *CH* index signifies greater success in clustering [21]. Here, k represents the number of clusters, n indicates the total number of elements in the dataset, x_{ij} denotes the j -th element of the i -th cluster, n_i signifies the number of elements in the i -th cluster, m_i represents the i -th cluster center, \bar{M} signifies the center of the entire dataset, B_k signifies the measure of similarity between cluster centers (as in Equation 4), and W_k signifies the measure of intra-cluster similarity (as in Equation 5). The calculation of the *CH* index is done as indicated in the Equation 6.

$$CH = \frac{B_k}{W_k} \frac{n-k}{k-1} \quad (4)$$

$$B_k = \sum_{i=1}^k n_i \|m_i - \bar{M}\|^2 \quad (5)$$

$$W_k = \sum_{i=1}^k \sum_{j=1}^{n_i} \|x_{ij} - m_i\|^2 \quad (6)$$

The Krzanowski-Lai (*KL*) index is a metric utilized to determine the optimal number of clusters based on the slope of the graph that emerges when the sum of squared distances of each data point within a cluster to its cluster center ($Z(k)$) is computed separately for different chosen numbers of clusters [22]. The *KL* index is formed by observing a rapid decrease in the value of $Z(k)$ until it reaches an appropriate number of clusters, followed by a slow change after reaching this optimal point. Here, denoting the number of clusters as k , the *KL* index is calculated as shown in Equation 7 [23]. Using the number of clusters (k) where the *KL* index reaches its maximum value is considered suitable for successful clustering. Simultaneously, in this study, it serves as a criterion for determining the success of clustering based on the initially chosen number of clusters by the user.

$$DIFF(k) = \left[(k-1)^{2/p} Z(k-1) - k^{2/p} Z(k) \right] \quad (6)$$

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \quad (7)$$

The Silhouette Global (*SG*) index indicates how homogeneous each clustered data point is within its cluster and how separated it is from other clusters. Silhouette scores are computed for each data point, and these scores' average yields the global score. Ranging between -1 and 1, higher values indicate successful clustering [23]. Here, n represents the number of data set elements, $S(i)$ denotes the Silhouette score calculated for the i -th data point (Equation 8), $a(i)$ represents the average distance of the i -th point to other points in its cluster, and $b(i)$ indicates the average distance of the i -th point to the nearest points in other clusters. The *SG* index is calculated according to Equation 10.

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (8)$$

$$a(i) = \frac{\sum d_{ik}}{n_r-1}, \quad b(i) = \min\left(\frac{\sum d_{ik}}{n_s}\right) \quad (9)$$

$$SG = \frac{1}{n} \sum_{i=1}^n S(i) \quad (10)$$

2.1. COOT Algorithm

The COOT optimization algorithm is designed by referencing the movements of coots on the water. This algorithm mimics four fundamental coot behaviors:

- 1- Random Movement: Coots explore different areas by expanding their search field. If the algorithm gets stuck in a local optimum, random movement helps the coot escape from this situation.
- 2- Chain Movement: The algorithm calculates the distance vector between two coots, and one coot moves towards the other halfway. This movement is based on the average position of the two coots.
- 3- Position Adjustment According to Group Leaders: The algorithm simulates the adjustment of coot positions based on group leaders. This is done by considering the average position of the group leaders.
- 4- Leader Movement: To direct the group towards a specific target, the positions of leaders need to be updated. These positions are calculated by seeking a better position around the current best position.

The pseudocode for the Coot Optimization algorithm is depicted in Figure 2.

