International Journal of Advanced Technology and Engineering Exploration, Vol 11(117)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10102298

Research Article

Enhancing clustering performance: an analysis of the clustering based on arithmetic optimization algorithm

Hakam Singh* and Ashutosh Kumar Dubey
Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India

Received: 08-September-2023; Revised: 12-August-2024; Accepted: 14-August-2024

Abstract
This study explored the clustering based on arithmetic optimization algorithm (CAOA) and its potential for addressing challenging clustering problems. CAOA is based on the arithmetic optimization algorithm (AOA), which utilizes arithmetic operators, including Addition, Subtraction, Multiplication, and Division, to optimize solutions. The performance of CAOA was investigated by applying it to diverse real-life datasets and meticulously analysing its clustering performance. Two primary evaluation metrics, namely the average distance among cluster members (intra-cluster distance) and the F-measure, were employed to gauge the clustering quality. Statistical validation was conducted using the Friedman test, ensuring robust and significant results. The results revealed substantial insights into CAOA's performance. In terms of average intra-cluster distance, CAOA consistently recorded the lowest values among all tested clustering algorithms. This outcome indicated CAOA's ability to form tightly packed, well-defined clusters, enhancing its suitability for applications like pattern recognition and data segmentation. Regarding F-measure, CAOA delivered competitive clustering quality. Notably, it achieved among the highest F-measure values, especially in datasets like "Cancer" and "LR," signifying its potential for accurate cluster identification, crucial in domains such as medical diagnosis and customer segmentation. This study indicated the effectiveness of CAOA in addressing real-world clustering challenges. The findings emphasized CAOA's consistent superiority over other algorithms in minimizing the average intra-cluster distance while also demonstrating competitive clustering quality as measured by the F-measure. Statistical validation through the Friedman test confirmed the distinctiveness of CAOA's performance.

Keywords
CAOA, AOA, F-measure, Average Intra-cluster Distance, Friedman test.

1.Introduction
The technological revolution, primarily driven by digitization, has led to a surge in the world of data. Data, in its raw form, consists of concealed patterns or valuable information that plays a significant role in both business and real-world environments [1–4]. Various online and offline techniques are employed for pattern extraction from extensive databases. This process of pattern extraction is commonly referred to as data mining or knowledge discovery [1]. Data mining primarily involves three tasks: data preprocessing, pattern extraction, and data visualization [2, 3].

In the preprocessing phase, essential information is processed for normalization, interpretation, distribution, and transformation.

The subsequent pattern extraction process employs diverse strategies, including prediction, rule mining, classification, and clustering [2, 3]. The results are then presented in tabular or graphical formats and validated using other techniques. Furthermore, data mining is categorized into descriptive and predictive analysis. Predictive analysis predicts values based on previously known facts or data, often referred to as supervised learning [3, 4]. Descriptive analysis identifies hidden patterns using dissimilarity measures and is known as unsupervised learning [2–6]. Clustering belongs to the category of unsupervised learning

Clustering techniques group data objects or instances into distinct clusters or groups. Data within the same cluster exhibit similarity and introduce diversity when compared to data in other clusters [7–10]. Several evolutionary meta-heuristic algorithms are applied to obtain optimal solutions in clustering [11–15]. Meta-

*Author for correspondence

heuristic techniques find broad utilization in addressing a diverse range of optimization issues. These algorithms incorporate multiple heuristic features and specific traits, such as proximity to the optimal solution and computational efficiency. They are inspired by natural phenomena and consist of various operators to achieve optimal solutions.

A wide range of metaheuristic algorithms, including particle swarm optimization (PSO), cat swarm optimization (CSO), teaching-learning-based optimization (TLBO), biogeography-based optimization (BBBC), ant colony optimization (ACO), bat algorithm (BH), artificial bee colony (ABC), and whale optimization algorithm (WOA), has been investigated and employed in various applications [14–21].

It is important to note that many clustering algorithms are either newly adopted or improved and hybridized. In any clustering technique, the choice of distance measure or approximation function is crucial for grouping the data entities within the system. Several distance measures are employed in clustering, with the Euclidean distance being the most used distance measure in the field, as shown in Equation 1.

$$D(Z_i, C_j) = \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{d} (Z_{ik}, C_{jk})^2} \qquad (1)$$

Where
Z_i = Object/data instance (i^{th})
C_j = Cluster center (j^{th} centroid)
n = Data points/instances (number)
d=dimension/features/attributes of the dataset (number)

In this study, arithmetic optimization algorithm (AOA) based approach has been implemented and developed for the clustering domain. This pioneering approach harnesses the inherent mathematical distribution behaviors of arithmetic operators. The driving force behind this algorithm lies in the profound significance of arithmetic operators, which serve as foundational components in number theory and are widely employed across various disciplines to identify sets of solutions, whether alternative or optimized.

The primary contributions of this study can be encapsulated as follows:

1. Implementation of the AOA in the clustering domain: The AOA was successfully applied to the field of clustering, suggesting a new framework called clustering based on arithmetic optimization algorithm (CAOA).
2. Mitigation of clustering challenges: The CAOA algorithm was designed to effectively address common clustering challenges, including but not limited to local optima and convergence rate issues. It offers a promising avenue for enhancing the performance and robustness of clustering techniques.

The structure of this paper is as follows: In Section 2, a comprehensive review of related work in the field has been conducted. Section 3 is dedicated to elucidating the intricacies of the CAOA algorithm. In Section 4, an extensive account of the simulation procedures and the resultant findings is elaborated. Finally, Section 5 encapsulates the discussion of the work, and Section 6 concluded it.

2. Literature review

An extensive survey was conducted, and it was divided into three subsections, namely: Foremost meta-heuristic clustering algorithms: In this section, leading meta-heuristic clustering algorithms were considered. Automated clustering algorithms: This section covers algorithms that incorporate automatic cluster number and population selection properties. Improved/hybridized clustering algorithms: These are advanced versions or variants of traditional clustering algorithms.

Foremost meta-heuristic clustering algorithms

Meta-heuristic algorithms draw inspiration from natural, biological, and physical principles, incorporating various methods and mechanisms to find optimal solutions. These algorithms are iterative and excel in efficiently tackling complex problems. Bezdek et al. [22] introduced a biologically inspired method for partitional clustering called a genetic algorithm (GA). This algorithm utilizes crossover and mutation operators to optimize partitional clustering problems.

In [23], a novel algorithm based on the behavior of ACO was introduced. ACO utilizes a constructive greedy heuristic, distributed computation, pheromone matrix, and optimistic feedback. Its efficiency is evaluated in terms of CPU utilization time and compared to GA, simulated annealing (SA) and tabu search (TS), demonstrating ACO's effectiveness as a clustering method. In [24], a TS-based algorithm was detailed to address clustering issues. It has been introduced for minimizing sum-of-squares clustering. It employs five improvement operations and three neighbourhood modes. It outperforms existing techniques on artificial and real datasets, showcasing its effectiveness. Mahdavi et al. [25] designed a robust method for partitional clustering based on musical harmony. This algorithm generates a new vector solution by computing other search space vectors. Additionally, it hybridizes the harmony search (HS) with the k-means algorithm to enhance convergence speed.

