A new hybridization strategy for krill herd algorithm and harmony search algorithm applied to improve the data clustering

Laith Mohammad Abualigah[1], Ahamad Tajudin Khader[1], Mohammed Azmi Al-Betar[2], and Essam Said Hanandeh[3]

[1] School of Computer Sciences, Universiti Sains Malaysia (USM), Pulau Pinang, Malaysia 11800
[2] Department of information technology, Al-Huson University College, Al-Balqa Applied University, Al-Huson, Irbid-Jordan
[3] Department of Computer Information System, Zarqa University, Zarqa-Jordan

Abstract. Krill herd (KH) is a stochastic nature-inspired algorithm, it has been successfully used to solve many complex optimization problems. The performance of krill herd algorithm (KHA) is effected by poor ex-ploitation capability. This paper proposes new data clustering algorithm based on a hybrid of krill herd algorithm (KHA) and harmony search (HS) algorithm (Harmony-KHA) in order to improve the data clustering technique. This hybrid strategy seeking to enhance the global search ca-pability of the KHA. The enhancement includes of adding global search operator from HS algorithm for exploration around the optimal solution in KH and thus kill individuals move towards the global best solution. The proposed method is applied to preserve the best krill individual during the krill position update. Experiments were conducted using four standard datasets from the UCI Machine Learning Repository, which is used in the domain of data clustering. The results showed that the pro-posed hybrid KHA and HS algorithm (Harmony-KHA) is produced very accurate clusters, especially in the large dataset.

Key words: krill herd algorithm, improvise a new solution, hybridiza-tion, data clustering

1 Introduction

Clustering is an important unsupervised learning technique used in data analysis applications. Data clustering technique is used for clustering similar or dissimilar data in a given dataset based on the distance between the data objects, which means that similar objects are partitioned in the same cluster and different objects partitioned in different clusters [10, 12, 18]. The aim of the clustering technique is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. Data clustering technique have been used in many areas such as decision-making, machine-learning, data mining, text document retrieval, text categorization, image segmentation and pattern classification [15, 16]. And large transactions for the customer are marketing analysis, sales management, and document management [9, 11].

Clustering technique attempts to partition a set of data objects into a subset of similar clusters based on minimizing the objective function without the hierar-chical structure, each cluster has one centroid. Where each cluster is represented by the centroid of the data objects and the square error function (objective func-tion) is the sum of the distance between each object with the cluster centroid [5, 10, 15].

Nature meta-heuristic algorithms have been used in many areas such as com-puter science, data mining, agriculture, computer vision, forecasting, medicine and biology, economy and engineering [8]. Recently, many meta-heuristic algo-rithms have been used to improve the data clustering technique such as k-mean [2], artificial bee colony (ABC) [6], bee colony optimization (BCO) [7], tabu search approach (TS) [4], particle swarm optimization (PSO) [8, 24] and harmony search (HS) algorithms [17, 19, 21, 25].

Hybridization of discrete KH algorithm and harmony search algorithm is not popular; therefore, the hybridization of KHA and HS has not yet been in-vestigated. This paper proposes an innovative hybrid method by inheriting the improvising solution of harmony search algorithm for enhancing KH .algorithm solutions. This hybridization is proposed in order to improve the data clustering technique, which is a difficult technique with continuous-valued variables. Data clustering technique divides a set of data objects into a subset of clusters with a small distance (intra-clusters) and a large distance (inter-clusters). The per-formance of the proposed method is verified on four common datasets that used in the domain of data clustering and the results are compared with some recent well-known algorithms.

The rest of the paper is organized as follows. Section 2 introduces more related works in the domain of data clustering using meta-heuristic algorithms. Section 3 describes the clustering technique. Section 4 explains the KH algorithm for data clustering and its procedures. Section 5 explains improvising of the harmony solution. Section 6 describes presents the architecture of our hybrid clustering algorithm. Section 7 illustrates experimental results. Finally, Section 8 makes conclusions.

