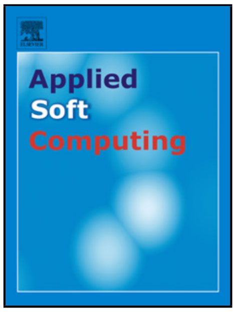


Title: A novel hybridization strategy for krill herd algorithm applied to clustering techniques

Author: Laith Mohammad Abualigah Ahamad Tajudin Khader Mohammed Azmi Al-Betar Amir Hossein Gandomi

PII: S1568-4946(17)30407-6
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2017.06.059
Reference: ASOC 4333

To appear in: *Applied Soft Computing*

Received date: 23-12-2016
Revised date: 2-6-2017
Accepted date: 30-6-2017

# A novel hybridization strategy for krill herd algorithm applied to clustering techniques

Laith Mohammad Abualigah[a,*], Ahamad Tajudin Khader[a], Mohammed Azmi Al-Betar[b], Amir Hossein Gandomi[c]

[a] *School of Computer Sciences, Universiti Sains Malaysia, 11800 Pinang, Malaysia*
[b] *Department of Information Technology, Al-Huson University College, Al-Balqa Applied University, P.O. Box 50, Al-Huson, Irbid, Jordan*
[c] *School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA*

## Abstract

Krill herd (KH) is a stochastic nature-inspired optimization algorithm that has been successfully used to solve numerous complex optimization problems. This paper proposed a novel hybrid of KH algorithm with harmony search (HS) algorithm, namely, H-KHA, to improve the global (diversification) search ability. The enhancement includes adding global search operator (Improvise a new solution) of the HS algorithm to the KH algorithm for improving the exploration search ability by a new probability factor, namely, Distance factor, thereby moving krill individuals toward the best global solution. The effectiveness of the proposed H-KHA is tested on seven standard datasets from the UCI Machine Learning Repository that are commonly used in the domain of data clustering, also six common text datasets that are used in the domain of text document clustering. The experiments reveal that the proposed hybrid KHA with HS algorithm (H-KHA) enhanced the results in terms of accurate clusters and high convergence rate. Mostly, the performance of H-KHA is superior or at least highly competitive with the original KH algorithm, well-known clustering techniques and other comparative optimization algorithms.

*Keywords:* Krill herd algorithm, Hybridization, Global exploration, Data clustering, Text clustering

[*] Corresponding author. Tel.:+60123053158
*Email address:* `lmqa15_com072@student.usm.my` (Laith Mohammad Abualigah)

## 1. Introduction

Clustering is an important unsupervised learning mode that is used in numerous analysis applications to group a set of objects into a subset of coherent groups [1]. Clustering algorithms are labeled into two classes: hierarchical clustering algorithms and partitional clustering algorithms[2]. In the hierarchical algorithms, each object can simultaneously belong to two clusters; each object distributes data objects within a series of partitions either from one cluster to another cluster, including all objects or vice versa. Hierarchical can be either agglomerative or divisive algorithms. Agglomerative algorithms begin with each object as a freelance cluster, merging them in large clusters respectively. By contrast, divisive algorithms begin with all dataset objects and proceed to partition these objects into analogous clusters, respectively. Partition clustering procedures are the primary concern of this paper. This clustering technique attempts to partition a set of data objects into a subset of similar clusters by minimizing the objective function without the hierarchical structure [3].

Data clustering is a common technique that is used for clustering similar or dissimilar data in a provided dataset based on the distance between data objects, thereby indicating that similar objects are partitioned into the same cluster and the dissimilar objects are partitioned into different clusters [3]. The aim of the data clustering technique is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. Data clustering techniques have been used in many areas such as decision-making, data compression, machine learning, data mining, text document retrieval, text categorization, image segmentation, pattern classification, and large transactions for the customer such as marketing analysis, sales management, and document management [4, 5].

Text clustering techniques are used to partition a set of considerably large text documents into a subset of document-related clusters; such a process implies that documents within the same cluster share similar text features whereas the documents within different clusters share dissimilar text features [6]. Subsequently, clustering technique is an important automatic task in text mining domains [7]. In addition, the clustering technique is also an unsupervised learning technique because it deals with a set of documents without any foreknowledge of the class label [8]. This technique is extensively used in numerous areas, such as text classification, information retrieval, search engine, image clustering, and others [9].

In the last few years, several metaheuristic algorithms have been proposed to play a pivotal role in solving several complex optimization problems [10, 7, 11]. Natural metaheuristic algorithms have been used in numerous areas, such as text clustering, data mining, agriculture, computer vision, forecasting, medicine and biology, as well as economics and engineering [12]. Recently, numerous metaheuristic algorithms have been used to improve the data clustering technique, such as k-mean, artificial bee colony (ABC), bee colony optimization (BCO), tabu search approach (TS), particle swarm optimization (PSO), bat algorithm (BA), and harmony search (HS) algorithm [5].

Krill herd algorithm (KHA), which is inspired by the individual krill behaviors of herding, is a novel evolutionary algorithm based on swarm intelligence and bacterial foraging algorithms. This algorithm was introduced by Gandomi and Alavi in 2012 [13]. The algorithm has attracted researchers and has been extensively used in solving numerous optimization problems because of its simple idea and concept, easy implementation, and suitable behavior for clustering techniques [14, 13, 15].

In this paper, a hybrid strategy is proposed to increase the diversity of the KH population by adding HS operator based on a new factor, namely, Distance factor, to make a fine-tune for each position before the updating. The aim of this hybrid strategy is increase its performance and convergence speed. KHA was tested and was not good enough and that is the reason for hybridization [16]. The HS serving (improvise operator) is used to produce a new solution iteratively. Eventually, according to the principle of KH and HS, we merge these two approaches to propose a new hybrid KHA for enhancing the search ability in order to obtain the optimal fitness function value. The proposed method is evaluated on several common standard benchmark datasets which are used in domains of data and text clustering. The results are compared with the other comparative techniques and algorithms. Experimental results illustrate that the hybrid KHA (H-KHA) performs more accurately and efficiently than well-known clustering techniques (i.e., K-mean, k-mean++, Spectral clustering, agglomerative clustering, and DBSCAN), original and hybrid optimization algorithms (i.e., GA, HS, PSO, ABC and CS).

The overall structure of this paper is constructed as follows. Section 2 introduces additional related works in the domain of clustering technique using metaheuristic algorithms. Section 3 describes the proposed H-KHA. Section 4 explains the hybridization architecture of KH algorithm and HS algorithm. Sections 5 and 6 explain the data clustering and text clustering techniques, respectively, along with their results. Finally, Section 7 provides the conclusion and future work.

## 2. Additional related works

In recent years, many scientists who work on data clustering proposed several methods such as metaheuristic algorithms, which they frequently use for solving complex optimization problems [17, 18]. Data clustering is a common data analysis technique that is necessary for many areas [19, 20]. The HS algorithm is proposed as a novel approach to address the problem of data clustering. The proposed algorithm depends on two steps. The first steps explores the available search space of the HS algorithm to identify the optimal clusters of the centroid. The optimal cluster of the centroid is after that evaluated using reformulated c-means as an objective function. In the second steps, the optimal clusters centers obtained are used the initial cluster centers for the c-means algorithm. The experiments were conducted using standard datasets. The results reveal that the proposed HS algorithm reduced the difficulty of selecting the initial clusters of the centroid [21].

