Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 03, September 2022.

RAT SWARM OPTIMIZER FOR DATA CLUSTERING

Ibrahim Zebiri, Djamel Zeghida and Mohammed Redjimi

(Received: 1-Jun.-2022, Revised: 27-Jul.-2022, Accepted: 18-Aug.-2022)

ABSTRACT

Rat Swarm Optimizer (RSO) is one of the newest swarm intelligence optimization algorithms that is inspired from the behaviors of chasing and fighting of rats in nature. In this paper, we will apply the RSO to one of the most challenging problems, which is data clustering. The search capability of RSO is used here to find the best cluster centers. The proposed RSO algorithm for clustering (RSOC) is tested on several benchmarks and compared to some other optimization algorithms for data clustering, including some well- known and powerful algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), as well as other recent algorithms, such as the Hybridization of Krill Herd Algorithm and harmony search (H-KHA), hybrid Harris Hawks Optimization with differential evolution (H-HHO) and Multi-Verse Optimizer (MVO). Results are validated through a bunch of measures: homogeneity, completeness, v-measure, purity and error rate. The computational results are encouraging, where they demonstrate the effectiveness of RSOC over other clustering techniques.

KEYWORDS

Rat swarm optimization (RSO), Swarm intelligence, Cluster analysis, Clustering.

1. INTRODUCTION

Data clustering is an important procedure in data mining [1]-[3]. It consists of dividing a given set of unlabeled data (objects) into finite groups of similar objects. Data clustering has been widely used in several fields such as: image processing, pattern recognition, intrusion detection, biology, medical fields, among others [1]-[6]. There are many categorizations of data-clustering techniques depending on some criteria [2], [7]-[8], such as categorizing data clustering into hard (crisp) and fuzzy clustering. In hard clustering, an object cannot be a part of more than one cluster. However, in fuzzy clustering, an object can be a part of multiple clusters with certain values that indicate the degree of membership to each cluster [2], [9]. Another well-known categorization is partitional and hierarchical clustering [2], [7]-[8]. Hierarchical clustering clusters data progressively making a clusters hierarchy, generally with each object as a cluster at the bottom stage and the whole dataset as a cluster at the top stage; between these two stages, there is a bunch of other stages, where in each stage there is a different number of clusters. Each stage can be used as the final clustering, where the choice of the final clustering (stage) may depend on the number of clusters or any other criterion, such as the distance between clusters. Partitional clustering, however, divides the dataset directly into a certain number of clusters [1]-[3], [8]. In this work, we are interested in partitional clustering. The most known technique of this type is k-means [1]-[2], 10].

Data clustering is considered as an optimization problem [2], as it is impossible in most cases to find the global optimal solution with exact methods. For a machine that can verify a million solutions per second, to test all possibilities of clustering a dataset of 50 objects in three clusters, it would take more than 3 billion years. Thus, the need of powerful (efficient and effective) methods that can find a good solution near to the best one in acceptable time is indispensable. Nature-inspired metaheuristics are optimal tools for such problems [2]. Mainly, they can be categorized into four general types: evolutionary-based algorithms, swarm intelligence-based algorithms, human-based algorithms and physical and chemical-based algorithms [11]. Swarm intelligence-based algorithms are methods inspired from the intelligence shown by swarms in nature. They mimic their collective intelligent behavior of finding food, fighting, defending, hunting, …etc. to explore and find solutions to optimization problems. Ant Colony Optimization (ACO) [12] and Particle Swarm Optimization (PSO) [13] are examples of swarm-intelligence techniques.

Metaheuristics has been widely applied to the clustering problem. Selim and Al-Sultan [14] applied simulated annealing (SA) to clustering. Al-Sultan [15] proposed a tabu-search (TS) [16]-[17] approach for data clustering. It was compared to SA and k-means and it outperformed them on almost all datasets.

I. Zebiri, D. Zeghida and M. Redjimi are with Department of Computer Science, Université 20 Août 1955, Skikda, Algeria. Emails: i.zebiri@univ-skikda.dz,dj.zeghida@gmail.com and m.redjimi@univ-skikda.dz

Genetic algorithm (GA) [18]-[19] is widely applied to this problem [2], [20]-[22]. Shelokar et al. [23] developed an ant-colony approach, where it was compared to SA, TS and GA and the results showed the power of the mechanisms of this approach. In [24], Jinchao et al. proposed a novel artificial bee colony (ABC) [25] based on k-modes (ABC-K-modes) for clustering of categorical data. The proposed algorithm was tested on several datasets and compared to some other popular algorithms for categorical data, where ABC-K-modes outperformed the algorithms compared with in all but few datasets. In [26]-[28], some applications of PSO to data clustering are demonstrated. In [29], authors proposed a hybrid PSO and grey wolf optimizer (GWO) [30] to take advantage of both mechanisms of PSO and GWO and applied it to data clustering. Kumar et al. [31] developed a grey wolf algorithm-based clustering (GWAC) technique, where GWO was applied to find the optimal center for each cluster and k-means to cluster data. The proposed algorithm was tested on both artificial and real datasets and compared with other algorithms. In [32], a magnetic optimization algorithm for data clustering (MOAC) was proposed. The algorithm was tested on eleven datasets and compared to five algorithms, where MOAC showed better results than other algorithms in general. The authors in [33] proposed an enhanced version of black hole algorithm (LBH) and applied it to data clustering. The proposed algorithm was tested on six real datasets and com-pared with nine other algorithms, where it outperformed them in all datasets. In [34], authors hybridized GWO with TS (GWOTS). TS was used to search for optimal solutions near the best ones. GWOTS was tested on several datasets and compared to other algorithms including, GWO and TS, where the results showed the effectiveness of the hybrid method. Aljarah et al. [35] applied multi-verse optimizer (MVO) [36] to data clustering and tested it on several datasets with four measures. MVO outperformed the other algorithms compared with in almost all datasets.