Santosa and Ningrum [26] implemented the CSO to optimize clustering problems, which comprises two modes: tracing and seeking. The tracing mode replicates the cats' resting behavior, serving as the basis for local search, whereas the seeking mode emulates their hunting instincts, facilitating global search. In [27], a novel ABC algorithm is introduced to obtain optimal solutions in the clustering domain. The Deb rules guide the search in the most favourable direction. Singh and Srivastava [28] presented an approach combining kernel fuzzy c-means clustering with TLBO. It improves clustering quality and compactness, outperforming GA and PSO on five datasets. Additionally, it surpasses TLBO with fuzzy c-means in clustering performance. In [29, 30] a novel clustering algorithm was introduced inspired by micro-bats' behavior. Their approach addressed issues like slow convergence, local optima, and search mechanism trade-offs, delivering significant results across various datasets. Niknam and Amiri [31] address k-means clustering's sensitivity to initial conditions by proposing a hybrid evolutionary algorithm, fuzzy adaptive PSO (FAPSO)-ACO-k-means (K), combining FAPSO, ACO, and k-means. Aggarwal et al. [32] proposed enhancing traditional k-means clustering by integrating nature-inspired optimization methods (Cuckoo, Bat, Krill Herd algorithms) with k-means++. Evaluation experiments validate their efficacy.

Automated clustering algorithms

Several issues in clustering techniques were addressed, including manual cluster number assignment, inadequate population initialization/selection, and slow convergence. Hartigan and Wong [33] developed a k-means variant that positions data around the sample's mean, effectively resolving initialization problems. Another approach combined traditional k-means with the HS algorithm, creating a novel initialization method called k-means-HS, outperforming k-means. K'-means [34] disperses cluster numbers initially, uniting them later using a minimal cost function. In [35], a novel initialization strategy combining neighborhood rough set theory with k-means improved the clustering results. Geng et al. [36] enhanced the k-means clustering algorithm by introducing ambiguity as a constraint, proposing a new membership equation, and optimizing with Gaussian distribution and fuzzy entropy. Results show improved clustering accuracy and efficiency. Tang et al. [37] introduced the rough set-based semi-supervised k-means (RSKmeans) algorithm to address high-dimensional sparse data. It calculates non-zero value proportions from labeled data, selects crucial attributes per cluster, and employs the approximation set and information gain to partition attribute values, iteratively updating clustering centers. Experimental results on text data demonstrate RSKmeans' effectiveness in attribute selection and performance improvement. Liu et al. [38] addressed the importance of chatter detection during metal cutting, highlighting limitations of current methods such as human interference, data labeling, and time consumption. They proposed an unsupervised chatter detection method using unlabeled dynamic signals. This method, independent of processing parameters and labeling, employs auto-encoders to compress signals into two dimensions. Ning et al. [39] introduced an enhanced version of clustering approach which is based on k-means. It is based on validation index(internal) and weighted distance approach. The weighted distance effectively captures global spatial correlations and local variable trends in high-dimensional data. It was observed that WeDIV demonstrated superior performance in both cluster number specification and overall clustering quality. Cheng et al. [40] discussed the impact of initial centers on k-means clustering and its limitation in identifying arbitrary-shaped clusters. They propose natural density peaks (NDP)-kmeans, using NDPs to represent local data and compute dissimilarity based on neighbor-based distance. NDP-kmeans effectively identifies arbitrary-shaped clusters, outperforming other algorithms in recognizing both spherical and manifold clusters.

Improved/hybridized clustering algorithms

This section focuses on recent advancements in partitional clustering techniques. The particle swarm optimization k-harmonic means (PSOKHM) algorithm, introduced by Yang et al. [41], presents a novel hybrid approach. It harnesses both PSO and k-harmonic-means (KHM) features, offering an integrated solution to optimize cluster formation, mitigate local optima, and enhance convergence speed. Sixu et al. [42] addressed the need for energy-efficient routing in wireless sensor networks, particularly in the context of the internet of things (IoT). They introduced a cluster routing protocol utilizing PSO for clustering and the ABC algorithm for mobile base station path planning. Employing software-defined network architecture results in energy-efficient sensor nodes and improved network lifetime, reducing control overhead. In [43], a hybrid algorithm combining GA and ABC methods is presented, using crossover operators to enhance bee information exchange. Huang et al. [44] hybridize ACO and PSO, introducing four search methods to achieve optimal clustering. A two-stage clustering approach was employed in [45], utilizing heuristic search and PSO. The clusters are initially generated with PSO, and the optimal cluster is selected using the heuristic search algorithm. Singh and Kumar [46] employed clustering with a cat-inspired meta-heuristic algorithm to optimize clustering problems. They enhanced the CSO algorithm, achieving optimal results when compared to other clustering algorithms on eight real-life datasets.

The literature provides insights into various clustering techniques and hybrid approaches to address clustering challenges. Meta-heuristic algorithms, inspired by nature, have been successfully applied to clustering problems. GA, ACO, TS, HS, and ABC are among these algorithms. They have shown effectiveness in optimizing cluster formation, enhancing convergence, and improving clustering results. Automated clustering algorithms aim to address issues such as manual cluster number assignment and slow convergence. Variants of k-means, including k-means++, k-means-HS, and k'-means, have been introduced to address initialization problems and improve clustering quality. Novel initialization strategies, such as neighborhood rough set theory-based initialization, have also contributed to better clustering results. Moreover, approaches incorporating ambiguity constraints, improved solution search equations, and neighborhood-based search mechanisms have enhanced clustering algorithms. Hybrid algorithms like PSOKHM combine different optimization methods to achieve better clustering results. The use of NDPs to represent local data and compute dissimilarity has led to the development of NDP-kmeans, which effectively identifies arbitrary-shaped clusters. Additionally, researchers have addressed clustering challenges in wireless sensor networks by proposing energy-efficient routing protocols with mobile base stations, leveraging optimization algorithms like PSO and ABC. Overall, these advances in clustering techniques offer improved solutions for various clustering problems. However, there is a need for further research to explore the applicability of these techniques in different scenarios, scalability to large datasets, and adaptability to evolving data types. Additionally, evaluating the robustness and efficiency of these algorithms across diverse domains and datasets remains a crucial area for research.

3. Methods

This section presents detailed background information about the AOA.