The COOT algorithm initiates with an initial population, evaluating the fitness of solutions using an objective function after determining each coot's position. Subsequently, leaders are selected, and coots update their positions through different movements. This algorithm attempts to solve optimization problems by combining random movements, chain movements, position adjustments according to group leaders, and leader movements. This method enhances the likelihood of reaching the global optimum. The population is initially created within the specified search area using Equation 11, generating random positions for the coots.

$$CootPos(i) = rand(1, d) * (ub - lb) + lb \quad (11)$$

Here, $CootPos(i)$ denotes the coot's position, d signifies the problem's dimensionality, and lb and ub represent the lower and upper bounds of the search space (Equation 12).

$$\begin{aligned} lb &= [lb_1, lb_2, \dots, lb_d] \\ ub &= [ub_1, ub_2, \dots, ub_d] \end{aligned} \quad (12)$$

Random Movement:

The coot updates its position in different parts of the search space by Equation 13.

$$CootPos(i) = CootPos(i) + A * R2 * (Q - CootPos(i)) \quad (13)$$

Random Movement involves selecting a random position Q (Equation 14) within the search space for the coots to move toward.

$$Q = rand(1, d) * (ub - lb) + lb \quad (14)$$

The value of $R2$ ranges between 0 and 1, while A is computed using Equation 15.

$$A = 1 - L \frac{1}{Iter} \quad (15)$$

Chain Movement:

The Chain Movement method determines a coot's new position based on the average position of two consecutive coots using Equation 16.

$$CootPos(i) = 0.5 (CootPos(i - 1) + CootPos(i)) \quad (16)$$

Position Adjustment According to Group Leaders:

Position Adjustment According to Group Leaders involves selecting a leader and updating a coot's position based on this leader using Equations 17 and 18, respectively.

$$CootPos(i) = LeaderPos(k) + 2 * R1 * \cos(2\pi R) * (LeaderPos(k) - CootPos(i)) \quad (17)$$

$$k = 1 + (i \text{ MOD } NL) \quad (18)$$

Here, NL represents the number of leaders, i signifies the current coot's index, k is the leader's index, and $R1$ ranges between 0 and 1.

Leader Movement:

Leader Movement adjusts the leader's position around the current optimal point using Equation 19.

$$LeaderPos(i) = \begin{cases} B * R3 * \cos(2\pi R) * (gBest - LeaderPos(i)) + gBest, & R4 < 0.5 \\ B * R3 * \cos(2\pi R) * (gBest - LeaderPos(i)) - gBest, & R4 \geq 0.5 \end{cases} \quad (19)$$

Where $gBest$ represents the best-found position, $R3$ and $R4$ are random numbers between 0 and 1, R ranges between -1 and 1, and B is computed using Equation 20.

$$B = 2 - L \frac{1}{Iter} \quad (20)$$

```

1  Set parameters:
   MaxIterations,
   Initial Population,
   Objective Function Create initial population by
   eq.11. Evaluate fitness of initial population,
   Best position in Initial Population,
   Iteration Count = 0
2  while Iteration Count < MaxIterations do
3    for each coot in Population do
4      Random Movement(each coot) by eq.13.
5      Chain Movement(each coot) by eq.16.
6      Pos Adj According to Group Leaders
       (each coot) by eq.17.
7      Leader Movement(each coot) by eq.19.
8      Calculate Fitness of Coot
9      if Fitness of Coot is better than Best Position then
10       Best Position = Coot's Position
11     end if
12   end for
13   Iteration Count += 1
14 end while
15 Solution: Best Position

```

Figure 2. Pseudo Code for Coot Optimization Algorithm

2.1.1. The Proposed COOT Clustering Algorithm (COOTC)

The suggested COOTC algorithm is designed to cluster the dataset using the specified number of clusters. This algorithm utilizes a population of candidate solution vectors containing cluster centroids and adopts the update principles of the COOT algorithm based on fitness values, aiming to determine the most suitable cluster centroids. It represents an unsupervised clustering method.

The pseudocode for the proposed metaheuristic clustering algorithm is depicted in Figure 3.