Rat Swarm Optimizer (RSO) [37] is a novel swarm intelligence-based algorithm, which mimics the behavior of rats in chasing and fighting prey in nature. It was applied to several optimization problems [38]-[42]. In this paper, we will apply this method to the clustering problem. The performance of Rat Swarm Optimizer for Clustering (RSOC) has been tested on several various real benchmarks to show its performance.

The remainder of this work is structured as follows: Section 2 introduces the data-clustering problem briefly. Section 3 describes the RSO. The adaptation of the proposed RSO to the clustering problem is presented in Section 4. Finally, experimental results and their discussion are provided in Section 5. Section 6 concludes the paper and opens some horizons for future research.

2. DATA CLUSTERING

Data Clustering is the task of grouping a set of unlabeled data D in k groups called clusters C = (C1, C2,..., Ck), based on some distance or similarity measurements, such as Euclidean and Manhattan distances. Each object should be a member of one and only one cluster and a cluster should at least have a member [2], [4], [10], [43]:

$$\forall i,j \in \{1,\dots,k\}\ and\ i \neq j, C_i \cap C_j = 0$$
$$U_{i=1}^{k} C_i = D$$
$$\forall i \in \{1,\dots,k\}, C_i \neq 0$$

Objects of the same cluster should be closer to each other or similar, while objects from different clusters should be dissimilar or distant. Thus, the problem of clustering can be reformulated as: minimizing the intracluster distances and maximizing the intercluster distances. The Euclidean distance between two objects *x* and *y* is defined as follows:

$$d_{Euc}(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \quad (1)$$

where, x_i and y_i are respectively the i^{th} attributes of *x* and *y*.

Clustering techniques need to be evaluated to reveal their efficacy. Algorithm efficacy is generally measured by two main measures: performance and effectiveness [2]. Performance measures are generally used to compare the efficiency (computational time) of algorithms, without caring about the quality of results. Algorithms to be compared should be applied on the same programming language, tested on the same benchmark and executed on the same machine. On the other hand, effectiveness measures are used to assess the quality of results. Generally, there are three main types of effectiveness measures: internal, external and relative measures [2], [44]-[46]. Internal indices (intrinsic indices)

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 03, September 2022.

measure the validity using the information intrinsic to data. Sum of intracluster distances is an example of this type. However, external indices (extrinsic indices) measure the validity of the clustering results using some external information (ground truth), such as the class distribution of the clustered dataset [1]-[2], [47]-[48]. Homogeneity, completeness, v-measure, purity and error rate are external indices, which are, respectively, defined as follows:

$$Homogeneity = 1 - \frac{H(C|L)}{H(C)} \tag{2}$$

$$Completness = 1 - \frac{H(L|C)}{H(L)} \tag{3}$$

where,

$$H(C|L) = -\sum_{i=1}^{k}\sum_{j=1}^{q}\frac{n_{ij}}{n}.\log(\frac{n_{ij}}{n_j})$$

$$H(C) = -\sum_{i=1}^{k}\frac{n_i}{n}.\log\left(\frac{n_i}{n}\right)$$

$$V - measure = 2.\frac{Homogeneity.Completness}{Homogeneity+Completness} \tag{4}$$

$$Purity = \frac{1}{n}\sum_{i=1}^{k} max_j(n_{ij}) \tag{5}$$

$$ErrorRate = \frac{Number\ of\ misplaced\ objects}{Total\ number\ of\ objects}.100 \tag{6}$$

k and *q* are, respectively, the number of clusters and true classes. *n* is the total number of objects (size of the dataset), n_{ij} is the number of objects that are from class *j* and clustered in cluster *i*. n_i and n_j are, respectively, the size of cluster *i* and class *j*.

Relative indices are different from the two aforementioned indices. They compare the results of different clustering algorithms or the same algorithm, yet with different parameters [10], [45].

3. RAT SWARM OPTIMIZER (RSO)

3.1 Inspiration

RSO [37] is a novel swarm intelligence technique inspired from two behaviors of rats in nature; chasing and fighting a prey. Black and brown rats are the two main species of rats. In general, rats show a social intelligence by nature. They contribute and help each other in different tasks. Rats live in groups and they are known by their aggressiveness in chasing and fighting prey, which is the fundamental motivation of the RSO algorithm.

In RSO, each rat represents a different solution. The RSO starts by initializing the set of solutions (rats) randomly and then evaluates them by an objective function, where the optimal solution is considered as the best rat $\overrightarrow{P_r}$ and so, the following processes are repeatedly executed a certain number of times (*T*), starting by firstly updating the position of each rat by the two behaviors chasing and fighting prey; secondly, the parameters are updated and any solution beyond the search space is adjusted and finally, the fitness of each rat is recalculated and the position of the best rat is updated if there is a better solution than $\overrightarrow{P_r}$. After completing that, the RSO returns the best solution $\overrightarrow{P_r}$. Algorithm 1 represents the pseudo-code of RSO.