An efficient method is introduced for improving the data clustering technique based on cuckoo optimization algorithm (COA) and fuzzy cuckoo optimization algorithm (FCOA). The COA is inspired by the natural life of a cuckoo bird to solve optimization problems. The authors have used COA to cluster a set of data objects into a subset of clusters. The fuzzy logic technique obtained optimal results. This algorithm generates a random solution to the cuckoo population that calculates the cost function for each solution. Finally, fuzzy logic aims to obtain the optimal solution. The performance of the proposed algorithms is evaluated and compared with other algorithms, thereby showing that the proposed algorithm has improved the performance of data clustering [17].

Four versions of hybridization ABC and PSO are used, which includes sequence, parallel, sequence with an enlarged pheromone-particle table, and global best exchange approaches, was proposed to improve the data clustering technique. These hybrid versions were investigated by the data clustering problem. The experimental results showed that the performances of the proposed methods are superior compared with the other standalone algorithms. These experiments were conducted using standard datasets from the UCI Machine Learning Repository. Among the versions of hybridization, the sequence approach is superior to all other approaches because the growth diversifies during the generation of new solutions, thereby preventing being stuck into the local optimum [4].

A hybrid ABC algorithm is proposed to improve the data clustering technique. The main goal of the hybrid ABC algorithm is to enhance the social learning between bees by adding the crossover operator of the genetic algorithm to an artificial bee colony. Ten benchmark functions and six datasets were used to investigate the data clustering technique from the machine learning repository (UCI). The results show that the proposed algorithm is better compared with other algorithms and obtained better results in the data clustering technique [19].

One of the main difficulties in data clustering algorithms is the sensitivity of tuning the initial clusters centroid, which has long elicited the attention of clustering researchers [5, 21]. The author proposed a novel evolutionary algorithm, namely, HS algorithm, as a new method aimed at solving the data clustering problem [21]. The proposed algorithm comprises two stages. In the first stage, the HS investigates the search space of the provided dataset to show the optimal clusters centroid. The center of clusters obtained by the harmony search is evaluated using an objective function of the c-means algorithm. In the second stage, the obtained optimal cluster centers are used as the initial cluster centers for the c-means. Experiments were conducted using standard benchmark data from the UCI Machine Learning Repository; the data showed that the proposed harmony search algorithm had reduced the challenges of choosing an initialization cluster centroid for the c-means clustering [21].

An efficient hybrid optimization algorithm, which is called Tabu-KM, is proposed for solving data clustering problem. This algorithm concurrently gathers the optimization characteristic of tabu search and the local search ability of the k-mean algorithm. The proposed algorithm creates tabu space to escape the trap of the local optimal solution and find optimal solutions. The proposed algorithm is also tested on several standard datasets and its performance is compared to popular algorithms in the domain of text clustering. The experimental results showed that the robustness and efficiency of the proposed algorithm are suitable for enhancing the data clustering problem [3].

A hybrid approach based on particle swarm optimization (PSO) and K- harmonic mean is proposed for the data clustering technique; the approach fully uses the merits of both algorithms. The proposed hybridization algorithm not only helps the K-harmonic mean clustering escape from local optima solution but also overcomes its limitations by tuning the convergence speed of the particle swarm optimization algorithm. The performance of the proposed hybridization algorithm is investigated by using seven datasets that are used for the data clustering technique from the UCI machine learning repository and compared with swarm optimization and K-harmonic mean clustering standalone. Experimental results show the superiority of the hybridization technique [22].

A new hybrid algorithm, namely, differential evolution KH (DEKH), are proposed for solving function optimization to overcome the poor intensification of the KH algorithm [23]. This improvement has been made by appending a hybrid differential evolution (HDE) into the KH to deal with complex optimization problems more efficiently. HDA inspires the intensification and encourages the krill to perform the intensification search within the defined region. Experiments were conducted on 26 optimization functions. The results clearly revealed that the proposed DEKH are adequate to find the accurate solution than the KH and the other comparative methods. As well, the robustness of the DEKH method and the control of the initial population volume on convergence and effectiveness are tested by a set of experiments.

A new hybrid strategy, namely, cuckoo search and krill herd (CSKH), is proposed to make KH more efficient [24]. The CSKH includes krill updating (KU) and abandoning (KA) operator introduced from CS through the process when the krill position was updating. Thus, as to significantly improve its performance and reliability dealing with function optimization problems. The KU operator encourages the exploitation search and allows the krill individuals perform a careful search, while KA operator is applied to improve the exploration search of the CSKH further in place of the poor krill by the end of each iteration. The performance of this strategy is tested by 14 standard optimization functions, and the results reveal that the proposed hybrid strategy of CSKH algorithms is more powerful and efficient than the basic KH and the other comparative methods.

Over the past few years, a large proportion of researchers applied metaheuristic algorithms to solve several optimization problems. However, a major drawback of these algorithms is that it provides a good exploration of the search space at the cost of exploitation [25]. The best results are obtained to solve these problems were by applying the hybrid strategies [23, 24, 3, 18]. However, these reasons can be justified by the no free lunch theorem.

## 3. The proposed hybrid krill herd algorithm

KH is a recent metaheuristic population-based algorithm. The inspiration of the algorithm is the herding behavior of krill individuals when looking for the nearest food; krill herd with high density based on communication with each other [13, 16, 26]. In this paper, the H-KHA is used for enhancing the clustering problem by testing GA, PSO, k-mean, k-mean++, HS, KH and others. The researchers proposed novel H-KHA to solve the problem of the basic KH algorithm. The proposed H-KHA algorithm has three stages as follows:

1. Motion calculation.
2. Genetic operators.
3. Improvising a new solution.

### *3.1. Motion calculation*

H-KH algorithm follows the Lagrangian model for efficacious search, which is calculated by Eq. (1) based on three factors as follows [27, 23]:

1. Movement motivated other individuals krill.
2. Foraging action.
3. Physical diffusion.

$$\frac{dx_i}{dt} = N_i + F_i + D_i, \tag{1}$$

where for krill $i$, $N_i$ is first part, which indicates to the motion induced by other krill individuals, $F_i$ indicates the forging motion, and $D_i$ indicates the physical diffusion of the $i_{th}$ krill individual [13]. The H-KHA factors discussed below.

#### *3.1.1. Movement induced by other krill individuals*

Based on certain theoretical arguments, each krill individual attempts to maintain a high density and closeness to the nearest food. The direction of the induced motion is derived from the local effect of each solution density, a target effect of the density of the individuals, and a repulsive effect of the individuals [28, 15]. Eq. (2) is used to calculate the motion induced by other krill individuals.

$$N_i^{new} = N^{max}\alpha_i + \omega_n N_i^{old} \tag{2}$$

where,

$$\alpha_i = \alpha_i^{local} + \alpha_i^{target} \tag{3}$$

$N^{max}$ is the parameter used to tune the part of the induced motion, $\omega_n$ is the array of random values in the range [0, 1] and $N^{old}$ is the current motion induced. For More details refer [15].

### *3.1.2. Foraging motion*

This factor has two affected parameters: the first is the food location and the second is the old food location. This action can be expressed for the $i_{th}$ krill individual by Eq. (4).

$$F_i = V_f \beta_i + \omega_f F_i^{old} \tag{4}$$

where,

$$\beta_i = \beta_i^{food} + \beta_i^{best}, \tag{5}$$

$V_f$ is the forging speed, $\omega_f$ is the intra weight used to balance the local exploitation and global exploration for each individual, $\beta_i^{food}$ is the food attraction, $\beta_i^{best}$ is the best food attraction so far. For More details refer [15].

$$x^{food} = \frac{\sum_{i=1}^{N} \frac{1}{K_i} x_i}{\sum_{i=1}^{N} \frac{1}{K_i}} \tag{6}$$

### *3.1.3. Physical diffusion*

In this factor, the krill individual is estimated as the random process which used two terms to express the physical diffusion: the first is the maximum diffusion speed and the second is the random directional vector [13]. The physical diffusion is determined by Eq. (7).

$$D_i = D^{max}\left(1 - \frac{I}{I_{max}}\right)\delta \tag{7}$$

where $D^{max}$ is the maximum diffusion speed and $\delta$ is the random values of the vector which has arrays containing random values between [-1, 1]. This action decreased the speed value of the krill individual [15].