1	Start $k = \text{number of clusters, } m = \text{total number of data points, } V_i = i\text{-th data vector,}$ $d = \text{dimensionality of data vector, } M_k = k\text{-th cluster centroid, } X = \text{population}$ $D_{ik}^2 = \text{square of the Euclidean distance between } i\text{-th data and } k\text{-th cluster centroid}$ $M(i,j)^{1..d} = j\text{-th candidate cluster centroid in the } d\text{-dimensional } i\text{-th solution vector,}$ $X_i \text{ is the } i\text{-th candidate solution vector.}$
2	Load the d-dimensional dataset to be clustered.
3	Determine the appropriate number of clusters, set COOT parameters, and enter the maximum iteration count.
4	Create an initial random COOT population (P) containing starting candidate solutions. The population consists of solution vectors containing cluster centroids, where each individual represents all cluster centroids.
5	Assign each data point to the closest cluster based on the distance between data points and cluster centroids. (Use the square of the Euclidean distance $D^2(V_m, M_k)$)
6	Calculate the fitness value of each candidate solution vector (each individual) for SSE (Sum of Squared Errors). (The fitness value of each solution vector measures the ability of cluster centroids to represent data points, evaluating how well an individual's clustering solution performs.)
	$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} M(1,1)^{1..d} & M(1,2)^{1..d} & \dots & M(1,k)^{1..d} \\ M(2,1)^{1..d} & M(2,2)^{1..d} & \dots & M(2,k)^{1..d} \\ \vdots & \vdots & \ddots & \vdots \\ M(N,1)^{1..d} & M(N,2)^{1..d} & \dots & M(N,k)^{1..d} \end{bmatrix}$
7	While (loop until termination condition is met, aiming for SSE minimization):
8	Select the optimal candidate solution vector (X_{best}).
9	Explore and exploit new candidate solutions based on COOT using X_{best} .
10	Distribute the data to the new candidate solution clusters.
11	Calculate the SSE fitness value for the new candidate solution vectors (individuals).
12	if (the SSE value of the new candidate solution vector is smaller than the previous)
13	Update the candidate solution vector in the population.
14	end(if)
15	end(while)
16	Assign all data points in the dataset to the k optimal cluster centers and display the clusters.
17	Finish.

Figure 3. Pseudo Code for CootC Clustering Algorithm**Table 1.** Comparison of Clustering Performances between COOTC and KM Methods

Datasets	Parameters	Algorithm	SI Index	DB Index	CH Index	KL Index	SSE Index
(SV-1)	Population:10 Iteration:50000	COOTC	0.8268	0.4041	990.1727	30.2769	7.5692
	Feature Count:2 Cluster Count=4	KM	0.8268	0.4590	990.1727	30.2769	7.5692
(SV-2)	Population:10 Iteration:50000	COOTC	0.8120	0.4382	2049.9039	46.5920	9.3184
	Feature Count:2 Cluster Count=5	KM	0.6783	0.6017	1252.1985	73.6037	14.7207
Iris	Population:30 Iteration:50000	COOTC	0.7357	0.5873	561.6278	13657.4703	7885.1441
	Feature Count:4 Cluster Count=3	KM	0.7344	0.5901	561.5937	13658.2020	7885.5666

3. RESULTS

In Table 1, the performances of COOTC and KM algorithms are compared by testing on different datasets. This comparison was performed using the IRIS dataset obtained from the UCI Repository and synthetically generated SV-1 and SV-2 datasets. The performance of COOTC and KM algorithms has been evaluated using different metrics, and the results are provided in Table 1. In the synthetic dataset SV-1 with distinct separable clusters, both the KM and COOTC algorithms demonstrate similar performances, effectively clustering the data. This observation is further supported by the evaluation metrics provided in Table 1.

Although with a slight difference, it's notable that in terms of the Davies-Bouldin metric, COOTC exhibits a lower value compared to KM. This suggests a potentially superior performance of COOTC concerning this evaluation criterion. Figure 4 depicts the data distribution and the cluster centers determined by the methods.

In the complexly distributed synthetic dataset SV-2, consisting of 5 clusters, KM's performance appears notably low as seen in Table 1. Evaluative metrics position COOTC as the most successful method. KM faces challenges in discerning data within clusters that exhibit low separability, potentially resulting in different cluster centers in each attempt. This aspect signifies a notable weakness in the method's performance compared to COOTC.