3.2 Mathematical Model and Optimization Algorithm

The two behaviors of chasing and fighting the prey are modeled as follows.

3.2.1 Chasing the Prey

Rats' chasing is generally a social task. The best search agent is considered as the rat which has knowledge about the prey's location. The rest of the group will update their positions according to the best-rat position as follows [37]:

$$\overrightarrow{P_r} = A.\overrightarrow{P_l}(t) + C.(\overrightarrow{P_r}(t) - \overrightarrow{P_l}(t)) \tag{7}$$

where $\vec{P_i}(t)$ represents the position of the i^{th} rat (solution) and t represents the number of the current iteration. $\vec{P_i}(t)$ is the position of the best soon. A is calculated as follows:

$$A = R - t.\left(\frac{R}{max\ Iteration}\right) \quad (8)$$

R and C are random numbers, respectively, in [1,5] and [0,2]. A and C are two parameters for exploration and exploitation mechanisms.

$$R = rand(1{,}5) \quad (9)$$

$$C = rand(0{,}2) \quad (10)$$

3.2.2 Fighting the Prey

The fighting behavior is mathematically modeled as follows:

$$\vec{P_i}(t+1) = \left|\vec{P_i}(t) - \vec{P}\right| \quad (11)$$

where $\vec{P_i}(t+1)$ is the next position of rat number i.

A and C parameters are used to make balance between exploration and exploitation mechanisms. A small value of A (such as 1) and a moderate value of C will lead to emphasise exploitation. Other distant values may lead to emphasise exploration. The objective function used to evaluate results quality is the sum of intra-cluster distances which is defined as:

$$\sum_{C_i \in C} \sum_{x \in C_i} d^2(x, \mu_i) \quad (12)$$

μ_i is the center of the cluster i and $d^2(...)$ is the squared Euclidean distance.

Algorithm 1: RSO [37]

Parameter Initialization:
Initialize $\vec{R}$, $\vec{A}$ and $\vec{C}$ and set $t = 0$
Population Initialization:
Initialize the group of rats $P_i (i = 1,...,n)$
Calculate the fitness value of each rat
The best solution is assigned to $\vec{P_r}$
while ($t < T$) **do**
  for each rat **do**
    Update the position of the current rat by Equation (11)
  Update $\vec{R}$, $\vec{A}$ **and** $\vec{C}$ by Equations (9, 8 and 10)
  Adjust the rat if it goes beyond the search space
  Calculate the fitness value of each rat
  If the best solution of the current iteration is better than $\vec{P_r}$ **then**
    The position of $\vec{P_r}$ is updated to the position of the best solution
  $t \leftarrow t+1$
Return: $\vec{P_r}$

4. Proposed RSO-based Clustering Method (RSOC)

In RSOC, the idea is to find the best cluster centers. Thus, each rat is represented by a vector of k cluster centers, where each cluster center is an object in a d−*dimentional* space (feature space). Hence, a solution can be represented in a $(k*d)$−*dimensional* space as follows:

$$P_i = ((\mu_{i,1,1}, \mu_{i,1,2}, \dots, \mu_{i,1,d}), (\mu_{i,2,1}, \mu_{i,2,2}, \dots, \mu_{i,2,d}), \dots, (\mu_{i,k,1}, \mu_{i,k,2}, \dots, \mu_{i,k,d}))$$

where, $\mu_{i,j,l}$ is the attribute number l of the center number $\boldsymbol{j}$ of the i^{th} rat.

The RSO process starts firstly by initializing each rat of the population by k random points from the dataset. The data is so clustered by each rat according to centers and each object is added to the cluster with the nearest center. After initializing parameters A, C and R, results are assessed by an objective function, where the best solution is saved in $\vec{P_r}$, then rats' positions are updated by Equation 11 and parameters R, A and C are so updated respectively by Equations (9, 8 and 10). If there is a rat beyond the search space, its position will be adjusted by reassigning the previous centers. The data is so clustered by each rat and the results are assessed by the objective function. If there is a better solution than $\overrightarrow{P_r}$, $\overrightarrow{P_r}$ is then updated to the position of the best solution. This process of rats' position updating continues until the end, where a max. number of iterations T are repeated. Finally, the data is clustered using the best cluster centers found ($\overrightarrow{P_r}$).

The pseudo-code (**Algorithm** 2) depicts the proposed RSOC.

Algorithm 2: RSOC

Parameter Initialization:
Number of clusters k, rats' group size, max. number of iterations T and the dataset
Population Initialization:
Data clustered with the best solution obtained $\overrightarrow{P_r}$
Initialize the group of rats $P_i(i=1,...,n)$
Initialize $\vec{R}$, $\vec{A}$ and $\vec{C}$ by Equations (9, 8 and 10) and set $t=0$
Cluster data by each rat
Assess results and the best solution is assigned to $\overrightarrow{P_r}$
while $(t < T)$ **do for**
each rat **do**
Update the position of the current rat by Equation (11)
Cluster data by the current rat
if *a solution is beyond the search space* **then**
The current rat centers are not updated to the new centers.
Update $\vec{R}$, $\vec{A}$ and $\vec{C}$ by Equations (9, 8 and 10)
Calculate the fitness of the current solution by Equation (12)
if *the best solution of the current iteration is better than* $\overrightarrow{P_r}$ **then**
The position of $\bar{P_r}$ is updated to the position of the best solution
$t \leftarrow t+1$
Cluster the dataset by $\overrightarrow{P_r}$ and return the result
Return: $\overrightarrow{P_r}$ and data clustered with it

5. Experimental Results and Discussion

In this section, the proposed RSOC approach is applied to several real datasets and compared to other optimization algorithms. The results were measured for the first comparison by four measures: homogeneity: Equation (2), completeness: Equation (3), v-measure: Equation (4) and purity: Equation (5). Results are measured by error rate: Equation (6) for the second comparison. Table 1 details the utilized benchmark datasets, which are obtained from UCI Machine Learning Repository [49].