### *3.1.4. Motion process of the KH algorithm*

The motion-inducing and foraging motion contained two local and two global strategies [29]. These strategies work in parallel to obtain a powerful algorithm. The physical diffusion generates random vectors [13, 15]. KH algorithm parameters are effective during the algorithm acts. The positions of krill individuals are updated in each iteration using the Langranging model by Eq. (8).

$$x_i(I+1) = x_i(I) + \Delta t\frac{dx_i}{dt} \tag{8}$$

where,

$$\Delta t = C_t\sum_{j=1}^{N}(UB_j - LB_j) \tag{9}$$

$x_i$ is the position $i$ in the search space, $(I+1)$ is the next iteration, $\Delta t$ is an important and more sensitive constant computed by Eq. (9); $N$ represents the total number of variables, the lower bounds $LB_j$, and the upper bounds $UB_j$ of the $i_{th}$ variables $(J = 1, 2, ...., N)$, respectively. $C_t$ is a constant value between [0,2] [13].

### 3.1.5. Genetic operators

Reproduction procedures are incorporated into H-KHA algorithm to improve its performance. Crossover and mutation operators are inspired from the classical differential evolutionary algorithm. For More details refer [13, 15].

### 3.2. Improvise a new solution of H-KHA

Improvising a new solution is the most influential part in HS algorithm, it is used to generate a new solution by global exploration strategy. HS is a stochastic population-based metaheuristic approach that was introduced in 2001 [30]. This algorithm has been successfully applied to solve numerous optimization problems, such as numerical function optimization [14], text clustering [31], and data clustering problems [32].

HS operator is based on the probability of the Distance factor (Def). It is determined by using the distance between each position with the best-achieved fitness function value. Def performs a fine-tune for each position towards the best global solution based on their objective functions using Eq. (10) to enhance the exploration search ability. If the *rand* less than probability value of (Def) that means the current position could be improved using Algorithm (1). This factor is developed as an operator improvement in order to adjust the solutions to prevent the premature convergence during the global (diversification) search. Note, Def value is between (0, 1) because the left part value multiplied by two values, each one less than one and the right part is the objective function of the current position which is between (0, 1).

$$Def(i,j) = \left[ X^{best} * (\frac{1}{S}\sum_{j=1}^{S} |X_j - X^{best}|) \right] + K_j \tag{10}$$

Where, $X^{best}$ is the fitness function of the best solution, $S$ is the number of all solutions, $X_j$ is the solution number $j$ and $\widehat{K}_{i,j}$ is the objective function of the position number $j$.

H-KHA is going to make logical decisions according to the objective function of each krill individual. After obtaining the new position of the krill, solution

**Algorithm 1** Improvising a new solution

```
Improvise a new solution
for each $j \in [1, n]$ do
    if $rand \leq Def_{i,j}$ then
        if $rand$ [0,1] $\leq HMCR$ then
            $x_{i,j}'=KHM[i][j]$ where $i \sim U(1, 2, ...., S)$
            if $rand$ [0,1] $\leq PAR$ then
                $x_{i,j}' = x_{i,j}' \pm rand\times$ bw, where r$\varepsilon$ U(0,1) and $bw$

                else $x_{i,j}' = LBj + rand \times (UBj - LBj)$
            end if
        end if
    end if
end for
```

fitness is evaluated. The selection technique is employed according to Def values for obtaining an optimal krill positions. Thus, after computing the Def value of $j_{th}$ krill position. This value is used to assist the decision making either to update the $j_{th}$ krill position by HS operator or let this position for the KH motion calculation. We conclude that if the objective function of $j_{th}$ krill position is high the Def value is increased, then the HS operator is applied. Otherwise, the old position is preserved in order to keep the current best position in the solution to the next iteration. Finally, HS operator encourages the good solutions (Solutions that have high fitness values) that made by KH motion calculation in order to improve the convergence behavior by preventing the premature convergence.

One of the main motivations to contribute in hybrid KH with HS operator is the HS algorithm one of the most powerful algorithms successfully utilized to solve the text clustering technique [10]. Hence, HS works depend on the HS operator, which is the main operator used to improve the solutions by exploration search. In terms of the KH, it is the unique metaheuristic algorithm owns behavior similar to the clustering technique, so we hybrid these two factors to obtain an effective hybrid algorithm in comparison with other original algorithms, well-known clustering techniques and the other comparative algorithms.

**Improvise a new solution**: Each krill position is updated based on three rules after got permission by $Def$. These rules are: memory consideration, pitch adjustment, and random selection using Eqs. (11) and (12). This operator is applied in krill herd memory (KHM) in order to improve the exploration search ability of the H-KHA.

$$PAR(I) = PAR_{min} + \left(\frac{PAR_{max} - PAR_{min}}{I_{max}}\right) * I, \tag{11}$$

where,

$$bw(I) = bw_{max} exp\left(\frac{In\left(\frac{bw_{min}}{bw_{max}}\right)}{I_{max}}\right) * I, \tag{12}$$

Where, $bw_I$ is the bandwidth for iteration $I$ and $bw_{min}$ and $bw_{max}$ are minimum and maximum adjusting memory consideration, respectively, $PAR(I)$ is the pitch adjusting rate for for iteration $I$, $PAR_{min}$ and $PAR_{max}$ are minimum and maximum pitch adjustment, respectively, $I$ is the number of the current iteration, and $I_{max}$ is the max number of iterations

## 4. H-KHA

The early sections represent the introduction of H-KHA. In this section, we explain the integration of two approaches to made the proposed hybrid krill herd with harmony search algorithm (H-KHA) as shown in Algorithm 2. It modified the KH solutions with poor fitness function to increase the diversity of the KH solutions for improving the search capability and speed up the convergence to the global optimal.

Sometimes, KH gets tripped in the local search or premature convergence in the global search because of the poor exploration. A novel algorithm, namely, H-KHA, which combines a KHA and HS operator is proposed for solve the KH weakness. The main improvement in H-KHA is to add HS operator (i.e., Improvise a new solution). Global search strategy in the HS algorithm (i.e., improvising a new solution) is combined into KH algorithm in order to enable the proposed H-KHA to reach the promising search region. H-KHA is performed to change each krill position into a new proper position by global strategy. After the local search stage, global search is executed with the improvised new solution to obtain the global best solution, then add this solution to the proposed H-KH memory.

