11th International Conference on Computer and Knowledge Engineering (ICCKE 2021), October 28-29, 2021, Ferdowsi University of Mashhad

Data Clustering Using Chimp Optimization Algorithm

1st Sayed Pedram Haeri Boroujeni
Department of Artificial Intelligence and Data Science
Istanbul Aydin University
Istanbul, Turkey
sayedpedramboroujeni@stu.aydin.edu.tr

2nd Elnaz Pashaei
Department of Software Engineering
Istanbul Aydin University
Istanbul, Turkey
elnazpashaei@aydin.edu.tr

2021 11th International Conference on Computer Engineering and Knowledge (ICCKE) | 978-1-6654-0208-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCKE54056.2021.9721483

***Abstract*— In recent decades, many successful meta-heuristic algorithms are widely applied to data clustering problems due to their powerful capabilities. The Chimp Optimization Algorithm (ChOA) is a recently proposed meta-heuristic search algorithm that is inspired by chimps' individual intelligence and sexual motivation in their group hunting. The good performance of ChOA in a variety of optimization problems, prove the superiority of this algorithm over other swarm-based intelligence. This study presents a novel approach according to ChOA for data clustering. ChOA, like other Swarm Intelligence-based Algorithms (SIAs), starts with a random population and then calculates an objective value for them. In data clustering problems the population of candidate solutions is the position vectors of the centroids, and the objective function is the sum of intra-cluster distances (SICDs) between each sample and its nearest centroid. Following that, ChOA attempts to optimize the population by finding the best position vectors of optimal centroids. There are four different kinds of chimps in this regard: driver, chaser, barrier, and attacker. Furthermore, four major hunting moves including driving, blocking, chasing, and attacking are considered to find the promising solution. In this research, four datasets of the UCI machine learning repository are used to evaluate the performance of the proposed new approach. Experimental results illustrate that the proposed work significantly outperforms other clustering algorithms regarding the objective value, Error Rate (ER), and some other statistical tests.**

Keywords—Meta-heuristic, Chimp Optimization Algorithm, Data clustering, Optimization problem.

I. Introduction

Due to the advances in database technology, data clustering has become one of the most challenging tasks which provide a significant impact on data analysis performance. Clustering is unsupervised learning [1] which means it does not make any prior information about data. It is a process of categorizing a set of data into different clusters, where the data within a cluster must be extremely similar to each other and the data within different clusters must be highly dissimilar to each other [2]. This similarity and dissimilarity can be defined by the SICD measure, which should be minimized to achieve better clustering. In recent years, many researchers and professionals from various fields of science have been attracted to this prominent area of research. Over the past years, clustering has become a popular challenge in different fields such as bioinformatics [3], text mining [4], face recognition [5], and medicine [6].

Recently, various kinds of nature-inspired algorithms are applied to different real-world problems because of their great capability in solving complex problems within a reasonable amount of time. Optimization problems can be assigned as a computational problem in which the goal is to maximize or minimize the objective function. Due to some important characteristics of data clustering such as non-linear objective function and a wide range of search domains, algorithms are faced with different difficulties in converging towards the best optimum solution. Therefore, data clustering is a challenging task where SIAs play a significant role in solving the clustering problems [7]. SIAs, imitating the collective (swarm) behavior of different entities, especially those species who rely on consensus decision-making in their processes. They offers new ways for solving clustering problems such as Particle Swarm Optimization (PSO) [8], Artificial Bee Colony (ABC) [9], and Genetic Algorithm (GA) [10].

Generally, there are two groups of algorithms including heuristic algorithms and meta-heuristic algorithms. Heuristic approaches are designed to search for the best solution without checking the entire search space. The main benefit of this technique is that they are able to discover the optimal solution in a reasonable time and faster than other methods [11]. The major drawback of this method is getting stuck in local minima, especially for optimizing high-dimensional problems. By contrast, meta-heuristic approaches are designed to escape from a local optimum problem by allowing flexible movements or random behaviors. They have two important parts which are exploration and exploitation, where the first one explores the entire search space by generating random solutions and the second one tries to discover the best optimal solution close to the current solutions [11]. Therefore, an appropriate balance should be chosen between exploration and exploitation to achieve the highest performance in meta-heuristic algorithms.

