




An enhanced whale optimization algorithm for clustering

Hakam Singh¹ · Vipin Rai² · Neeraj Kumar² · Pankaj Dadheech³ · Ketan Kotecha⁴ · Ganeshsree Selvachandran⁵  · Ajith Abraham⁶

Received: 8 March 2022 / Revised: 18 May 2022 / Accepted: 2 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Clustering is a technique of grouping the data objects into clusters. Many metaheuristic algorithms based on swarm intelligence, physic laws, and chemical reactions, among others, have been developed for clustering. In this study, an enhanced whale optimization algorithm (EWOA) is introduced to solve clustering problems. The whale optimization algorithm (WOA) is adapted and enhanced with two additional operational procedures. The position update equations from the water wave optimization algorithm are incorporated into the algorithm to improve the search space and accelerate the convergence rate. The tabu and neighbourhood search mechanisms were added to handle the local optima situation. The efficiency of the proposed EWOA is measured using a simulation-based experiment conducted on eight benchmark datasets, and the results obtained are then compared to seven existing clustering algorithms/techniques. The performance of each algorithm is compared and analyzed using the average intra-cluster distance and f-measure parameters. The experimental results demonstrated the applicability and feasibility of the enhancements that were made and proved the superiority of the proposed EWOA clustering algorithm.

Keywords Clustering · Metaheuristic · Tabu search · Neighbourhood search · Whale optimization

1 Introduction

Data mining is a procedure that pulls previously unknown patterns or information from databases. It is also characterised as a descriptive analytics approach known as clustering, which discovers patterns based on specified dissimilarity criteria [9]. Due to extensive applicability, clustering has captured the attention of research communities who have worked to develop several evolutionary metaheuristic algorithms to solve clustering

✉ Ganeshsree Selvachandran
Ganeshsree@ucsiuniversity.edu.my; ganeshsree86@yahoo.com

problems (Dorigo et al. [5]; Cura [4]; Kumar and Sahoo [18–20]; Karaboga and Ozturk [16]; Hatamlou et al. [13]; Hatamlou [11]). Clustering methods optimally divide a set of data objects and retain them in clusters (Nanda and Panda [26]; Mat et al. [22]). The clustering process is carried with the help of some dissimilarity measures. The Euclidean distance given in Eq. (1) is an extensively accepted similarity measure in the partitioned clustering techniques. It is described as a sum of the square root of the difference between data objects and the cluster centres. The data objects are tailored into the clusters according to the distance values.

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

where Z_i symbolizes the i^{th} data instance/object, C_j represents the j^{th} cluster centre/centroid, whereas n and d denote the number of instances/data objects and dimension/attribute in the dataset, respectively.

In this study, an enhanced version of the whale optimization algorithm (WOA) is proposed to find optimized cluster centres. The WOA is a nature-inspired method that simulates the foraging behaviour of humpback whales (Mirjalili and Lewis [24]). The original WOA suffers from various issues such as the ones listed below:

- **Convergence rate:** The convergence rate is concentric around the search space mechanism and the coordination among exploration and exploitation processes is lacking (Kumar and Sahoo [19]).
- **Local optima:** It is a situation when the candidate solution is not getting an update and primarily occurs due to the absence of a population diversification mechanism. It is observed that the original WOA suffers from local optima (Kumar and Kaur [17]).

To overcome these and other problems that are inherent in the original WOA, this study proposes an improved algorithm called the Enhanced Whale Optimization Algorithm (EWOA). The EWOA is adapted from the original WOA and enhanced with two additional operational procedures to accelerate the convergence rate and overcome the local optima situation. To accelerate the convergence rate, position update equations from the water wave optimization algorithm are incorporated into the algorithm to improve the search space, while the tabu and neighbourhood search mechanisms were incorporated to overcome the local optima situation. The efficiency of the proposed EWOA is measured using a simulation-based experiment conducted on eight benchmark datasets, namely the Iris, Cancer, CMC, Wine, Glass, Thyroid, LR and ISOLET datasets, and the results obtained are then compared to seven existing clustering algorithms/techniques, namely the Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Cat Swarm Optimization (CSO), Genetic Algorithm (GA), Advanced Chemical Reaction Optimization (ACRO), WOA, and K-means algorithms. The performance of each algorithm is compared and analyzed using the average intra-cluster distance and f-measure parameters. The applicability and feasibility of the proposed algorithm is demonstrated via the experimental study

that has been carried out in this paper. The experimental results highlighted the enhancements that were made and proved the superiority of the proposed EWOA clustering algorithm.

The core contributions of this study are summarized as follows:

- i). Introduced the EWOA as an improvement to the original WOA by incorporating some key enhancements to overcome the problems that are inherent in the the original WOA.
- ii). Incorporate tabu and neighbourhood strategy to handle local optima situations.
- iii). Incorporate the position update equations from the water wave optimization algorithm to improve the search space, minimize the intra-cluster distance and accelerate the convergence rate.
- iv). Demonstrate the feasibility and applicability of the proposed EWOA model by implementing the algorithm in solving cluster analysis problems using eight experimental benchmark datasets.
- v). Proved the superiority of the proposed EWOA model by comparing the results of the experimental study obtained via the proposed EWOA model and seven other well-known clustering algorithms.

This paper is divided into 6 sections. The literature review and related works is presented in Section 2, while Section 3 describes the background details of the algorithms and methods that are employed in this paper. The improvements and enhancements that were done to the algorithm are detailed in Section 4, while the experimental study and the results that were obtained are presented in Section 5. Concluding remarks and the future scope of this study is presented in Section 6, followed by the list of declarations, acknowledgments and the list of references.

2 Literature review

Several algorithms have been developed, hybridized, and improved in the past few decades to solve clustering problems. Some of them are expounded in this section. Premalatha and Natarajan [27] proposed a PSO algorithm with an enhanced discrete binary PSO, where the model was tested using three datasets and compared with the K-means clustering algorithm, and the outcome was an improved execution and more diversity in the swarm. Kao et al. [15] studied partitional clustering problems using a hybrid optimization solution, while Chang et al. [3] eliminated the local optima and premature convergence problems of the standard genetic algorithm with their proposed gene rearrangement strategy. Jiang and Wang [14] presented a cooperative coevolution framework for BPSO algorithms, in which cooperative coevolution was used to decompose the problem into K subproblems, while PSO was used to solve these problems. Wang et al. [38] proposed a chaotic KH hybridized clustering method which had a better convergence speed. Kumar and Sahoo [18] proposed a hybrid data clustering algorithm by combining CSO and K-harmonic means algorithms, tested it against various existing algorithms, and concluded that the proposed hybrid model has im-

proved convergence speed. Kumar and Sahoo [19] introduced another hybrid algorithm that combined the PSO and MCSS algorithms, in which this hybrid model employed the neighbourhood technique to improve the search process. Menéndez et al. [23] proposed a multi medoid-based ACO clustering algorithm that automatically determines the optimum number of clusters and works without predetermined criteria, i.e. the number of clusters. Hatamlou [12] hybridized the PSO and the big bang-big crunch (BB-BC) algorithms to overcome the local optima and premature convergence problems.