**Algorithm 2** Hybrid-krill herd algorithm (H-KHA)

```
Initialization of krill parameters: $N^{max}$, $D^{max}$, $Vf$, $\omega_n$ $etc.$
**for** $i$=1 to $S$ **do**
    **for** $j$=1 to $n$ **do**
        $x_{i,j} = LB_i + (UB_i - LB_i) * U(0,1)$ Initialization of krill memory.
    **end for**
    Evaluate the krill (i)
**end for**
Sort the krill and find $x^{best}$, where $best$ ∈ (1, 2, ..., S)
**while** $I \neq I_{Max}$ **do**
    **for** $i$=1 to $S$ **do**
        Perform the three motion calculation using Eq. (1)
        $x_i(I+1) = x_i(I) + \delta t \frac{dx_i}{dt}$
        Fine-tune $x_i + 1$ by using KH operators: Crossover and mutation.
        Evaluate each krill
    **end for**
    Replace the worst krill with the best krill
    **Improvise a new solution**
    **for** each $j \in [1, n]$ **do**
        **if** $rand \leq Def_{i,j}$ **then**
            **if** $rand$ [0,1] $\leq HMCR$ **then**
                $x_{i,j}'$=$KHM[i][j]$ where $i \sim U(1, 2, ...., S)$
                **if** $rand$ [0,1] $\leq PAR$ **then**
                    $x_{i,j}' = x_{i,j}' \pm rand\times$ bw, where r$\varepsilon$ U(0,1) and $bw$

                    **else** $x_{i,j}' = LB_j + rand \times (UB_j - LB_j)$
                **end if**
            **end if**
        **end if**
    **end for**
    Sort the krill and find $X^{best}$, where $best$ ∈ (1, 2, ..., S)
    $I = I + 1$
**end while**
Return $X^{best}$
```

The H-KHA randomly initializes a krill herd memory (KHM) of krill individuals of size $S * n$. These individuals may be regarded as a herd in the case of KH algorithm or as harmonies in the case of HS algorithm. Subsequently, the $S * n$ krill individuals are sorted by the fitness function and are fed into the HS to improve the H-KHA by generating a new solution and adding it to the population if it is better than the worst solution. Thus, H-KHA can use better individuals in each iteration to search for the optimum solution. Also, krill individuals with poor performance remain to avoid premature convergence. In this strategy, the krill positions are explored the search space effectively towards the best solution using the Def. Consequently, it converges quickly and the diversity problem is avoided. The mainframe of the hybrid krill herd with harmony search (H-KHA) is presented in Fig. 1.

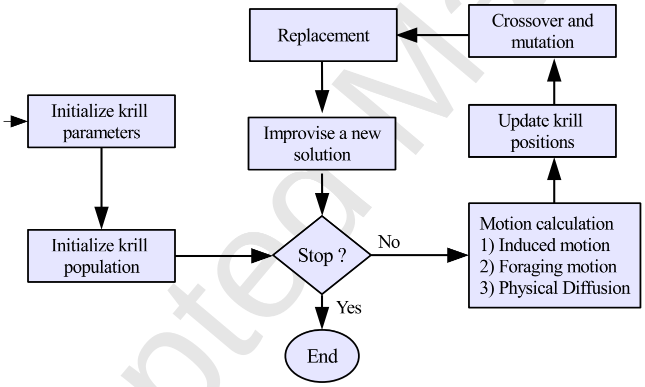


Figure 1: A flowchart of the hybrid krill herd algorithm (H-KHA)

## 5. Data clustering using H-KHA

Data clustering is a popular technique that is used to partition a set of data objects and statistical data analysis [33, 34], in which a cluster of data objects is distributed in such a way that the data objects within the same clusters are similar and the data objects in different clusters are dissimilar [14, 35]. The similarity between each data object with the cluster centroid is determined by a distance metric [14].

### *5.1. Data clustering formulation*

Data clustering technique is the process of partitioning the set of data objects into a subset of $K$ clusters based on certain distance measure [2, 36]. Let $D$ be a set of $n$ data objects $D = d_1, d_2, ..., d_i, ..., d_n$ to be distributed over $K$ clusters and each data object $d_i$, $i = 1....n$ is represented as vector $d_i = d_{i1}, d_{i2}, ..., d_{ij}, ..., d_{it}$, where $d_{ij}$ represents $j_{th}$ dimension value of the data object number $i$ and $t$ is the length of each object [3, 14]. The aim of clustering algorithm is to find a subset of $K$, where each cluster belongs to one centroid as $C = c_1, c_2, ..., c_k, ..., c_K$ and $c_k \neq \emptyset$. The measurement of these similarities are evaluated by certain optimization criterions, particularly distance measure and squared error function [14], which have been calculated by Eqs. (13) and (14).

$$FF = \sum_{i=1}^{n} \sum_{j=1}^{K} min(Des(d_i, c_j)), \tag{13}$$

where $c_j$ represents a $j_{th}$ cluster center, $d_j$ represents a $j_{th}$ data object, and $Des$ is the distance measure between the object $d_i$ and the cluster center $c_j$ . This criterion 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, Manhattan, Minkowski, Cosine, Pearson correlation coefficient, and Jaccard coefficient measures [37, 14]. In this paper, Euclidean distance is used as distance measure from numerous distance metrics used in the literature, which is defined by Eq. (14).

$$Des(d_i, c_j) = \sqrt{\sum_{j=1}^{t} (d_{1j}, c_{2j})^2}, \tag{14}$$

where $Des(d_i, c_j)$ is the 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 cluster 1 [14], which is calculated by Eq. (18).

$$c_j = \frac{1}{n_j} \sum_{d_i \epsilon c_j} d_i, \tag{15}$$

where $d_i$ represents the object $i$, $c_j$ represents the cluster centroid of cluster $j$, and $n_j$ is the total number of objects in cluster $j$. The centroid of each cluster is repeatedly computed based on data objects assigned to the cluster. This process proceed continuously until the maximum number of iterations is reached (stopping criteria) [2]. Note, each cluster has one centroid.

### *5.2. Experimental results of H-KHA for data clustering*

This section applies seven standard data benchmark datasets to validate the proposed H-KHA method in comparison with K-mean [14], K-mean++ [14], Spectral clustering [38], Agglomerative clustering [39], DBSCAN [40], GA [41], HS [21], PSO [42], CS [3], ABC [19] and other hybrid algorithms [14]. These comparative methods are programmed by the authors as discussed in their publications. A more detailed explanation is presented as follows.

#### *5.2.1. Benchmark Datasets*

In this paper, we conducted the results using seven datasets provided by the Machine Learning Repository (UCI) of the University of California[1], namely, (CMC, Iris, Vowel, Seeds, Cancer, Glass, and Wine). Table 1 presents the related information that given in each dataset, including the number of datasets, dataset name, number of features, and number of clusters in advance.

#### *5.2.2. Data preprocessing and parameter sitting*

The features in each dataset have variant value ranges. Accurate clusters cannot be obtained if these values are quite different [43]. Determining the parameter values are inevitable for the proposed H-KHA and other comparative

[1] http://archive.ics.uci.edu/ml/datasets.html

Table 1: Characteristic of the data clustering datasets

| Datasets Number | Datasets Name | Number of Features | Number of Clusters | Number of Objects |
|---|---|---|---|---|
| DS1 | CMC | 3 | 9 | 1473(629,334,510) |
| DS2 | Iris | 4 | 3 | 150(50,50,50) |
| DS3 | Vowel | 6 | 3 | 871(72,89,172,151,207,180) |
| DS4 | Seeds | 7 | 3 | 210(70, 70, 70) |
| DS5 | Cancer | 9 | 2 | 683(444,239) |
| DS6 | Glass | 9 | 6 | 214(70,76,17,13,9,29) |
| DS7 | Wine | 13 | 3 | 178(59,71,47) |

algorithms. The related information is summarized in Table 2. The KH algorithm parameters are chosen accounting to the standard values of $V_f$, $D_{max}$, 380 $N_{max}$, $HMCR$, and etc [13]. These parameters are recorded in Table 2.