Table 1. Used datasets.

Dataset	Number of instances	Number of features	Number of classes
Iris	150	4	3
Ecoli	336	7	8
Glass	214	9	6
Heart	270	13	2
Cancer	683	10	2
Seeds	210	7	3
Wine	178	13	3
CMC	1473	9	3

5.1 Comparison with MVO

The RSOC here was compared with: Deferential Evolution (DE), Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Multi-verse Optimizer (MVO). The results were validated through four measures: homogeneity: Equation (2), completeness: Equation (3), v-measure: Equation (4) and purity: Equation (5). Parameters of algorithms compared with are the same mentioned in [35], since the results are taken directly from [35]. For RSOC, maximum number of iterations and population size are the same as for MVO; 200 as max. number of iterations and 50 as population size. The results are gathered through 10 independent runs. The results are presented in Tables (2-6).

Table 2. Clustering results of Iris dataset.

	Homogeneity	Completeness	V-measure	Purity
DE	0.72778	0.75507	0.74096	0.86733
	(0.04379)	(0.04469)	(0.04293)	(0.03777)
PSO	0.65750	**0.82877**	0.72629	0.77133
	(0.07052)	**(0.09641)**	(0.02481)	(0.09270)
GA	0.60002	0.69056	0.64046	0.75333
	(0.09578)	(0.09905)	(0.09045)	(0.08433)
MVO	0.73642	0.74749	0.74191	0.88667
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
RSOC	**0.74847**	0.76149	**0.75492**	**0.89200**
	(0.00635)	(0.00738)	**(0.00686)**	**(0.00281)**

Table 3. Clustering results of Ecoli dataset.

	Homogeneity	Completeness	V-measure	Purity
DE	0.43868	0.56485	0.49188	0.69235
	(0.10838)	(0.12341)	(0.11438)	(0.07287)
PSO	0.22629	**0.74693**	0.31740	0.57187
	(0.14762)	**(0.17063)**	(0.20040)	(0.09071)
GA	0.44054	0.52583	0.47512	0.67890
	(0.08253)	(0.08403)	(0.06779)	(0.06717)
MVO	0.50214	0.71637	0.58060	0.72508
	(0.13705)	(0.04119)	(0.10298)	(0.07459)
RSOC	**0.69627**	0.54324	**0.61021**	**0.81815**
	(0.01868)	(0.02423)	**(0.02162)**	**(0.01559)**

Table 4. Clustering results of Glass dataset.

	Homogeneity	Completeness	V-measure	Purity
DE	0.18996	0.46231	0.26717	0.45047
	(0.05362)	(0.08873)	(0.06913)	(0.03201)
PSO	0.17044	0.46871	0.24495	0.44206
	(0.07987)	(0.11835)	(0.10986)	(0.04295)
GA	0.24416	0.40213	0.30203	0.48972
	(0.04901)	(0.08900)	(0.05786)	(0.03763)
MVO	0.24341	**0.50376**	0.32666	0.47804
	(0.03544)	**(0.07557)**	(0.04368)	(0.02136)
RSOC	**0.36172**	0.43519	**0.39355**	**0.56028**
	(0.02865)	(0.07616)	**(0.04263)**	**(0.02611)**

Table 5. Clustering results of Heart dataset.

	Homogeneity	Completeness	V-measure	Purity
DE	0.13902	0.13881	0.13890	0.68815
	(0.11303)	(0.11186)	(0.11245)	(0.10174)
PSO	0.17086	0.20987	0.18129	0.70148
	(0.11114)	(0.08492)	(0.11056)	(0.10612)
GA	0.14584	0.15514	0.15021	0.70519
	(0.09743)	(0.10498)	(0.10090)	(0.08145)
MVO	**0.25875**	**0.25761**	**0.25816**	**0.78222**
	(0.06571)	**(0.06283)**	**(0.06432)**	**(0.05627)**
RSOC	0.01881	0.01944	0.01912	0.59074
	(0.00086)	(0.00092)	(0.00089)	(0.00195)

Table 6. Clustering results of Seeds dataset.