The major contribution of this research article is to suggest a new method for solving clustering problems in which the recent ChOA [12] is employed to optimize the specified objective function. ChOA is applied to transform clustering into an optimization problem and select the appropriate cluster centers to find the best solution for our problem. After that, all of the data in a dataset are classified into different clusters according to the principle of the minimum intra-cluster distances. Eventually, the efficiency of the ChOA is compared against other clustering algorithms using different datasets. To the best of our knowledge, this is the first time that ChOA is considered as a clustering technique. The rest of this study has been structured as follows: Section II reviews the related works. In Section III, the details of the material and proposed methods are discussed. The experimental evaluation and analysis are illustrated in Section IV, Meanwhile, Section V outlines onclusion.

II. Review of the Related Literature

In recent years, many studies focused on data clustering not only as an important task of data mining but also as a dynamic way for testing optimization algorithms' efficiency. Classical clustering algorithms such as K-means are efficient techniques for solving clustering problems. Even though such a kind of these algorithms has their benefits, they are highly dependent on the initialization parameters which make the process hard to find the optimal solutions. The first and foremost disadvantage of this algorithm is trapping in local optima. The second one is that it is strongly sensitive to the cluster centers' initial values [13]. To address these issues, different optimization algorithms which are inspired by nature have been offered to enhance the efficiency of data clustering.

According to the available literature, there is a wide range of optimization algorithms for data clustering tasks. The Ant Colony Optimizer (ACO) which mimics the real ants' behavior [14], is a probabilistic technique that is able to find suitable solutions for solving clustering problems [15]. Another algorithm is Grey Wolf Optimizer (GWO) [16] that simulates the hunting behavior of gray wolves. It has been applied for solving clustering problems in [17]. Symbiotic Organism Search (SOS) [18] is another method that was simulated by the symbiotic interactions within a paired organism relationship and it is proposed for clustering analysis in [19]. The Black Hole Algorithm (BHA) is another optimization algorithm where the black hole phenomenon is the primary source of its inspiration [20]. It is a type of meta-heuristic approach based on physical phenomena, which is introduced for clustering tasks [21]. The Krill herd Algorithm (KHA) is another optimization algorithm [22] that tries to imitate the herding mechanism of krill. It is also suggested for solving clustering problems in [23]. Finally, the Whale Optimization Algorithm (WOA) [24] is another example of clustering algorithms [25], which the source of its inspiration is the behavior of humpback whales in their hunting mechanism. All of the optimization algorithms mentioned previously demonstrate their high efficiency in solving various optimization problems. This is the reason what motivated us to keep working on this field and propose our new approach according to the ChOA for data clustering.

III. Material and Proposed Methods

A. Data Clustering Analysis

The main aim of our paper is data clustering which refers to partitioning N $\{D_i=D_1, D_2, \ldots, D_N\}$ data samples into K $\{C_i=C_1, C_2, \ldots, C_K\}$ clusters, where N indicates the number of samples, D_i represents the position of each sample, K shows the number of clusters, and C_i is the position of each cluster's center. In order to have appropriate and efficient data clustering, the best cluster centers need to be identified. To be more specific, the objective is to minimize the SICDs between each data sample and the center of its cluster. The intra-cluster sum or SICD is defined according to Equation (1).

$$F(D,C) = \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} \left\|(D_i - C_j)\right\|^2 \quad (1)$$

Where F(D, C) represents the objective value that should be minimized, $\left\|(D_i - C_j)\right\|$ indicates the Euclidean distance between a data sample and the cluster center, and w_{ij} is the association weight which could be one (if sample i belongs to cluster j) or zero (if sample i does not belong to cluster j).