3.1Arithmetic optimization algorithm (AOA)

The AOA is a mathematical model that utilizes the characteristics of arithmetic operators (Multiplication (M "×"), Division (D "÷"), Subtraction (S "–"), and Addition (A "+")) to determine the optimal solution [47]. It has been developed by Abualigah et al. [47]. The execution of operations occurs in three distinct phases.

a) **Initialization phase** begins with the random selection of a population, followed by the exploration and exploitation processes. Equation 2 shows the iteration steps.

$$Z = \begin{bmatrix} Z_{1,1} \ldots\ldots Z_{1,j} \ldots . Z_{1,n-1} & Z_{1,n} \\ Z_{2,1} \ldots\ldots Z_{2,j} \ldots . Z_{2,n-1} & Z_{2,n} \\ \ldots \quad \ldots \quad \ldots & \ldots \\ Z_{N,1} \ldots\ldots Z_{N,j} \ldots . Z_{N,n-1} & Z_{N,n} \end{bmatrix}$$

$$\text{MOA\_Current\_Iteration} = \text{Minimum} + \text{Current\_Iteration} \times \frac{(Max - Min)}{\text{M}_{\text{Iteration}}}) \qquad (2)$$

The MOA is "Math optimizer accelarated", "M_Iteration" represents the maximum number of iterations. The terms "Min" and "Max" refer to the minimum and maximum values.

b) **Exploration phase** explore a broader range of potential solutions and prevent being trapped in local optima. By employing mathematical calculations that involve either the (D "÷") or (M "×") operators, produces highly distributed values. This characteristic facilitates the exploration search mechanism, allowing for more efficient discovery of optimal solutions (Equation 3).

$$Zi,j(Current\_Iteration + 1) = \begin{cases} best(Z_j) \div (MOP + \epsilon) \times \\ \left((UB_j - LB_j) \times \mu + LB_j\right), r2 < 0.5 \\ best(Z_j) \times MOP \times \\ \left((UB_j - LBj) \times \mu + LB_j\right), \quad otherwise \end{cases} \qquad (3)$$

In the context provided, $Z_{i,j}$(Current_Iteration + 1) represents the i_{th} solution in the subsequent iteration, while Zi,j(Current_Iteration) signifies current iteration (Equation 4). The term best (Zj) refers is best solution attained up to that point. The "ϵ" is an integer value, while UB_j (upprer bound) and LBj (lower bound) values. Furthermore, μ functions as a control parameter, influencing the modulation of the search process.

$$\text{MOP\_Current\_Iteration} = 1 - (\text{Current\_Iteration})^{\frac{1}{\alpha}} / (\text{Maximum\_Iteration})^{\frac{1}{\alpha}} \quad (4)$$

The MOP stands for "Math Optimizer Probability," while α is a sensitive parameter that varies depending on the experiment.

c) Exploitation phase

The exploitation strategy plays a significant role in numerous optimization algorithms, contributing to refining and enhancing existing solutions. In the context of the AOA algorithm, utilizing the (S "–"), and Addition (A "+") operators during this phase aids in facilitating the exploitation process. The exploitation phase of the MOA is conditioned on function value, "R_1" does not exceed the MOA(Current_Iteration).

$$Z_{i,j}(Current\_Iteration + 1) = \{(best(x_j) - MOP \times ((UB_j - LB_j) \times \mu + LB_j),\ R_3 < 0.5 best(Z_j) + MOP \times ((UB_j - LB_j) \times \mu + LB_j), \quad (5)$$

The rest of the tasks are repeated as in the previous phase. However, the operators (S "–") and (A "+") are specifically designed to minimize the chances of getting trapped in local search areas.

Clustering based on arithmetic optimization algorithm (CAOA)

This section provides a detailed walkthrough of applying the AOA to clustering, offering insights into the parameters used, convergence criteria, and adjustments tailored to the specific clustering task. The AOA is carefully designed to address clustering problems, focusing on the fundamental objective of grouping data points into distinct clusters [47].

The algorithm initiates by randomly selecting initial positions for optimal solutions, effectively setting the stage for potential cluster configurations. Following this initialization phase, it proceeds to calculate objectives, primarily aimed at assessing how well the existing clusters represent the underlying data patterns. To accomplish this, the AOA employs a series of arithmetic operations, assignments, and updates. These operations, denoted by operators such as (M "×"), (D "÷"), (S "−"), and (A "+"), play a pivotal role in refining the initial solutions, steadily enhancing their clustering performance. The unique feature of AOA is its capability to estimate feasible positions for achieving near-optimal solutions through these operators [47].

This distinctive approach enables the AOA to explore a wide range of cluster configurations and adapt its solutions in accordance with the specific requirements of the clustering problem under consideration. As the iterative process unfolds, the algorithm consistently updates the status of each solution, incorporating improvements based on the best solution discovered thus far. The iterative process persists until the algorithm fulfils a predefined criterion, serving as an indicator of a successful clustering solution.

In this section, a process for optimizing cluster centers using a metaheuristic optimization technique called AOA was introduced. This process is specifically referred to as CAOA. It starts by loading the dataset and specifying various user-defined parameters such as the number of clusters and maximum iterations. The algorithm then selects initial cluster centers through random sampling.

The optimization process iterates until the maximum number of iterations is reached. In each iteration, it evaluates the objective function and computes a fitness function to identify the best solution. It adjusts the parameters using MOA_Current_Iteration and MOP_Current_Iteration.

The core of the algorithm consists of a nested loop that iterates through each instance (row) and attribute (column) in the dataset. For each data point, it generates three random values that determine whether the algorithm enters the exploration or exploitation phase. In the exploration phase, positions are either divided or multiplied based on the random values. In the exploitation phase, positions are adjusted by adding or subtracting values. The behavior of this phase is governed by Equations 3 and 5.

The CAOA algorithm follows an iterative process in which the counter iteration is incremented after processing all data points. This process continues for a specified number of iterations to determine the optimal cluster centers. The algorithm utilizes a metaheuristic approach that combines exploration and exploitation phases to identify the best cluster centers based on fitness function evaluation. Its design enables the adaptation and optimization of cluster centers for complex datasets. The detailed algorithm provides a comprehensive overview of the approach, and *Figure*

1 displays the complete CAOA algorithm flowchart.

To employ the CAOA clustering algorithm, a dataset and user-defined parameters are inputted, enabling the discovery of optimal cluster centers through iterative adjustments. The primary goal is to adapt and optimize cluster centers to best suit the dataset, ensuring effective clustering in diverse and intricate data scenarios. The algorithm systematically refines cluster centers as it progresses, aiming to enhance its clustering performance. In essence, the CAOA algorithm is an iterative process tailored for uncovering the optimal cluster centers in complex datasets. It effectively combines exploration and exploitation phases within a metaheuristic framework to determine the most suitable cluster centers based on fitness function evaluation. The process starts with initializing the counter iteration, which is then incremented after processing each data point, continuing for a specified number of iterations.