Zhang et al. [41] and Karaboga and Ozturk [16] studied the use of the ABC algorithm that simulated the intelligent foraging behaviour of honey bee swarms in clustering problems and proved that the ABC algorithm is indeed efficient for solving multivariate data clustering problems. Yan et al. [40] introduced a hybrid variant of the ABC algorithm for solving clustering problems, in which the crossover operator of GA is integrated with the ABC algorithm to accelerate its convergence speed and achieve optimality in the solution faster. Alshamiri et al. [2] integrated the extreme learning machine (ELM) model into the ABC algorithm, in which the ELM model will project the input data into a high-dimensional feature space, while the ABC algorithm will perform the partitions. Kumar and Sahoo [20] proposed an efficient two-step ABC algorithm, in which the K-means algorithm is used to identify the initial seed points or food sources for the ABC algorithm.

Senthilnath et al. [31] implemented the firefly algorithm that simulates the social insects' behaviour and flash pattern of fireflies in solving clustering problems. Hatamlou [11] introduced a black hole (BH) phenomenon-based algorithm for clustering, in which the search space is defined in terms of the black hole, stars, and their absorption mechanism. The efficiency of the BH-based method was tested using standard datasets and was proven to be an effective clustering technique. Wang et al. [39] proposed a bee pollinator with a flower pollination algorithm to improve searchability and achieve faster convergence. Siddiqi et al. [32] introduced a new hybrid model that integrated the GA and SimE algorithms to automate the partitional clustering process. A greedy method is first applied to select the initial seed points, and optimization methods are then implemented to optimize them. Kushwaha et al. [21] proposed a magnetic force-based clustering algorithm in which a magnetic force-based search mechanism is implemented to find the optimal cluster centres. The data points are considered as particles and get rendered due to magnetic forces. The optimum position for the centroid particles is said to be achieved when the magnetic force applied by the data points approaches zero.

To automate the clustering process, Zhou et al. [43] projected the simplex method in the social spider algorithm. The simplex method is used to estimate and update the positions of the spiders. This stochastic variant strategy enhanced the population diversity and improved the local search ability of the traditional algorithm. Han et al. [10] hybridized the birds flock and gravitational search algorithm (BFGSA) to develop an efficient algorithm for partitional clustering that uses neighbourhood strategies to explore a broader range of search space. This hybrid model managed to overcome the local optima, handling of multidimensional data, and premature convergence problems. Ganguly [6] proposed a neighbour heuristic-based algorithm for cluster analysis, in

which a function was introduced to avoid the direct distance vectors computation and get the topmost similar vectors in this work. Singh et al. [35] introduced an artificial chemical reaction-based algorithm for partitional clustering problems, whereby neighbourhood and position-based operators were taught to overcome the deficiencies of traditional chemical reaction algorithms resulting in a more efficient clustering algorithm. Singh and Kumar [33] hybridized the ACRO algorithm with genetic operators, whereas Singh and Kumar [34] introduced a neighborhood search based on the CSO algorithm and applied this to solve clustering problems. Motwani et al. [25] developed three methods to generate the initial centroids for initial cluster selection and concluded that the farthest distributed centroid clustering algorithm produces quality clusters.

Santana-Velásquez et al. [30] focused on applying Machine Learning (ML) techniques as an alternative to DRG's traditional classification methods. The primary goal is to determine if ML techniques can categorize patients according to the DRGs criteria using information available during discharge. This data served as the foundation for subsequent research on the prediction of DRGs in the early phases of patients' hospitalization episodes. Stephan et al. [36] applied the HAW technique in an ANN model concurrently with feature selection (FS) and parameter optimization algorithms. Backpropagation learning was used to develop HAW in this study, which comprises robust backpropagation (HAW-RP), Levenberg–Marquart (HAW-LM), and momentum-based gradient descent (HAW-GD) methods. The accuracy, complexity, and computation time of this hybrid model was studied using several breast cancer datasets. Goyal et al. [8] applied various optimization algorithms such as the particle swarm optimization (PSO), cat swarm optimization (CSO), BAT, cuckoo search algorithm (CSA) optimization algorithm, and whale optimization algorithm (WOA) for load balancing, energy efficiency, and better resource scheduling to create an efficient cloud environment. The study found that the WOA beat all the other algorithms in response time, energy consumption, execution time, and throughput in the scenario of seven servers and eight server configurations.

Stephan et al. [37] proposed a novel hybrid Artificial Bee Colony (hybrid ABC) optimization algorithm where the strong explorative capabilities of the chemotaxis phase of the bacterial foraging optimization were integrated with a spiral model-based exploitative phase of the ABC algorithm. This enabled the proposed hybrid ABC algorithm to overcome the demerits of poor exploration procedures in the standard ABC algorithm and outperform the corresponding standalone ABC algorithm. Rahnema and Gharehchopogh [29] proposed an improved version of the ABC algorithm based on the swarm intelligence characteristic of whales and found that random memory and elite memory enhanced the convergence speed of the improved algorithm. Ghany et al. [7] combined the WOA with the tabu search method. The tabu search enabled the WOA to store multiple best solutions and utilize them to explore the solution space more effectively. Purushothaman et al. [28] combined the Gray wolf optimization and grasshopper algorithms for clustering. This hybridization improved reliability and reduced computational time. Ahmadi et al. [1] modified the Gray wolf optimization algorithm by introducing a balanced approach to exploration and exploitation and centers around the best solution, and showed that the proposed algorithm produced state-of-the-art results

with a higher accuracy rate. Kumar and Kaur [17] introduced three new variants of the bat algorithm that managed to resolve problems related to initial cluster selection, convergence rate, and local optima with the help of enhanced cooperative evolution, elitist, and neighbourhood search strategies. These enhancements resulted in a robust partitional clustering algorithm.