Table 2: Characteristics of H-KHA

| H-KHA | value |
|---|---|
| Number of solution | 20 |
| Number of generation | 1000 |
| $V_f$ | 0.02 |
| $D^{max}$ | 0.002 |
| $N^{max}$ | 0.01 |
| $N^{max}$ | 0.05 |
| HMCR | 0.90 |
| PARmin | 0.45 |
| PARmax | 0.90 |
| bwmin | 0.10 |
| bwmax | 1.00 |

### *5.2.3. Data clustering results and discussion*

In this paper, the experiments of data clustering problem are implemented in a Windows 7 environment using MATLAB (7.10.0) computer programming with different CPU and RAM capabilities using different standard benchmark 385 datasets. The performance of the proposed algorithms is evaluated and compared with other popular algorithms using two criteria include (1) The sum of the intra-cluster distances is considered as an internal quality measure. The distance value between each object and the cluster centroid of the corresponding cluster is computed, as defined in Eq. (13). A higher quality data clustering provides a small fitness function [12, 17, 44]. (2) Error Rate (ER) value is an external quality measure. The percentage of misplaced objects on the overall number of objects [12, 14, 45], as shown in Eq. (16).

$$ER = \frac{number of misplaced objects}{size of test dataset} * 100, \tag{16}$$

ER measure is assigned a class label and compared with the desired class label. The pattern is distributed as incorrect partitioning if these measures are dissimilar. The measure is calculated for all data objects in the provided dataset and the total incorrect number of partitioning pattern is a percentage of the size of all data objects in the dataset. A summary of the ER obtained by the data clustering algorithms is provided in Table 3. The values reported are worst, average, and best solutions over 20 independent runs [14]. The experimental results are provided in Table 3, which shows that the proposed hybrid H-KHA algorithm obtains near optimal solutions compared with the comparative algorithms. The proposed H-KHA achieves better results for almost all datasets with small final rankings. Note, the statistical ranking of each comparative algorithm is based on the average ER among the seven datasets.

Table 3 shows the ER of using five clustering techniques and seven metaheuristic optimization algorithm to enhance the data clustering technique. For the CMC dataset, H-KHA obtains the best value over (average ER), whereas k-means++ obtains the best value over (best ER) and spectral clustering obtains the best value over (worst ER). For the Iris dataset, H-PSO obtains the best value over (average ER), whereas H-KHA obtains the best value over (best ER) and DBSCAN clustering obtains the best value over (worst ER). For the Vowel dataset, H-KHA obtains the best value for the overall statistic measures (average, best, and worst ER). For the Seeds dataset, H-KHA obtains the best value

Table 3: Error rate results for seven datasets

| | | Clustering techniques | | | | | Optimization algorithms | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Statistics** | **K-mean** | **K-mean++** | **Spectral** | **Agglomerative** | **DBSCAN** | **GA** | **PSO** | **HS** | **KHA** | **H-GA** | **H-PSO** | **H-KHA** |
| **CMC** | Best | 54.660 | **52.003** | 53.541 | 52.391 | 54.280 | 54.656 | 54.101 | 55.430 | 53.936 | 53.124 | 53.201 | 52.213 |
| | Average | 55.470 | 56.258 | 55.120 | 54.944 | 56.544 | 56.697 | 55.899 | 56.001 | 56.056 | 55.142 | 54.204 | **53.656** |
| | Worst | 56.667 | 57.001 | **54.044** | 57.487 | 56.654 | 57.296 | 56.486 | 57.906 | 56.999 | 56.214 | 55.333 | 54.333 |
| | Rank | 6 | 10 | 4 | 3 | 11 | 12 | 7 | 8 | 9 | 5 | 2 | 1 |
| **Iris** | Best | 10.660 | 10.101 | 10.547 | 9.874 | 9.987 | 10.666 | 10.667 | 10.509 | 8.430 | 9.765 | 9.666 | **9.000** |
| | Average | 21.467 | 20.983 | 17.458 | 18.544 | 16.311 | 21.652 | 15.867 | 21.054 | 22.658 | 21.100 | **15.800** | 19.866 |
| | Worst | 56.667 | 54.274 | 55.541 | 48.397 | **43.111** | 43.333 | 43.447 | 44.286 | 42.548 | 44.667 | 44.333 | 43.333 |
| | Rank | 10 | 7 | 4 | 5 | 3 | 11 | 2 | 8 | 12 | 9 | 1 | 6 |
| **Vowel** | Best | 16.245 | 15.364 | 15.645 | 15.114 | 15.645 | 14.164 | 15.555 | 15.485 | 14.514 | 15.945 | 15.099 | **14.010** |
| | Average | 16.547 | 16.177 | 17.540 | 16.541 | 16.477 | 15.298 | 15.957 | 16.000 | 16.501 | 16.987 | 15.800 | **14.666** |
| | Worst | 18.551 | 17.560 | 18.006 | 17.450 | 17.499 | 16.254 | 19.547 | 18.562 | 18.659 | 19.245 | 17.568 | **17.154** |
| | Rank | 10 | 6 | 12 | 9 | 7 | 2 | 4 | 5 | 8 | 11 | 3 | 1 |
| **Seeds** | Best | 12.643 | 11.254 | 9.454 | 10.254 | 12.322 | 13.810 | 8.571 | 12.290 | 13.595 | 9.565 | 9.047 | **8.623** |
| | Average | 12.643 | **11.200** | 11.297 | 13.021 | 12.214 | 21.000 | 15.262 | 13.015 | 13.595 | 13.458 | 13.881 | 11.666 |
| | Worst | 12.643 | **14.111** | 18.451 | 20.214 | 19.397 | 25.714 | 36.190 | 16.021 | 20.321 | 19.525 | 19.238 | 14.523 |
| | Rank | 5 | 1 | 2 | 7 | 4 | 12 | 11 | 6 | 9 | 8 | 10 | 3 |
| **Cancer** | Best | 39.865 | 39.500 | **38.111** | 39.148 | 39.654 | 39.510 | 40.775 | 40.111 | 39.256 | 40.254 | 39.775 | 38.670 |
| | Average | 42.388 | 40.145 | 40.154 | 41.645 | 42.199 | 44.270 | 43.051 | 42.054 | 42.543 | 41.214 | 39.125 | **39.012** |
| | Worst | 45.970 | 44.965 | 44.685 | 46.699 | **44.021** | 47.753 | 45.455 | 45.640 | 47.191 | 46.214 | 46.758 | 44.154 |
| | Rank | 9 | 3 | 4 | 6 | 8 | 12 | 11 | 7 | 10 | 5 | 2 | 1 |
| **Glass** | Best | 42.262 | 45.123 | 38.541 | **32.001** | 33.717 | 42.991 | 43.925 | 41.162 | 38.318 | 35.249 | 41.589 | 32.242 |
| | Average | 46.154 | 44.566 | 46.614 | 43.222 | 44.984 | 51.028 | 46.262 | **42.054** | 43.925 | 44.219 | 47.617 | 42.219 |
| | Worst | 46.215 | **45.250** | 51.991 | 52.140 | 51.123 | 56.075 | 52.804 | 46.255 | 50.476 | 51.985 | 56.075 | 51.420 |
| | Rank | 8 | 6 | 10 | 3 | 7 | 12 | 9 | 1 | 4 | 5 | 11 | 2 |
| **Wine** | Best | 29.775 | 30.546 | **29.189** | 30.665 | 30.140 | 29.310 | 29.775 | 29.865 | 29.213 | 29.654 | 29.775 | 29.650 |
| | Average | 32.388 | 31.841 | 33.585 | 34.154 | 33.487 | 34.270 | 32.051 | 32.568 | 32.303 | 30.989 | **30.871** | 33.000 |
| | Worst | 43.820 | 43.534 | 43.137 | 42.688 | **42.009** | 47.753 | 44.449 | 44.467 | 47.191 | 44.001 | 43.888 | 42.134 |
| | Rank | 6 | 3 | 10 | 11 | 9 | 12 | 4 | 7 | 5 | 2 | 1 | 8 |
| **Mean rank** | | 7.71 | 5.14 | 6.57 | 6.28 | 7.00 | 10.42 | 6.85 | 6.00 | 8.14 | 6.42 | 4.28 | 3.14 |
| **Final rank** | | 10 | 3 | 7 | 5 | 9 | 12 | 8 | 4 | 11 | 6 | 2 | 1 |

Note: The lowest ranked algorithm is the best one.

over (best ER), whereas k-mean++ obtains the best value over two statistic measures (average, and worst ER). For the Cancer dataset, H-KHA obtains the best value over (average ER), whereas spectral clustering obtains the best value over (best ER) and DBSCAN clustering obtains the best value over (worst ER). For the Glass dataset, HS obtains the best value over (average ER), whereas agglomerative clustering obtains the best value over (best ER) and k-mean++ clustering obtains the best value over (worst ER). Finally, for the Wine dataset, H-PSO obtains the best value over (average ER), whereas spectral clustering obtains the best value over (best ER) and DBSCAN clustering obtains the best value over (worst ER). Thus, the proposed method able to tackle the clustering problem very well in comparison with other clustering techniques and other optimization algorithms.