1.1 Data clustering formulation

Data clustering technique is the process of partitioning D the set of data objects into a subset of K clusters based on some distance or similarity measure. Let D = d_1, d_2, ..., d_i, ..., d_n be a set of n data objects to be distributed over K and each data object d_i, $i = 1$ to n is represented as vector di= d_{i1}, d_{i2}, ..., d_{ij} , ..., d_{it} where d_{ij} represents jth dimension value of the data object number i [10, 12, 22]. The aim of clustering algorithm is to find a set of K partitions C= C_1, C_2, ..., C_k , C_K and Ck $\neq \emptyset$ in such a way that objects within the same cluster are more similar and but the objects within different clusters are dissimilar. These similarities measurement are evaluated by some optimization criterions, especially, distance measure and squared error function [15, 22, 27], which have been calculated as the following:

$$FF = \sum_{i=1}^{K} \sum_{j=1}^{K} min(D(d_i, c_j)), \tag{1}$$

Where, c_j represents a jth cluster center, d_j represents a jth data object, D is the distance measure between the object d_i and the cluster center c_j . This cri-terion is used as the objective function value to evaluate the algorithm solution. Different distance measurements have been used in the domain of data clustering techniques such as Euclidean distance, Manhatten distance, Minkowski metric, Cosine similarity, Pearson correlation coefficient, Jaccard coefficient, and so on [12, 22]. In this paper, Euclidean distance is used as distance measure from many distance metric used in literature which is defined as follows:

$$D(d_i, c_j) = \sqrt{\sum_{j=1}^{t} (d_{1j}, c_{2j})^2}, \tag{2}$$

Where, $D(d_i,\ c_j\ )$ is distance measure between the document i and the cluster j, d_{1j} represents the term j in document 1, c_{2j} represents the term j in cluster centroid 2 and c_1 the cluster center of the cluster 1 [12].

2 Krill herd algorithm for data clustering

In this paper, KH algorithm is used for enhancing the data clustering problem. Two modifications are original KH algorithm and hybrid with harmony search algorithm.

2.1 Data clustering solution representation

In the KH algorithm, each solution consists the clusters centroid number in a one-dimensional array with the length of n, where n the number of data objects in the given datasets. Table 1 shows the solution representation of KH algorithm for the data clustering problem. The length of each solution is n number of all objects, it represents data objects that belong to each cluster, where the cluster number 1 contains objects number (1,2 and 9), cluster number 2 contains objects number (3,5 and 8) and cluster number 3 contains objects number (4 and 10).

Table 1. Representation of a candidate solution.

X	1	1	2	3	2	3	3	2	1	3

2.2 Memory initialization

KH algorithm generates a feasible solution by a random uniform distribution, but it is not sensitive to the initialization of the krill herd memory (KHM). Each position (krill) corresponds to a specific number of the cluster. The value of i_{th} position in each solution (herd) is randomly selected between the range value of the problem's solutions. KH memory is a memory space of size S*n, where S
is the number of solutions, ${x^2}_1$ is the position *1* in the solution 2 and n is the length of each solution, and $f(\boldsymbol{X}_1)$ is the fitness function of the solution number
1. The population of solutions of KH algorithm is represented as given below:

$$\mathbf{KHM} = \begin{bmatrix} x_1^1 & \cdots & x_{n-1}^1 & x_n^1 \\ \vdots & \ddots & \vdots & \vdots \\ x_1^{S-1} & \cdots & x_{n-1}^{S-1} & x_n^{S-1} \\ x_1^S & \cdots & x_{n-1}^S & x_n^S \end{bmatrix} \begin{bmatrix} f(\boldsymbol{X}_1) \\ \vdots \\ f(\boldsymbol{X}_{S-1}) \\ f(\boldsymbol{X}_S) \end{bmatrix} \tag{3}$$