All these innovations to the existing bio-inspired algorithms were proven to have improved efficiency, faster convergence rate, shorter computation time, and higher accuracy when compared to the corresponding standard, standalone bio-inspired algorithms.

3 Methodology

This section gives the background description of the algorithms and methods that have been implemented in this work. The Enhanced Whale Optimization Algorithm (EWOA) has been successfully utilized in the field of clustering to produce optimal cluster centres. The dataset is first put into memory, and the fundamental parameters are then configured. Following that, other sequential processes, such as sampling or cluster centre selection, goal function computation, assignment of data items to appropriate clusters, and updating of points are performed.

3.1 Whale optimization algorithm

The whale optimization algorithm is a nature-inspired algorithm that simulated the foraging behaviour of humpback whales [24]. Although it was initially designed to solve numerical problems, it was soon applied to several other domains such as clustering, due to its self-exploratory nature and ability to achieve convergence at a faster rate. The formulated mathematical model stimulated the prey identification and hunting strategies of humpback whales. The prey finding and encircling processes are modelled using Eqs. (2) and (3):

$$\vec{D} = \left| \vec{C}_{cv} \vec{Z}^*(t) - \vec{Z}(t) \right| \quad (2)$$

$$\vec{Z}(t+1) = \vec{Z}^*(t) - \vec{A}_{cv} \cdot \vec{D} \quad (3)$$

where $\vec{A}_{cv} = 2\vec{a} \cdot r - \vec{a}$, $\vec{C}_{cv} = 2r$. The terms \vec{Z} and \vec{Z}^* denote the current position vector and global best position vector, respectively, \vec{C}_{cv} and \vec{A}_{cv} are coefficient vectors, r is a rand (0, 1) function, a is linearly decreased from 2 to 0 over the iterations.

The bubble-net attacking process is a combination of shrinking encircling, and spiral position update methods. In shrinking encircling, the coefficient vectors get varied to simulate the humpback whale behaviour. At the same time, in the spiral position update method, the formulated spiral equation is trailed to find the helix-shaped movement of whales as denoted

by Eqs. (4) and (5). The humpback whales perform shrinking encircling or spiral movements that can be calculated using Eq. (6).

$$\overrightarrow{D'} = \left| \overrightarrow{Z^*}(t) - \overrightarrow{Z}(t) \right| \quad (4)$$

$$\overrightarrow{Z}(t+1) = \overrightarrow{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \overrightarrow{Z^*}(t) \quad (5)$$

$$\overrightarrow{Z}(t+1) = \begin{cases} \overrightarrow{Z^*}(t) - \overrightarrow{A_{cv}} \cdot \overrightarrow{D} & \text{if } p < 0.5 \\ \overrightarrow{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \overrightarrow{Z^*}(t), & \text{if } p \geq 0.5 \end{cases} \quad (6)$$

Here, \overrightarrow{D} is a distance vector, b is a constant vector, and l is a rand $[-1,1]$, and p is a rand $(0,1)$ function. The humpback whale search preys randomly in search space. The movements of the whale results in the change in vector location as denoted by Eqs. (7) and (8).

$$\overrightarrow{D} = \left| \overrightarrow{C_{cv}} \cdot \overrightarrow{Z_{rand}} - \overrightarrow{Z} \right| \quad (7)$$

$$\overrightarrow{Z}(t+1) = \overrightarrow{Z_{rand}} - \overrightarrow{A_{cv}} \cdot \overrightarrow{D} \quad (8)$$

The term $\overrightarrow{Z}(t+1)$ represents a new position vector, while the term $\overrightarrow{Z_{rand}}$ denotes a randomly chosen vector.

3.2 Water wave optimization algorithm (WWOA)

Recently, a water wave theory-based optimization algorithm was introduced for solving global optimization problems (Zheng [42]). This algorithm inherits the propagation, refraction, and breaking phenomena of water waves for searching and optimization. In water wave propagation operations, the new water waves are generated using Eq. (9) while the wavelength, λ is calculated using Eq. (10).

$$\overrightarrow{Z}(t+1) = \overrightarrow{Z} + \text{rand}(-1, 1) \times \lambda \times L_d \quad (9)$$

$$\lambda = \lambda \times \alpha^{\frac{(f(x) - f_{min} + \epsilon)}{(f_{max} - f_{min} + \epsilon)}} \quad (10)$$

Here, L_d is the length of the search space ($1 \leq d \leq n$), λ is the wavelength, f_{min} and f_{max} are the minimum and maximum fitness values, respectively, α is the wavelength dropping factor, and ε is a fixed constraint.

3.3 Tabu search

Tabu search is an elite list-based global optimization technique. The starting solutions are stored in the list and iteratively compared with the upcoming solutions. If an improved solution is obtained, the previous/starting solution is updated/ replaced with a better solution. The implementation of tabu search avoids re-entering previously explored regions and uses a single point for exploration [42].

3.4 Neighbourhood strategy

The neighbourhood strategy is used to enhance the searchability of the algorithm and increases the probability of finding a new solution. This primary centers around the neighbouring solutions and uses them to generate new solutions [30].

4 Proposed work: An enhanced whale optimization algorithm (EWOA) for partitional clustering

This section detailed the EWOA for solving partitional clustering problems. In this study, two improvements are proposed: (i) The propagation method is incorporated into the whale optimization algorithm; (ii) An integrated strategy is proposed to handle the local optima situation. A detailed description is given below.

4.1 Improvements in search space mechanism

The whale optimization algorithm is incorporated with an additional exploration mechanism to enhance searchability. The random prey search operation of the whale optimization algorithm is replaced with the propagation method of the water wave optimization algorithm given in Eqs. (9) and (10). The explorative search mechanism of the water wave algorithm is utilized to generate the new location vector and diversify the solution.