H-KHA failed to reach the best value in all runs; however, it obtained the number one rank among all comparative algorithms. As regards Vowel data sets, H-KHA achieved the best optimum value of 14.666 overall runs and for Cancer data sets, H-KHA achieved the best optimum value of 39.012 overall runs and so on. Thus, H-KHA algorithm obtained the near best value in all runs. The statistical analysis is done based on average ER. The average rankings of the clustering algorithms are reported in Table 3. The proposed H-KHA is ranked the highest one, which is followed by H-KHA, H-PSO, k-mean++, HS, agglomerative clustering, H-GA, spectral clustering, PSO, DBSCAN, k-mean, KHA, and GA among the datasets.

Table 4 shows the objective function values obtained by two clustering techniques and all optimization algorithms for different data clustering datasets. Note, some clustering techniques are not dealing with an objective function like DBSCAN clustering. The proposed H-KHA obtained the best results among all the comparative algorithms. For the CMC, Iris, Vowel, and Wine datasets, H-KHA is better than the other values of algorithms based on its obtained best, average, and worst solutions. The results for the seeds and Glass datasets showed that the worst solution obtained by the H-KHA are 4399.254 and 217.355, respectively, which is much better than the worst solutions that were obtained

Table 4: The objective function values obtained by algorithms for different datasets

| **Dataset** | **Statistics** | **GA** | **PSO** | **K-mean** | **K-mean++** | **HS** | **KHA** | **H-GA** | **H-PSO** | **H-KHA** |
|---|---|---|---|---|---|---|---|---|---|---|
| **CMC** | Best | 5641.556 | 5609.369 | 5609.254 | 5710.915 | 5698.564 | 5688.854 | 5745.548 | 5699.964 | **5586.532** |
| | Average | 5790.448 | 5760.294 | 5689.299 | 5770.648 | 5781.854 | 5795.546 | 5714.287 | 5706.654 | **5601.681** |
| | Worst | 5896.101 | 5790.120 | 5879.119 | 5994.545 | 5814.861 | 5836.266 | 5847.214 | 5880.565 | **5666.943** |
| | Rank | 8 | 5 | 2 | 6 | 7 | 9 | 4 | 3 | **1** |
| **Iris** | Best | 113.986 | 96.148 | 97.325 | 97.325 | 98.648 | 96.744 | 96.989 | 96.590 | **96.154** |
| | Average | 125.197 | 97.997 | 104.576 | 99.569 | 98.447 | 96.914 | 96.752 | 97.235 | **96.524** |
| | Worst | 139.778 | 98.287 | 123.969 | 110.650 | 99.144 | 97.909 | 97.154 | 97.985 | **96.989** |
| | Rank | 9 | 5 | 8 | 7 | 6 | 4 | 3 | 2 | **1** |
| **Vowel** | Best | 152.648 | 149.540 | 149.499 | 161.154 | 156.155 | 155.684 | 150.145 | 149.348 | **149.123** |
| | Average | 153.697 | 150.154 | 149.990 | 161.990 | 156.489 | 156.244 | 152.145 | 149.954 | **149.565** |
| | Worst | 153.993 | 150.861 | 150.346 | 162.146 | 157.548 | 157.369 | 155.124 | 150.023 | **149.999** |
| | Rank | 6 | 4 | 3 | 9 | 8 | 7 | 5 | 2 | **1** |
| **Seeds** | Best | 4523.125 | 4532.142 | 4480.162 | 4471.654 | 4412.525 | 4490.815 | 4421.325 | 4395.445 | **4350.140** |
| | Average | 4550.910 | 4607.564 | 4488.542 | 4483.709 | 4450.848 | 4499.840 | 4521.648 | **4390.900** | 4391.546 |
| | Worst | 4621.021 | 4660.216 | 4499.649 | 4493.197 | 4460.464 | 4510.988 | 4596.145 | 4419.410 | **4399.254** |
| | Rank | 8 | 9 | 6 | 5 | 3 | 4 | 7 | **1** | 2 |
| **Cancer** | Best | 3010.325 | 3002.450 | 2989.258 | 3064.694 | 2988.856 | 3012.648 | 2998.259 | 2988.654 | **2975.191** |
| | Average | 3050.365 | 3012.394 | 2992.640 | 3086.159 | 2990.654 | 3045.542 | 2999.489 | 2990.263 | **2982.437** |
| | Worst | 3085.259 | 3090.910 | 3001.193 | 3094.369 | 2998.286 | 3089.654 | 3075.295 | 2993.372 | **2990.493** |
| | Rank | 8 | 6 | 4 | 9 | 3 | 6 | 5 | 2 | **1** |
| **Glass** | Best | 285.654 | 279.767 | 216.198 | 216.447 | 243.157 | 273.258 | 215.975 | **213.162** | 213.105 |
| | Average | 290.159 | 282.734 | 220.214 | 218.845 | 246.254 | 276.441 | 217.659 | 216.658 | **215.665** |
| | Worst | 294.565 | 285.299 | 222.015 | 221.125 | 251.556 | 279.982 | 225.154 | 218.851 | **217.355** |
| | Rank | 9 | 8 | 5 | 4 | 6 | 7 | 3 | 2 | **1** |
| **Wine** | Best | 16,854.654 | 16,764.901 | 16,642.141 | 17,142.158 | 16,759.449 | 16,954.901 | 16,765.456 | 16,980.297 | **16,350.154** |
| | Average | 17,900.154 | 17,964.390 | 16,976.691 | 17,556.149 | 16,945.698 | 17,015.148 | 16,765.456 | 17,587.145 | **16,410.147** |
| | Worst | 18,397.490 | 19,297.145 | 20,452.147 | 17,947.154 | 16,989.931 | 17,152.697 | 16,765.456 | 18,564.154 | **16,961.147** |
| | Rank | 8 | 9 | 5 | 7 | 4 | 6 | 3 | 2 | **1** |
| **Mean rank** | | 8.00 | 6.57 | 4.71 | 6.71 | 5.28 | 6.14 | 2.28 | 2.00 | **1.14** |
| **Final rank** | | 9 | 7 | 4 | 8 | 5 | 6 | 3 | 2 | **1** |

Note: The lowest ranked algorithm is the best one.

by the other algorithms. Finally, the average solution obtained by H-PSO as regards the Seeds dataset is 4390.900; by contrast, the best solution obtained by H-PSO for the Glass dataset is 217.355. These solutions are much better than those of the other algorithms. From these results, the proposed H-KHA was found to be superior in comparison with the other comparative algorithms because it can identify high-quality clusters.