	Homogeneity	Completeness	V-measure	Purity
DE	0.55015	0.64305	0.58691	0.77048
	(0.10567)	(0.03752)	(0.06628)	(0.09162)
PSO	0.54263	0.68222	0.59593	0.76095
	(0.11405)	(0.05097)	(0.06504)	(0.11586)

GA	0.54015 (0.06536)	0.61663 (0.05254)	0.57184 (0.03513)	0.76762 (0.08056)
MVO	0.61098 (0.09793)	0.67855 (0.03824)	0.63709 (0.05412)	0.82810 (0.10025)
RSOC	**0.69394 (0.00793)**	**0.69689 (0.00877)**	**0.69541 (0.00835)**	**0.89524 (0.00224)**

Tables (2-6) show the superiority of RSOC in most datasets. RSOC showed the best values outperforming all other techniques compared with in terms of homogeneity, v-measure and purity for all datasets, expect for Heart dataset, where it gave the worst values. MVO gave the best values on Heart dataset and on Glass dataset for completeness measure. PSO outperformed all other algorithms in terms of completeness for Iris and Ecoli datasets. However, for Seeds dataset, RSOC showed the best results in all measures. As presented in Tables (2-6), RSOC seems to find more homogeneous and pure clusters. To recapitulate, RSOC occupied the first place by outperforming other algorithms in 13 cases, 4 of which for homogeneity, 4 for purity, 4 for v-measure and one for completeness. MVO occupied the second place by outpassing other algorithms in 5 cases, 2 for completeness, one for homogeneity, one for v-measure and one for purity. At the third place, PSO outperformed other techniques in two cases for completeness.

5.2 Comparison with H-HHO

At the second comparison, RSOC was compared to a number of algorithms, namely: K-means++ (KM++) [52], Spectral, Agglomerative [53], DBSCAN [50], Genetic Algorithm (GA) [54], Particle Swarm Optimization (PSO) [55], Harmony Search (HS) [56], Krill Herd Algorithm (KHA) [57], Hybrid GA (H-GA) [50], Hybrid PSO (H-PSO) [51], H-KHA [50] and H-HHO [51]. Since the results were taken directly from [50]-[51], they are validated by error rate through five datasets: Iris, Wine, Cancer, CMC and Glass. Parameters of algorithms compared with are mentioned in [50]-[51]. Parameters of RSOC are set to be the same as for H-HHO, max number of iteration is set to (1000). Results are collected over 15 independent runs.

Table 7. Error-rate results.

	Criterion	Iris	Wine	Cancer	CMC	Glass	Rank
K-means	MEAN	21.467	32.388	42.388	55.470	46.154	
	BEST	10.660	29.775	39.865	54.660	42.262	12
	WORST	56.667	43.820	45.970	56.667	46.215	
KM++	MEAN	20.983	31.841	40.145	56.258	44.566	
	BEST	10.101	30.546	39.500	52.003	45.123	07
	WORST	54.274	43.534	44.965	57.001	45.250	
Spectral	MEAN	17.458	33.585	40.154	55.120	46.614	
	BEST	10.547	29.189	38.111	53.541	38.541	09
	WORST	55.541	43.137	44.685	54.044	51.991	
Agglomerative	MEAN	18.544	34.154	41.645	54.944	43.222	
	BEST	9.874	30.665	39.148	52.391	32.001	06
	WORST	48.397	42.688	46.699	57.487	52.140	
DBSCAN	MEAN	16.311	33.487	42.199	56.544	44.984	
	BEST	9.987	30.140	39.654	54.280	33.717	11
	WORST	43.111	42.009	44.021	56.654	51.123	
GA	MEAN	21.652	34.270	44.270	56.697	51.028	
	BEST	10.666	29.310	39.510	54.656	42.991	14
	WORST	43.333	47.753	47.753	57.296	56.075	
PSO	MEAN	15.867	32.051	43.051	55.899	46.262	
	BEST	10.667	29.775	40.775	54.101	43.925	10
	WORST	43.447	44.449	45.455	56.486	52.804	
HS	MEAN	21.054	32.568	42.054	56.001	43.054	
	BEST	10.509	29.865	40.111	55.430	41.162	08
	WORST	44.286	44.467	45.640	57.906	46.255	
KHA	MEAN	22.658	32.303	42.543	56.056	43.925	
	BEST	9.430	29.213	39.256	53.936	38.318	12

	WORST	42.548	47.191	47.191	56.999	50.476	
H-GA	MEAN	21.100	30.989	41.214	55.142	44.219	
	BEST	9.765	29.654	40.254	53.124	35.249	05
	WORST	44.667	44.001	46.214	56.214	51.985	
H-PSO	MEAN	15.800	30.871	42.125	54.204	51.617	
	BEST	9.666	29.775	39.775	53.201	41.589	04
	WORST	44.333	43.888	46.758	55.333	56.075	
H-KHA	MEAN	19.866	33.000	**39.012**	53.656	42.219	
	BEST	9.000	29.650	38.670	52.213	32.242	02
	WORST	43.333	42.134	44.154	54.333	51.420	
H-HHO	MEAN	20.866	33.564	39.470	54.109	44.002	
	BEST	9.332	29.653	39.119	53.165	34.242	03
	WORST	43.333	43.584	45.365	55.693	51.445	
RSOC	MEAN	**12.027**	**28.134**	45.737	**46.292**	**33.070**	
	BEST	12.027	28.134	45.737	46.208	31.587	**01**
	WORST	12.027	28.134	45.737	46.392	33.970	