B. Chimp Optimization Algorithm (ChOA)

In this study, a novel meta-heuristic method called ChOA has been used to handle complex clustering tasks. The concept of ChOA is introduced by the authors in [12]. This algorithm is inspired by the social behavior of chimps including their individual intelligence and sexual motivation while they are in a hunting group. Each group contains various types of chimps which are not similar to each other regarding ability and intelligence. Although they are all carrying out their responsibilities as a member of group, they have their own strategy to discover the solution. A chimp colony is a group of four types of chimps including drivers, barriers, chasers, and attackers. they have different skills and each skill can be useful in a specific situation. Drivers pursue the prey in order to look after it. Barriers construct a dam on top of the trees to prevent the prey from moving forward. Chasers are responsible for following the prey to catch up with it. Lastly, attackers attack the prey by closing its route to prevent it from escaping and force it back to the chasers or lead it into the trap. The attackers' performance has a close relationship with their age, intelligence, and physical ability.

generally, the chimps' hunting mecganism has two important parts namely exploration and exploitation. Exploration part refers to driving, chasing, and blocking the prey, and Exploitation part refers to attacking and hunting the prey. The key point in achieving the high performance of ChOA is that the appropriate balance should be chosen between exploration and exploitation. Preparing a mathematical model for the ChOA algorithm needs five independent parts which are formulated as follows.

1) Driving and Chasing Part: Equations (2) and (3) are the mathematical model of driving and chasing the prey:

$$d = \left|c.X_{prey}(t) - m.X_{chimp}(t)\right| \quad (2)$$

$$X_{chimp}(t+1) = X_{prey}(t) - a.d \quad (3)$$

Where t, X_{prey}, and X_{chimp} represent the current iteration, the position vector of prey, and the vector of chimp position, respectively. Parameters a , c , and m are the coefficient vectors that are calculated by Equations (4)-(6).

$$a = 2.f.r_1 - f \quad (4)$$

$$c = 2.r_2 \quad (5)$$

$$m = \text{Chaotic\_Value} \quad (6)$$

Where f indicates the boundary range of non-linearly that is declined from 2.5 to 0 over the course of iterations. r_1 and r_2 are the random vectors between 0 to 1 and m is a chaotic vector that shows the impact of chimp sexual motivation on the hunting process. As we mentioned before, different types of chimps have different behaviors in local search and global search so they using various strategies to update f. There are numerous kinds of continuous functions which can be used to update f. The only characteristic of these functions is that the value of f must be declined after each iteration. Table I demonstrates the dynamic coefficients of the f vector, where t is the number of current iteration and T is the maximum number of iterations.

TABLE I. THE DYNAMIC COEFFICIENT OF ***F*** VECTOR.

No.	**Function Detail**	
	Groups	***Functions***
1	First Group	$1.95 - 2\frac{t^{1/4}}{T^{1/3}}$
2	Second Group	$1.95 - 2\frac{t^{1/3}}{T^{1/4}}$
3	Third Group	$\left(-3\frac{t^{3}}{T^{3}}\right) + 1.5$
4	Fourth Group	$\left(-2\frac{t^{3}}{T^{3}}\right) + 1.5$

Various groups of dynamic coefficients which have been proposed in Table I are very effective because of some reasons. First of all, chimps with different abilities can explore the search space with different capabilities. Secondly, different kinds of strategies allow them to achieve an appropriate balance between local and global search. Finally, the non-linearity of these groups lets ChOA become more helpful in solving complex problems.

2) Exploitation Part: It refers to the attacking behavior of chimps in which the main task is conducted by attackers and the other chimps including drivers, barriers, and chasers assist them in the hunting process. As mentioned previously, this is an unsupervised searching where is no information available about the prey's position (best solution). Therefore, it can be supposed that the best solutions are obtained by the first driver, chaser, barrier, and attacker, then the rest of the chimps adjust their positions according to the best chimps positions. Equations (7)-(9) are the mathematical model of the attacking part:

$$\begin{aligned} d_{Driver} &= |c_4 x_{Driver} - m_4 x| \\ d_{Chaser} &= |c_3 x_{Chaser} - m_3 x| \\ d_{Barrier} &= |c_2 x_{Barrier} - m_2 x| \\ d_{Attacker} &= |c_1 x_{Attacker} - m_1 x| \end{aligned} \tag{7}$$