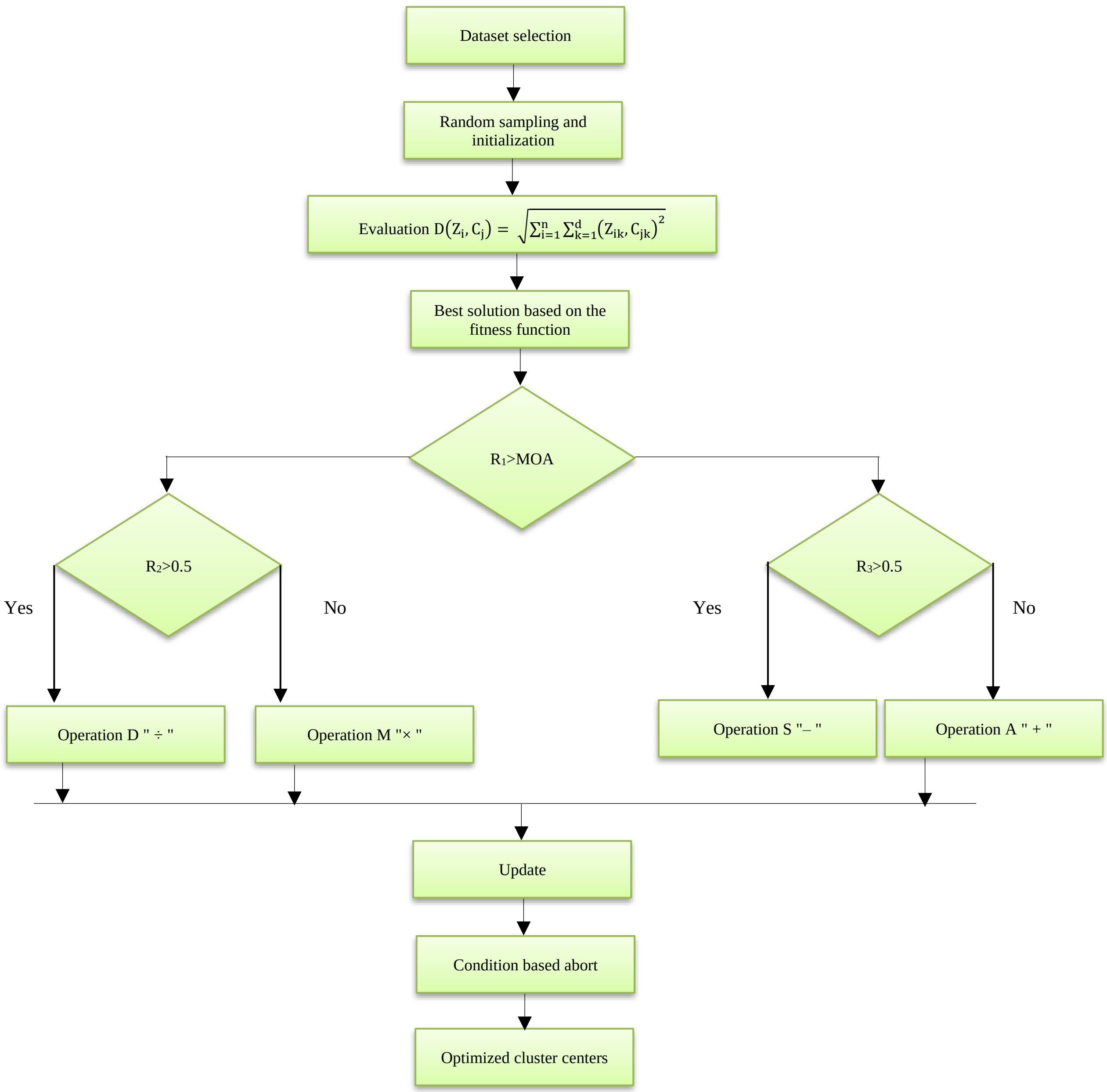

Figure 1 Flowchart of CAOA approach
R_1: Random_Value1, R2: Random_Value2, R3: Random_Value3, Math Optimizer Accelarated (MOA)

Algorithm: Clustering based on arithmetic optimization algorithm (CAOA)
Input: Dataset and user-defined values (parametric values)
Output: Optimal cluster centers
1. Load the dataset into memory and set the initial parametric values, including the number of clusters (K_i∈(i=1,2,...,n)), and the maximum iterations, etc.
2. Select initial cluster centers (K_i) through random sampling.
3. While (Current Iteration < Maximum number of iterations) do
   a. Evaluate the objective function values using Equation 1.
   b. Compute the fitness function.
   c. Identify the best solution.
   d. Adjust the MOA using Equation 2.
   e. Modify the MOP using Equation 4.
4. loop for i=1 to n, do
   a. loop for j= 1 to d, do
      i. Generate three random values (Random_Value1, Random_Value2, and Random_Value3) from a uniform distribution between 0 and 1.
      ii. If (Random_Value1 > MOA) then
         - "Exploration phase"
         iii. If (Random_Value2 > 0.5) then perform division (÷) operation
            Update the positions
            $Zi,j(Current\_Iteration + 1)$ [use Equation 3]
         iv. Else
            perform multiplication (×) operation
            Update the positions
            $Zi,j(Current\_Iteration + 1)$ [use Equation 3]
      v. End if; (inner if completed)
      vi. Else
         - "Exploitation phase"
         vii. If (Random_Value3 > 0.5) then perform subtraction (–) operation
            Update the positions
            $Zi,j(Current\_Iteration + 1)$ [use Equation 5]
         viii. Else
            Perform addition (+) operation
            Update the positions
            $Zi,j(Current\_Iteration + 1)$ [use Equation 5]
         ix. End if; (inner if completed)
      x. End if; (outer if completed)
   b. End for;
5. Increment the counter value: C_Iter = C_Iter + 1
6. Optimal Solution

Where
n = Data points/instances (number)
d = dimension/features/attributes of the dataset (number)

4. Results

The implementation was executed on a system configuration consisting of an Intel Core i5 processor, 8GB of RAM, and the Windows 10 operating system. The parametric values used in the CAOA algorithm and other algorithms for the experimental evaluation are listed in *Table 1*.

Table 1 The parametric values used in the CAOA algorithm and other algorithms for the experimental evaluation

CAOA		**ACO**		**PSO**		**BA**	
Population	"K × d"	Number of ants	50	Number of swarms	"10 × K × d"	Population	"K × d"
Random value	[0,1]	Threshold Probability	1	c 1 = c 2	2	A_0	0.9
		Searching probability	0	ω min	0.5	R	0.1
		Evaporation rate	0	ω max	1	α = γ	0.9
CSO		**GA**		**K-means**			
Population	"K × d"	Population	"K × d"	Population	"K × d"		
SMP	10	Crossover rate	0.8				
MR	0.5	Mutation rate	0.001				
C	2						
Maximum number of iterations = 200							

The CAOA algorithm's performance was assessed using eight real-life datasets, as displayed in *Table 2*. These datasets were sourced from the University of California, Irvine (UCI) Machine Learning Repository and were employed to gauge the performance and efficiency of the CAOA algorithm in comparison to various other clustering algorithms during the experimental evaluation. The number of instances (N), features (D), and the number of clusters (K) is depicted in *Table 2*.

Table 2 Information of the datasets used in this work

Datasets	N	D	K
Iris	150	4	3
Cancer	683	9	2
CMC	1,473	9	3
Wine	178	13	3
Glass	214	9	6
Statlog	58,000	9	7
LR	20,000	16	26
ISOLET	7797	617	26

The assessment of CAOA algorithm performance relies on two critical parameters: intra-cluster distance and F-measure. For each dataset, thirty runs, each consisting of 200 iterations, were conducted. *Table 3* presents a comparative analysis using intra-cluster distance. Results indicate that the CAOA algorithm generally achieves the lowest intra-cluster distance, signifying superior clustering quality compared to other algorithms. However, for the Glass datasets, K-means and CAOA algorithms yield similar values. Additionally, *Table 4* showcases the evaluation of CAOA algorithm efficiency through the F-measure parameter.