2.3 The basics of krill herd algorithm

The implementation of KH algorithm is based on three factors include movement influenced by other krill individuals, foraging action and physical diffusion [15, 18, 28]. KH algorithm follows Lagrangian model for efficacious search and it is calculated as:

$$\frac{dX_i}{dt} = N_i + F_i + D_i, \tag{4}$$

where, N_i is first part, which indicates to the motion induced by other krill individuals, F_i indicates to the forging motion, D_i is the last part, which indicates to the physical diffusion of the ith krill individual, these mathematical equations need to adjustment for data clustering problem [18]. The KH algorithm has two stages are the motion calculation and the genetic operators, which are as the following:

Movement induced by other krill individuals: Based on some theoretical arguments. Each krill individual tries to keep a high density and close to the nearest food. The direction of the motion induced is derived from the local effect of each solution density, a target effect of the individuals density, and a repulsive individuals effect [12, 15, 28, 26].

$$N_i^{new} = N^{max}\alpha_i + \omega_n N_i^{old} \tag{5}$$

where, N^{max} is the parameter used to tuning the part of motion induced, ω_n is the array of random values in range [0, 1], and N^{old} is the current motion induced, α_i local is the lead to the suitable class effect by the best individual [12]. In this study, the effect of the individual in the krill movement is determined as follows:

$$\alpha_i^{local} = \sum_{j=1}^{NN} \widehat{K}_{i,j} \widehat{X}_{i,j} \tag{6}$$

The known goal for each krill individual is to achieve highest fitness function, so the krill individual affected by the best fitness, that is taken from formula 7. This procedure leads to the global optimal solution values and is formulated as the following equation:

$$\alpha_i^{target} = C^{best} \widehat{K}_{i,best} \widehat{X}_{i,best} \tag{7}$$

where, C^{best} is the effective coefficient, α_i^{target} leads the solutions to the global optima values, it should be more near to the optimal solution. The *rand* is random number between *(0, 1)*, this value used for improve the exploration; I is the current iteration number; I_{max} is the maximum number of iteration [12, 15]..

Foraging motion: In this action has two affected parameters, the first one the food location (cluster centroid), the second one the old food location. This action can be expressed for the *ith* krill individual as the following formula:

$$F_i = V_f \beta_i + \omega_f F_i^{old} \tag{8}$$

where, V_f is the forging speed; ω_f is the intra weight used to balance the local exploitation and global exploration for each individual; β_i^{food} is the food attractive; β_i^{best} is the best food attraction so far [12, 15]. The food attraction each iteration is formulated as following equations [18]:

$$X^{food} = \frac{\sum_{i=1}^{N} \frac{1}{K_i} X_i}{\sum_{i=1}^{N} \frac{1}{K_i}} \tag{9}$$

Physical diffusion: Here in this action, the krill individual is estimated to be the random process that used two terms to express the physical diffusion, first one is maximum diffusion speed, and the second one is the random directional vector [12, 18]. The physical diffusion is determined by the following:

$$D_i = D^{max} \left(1 - \frac{I}{I_{max}}\right) \delta \tag{10}$$

Where D^{max} is the maximum diffusion speed, and δ is random values as vector and it has arrays contain random values between *[-1, 1]*. This action is decreased the speed value od the krill individual [12, 15].

Motion process of the KH algorithm: The motion induced and foraging motion is contained two local and two global strategies [15, 28]. These strategies are worked in parallel to get a powerful algorithm. The physical diffusion gen-erates random vectors [18]. KH algorithm parameters are effective during the algorithm acts. Positions of krill individual are updated in each iteration using the following equation:

$$X_i(t + \Delta t) = X_i(t) + \Delta t \frac{dX_i}{dt} \tag{11}$$

where, Δt is an important constants and more sensitive, which is consid-ered as $\mathit{0.5}$ in this study, NV is represent total number of variables, the lower bounds LB_j and the upper bounds UB_j of the ith variables ($J = 1, 2, ..., NV$), respectively. C_t is a constant value between $[\mathit{0,2}\,]$.

3 Harmony search algorithm

Harmony search (HS) algorithm is a stochastic new population-based meta-heuristic approach introduced in (2001) [21]. It imitates the music improvisation process where the musicians improvise their instruments pitch by searching for the optimal state of harmony. This algorithm has been success applied to solve many optimization problems, such as numerical function optimization [12, 23] and timetabling problems [19].