$$\begin{aligned} x_4 &= x_{Driver} - a_4(d_{Driver}) \\ x_3 &= x_{Chaser} - a_3(d_{Chaser}) \\ x_2 &= x_{Barrier} - a_2(d_{Barrier}) \\ x_1 &= x_{Attacker} - a_1(d_{Attacker}) \end{aligned} \tag{8}$$

$$x\,(t+1) = \frac{x_1+x_2+x_3+x_4}{4} \tag{9}$$

Where parameters a, m, and c represent the coefficient vectors which were calculated by Equations (4)-(6), x refers to the chimp positions, and d is the distance between each chimp and its prey.

3) Utilization Part: It refers to the last stage of hunting process in which the chimps attack the prey to get the meat. In the mathematical models for formulating the attacking part, a is a random number in [-2f, 2f] and f is a boundary range of [0, 2.5] which is reduced after each iteration. As mentioned before, the algorithm may be trapped in local minima because the chimps update their positions based on the driver, chaser, barrier, and attacker positions. So, the algorithm needs more focusing on the exploration part to prevent this issue.

4) Exploration Part: It refers to the searching task for finding the prey's location in order to finish the hunting process. In the mathematical models, two key parameters are existing that can affect the performance of the exploration part. The first parameter is a, which represents a random value that can be greater than 1 or less than -1. The inequality |a|<1 means that the chimps should converge to the prey, meanwhile, |a|>1 shows that the chimps should diverge from the prey. The second parameter is c, which is a random vector between the range of 0 and 2. The inequalities c>1 and c<1 are able to strengthen and weaken the effect of prey location on distance calculation, respectively. Furthermore, these parameters help to avoid trapping in the local minima issue because of their requirement for the generation of random values over the course of iterations.

5) Social Motivation Part: It refers to the sexual incentive of chimps in which they try to get meat for exchanging in social needs including sex and grooming in the final part. So, this chaotic behavior assists them to solve two important issues including slow convergence rate and trap in the local minima. Equation (10) represents the chaotic map that is used in ChOA algorithms.

$$x_{i+1} = \begin{cases} 1 & x_i = 0 \\ \frac{1}{mod(x_i,1)} & \text{otherwise} \end{cases} \tag{10}$$

Where this function is called Gauss/mouse map and the range of x_{i+1} is in the interval of [0, 1]. Also, the value 0.7 has been assumed as the initial value (x_i) for the chaotic map. the mathematical model of updating positions is defined in Equation (11). In this model, the chance of selecting between the normal updating position and the chaotic mechanism is set to 0.5.

$$x_{Chimp}(t+1) = \begin{cases} x_{prey}(t) - a.d & \text{if } \mu < 0.5 \\ \text{Chaotic} & \text{if } \mu \geq 0.5 \end{cases} \tag{11}$$

Where μ is a number in [0, 1]. Eventually, Algorithm (I) presents the pseudo-code of ChOA for solving data clustering problems.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this research, each experiment is repeated fifty independent times with random seed which leads to obtaining more reliable and confident results. Moreover, all of the algorithms are implemented using MATLAB R2018b with Intel Core i5 processor, 3.2 GHz CPU, and 6 GB of RAM.

A. Datasets and Parameters Setting

Table II summarizes the details about four benchmark datasets namely Iris, Wine, Wisconsin Breast Cancer, and Contraceptive Method Choice (CMC). Datasets are obtained from the UCI repository [26] to evaluate the efficiency of our proposed approach. They are different from each other with regards to the number of sizes, attributes, classes, clusters. Table II demonstrates the value of different parameters for ChOA.