The lower the value, the better the clustering. CAOA consistently outperforms the other algorithms in terms of intra-cluster distance, indicating it achieves more compact clusters. Notably, in datasets like "Iris," "Cancer," and "Wine," CAOA yields the lowest intra-cluster distances, demonstrating its ability to form tight, well-defined clusters. F-measure measures the quality of clustering. Higher values indicate better clustering quality. CAOA achieves competitive F-measure values, especially in datasets like "Cancer" and "LR." It showcases its effectiveness in creating clusters that closely match the ground truth.

Table 3 Comparative analysis using intra-cluster distance

S. No.	Algorithms						
	CAOA	ACO	PSO	BA	CSO	GA	K-means
Iris	95.76	98.36	98.73	97.53	97.64	125.19	113.56
Cancer	3047.56	3178.09	3116.64	3098.93	3124.15	3249.46	3248.25
CMC	5718.2	5831.25	5846.63	5778.14	5804.52	5756.59	5912.46
Wine	16398.82	16526.12	16491.52	16556.89	16486.21	16530.53	18059.91
Glass	259.85	281.46	278.71	269.61	264.44	282.32	246.51
Statlog	450676405	542216190	522200928	450769448	513208164	793000994	812558906
LR	607645.31	608495.87	608470.77	613775.68	611102.88	611731.68	624765.58
ISOLET	441825.51	455837.78	451718.88	442361.25	447733.55	460851.88	446502.65

Table 4 Comparative analysis using F-measure

S. No.	Algorithms						
	CAOA	ACO	PSO	BA	CSO	GA	K-means
Iris	0.784	0.778	0.78	0.782	0.781	0.774	0.781
Cancer	0.831	0.829	0.826	0.833	0.831	0.819	0.832
CMC	0.336	0.332	0.333	0.336	0.334	0.324	0.337
Wine	0.548	0.521	0.517	0.523	0.522	0.515	0.52
Glass	0.47	0.402	0.412	0.431	0.416	0.333	0.426
Statlog	0.329	0.328	0.322	0.316	0.312	0.314	0.262
LR	0.481	0.427	0.412	0.439	0.416	0.488	0.461
ISOLET	0.378	0.301	0.392	0.369	0.311	0.332	0.361

The clustering outcomes of the proposed algorithm across various datasets are illustrated in *Figures 2(a-g)*. In *Figure 2(a),* the iris dataset's clustering is depicted, showing the distinct presence of three clusters: "setosa," "versicolour," and "virginica." *Figure 2(b)* presents the clustering results of the cancer dataset, revealing clusters based on attributes such as "cell size," "cell shape," and "bare nuclei." *Figure 2(c)* displays the clustering of the CMC dataset, where three clusters, namely "Cluster No use1," "Cluster Long Term2," and "Cluster Short Term3," are evident. Despite the non-linear nature of data objects in the CMC dataset, the CAOA algorithm effectively assigns them to their respective clusters. Moving on to *Figure 2(d),* it illustrates the clustering of the wine dataset, featuring three clusters: "wine type 1," "wine type 2," and "wine type 3." Finally, *Figure 2(e)* represents the clustering of the glass dataset, which comprises six clusters with non-linear data patterns. The CAOA algorithm adeptly assigns data objects to their respective clusters. Moving on, *Figure 2(f)* exhibits the clustering of the statlog (shuttle) dataset. In *Figure 2(g),* the CAOA is applied to the LR dataset, which is divided into 26 clusters denoted by letters A to Z. These clusters are observed to be linearly inseparable. Shifting focus to *Figure 2(h)*, it presents the application of the proposed algorithm to the ISOLET dataset, which also comprises twenty-six clusters, and similarly, these clusters are found to be non-linearly separable from one another.