Algorithm 1 :Improvise a new solution

```
Input: Harmony memory $HM$ solutions
Output: A new solution as vector represented
for each $j \in [1, t]$ do
    if Rand (0,1) $\leq$ HMCR then
        $x_j{}' = HM[i][j]wherei \sim U(1, 2, ...., HMS)$
        if Rand(0,1) $\leq$ PAR then
            $x_j{}' = x_j{}' \pm rand\times$ bw, where r$\sim$ U(0,1) and $bw$ is distance bandwidth

            else $x_i{}' = LBj + rand \times (UBj - LBj)$
        end if
    end if
end for
```

HS solutions generated according to the $PAR(I)$ and $bw(I)$ parameters. If generated a random value between [0, 1] is less or equal PAR [19, 21], than the new decision variable (x) is determine as follows:

$$X^{'} = (x^{'}_1, x^{'}_2, ..., x^{'}_t) \tag{12}$$

Improvise a new solution: is generated a new solution based on three rules in Algorithm 1: memory consideration, pitch adjustment, and random selection.

Table 2. The characteristic of the datasets

Datasets Number	Datasets Name	Number of Features	Number of Clusters	Number of Objects
DS1	Iris	4	3	150(50,50,50)
DS2	Seeds	7	3	210(70, 70, 70)
DS3	Glass	9	6	214(70,76,17,13,9,29)
DS4	Wine	13	3	178(59,71,47)

4 Hybridization architecture of krill herd algorithm and harmony search algorithm

HS, like KHA, is a population-based stochastic search algorithm. The algorithm maintains a population of solutions, where each solution represents a candidate solution to an optimization problem. Thus, we propose a new method, a hybrid of KH algorithm and harmony search algorithm(Harmony-KHA) for data clustering problems.

The Harmony-KHA randomly initializes a population of size S*n. These in-dividuals may be regarded as a herd in the case of KHA, or as harmonies in the case of HS algorithm. S*n krill individuals are fed into the real-coded KH to create S*n new krill individuals by reproduction, KH acts. Then, the S*n krill individuals are sorted by the fitness function, and fed into harmony search to improve Harmony-KHA by improves a new solution and add it to the population if it is better than the worst solution by Algorithm 1. Thus, Harmony-KHA can use better individuals each iteration to search for the optimum solution. In addi-tion, krill individuals with poor performance remain in order to avoid premature convergence.

5 Experimental results

This section applies four standard benchmark datasets to validate the proposed method, Harmony-KHA, in comparison with k-mean, HS, PSO, KHA, Harmony-PSO and Harmony-KHA. In this paper, we conducted the results using four datasets, namely, (Iris, Seeds, Glass and Wine), provided by UCI Machine Learn-ing Repository of the University of California. Table 2 lists the related informa-tion including the number of datasets, dataset name, number of features, and number of clusters.

5.1 Results and discussion

The performance of the proposed algorithms is evaluated and compared with other well-known algorithm using two criteria: (1) The sum of intra-cluster dis-tances, it is considered as an internal quality measure: The distance value be-tween each object and the cluster centroid of the corresponding cluster is com-puted as defined in Eq. 1. The higher the quality of the data clustering which

has small fitness function [8]. (2) Error Rate (ER) value, it is as an external quality measure: the percentage of misplaced objects overall number of objects [8, 15, 12], as
$ER = \frac{numberofmisplacedobjects}{sizeoftestdataset} * 100$.