Algorithm I. Pseudo-code of ChOA

```
Define the objective function by Equation (1)
Initialize the population
Initialize ChOA's parameters
Obtain the position of each chimp
Divide chimps randomly into different groups
Calculate the fitness of each chimp
$X_{Attacker}$ = The best chimp
$X_{Chaser}$ = The second-best chimp
$X_{Barrier}$ = The third-best chimp
$X_{Driver}$ = The fourth-best chimp
 While (Iter < $Iter_{max}$)
     for each chimp:
         Extract the chimp's group
         Update f, m, and c
         Use f, m, and c to calculate a and d
     end for
         for each search chimp
             if (μ < 0.5)
                 if (|a| < 1)
         Update the position of the current chimp by Equation (3)
                 else if (|a| > 1)
                     Select a random chimp
                 end if
             else if (μ > 0.5)
         Update the position of the current chimp by Equation (11)
             end if
     end for
   Update f, m, a, and c
   Update $X_{Attacker}$, $X_{Chaser}$, $X_{Barrier}$, and $X_{Driver}$
   Iter = $Iter_{max}$ + 1
 end while
return $X_{Attacker}$
Set $X_{Attacker}$ as the best set of cluster centers
```

TABLE II. BENCHMARK DATASETS DESCRIPTION

No.	Datasets Detail			
	Dataset	*Number of clusters*	*Number of features*	*Number of data objects*
1	Iris	3	4	150 (50, 50, 50)
2	CMC	3	9	1473 (629, 334, 510)
3	Wine	3	13	178 (59, 71, 48)
4	Cancer	2	9	683 (444, 239)

TABLE III. PARAMETERS TUNING FOR THE PROPOSED METHOD

No.	Parameters Detail	
	Parameter	*Value*
1	f	Table I
2	r_1, r_2	[0, 1]
3	m	Chaotic
4	Population size	60
5	Maximum iterations	1000

B. Evaluation Criteria

In the proposed paper, SICD value is considered as the primary measurement, which is defined in Equation (1). It is obvious that the best clusters' centers have the minimum value of SICD. Moreover, Error Rate (ER) is the second measurement that is employed to assess the efficiency of our proposed work, as specified in Equation (12). Clearly, the smaller value of ER shows the good performance of the data clustering method.

$$\text{Error Rate} = \frac{\text{Number of misclassified data}}{\text{Total number of data}} \times 100 \qquad (12)$$

C. Result Analysis and Statistical Comparison

In this part, the experimental results obtained by ChOA clustering algorithm are compared against six popular and some recent optimization algorithms in terms of the SICD and ER measures. These algorithms are: K-means [27], PSO [8], Gravitational Search Algorithm (GSA) [28], Big Bang-Big Crunch (BB-BC) [29], Black Hole (BH) [30], and Improved Krill Herd (IKH) [31]. It should be mentioned that we tried to keep the values of essential parameters constant for all of the optimization algorithms to ensure a fair comparison. Tables IV-VII demonstrate the best, mean, worst, and standard deviation (STD) results of the SICD, which are recorded by six optimization algorithms for four benchmark datasets over fifty runs. It should be noted that we set the same value for all the common parameters to ensure fair comparisons with other algorithms.

TABLE IV. THE SICD VALUE OBTAINED BY VARIOUS CLUSTERING ALGORITHMS ON IRIS DATASET

Optimization Algorithm	Criteria Detail			
	Best	*Mean*	*Worst*	*STD*
K-means	96.986118	110.68215	131.39516	11.45546
BB-BC	96.707392	97.018237	97.284314	0.193931
GSA	96.716155	97.294381	97.915288	0.154373
IKH	96.683142	96.693924	96.703912	0.035193
PSO	96.778944	97.783955	98.341171	0.914421
BH	96.661767	96.677994	96.682173	0.002987
ChOA	**96.61130**	**96.64140**	**96.70710**	**0.02430**

TABLE V. THE SICD VALUE OBTAINED BY VARIOUS CLUSTERING ALGORITHMS ON WINE DATASET

Optimization Algorithm	Criteria Detail			
	Best	*Mean*	*Worst*	*STD*
K-means	16,447.91	18,571.16	22,688.13	993.79361
BB-BC	16,297.45	16,299.37	16,309.54	3.7746255
GSA	16,321.44	16,368.29	16,399.37	29.558863
IKH	16,291.98	16,311.28	16,472.87	29.488158
PSO	16,311.51	16,322.93	16,353.46	20.131319
BH	16,294.05	16,296.55	16,299.43	1.9851822
ChOA	**16,293.11**	**16,301.24**	**16,304.27**	**2.1754100**