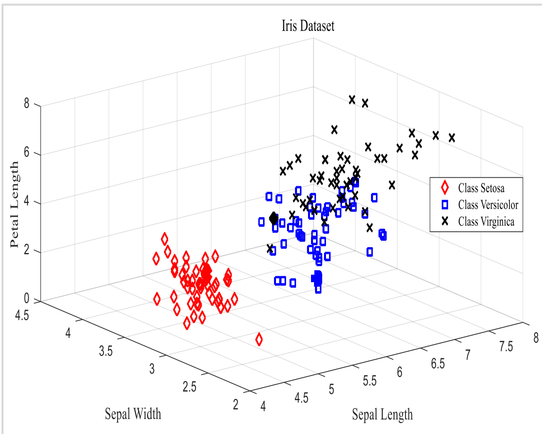

Figure 2(a) Iris

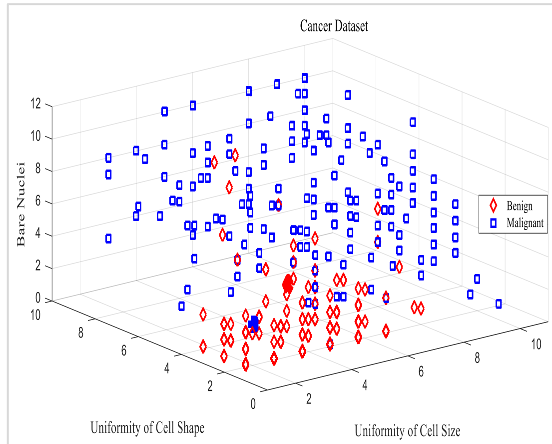

Figure 2(b) Cancer

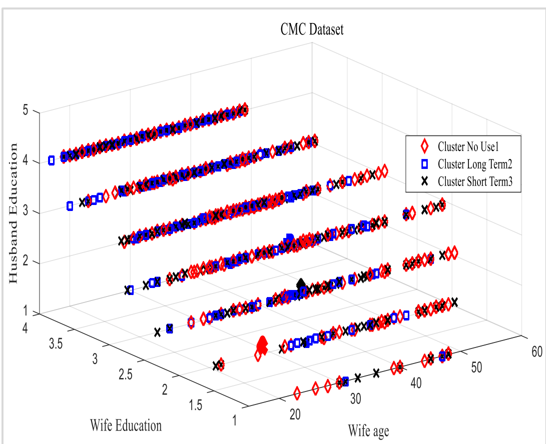

Figure 2(c) CMC

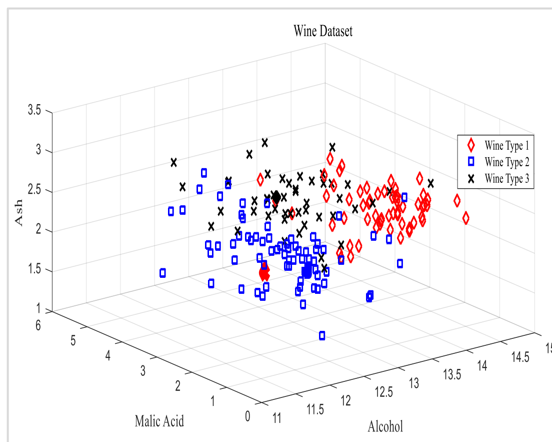

Figure 2(d) wine

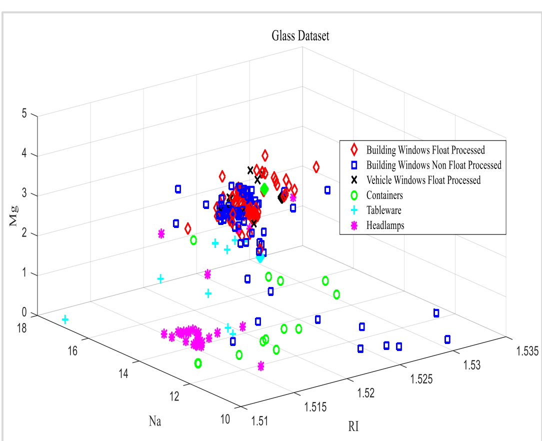

Figure 2(e) Glass

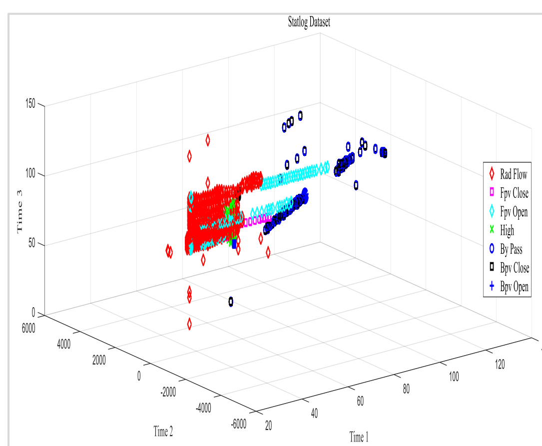

Figure 2(f) Statlog

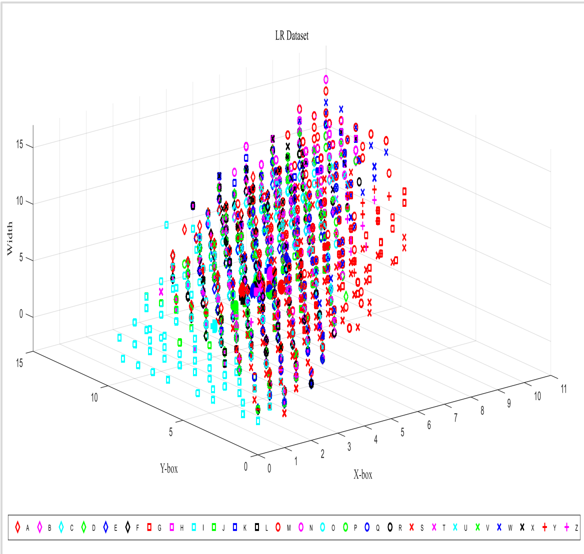

Figure 2(g) LR

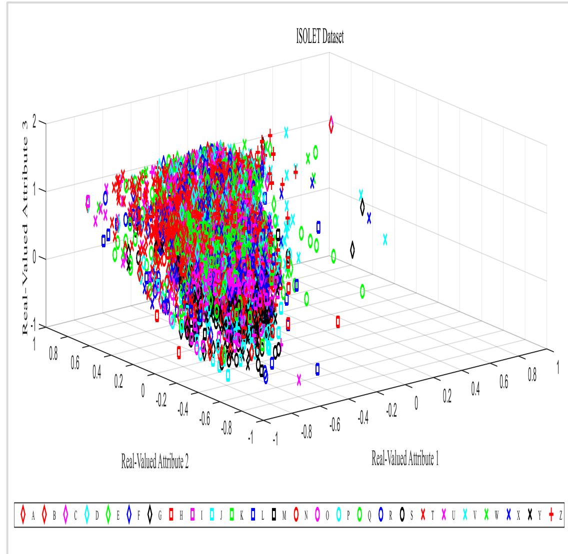

Figure 2(h) ISOLET

4.1 Statistical test

The Friedman test was performed on the average distance among cluster members (intra-cluster distance) and the F-measure parameters, with two hypotheses:

Hypothesis H0: Assumes that the considered algorithms have similar performance.

Hypothesis H1: Assumes that the algorithms demonstrate dissimilar performance.

Based on the test results, CAOA achieved the highest rank (1.3), critical value (12.591), and a p-value of 0.000263 (*Table 5*). With a significance level of 0.05, hypothesis H0 is decisively rejected, indicating significant differences among the algorithms. In other words, the CAOA algorithm exhibits performance dissimilar to the compared algorithms. The Friedman test statistic is 25.60, with a critical value of 12.5915 and a p-value of 0.000263. The sum of squares of rank sums is 8124. This result leads to the rejection of hypothesis H0, indicating that the CAOA algorithm exhibits significantly different performance compared to other algorithms. CAOA achieves the highest rank of 1.3, suggesting its effectiveness in minimizing intra-cluster distance.

The statistical results, using the "F-measure" parameter, are displayed in *Table 6*. The proposed AOAC algorithm secures the highest rank (1.75), critical value (12.5915), and p-value (0.001488). Consequently, hypothesis H0 is rejected, indicating that the AOAC algorithm exhibits dissimilar performance in comparison to the others. The statistical analysis depicted the effectiveness and promising outcomes of the AOAC algorithm. The Friedman test statistic is 22.50, with a critical value of 12.5915 and a p-value of 0.001488. The sum of squares of rank sums is 7965.5. Similarly, hypothesis H0 is rejected, indicating that the CAOA algorithm shows significantly different performance. CAOA attains the highest rank of 1.75, emphasizing its strong performance in terms of F-measure.

Table 5 The statistical results using "avg. Intra-cluster distance" parameter

Clustering Algorithms	CAOA	ACO	PSO	BA	CSO	GA	K-means
Sum	9	38	33	27	27	46	44
Rank	1.3	4.75	4.13	3.38	3.38	5.75	5.5
Friedman test statistic: 25.60			Correction factor: 896		The sum of squares of rank sums: 8124		
Critical value: 12.5915			p-value: 0.000263		Degree of freedom: 6		

Table 6 The statistical results using " F-measure " parameter

Clustering Algorithms	CAOA	ACO	PSO	BA	CSO	GA	K-means
Sum	14	41	38	20.5	36	46	28.5
Rank	1.75	5.13	4.75	2.56	4.5	5.75	3.56
Friedman test statistic: 22.50			Correction factor: 896		Sum of squares of rank sums: 7965.5		
Critical value: 12.5915			p-value: 0.001488		Degree of freedom: 6		

5. Discussion

This study constitutes a thorough investigation of the CAOA algorithm's performance, with a particular focus on its application to real-life datasets, assessed through the "avg. Intra-cluster distance" and "F-measure" parameters. The analysis provides valuable insights into the algorithm's capabilities, its comparative effectiveness against other clustering methods, and the significance of its performance.

The evaluation strategy employed in this study is based on two fundamental parameters: intra-cluster distance and F-measure. These parameters serve as robust indicators of clustering quality, with intra-cluster distance measuring the compactness of clusters and F-measure quantifying the overall quality of clustering outcomes.