Table 3. Error rate results for four datasets using ER

Dataset	Error Rate	KHA	PSO	K-mean	HS	Harmony-PSO	Harmony-KHA
Iris	Best	10.666	10.667	10.660	8.430	9.666	**9.000**
	Average	21.652	15.867	21.467	21.100	**15.800**	19.866
	Worst	43.333	43.447	56.667	44.667	44.333	**43.333**
	Rank	6	2	5	4	1	3
Seeds	Best	13.810	8.571	12.643	13.595	9.047	**8.623**
	Average	21.000	15.262	12.643	13.595	13.881	**11.666**
	Worst	35.714	36.190	**12.643**	19.525	45.238	20.523
	Rank	6	5	2	3	4	1
Glass	Best	42.991	43.925	46.262	38.318	41.589	**32.242**
	Average	51.028	46.262	46.262	43.925	47.617	**42.219**
	Worst	56.075	52.804	46.262	50.476	56.075	**51.420**
	Rank	6	3	3	2	5	1
Wine	Best	**27.310**	29.775	29.775	29.213	29.775	28.650
	Average	34.270	32.051	32.388	32.303	**30.871**	32.000
	Worst	47.753	44.449	43.820	47.191	43.888	**42.134**
	Rank	6	3	5	4	1	2
Mean rank		6.00	3.25	3.75	3.25	2.75	1.75
Final rank		6	3	5	3	2	1

This measure assigned a class label and compared with the desired class label. If they are not the same, the pattern is distributed as incorrect partitioning. It is calculated for all data objects in the given dataset and the total incorrect number of partitioning pattern is a percentage to the size of all data objects in the dataset. A summary of the error rate obtained by the data clustering algorithms is given in Table 3. The values reported are worst, average, and best solutions over 20 independent run [12, 15]. The experimental results are given in Table 3. It shows that proposed hybrid algorithm, namely, Harmony-KHA obtains near optimal solutions in compare well-known algorithm. Our proposed Harmony-KHA achieves better results for almost all datasets with small final ranking.

Table 3 shows the Error rate of using the meta-heuristic algorithm to enhance the data clustering technique. For Iris dataset, Harmony-PSO obtains the best value over (average ER) and Harmony-KHA obtains the best value over (best and worst ER). For Seeds dataset, Harmony-KHA obtains the best value over (best and average ER). For Glass dataset,Harmony-KHA obtains the best value over (best and average ER). For Wine dataset, Harmony-PSO obtains the best value over (average ER) and KHA obtains the best value over (best ER). Harmony-KHA fails to reach that best value in all runs, but it gets the one rank among all others comparative algorithm. As for Seeds data set, Harmony-KHA achieves the best optimum value of 11.666 overall suns and for Glass data set, it achieves the best optimum value of 42.219 overall runs. Thus, Harmony-KHA algorithm reaches near best value in all runs.

6 Conclusion

KH is a new optimization algorithm for solving many hard global optimiza-tion problems. In this paper, we presented an enhanced KH algorithm to solve the data clustering problems. The original KH saturated fast and subsequently trapped in the local optimum. To relieve the premature convergence of KH, en-hanced KH was invented by introducing global exploration operator of harmony search. Using this hybridization, Harmony-KHA converges to optimal solutions quickly under the harmony control. The results show that the proposed hy-bridization are fast and efficient for solving the data clustering problems. The experimental results are compared with other well-known algorithms in the liter-ature of the data clustering and indeed, it reveals that the proposed hybridization of KHA is suitable for solving data clustering problems.