Table 5: Identifying the clusters centroid for the Iris dataset

| Centroids | calculated centroid | | | |
|---|---|---|---|---|
| | Att.1 | Att.2 | Att.3 | Att.4 |
| Centroid1 | 5.624 | 3.102 | 3.210 | 1.355 |
| Centroid2 | 5.325 | 3.001 | 3.749 | 1.296 |
| Centroid3 | 5.623 | 2.998 | 3.447 | 1.191 |

Table 6: Identifying the clusters centroid for the Vowel dataset

| Centroids | calculated centroid | | |
|---|---|---|---|
| | Att.1 | Att.2 | Att.3 |
| Centroid1 | 461.4 | 1430.5 | 2651.5 |
| Centroid2 | 432.1 | 1490.3 | 2690.4 |
| Centroid3 | 471.2 | 1502.9 | 2669.4 |
| Centroid4 | 448.5 | 1490.5 | 2544.6 |
| Centroid5 | 482.4 | 1477.6 | 2539.1 |
| Centroid6 | 475.6 | 1409.8 | 2691.3 |

The proposed hybrid KHA and PSO have been demonstrated to be superior compared to an original standalone KH algorithm in this paper. In addition , in Table 4, H-KHA might lead to achieving superior solutions after a specific number of iterations. As an example, Tables 5 and 6 describe the calculated cluster centers for two datasets (Iris and wine) to identify the best solution. The best cluster centroid is displayed to validate the sum of the objective function (intra-cluster distances) values in Table 4. Each data object is assigned to the closest centroid in Tables 6 to reach the best values. For example, all of the 871 objects are assigned within the Vowel dataset to the closest centroid among the six cluster centers, which is presented in Table 6. Moreover, a new position update approach is proposed by H-KHA to enhance the ability of the search space. Subsequently, we conclude based on all of these results that the H-KHA obtained the best results almost on all the tested datasets.

The statistical analysis is done based on two measures include error rate and objective function. The average rankings of the clustering algorithms are reported according to error rate measure in Table 3. The proposed H-KHA is ranked the highest one among the datasets. Also, in Table 4, the average rankings of the clustering algorithms are reported according to objective function measure. Again, the proposed H-KHA is ranked the highest one, which is followed by H-PSO, H-GA, k-mean, HS, KHA, PSO, k-mean++ and GA, among the datasets.

This section explains the experimental results to empirically investigate the validation of the proposed H-KHA for data clustering problem and compare with the successful comparative algorithms. We concluded from Tables 3 and 4 that the proposed H-KHA obtained the best performance according to error rate and objective function measures compared with the other comparative algorithms. The H-KHA performed better on all seven data clustering datasets compared with the clustering techniques and the comparative algorithms. A proper balance between exploitation and exploration improves the performance of the proposed hybrid KH algorithms. It passes by obtaining the best results on almost all dataset and in comparison with all the other comparative algorithms to prove that the proposed hybrid strategy (H-KHA) is very effective to solve complex optimization problem by Adding new operators (HS operator) to the main structure of the KHA using a new probability value that reproduction by $Def$.

## 6. Text clustering using H-KHA

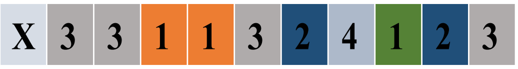
The text clustering technique aims to generate optimal text document clusters that contain relevant (similar) documents. This technique is based on partitioning a set of text documents into a subset of related clusters, in which each cluster contains a set of similar text documents, whereas different clusters contain dissimilar text documents [46, 47, 18].

### *6.1. Text clustering problem descriptions and formulations*

A combination of text documents $D$ is partitioned into $K$ clusters, where $D$ refers to a vector of documents $D = (d_1, d_2, d_i, ....., d_n)$, $d_1$ is the document number 1, $i$ is the document number $i$, and $n$ is the number of all documents in $D$ [31, 48, 49]. Each cluster has a cluster centroid $c_k$, which stands as a vector of terms weight $c_k = (c_{k1}, c_{k2}, ..., c_{kj}, ..., c_{kt})$, where $c_k$ is the $k_{th}$ cluster centroid, $c_{k1}$ is the value of position 1 in the centroid of cluster $k$, and $t$ is the number of all unique centroid terms (length). Note that as previously mentioned, each cluster centroid represents as a vector such as any document with the same dimension. The similarity measure is used to assign each text document to the similar cluster centroid [1, 10, 50].

### *6.2. Text clustering solution representation*

The text clustering solution is represented by a vector $X$= $(x_1, x_2, ..., x_i, ..., x_n)$, where $X$ represents one solution and $x_i$ is the value of the position $i$, this value represents that the document number $i$ belong to which cluster. The available range for each document is $[1, 2, ..., K]$, where $K$ is the number of all clusters [16, 51]. Fig 2 shows the representation of the document-clustering solutions in H-KHA. In this case, cluster 1 contains three documents (i.e., 3, 4, and 8), cluster 2 contains two documents (i.e., 6 and 9), cluster 4 contains four documents (i.e., 1, 2, 5, and 10), and cluster 4 contains one document (i.e., 7).



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 3 | 3 | 1 | 1 | 3 | 2 | 4 | 1 | 2 | 3 |

Figure 2: Representation of the text clustering solution.

### *6.3. Similarity measure*

Cosine measure is a common similarity gauge used in the unsupervised text document clustering technique to compute the similarity between two vectors by Eq. (17). Where $d_1$ is the document number 1 which represents as a vector of term weights $d_1 = (w_{11}, w_{12}, w_{13}, ....., w_{1t})$, and $c_1$ is the centroid of cluster 2, which represents as a vector of terms weight $c_2 = (w_{21}, w_{22}, w_{23}, ....., w_{2t})$, as defined below [52, 7, 53].

$$Cos(d_1, c_2) = \frac{\sum_{j=1}^{t} w(t_j, d_1) \times w(t_j, c_2)}{\sqrt{\sum_{j=1}^{t} w(t_j, d_1)^2} \sqrt{\sum_{j=1}^{t} w(t_j, c_2)^2}}, \tag{17}$$

where $w(t_j, d_1)$ is the weight of the term $j$ in document 1, and $w(t_j, c_2)$ is the weight of the term $j$ in the centroid of cluster 2. This measure returns to *one* if the document and the centroid are conformable and returns to *zero* if the document and the centroid are different.

### *6.4. Fitness function*

Fitness function (FF) is calculated for evaluating each solution based on its current position using Eq. 19, sorting it in an ascending order. KH memory contains several solutions to solve the text clustering problem. Each solution in the KH memory illustrates the candidate solution to solve the problem of documents clustering. Each solution has a set of $K$ centroid $C = (c_1, c_2, ...., c_k, c_K)$, where $c_k$ is the centroid of cluster $k$, which represents a vector $c_k = (c_{k1}, c_{k2}, ..., c_{kj}, ..., c_{kt})$ and is computed by Eq. (18) [31].

$$c_{kj} = \frac{\sum_{i=1}^{n} (A_{kj}) d_j}{r_i}, \tag{18}$$

where $d_i$ is the document number $i$ that belongs to centroid $c_j$ of the cluster $j$; $A_{kj}$ represents that document $k$, which belongs to cluster $j$ and $r_i$, is the number of documents in each cluster. The cosine similarity is used as an objective function to evaluate each solution position. The fitness function for each solution in the KH memory is determined by the average similarity of documents to the cluster centroid ($ASDC$) as represented by Eq. (19).

$$ASDC = \left[ \frac{\sum_{i=1}^{k} (\frac{\sum_{j=1}^{r_i} Cos(c_i, d_j)}{r_i})}{K} \right], \tag{19}$$

.

where $K$ is the number of all clusters, $r_i$ is the number of documents in cluster number $i$, and $Cos(c_i, d_i)$ is the similarity measure between document number $i$ and cluster centroid number $i$.