TABLE VI. THE SICD VALUE OBTAINED BY VARIOUS CLUSTERING ALGORITHMS ON CANCER DATASET

Optimization Algorithm	Criteria Detail			
	Best	***Mean***	***Worst***	***STD***
K-means	2979.5277	3119.5566	4978.1573	299.2799
BB-BC	2964.4183	2964.4711	2964.4972	0.001932
GSA	2964.9978	2969.9155	2988.7728	9.762231
IKH	2964.4019	2967.2254	2969.1471	6.419805
PSO	2969.5122	2978.3100	3048.1799	9.397221
BH	2964.3901	2964.3998	2964.4082	0.010982
ChOA	**2964.3867**	**2964.3868**	**2964.3872**	**0.000351**

TABLE VII. THE SICD VALUE OBTAINED BY VARIOUS CLUSTERING ALGORITHMS ON CMC DATASET

Optimization Algorithm	Criteria Detail			
	Best	***Mean***	***Worst***	***STD***
K-means	5539.97662	5541.1176	5549.7711	2.492219
BB-BC	5540.17720	5580.2460	5621.6933	37.97103
GSA	5539.98664	5579.8244	5599.2436	39.69221
IKH	5689.33901	5693.8832	5699.7105	3.224290
PSO	5538.29660	5544.1196	5559.2288	6.995361
BH	5532.86965	5533.6853	5535.6651	0.601711
ChOA	**5532.77291**	**5533.4806**	**5534.9221**	**0.117210**

Results in Tables IV-VII demonstrate the strength of our proposed approach in solving clustering problems. By inspecting the numerical results, we can conclude that the solutions which are obtained by the ChOA technique are much better than the other methods in three datasets including Iris, Cancer, and CMC. However, only in the Wine dataset, the SICD value achieved by ChOA is greater than that of the BH algorithm. It is necessary to mention that our proposed work provides all of the results with a small STD which is shows the stable behavior of the ChOA algorithm. Table VIII represents the average ER of our proposed method and the other six clustering algorithms. Obviously, a minimum average ER is obtained by ChOA in all the datasets.

TABLE VIII. THE AVERAGE ER VALUE OF DIFFERENT CLUSTERING ALGORITHMS ON FOUR DATASETS

Optimization Algorithm	Dataset			
	Iris	***Wine***	***Cancer***	***CMC***
K-means	14.11	32.25	4.41	54.47
BB-BC	10.05	28.68	3.69	54.52
GSA	10.03	29.28	3.72	55.67
IKH	9.35	28.29	4.91	54.38
PSO	10.11	28.81	3.77	54.50
BH	10.01	28.51	3.69	54.39
ChOA	**9.31**	**27.94**	**3.61**	**53.95**

D. Statistical Analysis

There are several statistical methods for comparing the performance of various clustering techniques. For this purpose, the algorithms are ranked according to their SICD results and ER values. After that, Equation (13) calculates the Average Rank (AR) of each method. Table IX represents the rank of each algorithm according to the SICD and ER values. As seen from the results, our proposed method is ranked first among all of the approaches. Finally, Table X demonstrates the results of the proposed method and six recent clustering algorithms [32-37] in terms of the SICD value.

$$AR = \frac{\text{The sum of algorithms' ranks for each dataset}}{\text{Total number of datasets}} \tag{13}$$

TABLE IX. THE RANKING OF SEVEN CLUSTERING ALGORITHMS BASED ON THE SICD AND ER VALUES

Method	Measure	Dataset				AR
		Iris	***Wine***	***Cancer***	***CMC***	
K-means	SICD	7	7	7	3	6
	ER	7	7	5	4	5.75
BB-BC	SICD	4	3	2	5	3.5
	ER	5	4	2	6	4.25
GSA	SICD	5	6	5	6	5.5
	ER	4	6	3	7	5
IKH	SICD	3	4	4	7	4.5
	ER	2	2	6	2	3
PSO	SICD	6	5	6	4	5.25
	ER	6	5	4	5	5
BH	SICD	2	1	3	2	2
	ER	3	3	2	3	2.75
ChOA	SICD	**1**	**2**	**1**	**1**	**1.25**
	ER	**1**	**1**	**1**	**1**	**1**