4.2 Integration of tabu and neighbourhood search strategies

In the second improvement, an integrated strategy based on tabu, and neighbourhood search is designed and implemented to solve local optima and nullify premature convergence problems. Here, the tabu list is extended to store N number of global best $Z_{N, gbest}$ positions in it. These $Z_{N, gbest}$ best positions are used as neighbouring points in the neighbourhood search strategy. Afterward, to generate a single point, the harmonic means of $Z_{N, gbest}$ point is calculated. To understand in a better way, assume that, $Z_{tabu, gbest}$ are a tabus list that stores (N) number of global best data points ($Z_{N, gbest}$). These data points are used as neighbouring points $Z_{i, neigh} =$

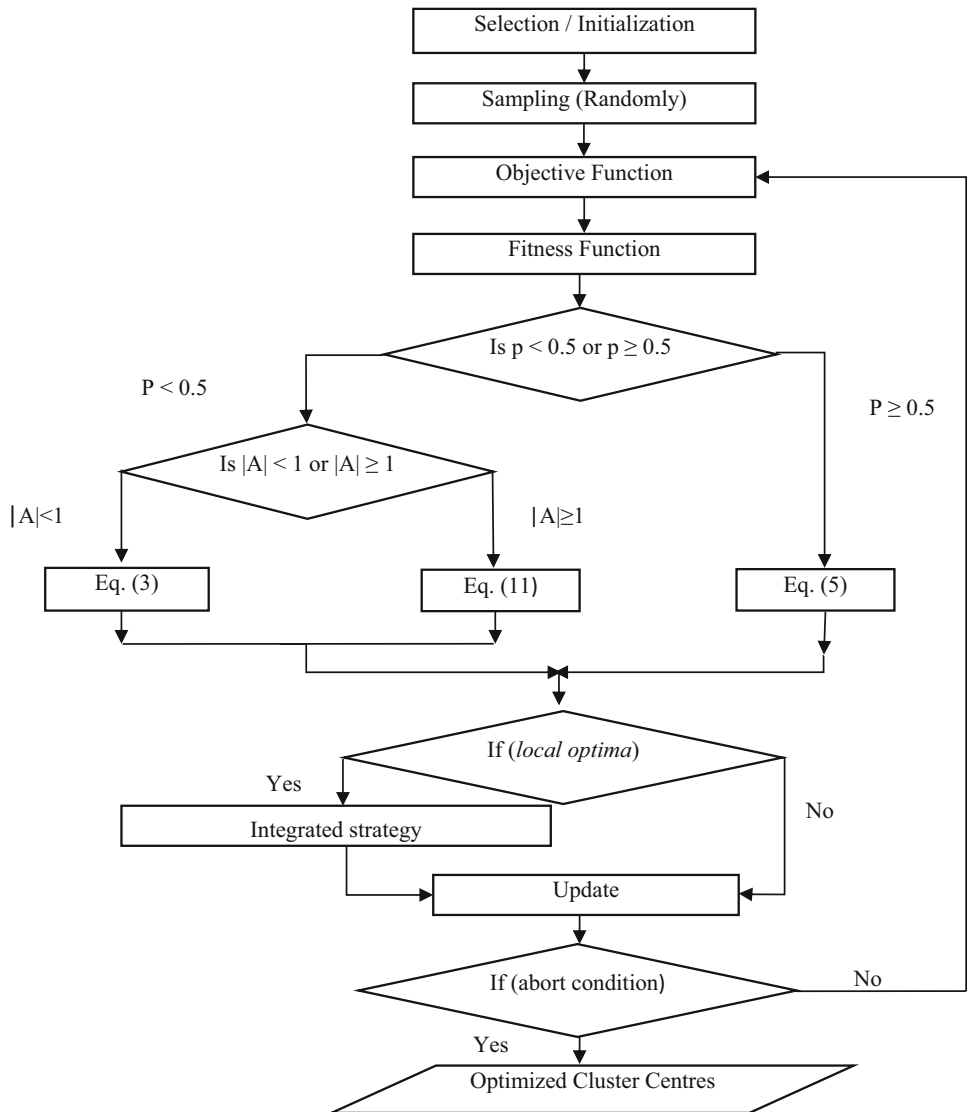


Fig. 1 Flow chart of EWOA

$\{Z_{1, \text{gbest}}, Z_{2, \text{gbest}}, \dots, Z_{N, \text{gbest}}\}$, where $N = 1, 2, \dots, 9$, and the harmonic mean of these neighbouring data points is calculated, $Z_{\text{new}} = \text{Harmonic mean of } (Z_{N, \text{neigh}})$ is used to generate a new data point.

4.3 Proposed EWOA model in solving clustering problems

The enhanced whale optimization algorithm is successfully implemented in the clustering field to achieve the optimal cluster centres. Initially, the dataset is loaded in memory, and basic parameters are initialized. Afterward, the different consecutive operations, sampling or cluster centre selection, objective function computation, assignments of data objects to respective

clusters, updates, and others are followed. The pseudo-code of the proposed algorithm is detailed in Algorithm 1 and graphically presented in Fig. 1.

Algorithm 1: The pseudo-code of EWOA for solving clustering problems	
Input: Dataset and predefined parameters.	
Output: Optimized cluster centres.	
1:	Select the dataset, initialize the basic parameters like the number of clusters $K_i \in (i = 1, 2, \dots, n)$, iterations, and others.
2:	Select the (K_i) initial centroids or population (K samples) from a dataset (Random fashion).
3:	Compute the objective function Eq. (1).
4:	Allocate data objects to clusters aggregating the minimum objective function values and compute their fitness Eq. (11). $\text{Fitness}(\vec{X}) = \sum_{j \in 1}^K \frac{SSE(\vec{Z}\vec{C}_j)}{\sum_{j=1}^K SSE(\vec{Z}\vec{C}_j)} \quad (11)$
5:	Generate new vector solution
	<p>If ($p < 0.5$)</p> <p style="padding-left: 40px;">If $A < 1$ then</p> <p style="padding-left: 80px;">Update the search vector using Eq. (3).</p> <p style="padding-left: 40px;">Else if $A \geq 1$ then</p> <p style="padding-left: 80px;">Update the search vector using Eq. (9).</p> <p>End if</p> <p style="padding-left: 40px;">Else if ($p \geq 0.5$)</p> <p style="padding-left: 80px;">Update the search vector using Eq. (5).</p>
6:	<p>Check for local optima</p> <p>If (local optima)</p> <p style="padding-left: 40px;">Apply integrated strategy</p> <p>Else continue</p>
7:	Update the search vector.
8:	Check the induced abort condition, i.e., iteration number, if fulfilled, abort execution, else carry-on steps 3-8.
9:	Optimal solution
Here SSE is the “sum of squared intra-cluster Euclidean distance”; C_j is the j^{th} cluster centre.	

4.4 Toy example

The working of EWOA algorithm in the clustering field is exemplified using an artificial dataset. The artificial dataset (9,3,4) contains 9 data instances, 3 classes, and 4 attributes.

Step 1. Load dataset and specify number of clusters ($K = 3$), total population = 9, no of iterations = 10.