One of the striking findings of this analysis is the consistent outperformance of the CAOA algorithm in terms of distance among clusters. The results are particularly remarkable in datasets such as "Iris," "Cancer," and "Wine," where CAOA achieves the lowest intra-cluster distances. This signifies that CAOA excels in creating well-defined, tightly packed clusters, which is crucial for various applications, such as pattern recognition and data segmentation.

Additionally, the F-measure analysis reaffirms the CAOA algorithm's competitive clustering quality. In datasets like "Cancer" and "LR," CAOA achieves F-measure values that are among the highest, indicating its ability to create clusters that closely correspond to the ground truth. This is crucial in applications where the accurate identification of clusters is of paramount importance, such as medical diagnosis and customer segmentation.

To ascertain the statistical significance of these results, the study employs the Friedman test. The test revealed that the CAOA algorithm exhibits significantly different performance compared to other clustering algorithms. The rejection of hypothesis H0 indicates that CAOA stands out and delivers distinct clustering outcomes, which is a valuable finding for practitioners seeking the most effective clustering solution for their specific needs.

The emphasis on both intra-cluster distance and F-measure provides a comprehensive evaluation of the algorithm's clustering quality. The statistical tests strengthen the findings, assuring that the algorithm's superior performance is not a random occurrence but a consistent trend. However, it is important to recognize that while the study provides valuable insights, it also has limitations, such as the limited dataset diversity and the need for a more detailed exploration of parameter sensitivity and algorithm comparison. Addressing these limitations in future research can further enhance the understanding and applicability of the CAOA algorithm in practical clustering scenarios.

A complete list of abbreviations is listed in *Appendix I*.

6. Conclusion and future work

This study thoroughly investigated the performance of the CAOA approach when applied to various datasets. The study focused on two key evaluation parameters, "average intra-cluster distance" and "F-measure," which serve as robust indicators of clustering quality. CAOA consistently outperformed other clustering algorithms, particularly in datasets like "Iris," "Cancer," and "Wine," where it achieved the lowest "average intra-cluster distance." This result signifies CAOA's ability to create well-defined, tightly packed clusters, which is crucial in applications like pattern recognition and data segmentation.

Furthermore, CAOA demonstrated competitive clustering quality in terms of "F-measure," particularly excelling in datasets like "Cancer" and "LR." This highlights its potential for precise cluster identification in domains such as medical diagnosis and customer segmentation. The application of the Friedman test confirmed the statistical significance of CAOA's performance.

This research showcases the potential and effectiveness of CAOA in addressing real-world clustering challenges. These findings offer valuable insights for practitioners looking for robust clustering solutions. For future work, it would be beneficial to expand the dataset diversity and conduct a more detailed exploration of parameter sensitivity. Moreover, a broader comparison with various state-of-the-art clustering algorithms could provide a deeper understanding of CAOA's performance. These efforts will contribute to further enhancing the practical applicability of CAOA in diverse clustering scenarios.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

All the datasets used in the experimentation in this study are publicly available at https://archive.ics.uci.edu/.

Author's contribution statement

Hakam Singh: Conceptualization, investigation, writing – original draft, writing – review and editing. **Ashutosh Kumar Dubey:** Conceptualization, writing – original draft, analysis and interpretation of results.

References

[1] Azevedo A. Data mining and knowledge discovery in databases. In advanced methodologies and technologies in network architecture, mobile computing, and data analytics 2019 (pp. 502-14). IGI Global.

[2] Hu C, Wu T, Liu S, Liu C, Ma T, Yang F. Joint unsupervised contrastive learning and robust GMM for text clustering. Information Processing & Management. 2024; 61(1):103529.

[3] Wang L, Yan J, Mu L, Huang L. Knowledge discovery from remote sensing images: a review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2020; 10(5):e1371.

[4] Chaouch S, Yvonnet J. An unsupervised machine learning approach to reduce nonlinear FE2 multiscale calculations using macro clustering. Finite Elements in Analysis and Design. 2024; 229:104069.

[5] Nanda SJ, Panda G. A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary Computation. 2014; 16:1-8.

[6] Thakur B, Kumar N, Gupta G. Machine learning techniques with ANOVA for the prediction of breast cancer. International Journal of Advanced Technology and Engineering Exploration. 2022; 9(87):232-45.

[7] Bouguettaya A, Yu Q, Liu X, Zhou X, Song A. Efficient agglomerative hierarchical clustering. Expert Systems with Applications. 2015; 42(5):2785-97.

[8] Shambhu S, Koundal D, Das P. Deep learning-based computer assisted detection techniques for malaria parasite using blood smear images. International Journal of Advanced Technology and Engineering Exploration. 2023; 10(105):990-1015.

[9] Amini A, Wah TY, Saboohi H. On density-based data streams clustering algorithms: a survey. Journal of Computer Science and Technology. 2014; 29:116-41.

[10] Song XF, Zhang Y, Gong DW, Gao XZ. A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data. IEEE Transactions on Cybernetics. 2021; 52(9):9573-86.

[11] Alswaitti M, Albughdadi M, Isa NA. Variance-based differential evolution algorithm with an optional crossover for data clustering. Applied Soft Computing. 2019; 80:1-7.

[12] Sharma R, Vashisht V, Singh U. EEFCM-DE: energy-efficient clustering based on fuzzy C means and differential evolution algorithm in WSNs. IET Communications. 2019; 13(8):996-1007.

[13] Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, et al. A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence. 2022; 110:104743.

[14] Rajwar K, Deep K, Das S. An exhaustive review of the metaheuristic algorithms for search and optimization: taxonomy, applications, and open challenges. Artificial Intelligence Review. 2023; 56(11):13187-257.

[15] Darvishpoor S, Darvishpour A, Escarcega M, Hassanalian M. Nature-inspired algorithms from oceans to space: a comprehensive review of heuristic and meta-heuristic optimization algorithms and their potential applications in drones. Drones. 2023; 7(7):1-134.

[16] Dorigo M, Birattari M, Stutzle T. Ant colony optimization. IEEE Computational Intelligence Magazine. 2006; 1(4):28-39.

[17] Kumar Y, Sahoo G. A charged system search approach for data clustering. Progress in Artificial Intelligence. 2014; 2(2):153-66.

[18] Dubey A, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. Computers, Materials and Continua. 2021; 70(3):4523-43.

[19] Harshavardhan A, Boyapati P, Neelakandan S, Abdul-rasheed AAA, Singh PAK, Walia R. LSGDM with biogeography-based optimization (BBO) model for healthcare applications. Journal of Healthcare Engineering. 2022; 2022(1):2170839.

[20] Nadimi-shahraki MH, Zamani H, Mirjalili S. Enhanced whale optimization algorithm for medical feature selection: a COVID-19 case study. Computers in Biology and Medicine. 2022; 148:105858.

[21] Sharma SK, Ghai W. Artificial bee colony optimized VM migration and allocation using neural network architecture. International Journal of Advanced Technology and Engineering Exploration. 2023; 10(102):590-607.