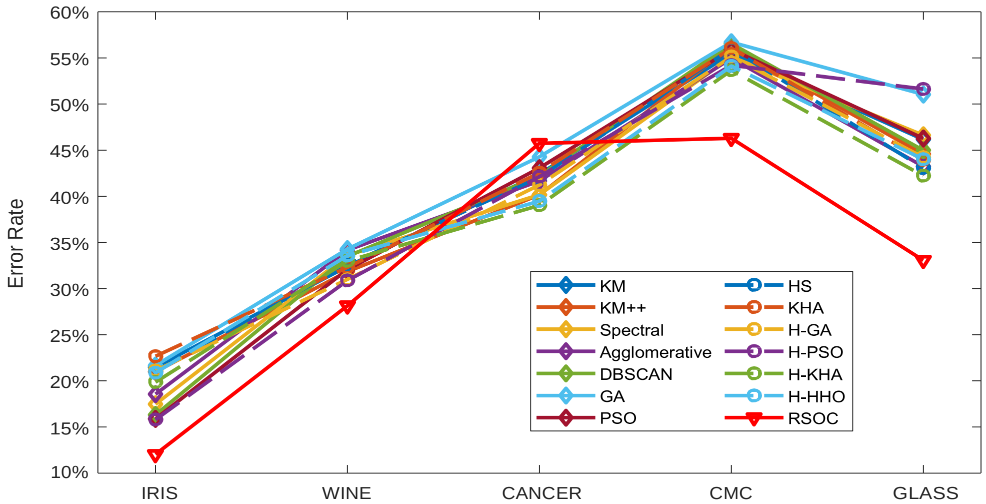

Figure 1. Visual comparison of error-rate results.

Table 8. Ranks of algorithms.

	Iris	Wine	Cancer	CMC	Glass	Sum
K-means	12	7	10	8	10	47
KM++	9	4	3	12	8	36
Spectral	5	12	4	6	12	39
Agglomerative	6	13	6	5	4	34
DBSCAN	4	10	9	13	9	45
GA	13	14	13	14	13	67
PSO	3	5	12	9	11	40
HS	10	8	7	10	2	37
KHA	14	6	11	11	5	47
H-GA	11	3	5	7	6	32
H-PSO	2	2	8	4	14	30
H-KHA	7	9	1	2	3	22
H-HHO	8	11	2	3	5	29
RSOC	1	1	14	1	1	18

Tables (7-8) and Figure 1 show impressive results, where RSOC ranked the first among other algorithms. It outperformed all other algorithms showing the least error rate on all datasets, expect for Cancer dataset, where it unexpectedly occupied the last place, which calls for no free lunch theorem (no algorithm is suitable for all problems). Next to RSOC, comes H-KHA occupying the second place; first place on Cancer dataset and second place on CMC and Glass datasets. The third place went to H-HHO, which got the second place on Cancer dataset and the third place on CMC. The rest of algorithms are ordered as follows: H-PSO, H-GA, agglomerative clustering, k-means++, HS, spectral clustering, PSO, DBSCAN, k-means and KHA sharing the same rank and finally GA. RSOC showed a small deviation compared to other algorithms with CMC and Glass datasets and no deviation for the rest of datasets.

6. CONCLUSION AND FUTURE WORKS

In this work, we applied RSO technique for the problem of data clustering, where the number of clusters is known *a priori*. The proposed technique was compared to other algorithms and the quality of results was measured in terms of five measures in two comparisons: homogeneity, completeness, v-measure and purity for the first comparison and error rate for the second. Results and analysis showed the superiority of RSOC. However, this technique is still showing a weakness, such as on Heart and Cancer datasets, where it gave the worst values. As a future work, we will try to improve this technique and apply it to solve other problems, such as feature selection. We will also try to compare this metaheuristic to grey wolf optimizer, since they are very similar.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd Edn., San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[2] A. M. Bagirov, N. Karmitsa and S. Taheri, Partitional Clustering *via* Nonsmooth Optimization: Clustering *via* Optimization, Springer Nature, 2020.

[3] P. Berkhin, "A Survey of Clustering Data Mining Techniques," Proc. of Grouping Multidimensional Data, pp. 25–71, Springer, 2006.

[4] B. Everitt, S. Landau, M. Leese and D. Stahl, Cluster Analysis, ser. Wiley Series in Probability and Statistics, Wiley, [Online], Available: https://books.google.dz/books?id=WSayDAEACAAJ, 2011.

[5] J. Hartigan, Clustering Algorithms, John Wiley and Sons, New York, 1975.

[6] K. Krippendorff, "Clustering," Book Chapter, Multivariate Techniques in Human Communication Research, pp. 259-308, Elsevier, 1980.

[7] K. Bailey, "Cluster Analysis," Book Chapter, Sociological Methodology, pp. 59-128, DOI: 10.2307/270894, 1975.

[8] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, USA: Prentice-Hall, Inc., 1988.

[9] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach, DOI:10.1201/9781420034912, 2005.

[10] R. Xu and D. Wunsch, Clustering, vol. 10, John Wiley & Sons, 2008.

[11] A. Naik and S. C. Satapathy, "Past Present Future: A New Human-based Algorithm for Stochastic Optimization," Soft Computing, vol. 25, no. 20, pp. 12 915–12 976, 2021.

[12] M. Dorigo, V. Maniezzo and A. Colorni, "Ant System: Optimization by a Colony of Cooperating Agents," IEEE Trans. on Systems, Man and Cybernetics, Part B (Cybernetics), vol. 26, no. 1, pp. 29–41, 1996.

[13] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," Proc. of the International Conference on Neural Networks (ICNN'95), vol. 4, , pp. 1942–1948, 1995.

[14] S. Z. Selim and K. Alsultan, "A Simulated Annealing Algorithm for the Clustering Problem," Pattern Recognition, vol. 24, no. 10, pp. 1003–1008, 1991.