5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.9	3.1	4.9	1.5
6.3	3.3	6	2.5
5.8	2.7	5.1	1.9
7.1	3	5.9	2.1

Step 2. Randomly selected initial cluster centres.

4.7000	3.2000	1.3000	0.2000
6.9000	3.1000	4.9000	1.5000
5.8000	2.7000	5.1000	1.9000

Step 3. Evaluate the objective function.

0.5099	4.1641	4.2083
0.3000	4.2367	4.1809
0.0000	4.4159	4.3347
4.2767	0.2646	1.4491
3.8497	0.6481	1.063
4.4159	0.000	1.253
5.4727	1.6155	1.3342
4.3347	1.253	0.000
5.529	1.1874	1.5684

Step 4. Assign data objects to clusters according to minimum objective function values.

0.5099	4.1641	4.2083
0.3000	4.1809	4.2367
0.0000	4.3347	4.4159
0.2646	1.4491	4.2767
0.6481	1.063	3.8497
0.0000	1.253	4.4159
1.3342	1.6155	5.4727
0.0000	1.253	4.3347
1.1874	1.5684	5.529

The index values of the clusters are:

1	2	3
1	3	2
1	3	2
2	3	1
2	3	1
2	3	1
3	2	1
3	2	1
2	3	1

Step 5. Generated cluster centres in 9th iteration.

5.1000	3.5000	1.4000	0.2000
6.3000	3.3000	6.0000	2.5000
7.1000	3.0000	5.9000	2.1000

Step 6. Check for local optima.

Step 7. Update the candidate solution.

Step 8. Check the ‘Stop’ criteria. If requirements are met, ‘Stop’, else repeat steps 3–8.

Step 9. Optimal solution.

5.1000	3.5000	1.4000	0.2000
6.3000	3.3000	6.0000	2.5000
7.1000	3.0000	5.9000	2.1000

5 Experimental results and analysis

This section provides a detailed description of simulation results and parameter settings for the EWOA. The simulation is performed in the MATLAB 2016 environment, configured on a Windows 10 OS, processor intel i3, 8 GB RAM equipped machine. The performance of the proposed EWOA is measured on eight datasets, and the characteristics are detailed in Table 1. The results are compared with seven clustering algorithms, namely the PSO, ACO, CSO, GA, ACRO, WOA, and K-means clustering algorithms. The user-defined parameters setting of

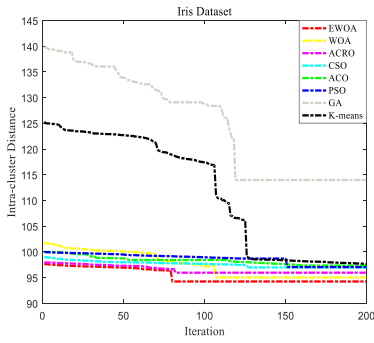
Table 1 Description of datasets

Datasets [Description]	D (Dimension/ attributes)	N (Instances)	K(Centroids/ Classes)
Iris [Fisher’s iris data]	4	150	3
Cancer [Breast cancer data]	9	683	2
CMC [Contraceptive method choice data]	9	1473	3
Wine [Wine data]	13	178	3
Glass [Glass identification data]	9	214	6
Thyroid [Thyroid disease data]	5	215	3
LR [Letter-Recognition data]	16	20,000	26
ISOLET [Isolated letter speech recognition data]	617	7797	26

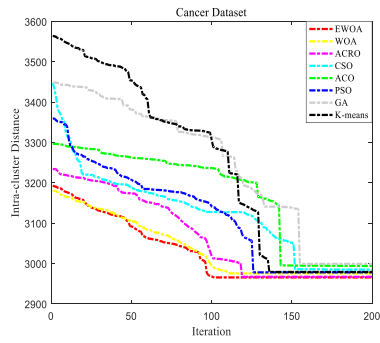
Table 2 Performance comparison of EWOA and other well-known clustering algorithms

Sr No	Parameters	Algorithms							
		K-means	GA	PSO	ACO	CSO	ACRO	WOA	EWOA
Iris	ICD_Avg	113.56	125.19	98.73	98.36	97.64	96.73	96.79	95.8
	FM_Avg	0.781	0.774	0.78	0.778	0.781	0.785	0.784	0.786
Cancer	ICD_Avg	3248.25	3249.46	3116.64	3178.09	3124.15	3063.34	3036.12	3034.53
	FM_Avg	0.832	0.819	0.826	0.829	0.831	0.835	0.822	0.835
CMC	ICD_Avg	5912.46	5756.59	5846.63	5831.25	5804.52	5746.32	5539.72	5587.14
	FM_Avg	0.337	0.324	0.333	0.332	0.334	0.339	0.337	0.337
Wine	ICD_Avg	18,059.91	16,530.53	16,491.52	16,526.12	16,486.21	16,334.85	16,295	16,293.391
	F-measure	0.52	0.515	0.517	0.521	0.522	0.526	0.522	0.526
Glass	ICD_Avg	246.51	282.32	278.71	281.46	264.44	266.23	231.29	229.29
	FM_Avg	0.426	0.333	0.412	0.402	0.416	0.428	0.419	0.429
Thyroid	ICD_Avg	1995.189	1888.209	1890.2071	1990.02	1960.06	1899.76	1870.93	1866.67
	FM_Avg	0.701	0.774	0.768	0.782	0.781	0.785	0.786	0.788
LR	ICD_Avg	624,765.58	611,731.68	608,470.77	608,495.87	611,102.88	604,612.52	604,602.34	604,600.42
	FM_Avg	0.461	0.488	0.412	0.427	0.416	0.441	0.439	0.452
ISOLET	ICD_Avg	446,502.65	460,851.88	451,718.88	455,837.78	447,733.55	441,268.61	441,361.25	441,240.53
	FM_Avg	0.361	0.332	0.392	0.301	0.311	0.408	0.405	0.411

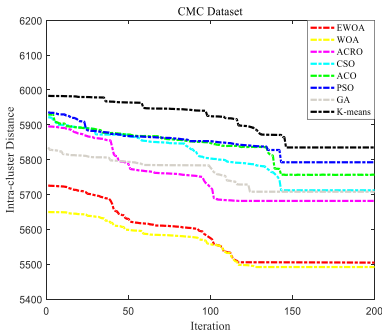
ICD_Avg (Intra-cluster distance average case) and FM_Avg (F-measure average case)