[22] Bezdek JC, Boggavarapu S, Hall LO, Bensaid A. Genetic algorithm guided clustering. In proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence 1994 (pp. 34-9). IEEE.

[23] Shelokar PS, Jayaraman VK, Kulkarni BD. An ant colony approach for clustering. Analytica Chimica Acta. 2004; 509(2):187-95.

[24] Liu Y, Yi Z, Wu H, Ye M, Chen K. A tabu search approach for the minimum sum-of-squares clustering problem. Information Sciences. 2008; 178(12):2680-704.

[25] Mahdavi M, Chehreghani MH, Abolhassani H, Forsati R. Novel meta-heuristic algorithms for clustering web documents. Applied Mathematics and Computation. 2008; 201(1-2):441-51.

[26] Santosa B, Ningrum MK. Cat swarm optimization for clustering. In international conference of soft computing and pattern recognition 2009 (pp. 54-9). IEEE.

[27] Zhang C, Ouyang D, Ning J. An artificial bee colony approach for clustering. Expert Systems with Applications. 2010; 37(7):4761-7.
[28] Singh S, Srivastava S. Kernel fuzzy C-means clustering with teaching learning based optimization algorithm (TLBO-KFCM). Journal of Intelligent & Fuzzy Systems. 2022; 42(2):1051-9.
[29] Kaur A, Kumar Y. Neighborhood search based improved bat algorithm for data clustering. Applied Intelligence. 2022; 52(9):10541-75.
[30] Kumar Y, Kaur A. Variants of bat algorithm for solving partitional clustering problems. Engineering with Computers. 2022; 38(Suppl 3):1973-99.
[31] Niknam T, Amiri B. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Applied Soft Computing. 2010; 10(1):183-97.
[32] Aggarwal S, Singh P. Cuckoo, bat and krill herd based k-means++ clustering algorithms. Cluster Computing. 2019; 22(Suppl 6):14169-80.
[33] Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1979; 28(1):100-8.
[34] Zhang B, Hsu M, Dayal U. K-harmonic means-a spatial clustering algorithm with boosting. In international workshop on temporal, spatial, and spatio-temporal data mining 2000 (pp. 31-45). Berlin, Heidelberg: Springer Berlin Heidelberg.
[35] Žalik KR. An efficient k′-means clustering algorithm. Pattern Recognition Letters. 2008; 29(9):1385-91.
[36] Geng X, Mu Y, Mao S, Ye J, Zhu L. An improved K-means algorithm based on fuzzy metrics. IEEE Access. 2020; 8:217416-24.
[37] Tang LY, Wang ZH, Wang SD, Fan JC, Yue GW. A novel rough semi-supervised k-means algorithm for text clustering. International Journal of Bio-Inspired Computation. 2023; 21(2):57-68.
[38] Liu B, Liu C, Zhou Y, Wang D, Dun Y. An unsupervised chatter detection method based on AE and merging GMM and K-means. Mechanical Systems and Signal Processing. 2023; 186:109861.
[39] Ning Z, Chen J, Huang J, Sabo UJ, Yuan Z, Dai Z. WeDIV–an improved k-means clustering algorithm with a weighted distance and a novel internal validation index. Egyptian Informatics Journal. 2022; 23(4):133-44.
[40] Cheng D, Huang J, Zhang S, Xia S, Wang G, Xie J. K-means clustering with natural density peaks for discovering arbitrary-shaped clusters. IEEE Transactions on Neural Networks and Learning Systems. 2023; 35(8):11077-90.
[41] Yang F, Sun T, Zhang C. An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. Expert Systems with Applications. 2009; 36(6):9847-52.
[42] Sixu L, Muqing W, Min Z. Particle swarm optimization and artificial bee colony algorithm for clustering and mobile based software-defined wireless sensor networks. Wireless Networks. 2022; 28(4):1671-88.
[43] Yan X, Zhu Y, Zou W, Wang L. A new approach for data clustering using hybrid artificial bee colony algorithm. Neurocomputing. 2012; 97:241-50.
[44] Huang CL, Huang WC, Chang HY, Yeh YC, Tsai CY. Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering. Applied Soft Computing. 2013; 13(9):3864-72.
[45] Hatamlou A, Hatamlou M. PSOHS: an efficient two-stage approach for data clustering. Memetic Computing. 2013; 5(2):155-61.
[46] Singh H, Kumar Y. An enhanced version of cat swarm optimization algorithm for cluster analysis. International Journal of Applied Metaheuristic Computing. 2022; 13(1):1-25.
[47] Abualigah L, Diabat A, Mirjalili S, Abd EM, Gandomi AH. The arithmetic optimization algorithm. Computer Methods in Applied Mechanics and Engineering. 2021; 376:113609.

Dr. Hakam Singh is an Assistant Professor in the Department of Computer Science and Engineering at Chitkara University, Himachal Pradesh, India. He obtained his Ph.D. from Jaypee University of Information Technology, with a thesis focused on designing new meta-heuristic algorithms for partitional clustering problems. Dr. Singh's research is at the intersection of machine learning, clustering algorithms, and data analysis, and he has made significant contributions to these fields through numerous publications in prestigious journals and conferences. His innovative work is further evidenced by several patents he holds, reflecting his commitment to advancing technology and addressing complex computational challenges.
Email: hakam.singh@chitkarauniversity.edu.in

Dr. Ashutosh Kumar Dubey is an Associate Professor in the Department of Computer Science at Chitkara University School of Engineering and Technology, situated in Himachal Pradesh, India. He is a Postdoctoral Fellow at the Ingenium Research Group Lab, Universidad de Castilla-La Mancha, Ciudad Real, Spain. He is a Senior Member of both IEEE and ACM and possesses more than 16 years of teaching experience. Dr. Dubey has authored and edited twenty books and has published over 80 articles in peer-reviewed international journals and conference proceedings. His research interests encompass Machine Learning, Renewable Energy, Health Informatics, Nature-Inspired Algorithms, Cloud Computing, and Big Data.
Email: ashutosh.dubey@chitkara.edu.in

Appendix I

S. No.	Abbreviation	Description
1	ABC	Artificial Bee Colony
2	ACO	Ant Colony Optimization
3	AOA	Arithmetic Optimization Algorithm
4	BBBC	Biogeography-Based Optimization
5	BH	Bat Algorithm
6	CAOA	Clustering Based on Arithmetic Optimization Algorithm
7	CSO	Cat Swarm Optimization
8	FAPSO	Fuzzy Adaptive PSO
9	GA	Genetic Algorithm
10	HS	Harmony Search
11	IoT	Internet of Things
12	KHM	k-Harmonic-Means
13	MOA	Math Optimizer Accelarated
14	MOP	Math Optimizer Probability
15	NDP	Natural Density Peaks
16	PSO	Particle Swarm Optimization
17	PSOKHM	Particle Swarm Optimization k-Harmonic Means
18	RSKmeans	Rough Set-Based Semi-Supervised k-Means
19	SA	Simulated Annealing
20	TS	Tabu Search
21	TLBO	Teaching-Learning-Based Optimization
22	UCI	University of California, Irvine
23	WOA	Whale Optimization Algorithm