Figure 4 illustrates COOTC accurately determining the optimal cluster centers, while KM tends to represent a larger data group with a single cluster and a smaller data group with 2 clusters.

In Table 1, the COOTC method's performance metrics, derived from the Iris dataset, are compared to those of KM. Considering these metrics concerning the multidimensional

Iris dataset, success is demonstrated by the proposed COOTC method in terms of SI, DB, and SSE metrics in comparison to KM.

Data distributions and cluster centers identified by the COOTC and KM methods across different combinations of the four dimensions (sepal length, sepal width, petal length, and petal width) are visualized in Figure 5 and concerning the determination of cluster centers, it is observed that COOTC is performed at least as effectively as the KM method.

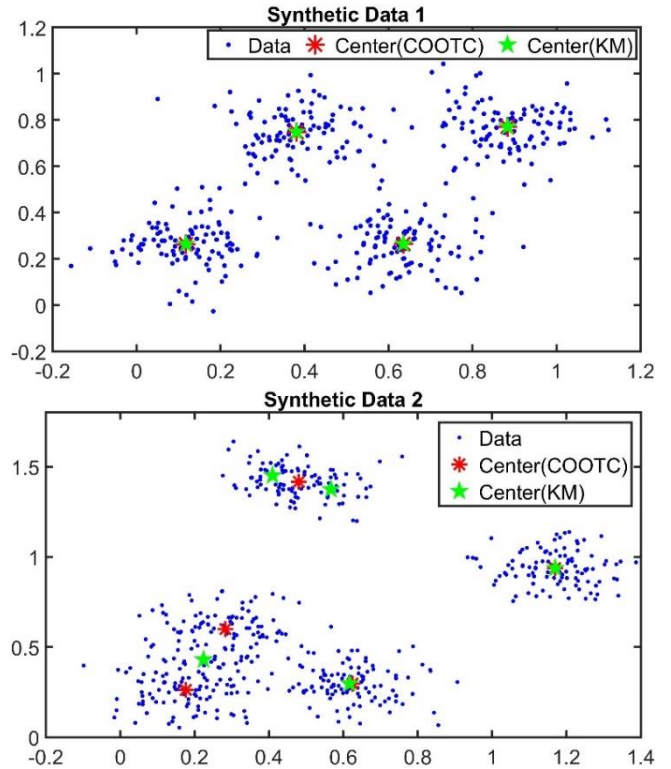


Figure 4. Clustering Synthetic Data 1 and Synthetic Data 2

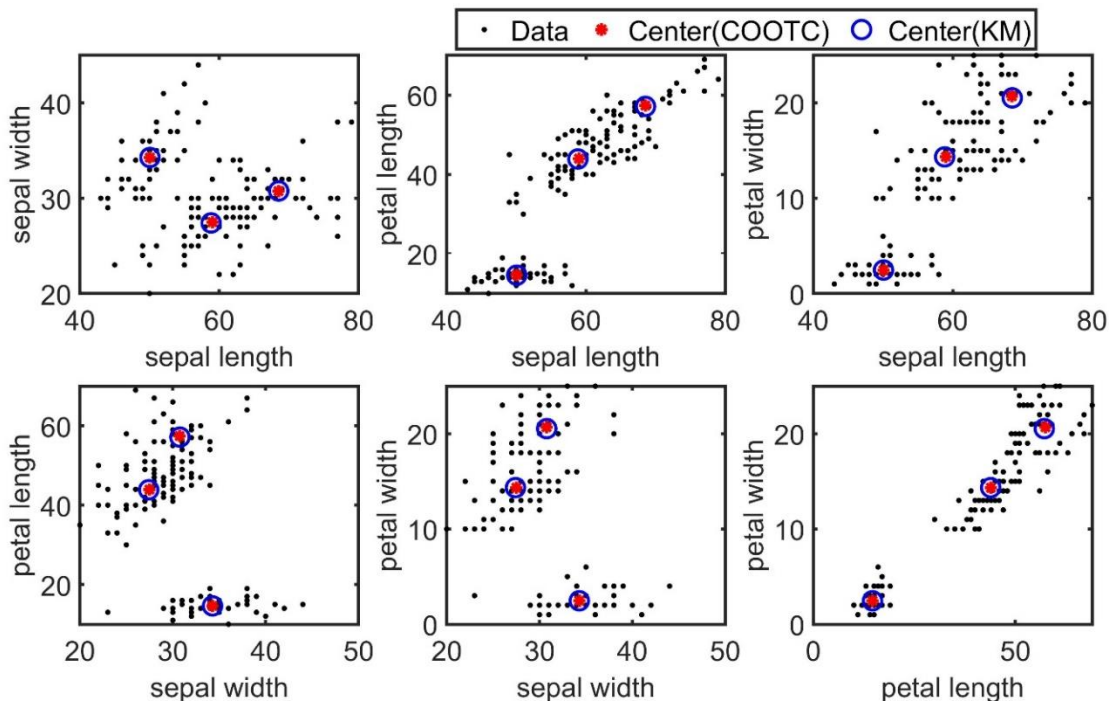


Figure 5. Clustering for Iris Data

4. CONCLUSION

This study introduces a novel clustering approach utilizing the metaheuristic Coot Optimization algorithm for the organization and analysis of numerical data. The experimental results provide compelling evidence that the proposed COOTC method exhibits comparable or superior performance compared to the traditional K-Means (KM) algorithm across various datasets, including the Iris plant science dataset and synthetic datasets. Moreover, while the COOTC method demonstrates similar performance to KM when applied to low-dimensional and well-separated datasets, it notably outperforms KM in the context of the low-complexity Iris dataset. This dataset, characterized by its high dimensionality yet homogeneous and distinct structure, poses a significant challenge for traditional clustering algorithms. However, the COOTC method effectively addresses this challenge, showcasing enhanced performance and providing promising insights for future research endeavors in similar domains.

The findings of this study suggest that the COOTC method represents a promising avenue for further exploration and development in the field of clustering algorithms. Future research efforts may focus on conducting comparative analyses with additional metaheuristic clustering algorithms, exploring hybridization strategies with various clustering techniques, and evaluating the performance of the COOTC method in real-world applications across diverse domains and datasets. Overall, this study contributes to the advancement of clustering methodologies by introducing a novel approach that demonstrates efficacy and potential for further refinement and application in practical settings.

Author contributions: All authors have contributed equally to the work.

Conflict of Interest: The authors declared that there is no conflict of interest.