[15] K. S. Al-Sultan, "A Tabu Search Approach to the Clustering Problem," Pattern Recognition, vol. 28, no. 9, pp. 1443–1451, 1995.

[16] F. Glover, "Future Paths for Integer Programming and Links to Artificial Intelligence," Computers & Operations Research, vol. 13, no. 5, pp. 533–549, 1986.

[17] F. Glover, "Tabu Search—Part i," ORSA Journal on Computing, vol. 1, no. 3, pp. 190–206, 1989.

[18] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, 1st Edn., USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[19] J. H. Holland, Adaptation in Natural and Artificial Systems, Ann Arbor, MI: University of Michigan Press, 1975, 2nd Edition, 1992.

[20] M. C. Cowgill, R. J. Harvey and L. T. Watson, "A Genetic Algorithm Approach to Cluster Analysis," Computers & Mathematics with Applications, vol. 37, no. 7, pp. 99–108, 1999.

[21] E. Falkenauer, Genetic Algorithms and Grouping Problems, John Wiley & Sons, Inc., 1998.

[22] E. R. Hruschka, R. J. Campello, A. A. Freitas et al., "A Survey of Evolutionary Algorithms for Clustering," IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), vol. 39, no. 2, pp. 133–155, 2009.

[23] P. Shelokar, V. K. Jayaraman and B. D. Kulkarni, "An Ant Colony Approach for Clustering," Analytica Chimica Acta, vol. 509, no. 2, pp. 187–195, 2004.

[24] J. Ji, W. Pang, Y. Zheng, Z. Wang and Z. Ma, "A Novel Artificial Bee Colony Based Clustering Algorithm for Categorical Data," PloS One, vol. 10, no. 5, p. e0127125, 2015.
[25] D. Karaboga and B. Basturk, "An Artificial Bee Colony (ABC) Algorithm for Numeric Function Optimization," Proc. of the IEEE Swarm Intelligence Symposium, pp. 181–184, Indianapolis, USA, 2006.
[26] D. Van der Merwe and A. P. Engelbrecht, "Data Clustering Using Particle Swarm Optimization," Proc. of the IEEE Congress on Evolutionary Computation (CEC'03), vol. 1, pp. 215–220, 2003.
[27] Y.-T. Kao, E. Zahara and I.-W. Kao, "A Hybridized Approach to Data Clustering," Expert Systems with Applications, vol. 34, no. 3, pp. 1754–1762, 2008.
[28] T. Cura, "A Particle Swarm Optimization Approach to Clustering," Expert Systems with Applications, vol. 39, no. 1, pp. 1582–1588, 2012.
[29] X. Zhang, Q. Lin, W. Mau, Z. Dou and G. Liu, "Hybrid Particle Swarm and Grey Wolf Optimizer and Its Application to Clustering Optimization, " Applied Soft Computing, vol. 101, p. 107061, 2021.
[30] S. Mirjalili, S. M. Mirjalili and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, 2014.
[31] V. Kumar, J. K. Chhabra and D. Kumar, "Grey Wolf Algorithm-based Clustering Technique," Journal of Intelligent Systems, vol. 26, no. 1, pp. 153–168, 2017.
[32] N. Kushwaha, M. Pant, S. Kant and V. Jain, "Magnetic Optimization Algorithm for Data Clustering," Pattern Recognition Letters, vol. 115, pp. 59-65, [Online], Available: 10.1016/j.patrec.2017.10.031, 2018.
[33] H. A. Abdulwahab, A. Noraziah, A. A. Alsewari and S. Q. Salih, "An Enhanced Version of Black Hole Algorithm *via* Levy Flight for Optimization and Data Clustering Problems," IEEE Access, vol. 7, pp. 142085-142096, DOI: 10.1109/ACCESS.2019.2937021, 2019.
[34] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris and S. Mirjalili, "Clustering Analysis Using a Novel Locality-informed Grey Wolf-inspired Clustering Approach," Knowledge and Information Systems, vol. 62, no. 2, pp. 507–539, 2020.
[35] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris and S. Mirjalili, "Multi-verse Optimizer: Theory, Literature Review and Application in Data Clustering," Nature-inspired Optimizers, vol. 811, pp. 123–141, 2020.
[36] S. Mirjalili, S. M. Mirjalili and A. Hatamlou, "Multi-verse Optimizer: A Nature-inspired Algorithm for Global Optimization," Neural Computing and Applications, vol. 27, no. 2, pp. 495–513, 2016.
[37] G. Dhiman, M. Garg, A. Nagar, V. Kumar and M. Dehghani, "A Novel Algorithm for Global Optimization: Rat Swarm Optimizer," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 8, pp. 8457–8482, 2021.
[38] M. Dhas and N. Singh, "Blood Cell Image Denoising Based on Tunicate Rat Swarm Optimization with Median Filter," Evolutionary Computing and Mobile Sustainable Networks, vol. 116, pp. 33–45, 2022.
[39] A. Tamilarasan, A. Renugambal and V. Dharanendran, "Parametric Estimation for AWJ Cutting of TI-6AL-4V Alloy Using Rat Swarm Optimization Algorithm," Materials and Manufacturing Processes, pp. 1–11, DOI: 10.1080/10426914.2022.2065011, 2022.
[40] M. Eslami, E. Akbari, S. T. Seyed Sadr and B. Ibrahim, "A Novel Hybrid Algorithm Based on Rat Swarm Optimization and Pattern Search for Parameter Extraction of Solar Photovoltaic Models," Energy Science and Engineering, DOI: 10.1002/ese3.1160, 2022.
[41] R. Ghadge and S. Prakash, "Investigation and Prediction of Hybrid Composite Leaf Spring Using Deep Neural Network Based Rat Swarm Optimization," Mechanics Based Design of Structures and Machines, pp. 1–30, DOI: 10.1080/15397734.2021.1972309, 2021.
[42] A. Vasantharaj, P. Rani, S. Huque, K. Raghuram, R. Ganeshkumar and S. Shafi, "Automated Brain Imaging Diagnosis and Classification Model Using Rat Swarm Optimization with Deep Learning Based Capsule Network," International Journal of Image and Graphics, p. 2240001, DOI: 10.1142/S0219467822400010, 2021.
[43] G. Gan, C. Ma and J. Wu, Data Clustering: Theory, Algorithms and Applications, SIAM, 2020.
[44] M. Halkidi, Y. Batistakis and M. Varzigiannis, "Cluster Validity Methods: Part I," ACM Sigmod Record, vol. 31, no. 2, pp. 40-45, 2002.
[45] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Clustering Validity Checking Methods: Part II,"ACM Sigmod Record, vol. 31, no. 3, pp. 19–27, 2002.
[46] J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," Proc. of the 19th IEEE Int. Conf. on Computational Science and Its Applications (ICCSA), pp. 71–81, St. Petersburg, Russia, 2019.
[47] R. Zafarani, M. A. Abbasi and H. Liu, Social Media Mining: An Introduction, Cambridge University Press, ISBN-10: 1107018854, 2014.
[48] J.-O. Palacio-Niño and F. Berzal, "Evaluation Metrics for Unsupervised Learning Algorithms," arXiv preprint arXiv:1905.05667, 2019.
[49] D. Dua and C. Graff, "UCI Machine Learning Repository," [Online], Available: http://archive.ics.uci.edu/ml, 2017.
[50] L. Abualigah, A. Khader, E. Hanandeh and A. Gandomi, "A Novel Hybridization Strategy for Krill Herd Algorithm Applied to Clustering Techniques", Applied Soft Computing, vol. 60, pp. 423-435, 2017.
[51] L. Abualigah et al., "Hybrid Harris Hawks Optimization with Differential Evolution for Data Clustering,"