TABLE X. COMPARISON OF THE PROPOSED ALGORITHM AND SIX RECENT CLUSTERING METHODS ACCORDING TO THEIR SICD VALUES

Method	Dataset			
	Iris	***Wine***	***Cancer***	***CMC***
MVO [32]	98.47380	16380.60	2965.2314	5835.0506
SOS [33]	96.65550	16293.05	2964.3870	5693.7253
FPA [34]	99.35051	16357.00	2984.6225	5910.7652
QALO-K [35]	96.73101	16453.16	2977.9322	5543.4400
KICS [36]	96.95257	16341.46	2973.3878	5540.6524
HCSDE [37]	101.90691	16428.542	4390.9237	5844.7981
ChOA	**96.64140**	**16,301.24**	**2964.3868**	**5533.4806**

V. CONCLUSION

This research suggested a novel technique for solving clustering problems based on the Chimp optimization algorithm. This is the first time that the main characteristics of ChOA are discovered and it is used as a clustering algorithm. In this study, four various kinds of datasets are considered to evaluate the proposed method and the results are compared against six well-known algorithms and six recent clustering methods. Experimental results represent that the suggested algorithm not only increases data clustering efficiency but also reduces the number of misclassified data objects. Statistical analysis also demonstrates the superior performance of our clustering approach, which is rated first for three of the four datasets based on the SICD value. Additionally, it is ranked first based on the ER measure in all of the datasets. In the future, our proposed research could be considered in various fields for solving more challenging problems.

REFERENCES

[1] S. Kant and I. A. Ansari, "An improved K means clustering with Atkinson index to classify liver patient dataset," International Journal of System Assurance Engineering and Management, vol. 7, no. 1, pp. 222-228, 2016.

[2] H. Deeb, A. Sarangi, D. Mishra, and S. K. Sarangi, "Improved Black Hole optimization algorithm for data clustering," Journal of King Saud University-Computer and Information Sciences, 2020.

[3] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," Briefings in bioinformatics, vol. 21, no. 1, pp. 1-10, 2020.

[4] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," 2017.

[5] Y. Zhang, Z. Tang, B. Wu, Q. Ji, and H Lu, "A coupled hidden conditional random field model for simultaneous face clustering and naming in videos," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5780-5792, 2016.

[6] I. E. Kaya, A. Ç. Pehlivanlı, E. G. Sekizkardeş, and T. Ibrikci, "PCA based clustering for brain tumor segmentation of T1w MRI images," Computer methods and programs in biomedicine, vol. 140, pp. 19-28, 2017.

[7] T. Singh, "A chaotic sequence-guided Harris hawks optimizer for data clustering," Neural Computing and Applications, vol. 32, pp. 17789-17803, 2020.

[8] J. Kennedy, and R. Eberhart, "Particle swarm optimization." In Proceedings of ICNN'95-international conference on neural networks, vol. 4, pp. 1942-1948, 1995.

[9] D. Karaboga, and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm," Applied soft computing, vol. 11, no. 1, pp. 652-657, 2011.

[10] B. Haddow, and G. Tufte, "Goldberg DE Genetic Algorithms in Search, Optimization and Machine Learning." In Proceedings of the 2000 Congress on, 2010.

[11] L. Das, and R. M. Singari, "Electrical Power Production Engineering Analysis and Heuristic Decisive Electrical Energy Tariff Determination," International Journal of Advanced Production and Industrial Engineering, pp. 42-46, 2017.

[12] M. Khishe and M. R. Mosavi, "Chimp optimization algorithm," Expert systems with applications, vol. 149, pp. 113338, 2020.