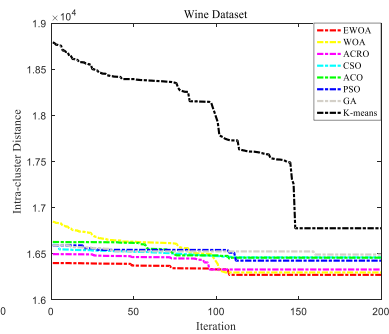
a) Convergence on Iris dataset



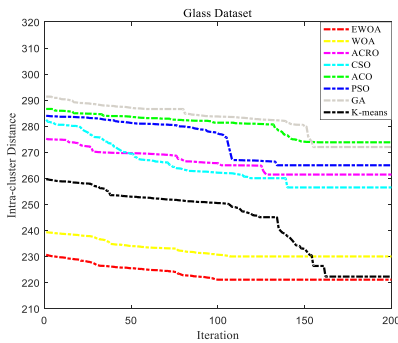
b) Convergence on Cancer Dataset



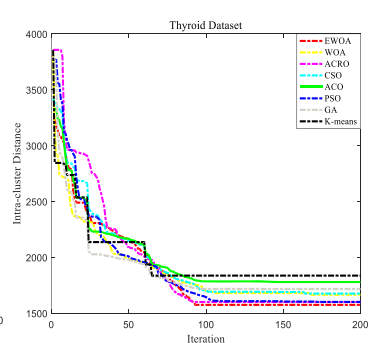
c) Convergence on CMC dataset



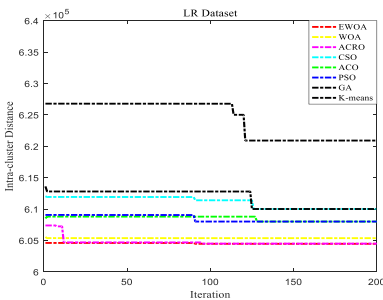
d) Convergence on Wine Dataset



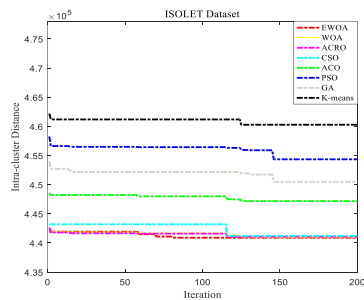
e) Convergence on Glass Dataset



f) Convergence on Thyroid Dataset



g) Convergence on LR Dataset



h) Convergence on ISOLET Dataset

EWOA are defined as population = $K \times d$, number of clusters or groups = K , $A = [-1, 1]$, random function (0,1) and length of search space ($1 \leq d \leq n$), iterations = 200. The algorithms run thirty times individually, and the results are evaluated as an average case of performance parameters (intra-cluster distance and f-measure).

5.1 Results and discussion

This subsection presents a comparative analysis and convergence behaviour of EWOA and other clustering algorithms. Table 2 presents the performance comparison of K-means, GA, PSO, ACO, CSO, ACRO, WOA, and EWOA using average intra-cluster distance and f-measure parameters. From simulation outcomes, it is observed that the EWOA obtain minimum intra-cluster distance values except for the CMC datasets. Further, the f-measure is also computed to assess the classification of data objects to corresponding clusters. The EWOA attained a healthy f-measure rate except for CMC and LR datasets. For the CMC dataset, ACRO algorithm, and LR dataset, GA has superior results.

The convergence behavior of AWOA, WOA, ACRO, CSO, ACO, PSO, GA, and K-means clustering algorithms are depicted in Fig. 2a-h. The x -axis shows the number of iterations, and the y -axis shows the intra-cluster distance. From graphs, it is revealed that EWOA converges on a minor level except for the CMC dataset. Although in most aspects, the EWOA provides a better convergence rate.

Except for the CMC and LR datasets, the EWOA achieved a better f-measure rate. GA outperforms the CMC dataset, the ACRO method, and the LR dataset.

Figure 2a-h shows the convergence behavior of EWOA and other clustering algorithms. The convergence on ISOLET dataset is described with the Number of Iterations v/s Intra Cluster Distance, compared with the EWOA, WOA, ACRO, CSO, ACO, PSO, GA & K-means.

5.2 Statistical analysis

The Friedman statistical test is carried out to prove the significance of the results and verify the feasibility of the newly proposed algorithm. Here two hypotheses (null hypothesis (H_0) and alternative hypothesis (H_1)) are projected; the H_0 expresses that the algorithms have similar performance; the H_1 expressese that algorithms have dissimilar performance. Table 3 shows the statistical analysis using the intra-cluster distance parameter. The test shows that the critical value is 14.067144, and the p value is 7.12E-07 at a significance level from 0.5. These values are consistent with the test which showed that the null hypothesis (H_0) is rejected, hence proving that the algorithms have dissimilar performances. The EWOA was also found to have significantly distinct performances compared to the other algorithms that are compared in this study.

Table 4 shows the statistical analysis using the f-measure parameter. The EWOA gets the first rank except for CMC and LR datasets. However, for cancer, wine and balance, it is

Fig. 2 **a** Convergence on Iris dataset **b** Convergence on Cancer Dataset **c** Convergence on CMC dataset **d** Convergence on Wine Dataset **e** Convergence on Glass Dataset **f** Convergence on Thyroid Dataset **g** Convergence on LR Dataset **h** Convergence on ISOLET Dataset

Table 3 Statistical analysis using intra-cluster distance

Datasets	Clustering Algorithms							
	K-means	GA	PSO	ACO	CSO	ACRO	WOA	EWOA
Iris	7	8	6	5	4	2	3	1
Cancer	7	8	4	6	5	3	2	1
CMC	8	4	7	6	5	3	1	2
Wine	8	7	5	6	4	3	2	1
Glass	3	8	6	7	4	5	2	1
Thyroid	8	3	4	7	6	5	2	1
LR	8	7	4	5	6	3	2	1
ISOLET	4	8	6	7	5	2	3	1
Sum	53	53	42	49	39	26	17	9
Rank	6.63	6.63	5.25	6.13	4.88	3.25	2.13	1.13
NSs: 64			NP: 08		NA: 8			
SSRS: 12350			CF: 1296		FDS: 41.2916			
DF: 7			p value: 7.12E-07		CV: 14.067144			

NSs Number of observations, NPs Number of problems, NAs Number of algorithms, SSRS Sum of squares of ranks sums, CF Correction factor, FTS Friedman test statistic, DF Degree of freedom and CV Critical value.

approximately equal to the ACRO algorithm. The critical value is 14.0671 that shows the significant difference among algorithms.