Metaheuristics in Machine Learning: Theory and Applications, vol. 967, pp. 267-299, 2021.

[52] D. Arthur and V. Sergei, "K-Means++: The Advantages of Careful Seeding," Proc. of the 18th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA '07), pp. 1027-1035, 2007.

[53] I. Davidson and S.S. Ravi, "Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results," Knowledge Discovery in Databases: PKDD 2005, pp. 59-70, [Online], Available: 10.1007/11564126_11, 2005.

[54] U. Maulik and S. Bandyopadhyay, "Genetic Algorithm-based Clustering Technique," Pattern Recognition, vol. 33, no. 9, DOI: 10.1016/S0031-3203(99)00137-5, 2000.

[55] S. Rana, S. Jasola and R. Kumar, "A Review on Particle Swarm Optimization Algorithms and Their Applications to Data Clustering," Artificial Intelligence Review, vol. 35, no. 3, pp. 211-222, 2010.

[56] O. Alia, M. Al-Betar, R. Mandava and A. Khader, "Data Clustering Using Harmony Search Algorithm," Proc. of Int. Conf. on Swarm, Evolutionary and Memetic Computing (SEMCCO 2011), pp. 79-88, DOI: 10.1007/978-3- 642-27242-4_10, 2011.

[57] L. Abualigah, A. Tajudin Khader, M. Azmi AlBetar and E. Said Hanandeh, "A New Hybridization Strategy for Krill Herd Algorithm and Harmony Search Algorithm Applied to Improve the Data Clustering," Proc. of the 1st EAI Int. Conf. on Computer Science and Engineering, DOI: 10.4108/eai.27-2-2017.152255, 2017.

ملخص البحث:

تعدّ عملية الأمْثَلَة باستخدام "سِرب الجرذان" من أحدث تطبيقات الأمْثَلَة اعتماداً على خوارزميات يتمّ استلهامها من سلوك أسراب الجرذان المتمثّل في مطاردة الضّحية والانقضاض عليها.

في هذا البحث، نعمل على تطبيق نظام أمْثَلَة يعتمد سلوك "سِرب الجرذان" على مشكلة هي من أبرز التّحدّيات تتمثّل في عَنْقَدة البيانات. وتعمل قدرة هذا النّظام على البحث على إيجاد أفضل مراكز عناقيد البيانات.

وقد جرى فحص النّظام المقترح بناءً على عدّة علامات مرجعية ومقارنته مع عدد من الأنظمة الأخرى المستخدمة في عَنْقَدة البيانات المعتمدة على خوارزميات قوية معروفة جيداً. وتمّ تقييم النتائج بواسطة حزمة من المقاييس، مثل: التّجانس، والاكتمال، ومقياس (v)، والنّقاء، ومعدّل الخطأ. وقد أسفرت نتائج الحسابات على استنتاجاتٍ مشجّعة، ممّا أثبت فعالية التّقنية المقترحة وتفوّقها بشكلٍ لافت على التّقنيات الأخرى المستخدمة في عَنْقَدة البيانات.