[13] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 36, no. 2, pp. 451-461, 2003.

[14] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," IEEE computational intelligence magazine, vol. 1, no. 4, pp. 28-39, 2006.

[15] Y. Kao and K. Cheng, "An ACO-based clustering algorithm." In International Workshop on Ant Colony Optimization and Swarm Intelligence, pp. 340-347, 2006.

[16] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in Engineering Software, vol. 69, pp. 46-61, 2014.

[17] S. Zhang and Y. Zhou, "Grey wolf optimizer based on Powell local optimization method for clustering analysis," Discrete Dynamics in Nature and Society, 2015.

[18] M.-Y. Cheng and D. Prayogo, "Symbiotic organisms search: a new metaheuristic optimization algorithm," Computers & Structures, vol. 139, pp. 98-112, 2014.

[19] Y. Zhou, H. Wu, Q. Luo, and M. Abdel-Baset, "Automatic data clustering using nature-inspired symbiotic organism search algorithm," Knowledge-Based Systems, vol. 163, pp. 546-557, 2019.

[20] S. Kumar, D. Datta, and S. K. Singh, "Black hole algorithm and its applications," Computational intelligence applications in modeling and control, pp. 147-170: Springer, 2015.

[21] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," Information sciences, vol. 222, pp. 175-184, 2013.

[22] G.-G. Wang, A. H. Gandomi, and A. H. Alavi, "Stud krill herd algorithm," Neurocomputing, vol. 128, pp. 363-370, 2014.

[23] L. M. Abualigah, A. T. Khader, E. S. Hanandeh, and A. H. Gandomi, "A novel hybridization strategy for krill herd algorithm applied to clustering techniques," Applied Soft Computing, pp. 423-435, 2017.

[24] S. Mirjalili, and A. Lewis, "The whale optimization algorithm," Advances in engineering software, vol. 95, pp. 51-67, 2016.

[25] J. Nasiri and F. Khiyabani, "A whale optimization algorithm (WOA) approach for clustering," Cogent Mathematics & Statistics, vol. 5, no. 1, pp. 1483565, 2018.

[26] P. Fränti, and S. Sieranoja, "K-means properties on six clustering benchmark datasets," Applied Intelligence, vol. 48, no. 12, pp. 4743-4759, 2018.

[27] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[28] A. Hatamlou, S. Abdullah, and H. Nezamabadi-Pour, "Application of gravitational search algorithm on data clustering." In International conference on rough sets and knowledge technology, pp. 337-346, 2011.

[29] A. Hatamlou, S. Abdullah, and M. Hatamlou, "Data clustering using big bang–big crunch algorithm." In International conference on innovative computing technology, pp. 383-388, 2011.

[30] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," Information sciences, vol. 222, pp. 175-184, 2013.

[31] Q. Li, and B. Liu, "Clustering using an improved krill herd algorithm," Algorithms, vol. 10, no. 2, pp. 56, 2017.

[32] S. Mirjalili, SM Mirjalili, and A. Hatamlou, "Multi-verse optimizer: a nature-inspired algorithm for global optimization," Neural Computing and Applications, pp. 495-513, 2016.

[33] Y. Zhou, H. Wu, Q. Luo, and M. Abdel-Baset, "Automatic data clustering using nature-inspired symbiotic organism search algorithm," Knowledge-Based Systems, pp. 546-57, 2019.

[34] X. S. Yang, "Flower Pollination Algorithm for Global Optimization. Unconventional Computation and Natural Computation," Springer Berlin Heidelberg, pp. 242 – 243, 2012.

[35] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," Knowledge-Based Systems, pp. 106-167, 2020.

[36] AC. Pandey, DS. Rajpoot, and M. Saraswat, "Data clustering using hybrid improved cuckoo search method," In 2016 Ninth International Conference on Contemporary Computing (IC3), pp. 1-6, 2016.

[37] A. Bouyer, H. Ghafarzadeh, and O. Tarkhaneh, "An efficient hybrid algorithm using cuckoo search and differential evolution for data clustering," Indian Journal of Science and Technology, 2015.