Table 4 shows the statistical analysis using the f-measure parameter. The EWOA gets the first rank for all the datasets except for the CMC and LR datasets. However, for the cancer, wine and balance datasets, the results obtained from the EWOA was found to be approximately equal to the results from the ACRO algorithm. The critical value of 14.0671 indicates that there is a significant difference in the performance of the algorithms.

Table 4 Statistical analysis using F-measure

Datasets	Clustering Algorithms							
	K-means	GA	PSO	ACO	CSO	ACRO	WOA	EWOA
Iris	4.5	8	6	7	4.5	2	3	1
Cancer	3	8	6	5	4	1.5	7	1.5
CMC	3	8	6	7	5	1	3	3
Wine	6	8	7	5	3.5	1.5	3.5	1.5
Glass	3	8	6	7	5	2	4	1
Thyroid	8	6	7	4	5	3	2	1
LR	2	1	8	6	7	4	5	3
ISOLET	5	6	4	8	7	2	3	1
Sum	34.5	53	50	49	41	17	30.5	13
Rank	4.31	6.63	6.25	6.13	5.13	2.13	3.81	1.63
NSs: 64			NP: 08			NA: 8		
SSRS: 11969.5			CF: 1296			FTS: 33.7665		
DF: 7			p value: 1.90E-05			CV: 14.067144		

6 Conclusion and future work

In this study, an improvement to the original WOA called the Enhanced Whale Optimization Algorithm (EWOA) has been developed for solving clustering problems. This improved algorithm has proven to be able to overcome the problems that are inherent in the original WOA, namely the slower convergence rate due to the convergence being concentric around the search space mechanism and the local optima situation. To overcome the problems that are inherent in the WOA, the EWOA is enhanced with two additional operational procedures to accelerate the convergence rate and overcome the local optima situation. Minimum intra-cluster distance and an accelerated convergence rate was achieved through the implementation of position update equations from the water wave optimization algorithm that were incorporated into the algorithm to improve the search space, whereas the local optima situation was overcome by implementing the tabu and neighbourhood search strategies into the algorithm. The efficiency of the proposed EWOA was measured using a simulation-based experimental study that was conducted on eight benchmark datasets, namely the Iris, Cancer, CMC, Wine, Glass, Thyroid, LR and ISOLET datasets. The results obtained were then compared to the results obtained via seven existing clustering algorithms/techniques, namely the PSO, ACO, CSO, GA, ACRO, WOA, and K-means algorithms. The performance of each algorithm was compared and analyzed using the average intra-cluster distance and f-measure parameters. The results obtained clearly showed the applicability and feasibility of the enhancements that were made to the EWOA and the superiority of the proposed EWOA model in solving clustering problems compared to the existing models/methods. The future scope of this work involves the application of the proposed EWOA model in solving problems related to vehicular networks for cluster head formation and load balancing.

Acknowledgments The authors would like to thank the Editors and the anonymous reviewers for their valuable comments and suggestions which has helped to improve the quality and clarity of the paper. The authors would also like to acknowledge the assistance rendered by Dr. Cherry Bhargava for the general supervision of the research group and general administrative support.

Data Availability The data that support the findings of this study are available upon request from the corresponding authors.

Author contributions All authors contributed to the conception and design of the study. Material preparation, data collection, data visualization and data analysis were performed by Hakam Singh, Vipin Rai, Neeraj Kumar, and Pankaj Dadheech. Advanced data analysis and validation were done by Ketan Kotecha, Ganeshsree Selvachandran and Ajith Abraham. The first draft of the manuscript was written by Hakam Singh, Vipin Rai, Neeraj Kumar, and Pankaj Dadheech. The second draft was prepared and edited by Ganeshsree Selvachandran and Ajith Abraham. All authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

Funding This work was supported by the Ministry of Education, Malaysia under grant no. FRGS/1/2020/STG06/UCSI/02/1.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethical compliance Authors' declaration: This manuscript is the authors' original work and has not been published elsewhere. All authors have checked the manuscript and have agreed to this submission.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References


1. Ahmadi R, Ekbatanifard G, Bayat P (2021) A modified grey wolf optimizer based data clustering algorithm. *Appl Artif Intell* 35(1):63–79
2. Alshamiri AK, Singh A, Surampudi BR (2016) Artificial bee colony algorithm for clustering: an extreme learning approach. *Soft Comput* 20(8):3163–3176
3. Chang DX, Zhang XD, Zheng CW (2009) A genetic algorithm with gene rearrangement for K-means clustering. *Pattern Recogn* 42(7):1210–1222
4. Cura T (2012) A particle swarm optimization approach to clustering. *Expert Syst Appl* 39(1):1582–1588
5. Dorigo M, Birattari M, Stutzle T (2006) Ant colony optimization. *IEEE Comput Intell Mag* 1(4):28–39
6. Ganguly D (2018) A fast partitional clustering algorithm based on nearest neighbours heuristics. *Pattern Recogn Lett* 112:198–204
7. Ghany KKA, AbdelAziz AM, Soliman THA, Sewisy AAEM (2022) A hybrid modified step whale optimization algorithm with Tabu search for data clustering. *Journal of King Saud University-Computer and Information Sciences* 34(3):832–839
8. Goyal S, Bhushan S, Kumar Y, Rana AUHS, Bhutta MR, Ijaz MF, Son Y (2021) An optimized framework for energy-resource allocation in a cloud environment based on the whale optimization algorithm. *Sensors* 21(5):1583
9. Han J, Pei J, Kamber M (2011) *Data mining: concepts and techniques*. Elsevier
10. Han X, Quan L, Xiong X, Almeter M, Xiang J, Lan Y (2017) A novel data clustering algorithm based on modified gravitational search algorithm. *Eng Appl Artif Intell* 61:1–7
11. Hatamlou A (2013) Black hole: A new heuristic optimization approach for data clustering. *Inf Sci* 222:175–184
12. Hatamlou A (2017) A hybrid bio-inspired algorithm and its application. *Appl Intell* 47:1059–1067
13. Hatamlou A, Abdullah S, Hatamlou M (2011) Data clustering using big bang–big crunch algorithm. In: Pichappan P, Ahmadi H, Ariwa E (eds) *Innovative computing technology*. INCT 2011. Communications in Computer and Information Science, vol 241. Springer, Berlin, Heidelberg, pp 383–388. https://doi.org/10.1007/978-3-642-27337-7_36
14. Jiang B, Wang N (2014) Cooperative bare-bone particle swarm optimization for data clustering. *Soft Comput* 18(6):1079–1091
15. Kao YT, Zahara E, Kao IW (2008) A hybridized approach to data clustering. *Expert Syst Appl* 34(3):1754–1762
16. Karaboga D, Ozturk C (2011) A novel clustering approach: artificial bee colony (ABC) algorithm. *Appl Soft Comput* 11(1):652–657
17. Kumar Y, Kaur A (2021) Variants of bat algorithm for solving partitional clustering problems. *Eng Comput*. <https://doi.org/10.1007/s00366-021-01345-3>
18. Kumar Y, Sahoo G (2015) A hybrid data clustering approach based on improved cat swarm optimization and K-harmonic mean algorithm. *AI Commun* 28(4):751–764
19. Kumar Y, Sahoo G (2015) Hybridization of magnetic charge system search and particle swarm optimization for efficient data clustering using neighborhood search strategy. *Soft Comput* 19(12):3621–3645
20. Kumar Y, Sahoo G (2015) A two-step artificial bee colony algorithm for clustering. *Neural Comput & Applic* 28(3):537–551
21. Kushwaha N, Pant M, Kant S, Jain VK (2018) Magnetic optimization algorithm for data clustering. *Pattern Recogn Lett* 115:59–65
22. Mat AN, İnan O, Karakoyun M (2021) An application of the whale optimization algorithm with levy flight strategy for clustering of medical datasets. *International Journal of Optimization and Control: Theories & Applications* 11(2):216–226
23. Menéndez HD, Otero FE, Camacho D (2016) Medoid-based clustering using ant colony optimization. *Swarm Intelligence* 10(2):123–145
24. Mirjalili S, Lewis A (2016) The whale optimization algorithm. *Adv Eng Softw* 95:51–67
25. Motwani M, Arora N, Gupta A (2019) A study on initial centroids selection for partitional clustering algorithms. In: Hoda M, Chauhan N, Quadri S, Srivastava P (eds) *Software engineering*. Advances in