Financial Disclosure: All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

REFERENCES

- [1] R. Dash and R. Dash, "Comparative Analysis of K-Means and Genetic Algorithm Based Data Clustering," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 2, pp. 257-265, 2012.
- [2] V. Kumar, J. K. Chhabra, and D. Kumar, "Grey Wolf Algorithm-Based Clustering Technique," *Journal of Intelligent Systems*, vol. 26, no. 1, pp. 153-168, 2017.
- [3] L. Abualigah, D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. Al-Qaness, and A. H. Gandomi, "Aquila Optimizer: A Novel Meta-heuristic Optimization Algorithm," *Computers & Industrial Engineering*, vol. 157, p. 107250, 2021.
- [4] S. Çınaroğlu and H. Bulut, "K-Means and Particle Swarm Optimization-Based Novel Initialization Approaches for Clustering Algorithms," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 33, no. 2, pp. 413-424, 2018.
- [5] Ö. Demirkale and Ç. Özarı, "Evaluation of Fundamental Macroeconomic and Financial Indicators with K-Means Clustering Method: The Case of Fragile Five Countries," *Finans Ekonomi ve Sosyal Araştırmalar Dergisi*, vol. 5, no. 1, pp. 22-32, 2020. A.
- [6] Sancho, J. C. Ribeiro, M. S. Reis, and F. G. Martins, "Cluster Analysis of Crude Oils with K-Means Based On Their Physicochemical Properties," *Computers & Chemical Engineering*, vol. 157, p. 107633, 2022.
- [7] O. S. Faragallah, H. M. El-Hoseny, and H. S. El-Sayed, "Efficient Brain Tumor Segmentation using OTSU and K-Means Clustering in Homomorphic Transform," *Biomedical Signal Processing and Control*, vol. 84, p. 104712, 2023.
- [8] E. A. Pambudi, A. Y. Badharudin, and A. P. Wicaksono, "Enhanced K-Means By Using Grey Wolf Optimizer for Brain MRI Segmentation," *ICTACT Journal on Soft Computing*, vol. 11, no. 3, 2021.
- [9] L. Khriissi, N. El Akkad, H. Satori, and K. Satori, "Clustering Method and Sine Cosine Algorithm for Image Segmentation," *Evolutionary Intelligence*, pp. 1-14, 2022.
- [10] H. Arslan and M. Toz, "Hybrid FCM-WOA Data Clustering Algorithm," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1-4, IEEE, 2018.
- [11] R. R. Mostafa, A. G. Hussien, M. A. Khan, S. Kadry and F. A. Hashim, "Enhanced COOT optimization algorithm for Dimensionality Reduction," *2022 Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, Riyadh, Saudi Arabia, pp. 43-48, 2022.
- [12] A. Ozden., I. İseri, "COOT optimization algorithm on training artificial neural networks," *Knowledge and Information Systems*, vol. 65, pp.3353–3383, 2023.
- [13] D.S. Irene, M. Lakshmi, A.M.J. Kinol, et al. "Improved deep convolutional neural network-based COOT optimization for multimodal disease risk prediction.", *Neural Comput & Applic*, vol. 35, pp.1849–1862, 2023.
- [14] M. Aslan, İ. Koç, "Modified Coot bird optimization algorithm for solving community detection problem in social networks.", *Neural Comput & Applic*, vol. 36, pp.5595–5619, 2024.
- [15] X. You, G. Yan, M. Thwin, "Applying modified coot optimization algorithm with artificial neural network meta-model for building energy performance optimization: A case study", *Heliyon*, vol.9, no 6, p.e16593, 2023.

- [16] I. Koc, "A fast community detection algorithm based on coot bird metaheuristic optimizer in social networks", *Engineering Applications of Artificial Intelligence*, vol. 14, p. 105202, 2022.
- [17] E. Pashaei, E. Pashaei, "Hybrid binary COOT algorithm with simulated annealing for feature selection in high-dimensional microarray data", *Neural Comput & Applic*, vol. 35, pp.353–374, 2023.
- [18] D. Jabbar Luaibi, "Precise Classification of Brain Magnetic Resonance Imaging (MRIs) using COOT optimization.", *Texas Journal of Engineering and Technology*, vol.26, pp.57–71, 2023.
- [19] D.S. Irene, M. Lakshmi, A.M.J. Kinol, et al. "Improved deep convolutional neural network-based COOT optimization for multimodal disease risk prediction.", *Neural Comput & Applic* vol.35, pp.1849–1862, 2023.
- [20] R. A. Fisher, "Iris," UCI Machine Learning Repository, 1988. <https://doi.org/10.24432/C56C76>.
- [21] Y. Liu, Z. Li, H. Xiong, X. Gao, ve J. Wu, "Understanding of Internal Clustering Validation Measures," in 2010 IEEE International Conference On Data Mining, pp. 911-916, December 2010.
- [22] W.J. Krzanowski ve Y.T. Lai, "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering," *Biometrics*, vol. 44, s. 23, 1988.
- [23] A. A. R. Fernandes, F. U. Solimun, A. Aryandani, A. Chairunissa, A. Alifa, E. Krisnawati, ..., F. L. N. Rasyidah¹², "Comparison of Cluster Validity Index using Integrated Cluster Analysis with Structural Equation Modeling the War-Pls Approach," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 18, 2021.