References

1. Kao, Yi-Tung, Erwie Zahara, and I-Wei Kao. "A hybridized approach to data clus-tering." Expert Systems with Applications 34.3 (2008): 1754-1762.
2. Yang, Fengqin, Tieli Sun, and Changhai Zhang. "An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization." Expert Systems with Applications 36.6 (2009): 9847-9852.
3. Liu, Xiao-yong, and Hui Fu. "An effective clustering algorithm with ant colony." Journal of Computers 5.4 (2010): 598-605.
4. Yaghini, M., and N. Ghazanfari. "Tabu-KM: a hybrid clustering algorithm based on tabu search approach." International Journal of Industrial Engineering 21.2 (2010).
5. Yaghini, M., and N. Ghazanfari. "Tabu-KM: a hybrid clustering algorithm based on tabu search approach." International Journal of Industrial Engineering 21.2 (2010).
6. Karaboga, Dervis, and Celal Ozturk. "A novel clustering approach: Artificial Bee Colony (ABC) algorithm." Applied soft computing 11.1 (2011): 652-657.
7. Yan, Xiaohui, et al. "A new approach for data clustering using hybrid artificial bee colony algorithm." Neurocomputing 97 (2012): 241-250.
8. Mizooji, K. K., A. T. Haghighat, and R. Forsati. "Data clustering using bee colony optimization." 7th International Multi-Conference on Computing in the Global IT. 2012.
9. Hatamlou, Abdolreza. "Black hole: A new heuristic optimization approach for data clustering." Information sciences 222 (2013): 175-184.
10. Saida, Ishak Boushaki, Kamel Nadjet, and Bendjeghaba Omar. "A new algorithm for data clustering based on cuckoo search optimization." Genetic and Evolutionary Computing. Springer International Publishing, 2014. 55-64.
11. Nanda, Satyasai Jagannath, and Ganapati Panda. "A survey on nature inspired metaheuristic algorithms for partitional clustering." Swarm and Evolutionary com-putation 16 (2014): 1-18.

12. Jensi, R., and G. Wiselin Jiji. "An improved krill herd algorithm with global exploration capability for solving numerical function optimization problems and its application to data clustering." Applied Soft Computing (2016).
13. Agarwal, Parul, and Shikha Mehta. "Comparative analysis of nature inspired algorithms on data clustering." 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEE, 2015.
14. Jensi, R., and G. Wiselin Jiji. "MBA-LF: A NEW DATA CLUSTERING METHOD USING MODIFIED BAT ALGORITHM AND LEVY FLIGHT." ICTACT Journal on Soft Computing 6.1 (2015).
15. Guo, Lihong, et al. "A new improved krill herd algorithm for global numerical optimization." Neurocomputing 138 (2014): 392-402.
16. Amiri, Ehsan, and Shadi Mahmoudi. "Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm." Applied Soft Computing 41 (2016): 15-21.
17. Abualigah, Laith Mohammad Qasim, and Essam S. Hanandeh. "APPLYING GENETIC ALGORITHMS TO INFORMATION RETRIEVAL USING VECTOR SPACE MODEL." International Journal of Computer Science, Engineering and Applications 5.1 (2015): 19.
18. Gandomi, Amir Hossein, and Amir Hossein Alavi. "Krill herd: a new bio-inspired optimization algorithm." Communications in Nonlinear Science and Numerical Simulation 17.12 (2012): 4831-4845.
19. Al-Betar, Mohammed Azmi, and Ahamad Tajudin Khader. "A harmony search algorithm for university course timetabling." Annals of Operations Research 194.1 (2012): 3-31.
20. Geem, Zong Woo, Joong Hoon Kim, and G. V. Loganathan. "A new heuristic optimization algorithm: harmony search." Simulation 76.2 (2001): 60-68.
21. Mohd Alia, Osama, et al. "Data clustering using harmony search algorithm." Swarm, Evolutionary, and Memetic Computing. Springer Berlin Heidelberg, 2011. 79-88.
22. Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." Neural Networks, IEEE Transactions on 16.3 (2005): 645-678.
23. Kuo, R. J., and L. M. Lin. "Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering." Decision Support Systems 49.4 (2010): 451-462.
24. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Unsupervised Feature Selection Technique Based on Genetic Algorithm for Improving the Text Clustering.
25. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Unsupervised Feature Selection Technique Based on Harmony Search Algorithm for Improving the Text Clustering.
26. Bolaji, A. L. A., Al-Betar, M. A., Awadallah, M. A., Khader, A. T., Abualigah, L. M. (2016). A comprehensive review: Krill Herd algorithm (KH) and its applications. Applied Soft Computing.
27. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Multi-objectives-based text clustering technique using K-mean algorithm.
28. Abualigah, L. M., Khader, A. T., Al-Betar, M. A., Awadallah, M. A. (2016, May). A krill herd algorithm for efficient text documents clustering. In Computer Applica-tions and Industrial Electronics (ISCAIE), 2016 IEEE Symposium on (pp. 67-72). IEEE.