- intelligent systems and computing, vol 731. Springer, Singapore, pp 211–220. https://doi.org/10.1007/978-981-10-8848-3_21
26. Nanda SJ, Panda G (2014) A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation* 16:1–18
 27. Premalatha K, Natarajan AM (2008) A new approach for data clustering based on PSO with local search. *Computer and Information Science* 1(4):139–145
 28. Purushothaman R, Rajagopalan SP, Dhandapani G (2020) Hybridizing gray wolf optimization (GWO) with grasshopper optimization algorithm (GOA) for text feature selection and clustering. *Appl Soft Comput* 96: 106651
 29. Rahnama N, Gharehchopogh FS (2020) An improved artificial bee colony algorithm based on whale optimization algorithm for data clustering. *Multimed Tools Appl* 79(43):32169–32194
 30. Santana-Velásquez, A., John Freddy Duitama, M., & Arias-Londoño, J.D. (2020). Classification of diagnosis-related groups using computational intelligence techniques. *Proceedings of the 2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020)*, 2020, pp. 1–6, <https://doi.org/10.1109/ColCACI50549.2020.9247889>.
 31. Senthilnath J, Omkar SN, Mani V (2011) Clustering using firefly algorithm: performance study. *Swarm and Evolutionary Computation* 1(3):164–171
 32. Siddiqi UF, Sait SM (2017) A new heuristic for the data clustering problem. *IEEE Access* 5:6801–6812
 33. Singh H, Kumar Y (2020) Hybrid artificial chemical reaction optimization algorithm for cluster analysis. *Procedia Computer Science* 167:531–540
 34. Singh H, Kumar Y (2020) A neighborhood search based cat swarm optimization algorithm for clustering problems. *Evol Intel* 13(4):593–609
 35. Singh H, Kumar Y, Kumar S (2019) A new meta-heuristic algorithm based on chemical reactions for partitional clustering problems. *Evol Intel* 12(2):241–252
 36. Stephan P, Stephan T, Kannan R, Abraham A (2021) A hybrid artificial bee colony with whale optimization algorithm for improved breast cancer diagnosis. *Neural Comput & Applic* 33:13667–13691
 37. Stephan P, Stephan T, Gandomi AH (2022) A novel breast cancer diagnosis scheme with intelligent feature and parameter selections. *Comput Methods Prog Biomed* 214:106432
 38. Wang GG, Guo L, Gandomi AH, Hao GS, Wang H (2014) Chaotic krill herd algorithm. *Inf Sci* 274:17–34
 39. Wang R, Zhou Y, Qiao S, Huang K (2016) Flower pollination algorithm with bee pollinator for cluster analysis. *Inf Process Lett* 116(1):1–14
 40. Yan X, Zhu Y, Zou W, Wang L (2012) A new approach for data clustering using hybrid artificial bee colony algorithm. *Neurocomputing* 97:241–250
 41. Zhang C, Ouyang D, Ning J (2010) An artificial bee colony approach for clustering. *Expert Syst Appl* 37(7): 4761–4767
 42. Zheng YJ (2015) Water wave optimization: A new nature-inspired metaheuristic. *Comput Oper Res* 55:1–11
 43. Zhou Y, Zhou Y, Luo Q, Abdel-Basset M (2017) A simplex method-based social spider optimization algorithm for clustering analysis. *Eng Appl Artif Intell* 64:67–82

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Affiliations

Hakam Singh¹ • Vipin Rai² • Neeraj Kumar² • Pankaj Dadheech³ • Ketan Kotecha⁴ • Ganeshsree Selvachandran⁵  • Ajith Abraham⁶

Hakam Singh
hakam.singh@chitkarauniversity.edu.in

Vipin Rai
vipin.raai@chitkara.edu.in

Neeraj Kumar
kumar.neeraj@chitkara.edu.in

Pankaj Dadheech
pankajdadheech777@gmail.com

Ketan Kotecha
director@sitpune.edu.in

Ajith Abraham
ajith.abraham@ieee.org

¹ Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India

² Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

³ Swami Keshvanand Institute of Technology, Management & Gramothan (SKIT), Jaipur, Rajasthan 302017, India

⁴ Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, MH 412115, India

⁵ Faculty of Business and Management, UCSI University, Jalan Menara Gading, 56000 Cheras, Kuala Lumpur, Malaysia

⁶ Machine Intelligence Research Labs, Auburn, WA 98071, USA

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com