### *6.5. Experimental results of H-KHA for text document clustering*

Also, the experiments of this section are implemented in a Windows 7 environment using MATLAB (7.10.0) computer programming with different CPU and RAM capabilities using different standard benchmark text datasets in the text clustering domain. The proceeding subsections explain datasets in detail, illustrate the evaluation criteria, and present the results of experiments and discussion. Note, clustering techniques[2] are available at Scikit-Learn Machine Learning in Python.

#### *6.5.1. Standard text document datasets*

Table 7 shows six standard benchmark text datasets that are used to analyze and compare the performance of the proposed hybrid algorithm [54]. Text clustering standard datasets[3] are available at Laboratory of Computational Intelligence (LABIC) by numerical form after the terms extraction [8]. More than 20 experimental runs were performed for statistical comparisons. This number, which was selected based on the literature, can sufficiently validate the proposed method. Local-based algorithms for clustering technique run 100 iterations in each run time. Experimentally, 100 iterations are adequate for the convergence of intensification search algorithm and 1000 iterations are adequate

[2] http://scikit-learn.org/stable/modules/clustering.html

[3] http://sites.labic.icmc.usp.br/text_collections/

for the convergence of diversification search algorithm for clustering techniques [55].

Table 7: Text document datasets characteristics

| Datasets | Number of Documents (d) | Number of Terms (t) | Number of Clusters (K) |
|---|---|---|---|
| Classic4 | 500 | 1800 | 2 |
| Classic4 | 2000 | 6500 | 4 |
| Reuters21578 | 3000 | 10150 | 7 |
| 20Newsgroup | 5000 | 20140 | 9 |
| Reuters21578 | 2000 | 7481 | 10 |
| 20Newsgroup | 2000 | 9560 | 20 |

The main characteristics of these text document datasets are as follows. The first dataset (DS1), Classic4, contains randomly 500 documents on two topics: cacm and cran. The second dataset (DS2), Classic4, contains randomly 2000 documents on four topics: cacm, cisi, med and cran. The third dataset (DS3), Reuters21578, contains randomly 3000 documents on seven topics: gold, tin, sunseed, nkr, lei, hog and citruspulp. The fourth dataset (DS4), 20Newsgroup, contains randomly 5000 documents on nnine topics: comp-windows-x, rec-autos, talk-politics-misc, comp-sys-mac-hardware, talk-religion-misc, misc-forsale, sci-crypt, sci-med and rec-motorcycles. The fifth dataset (DS5), Reuters21578, contains randomly 2000 documents on ten topics: gold , tin, sunseed , nkr, lei, hog, citruspulp, gas, peseta andnzdlr. The sixth dataset (DS6), 20Newsgroup, contains randomly 2000 documents on twenty topics: comp-windows-x, rec-autos, talk-politics-misc, comp-sys-mac-hardware, talk-religion-misc, misc-forsale, sci-crypt, sci-med, rec-motorcycles, alt-atheism, sci-electronics, soc-religion-christian, rec-sport-baseball, talk-politics-mideast, rec-sport-hockey, sci-space, comp-sys-ibm-pc-hardware, comp-os-ms-windows-misc, comp-graphics and talk-politic-guns.

*6.5.2. Evaluation measures*

Four external evaluation measures were conducted for comparative evaluations: Accuracy measure ($Ac$), Precision measure ($P$), Recall measure ($R$), and F-measure ($F$). These measures are popular evaluation criteria used to accurately identify the clusters and compare the different clustering methods [55, 56, 51].

*Precision and recall measures*

The precision and recall measurements are used together to calculate the F-measure score for cluster $j$ and class $i$ [1, 57, 49]. Eq. (20) and (21) are used to calculate the precision and recall measures, respectively.

$$P(i,j) = \frac{n_{i,j}}{n_j}, \tag{20}$$

$$R(i,j) = \frac{n_{i,j}}{n_i}, \tag{21}$$

where $n_{ij}$ is the number of true documents of class $i$ in cluster $j$, $n_j$ is the number of all documents of cluster $j$, and $n_i$ is the number of all documents of class $i$.

*F-measure evaluation*

F-measure is a common external measurement that is used in particular on the text clustering domain. F-measure calculates the percentage of the matched clusters depending on Precision and Recall measurements by Eq. (23) [57].

$$F(i,j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)}, \tag{22}$$

where $P(i,j)$ is the precision of the true documents of class $i$ in cluster $j$, $R(i,j)$ is the recall of the class $i$ in cluster $j$, and F-measure for all clusters is calculated by Eq. (23) according to the number of all documents $n$.

$$F = \sum_{j} \frac{n_j}{n} \max_{i}\{n(i,j)\}, \tag{23}$$

*Accuracy evaluation*

The accuracy is one of the popular external measurements that is used to precisely compute the percentage of true assigned text documents to each cluster by using Eq. (24) [8].

$$AC = \frac{1}{n} \sum_{i=1}^{K} n_{i,i} \tag{24}$$

Where, $n_{i,j}$ is the number of documents of the class $i$ in cluster $i$, $n$ is the number of all documents and $K$ is the number of all clusters are given in the dataset.

*6.5.3. Text clustering results and discussion*

Table 8 shows that the performance of the H-KHA is better than the original KH algorithm and the other comparative algorithms with regard to the five clustering measures include ASDC, Accuracy, Precision, Recall and F-measure. The proposed H-KHA realizes the best performance as regards the average F-measure. In particular, The performance of H-KHA according to the F-measure is better than the other comparative algorithms in all text datasets. The proposed algorithm also performs better than the original KH algorithm over the whole datasets.

Table 8 shows the performance of the H-KHA based on the quality of clusters using six benchmark standard text datasets. The proposed algorithm apparently performed very well and exceeded the popular algorithms. The proposed algorithm also scored a better performance based on Accuracy Precision, Recall, and F-measure as an external measurement in two dataset (i.e., DS1, and DS2) and almost it obtained near optimal clusters in the other datasets (i.e., DS3, DS4, DS5, and DS6). Subsequently, the proposed H-KHA significantly improved in comparison with the original KH algorithm. Notably, all experiments show that the proposed H-KHA obtained better results in comparison with the comparative algorithms over all datasets, which can be attributed to the high quality of clusters. The performance measures (i.e., precision, recall, accuracy and F-measure) are obtained by using the proposed H-KHA on all the datasets and comparable with that obtained from the comparative algorithms in all datasets.

A novel hybrid algorithm, namely, H-KHA with improved global search ability, which combines the extended search ability of HS operator and uses the search ability of the hybrid computation algorithm and the global increase capacity of the KH algorithm, is proposed. Besides, a new position update approach is proposed to enhance the ability of the search space. In conclusion, H-KHA obtained the best results for all the tested datasets.

The statistical analysis is done based on F-measure evaluation. The average rankings of the clustering algorithms are reported in Table 8. The proposed H-KHA is ranked the highest one, which is followed by agglomerative clustering, H-CS, H-ABC, H-PSO, Spectral, KHA, DBSCAN, PSO, GA, CS, HS, k-mean, H-GA, k-mean++ and ABC, among the datasets.

This section demonstrates the experimental results to empirically investigate the effectiveness of the proposed H-KHA and compare them with the successful comparative clustering algorithms. We concluded that the proposed H-KHA obtained the best performance according to F-measure evaluation compared with the other comparative algorithms and the clustering techniques. The H-KHA performed better on all seven data clustering datasets, also on all six text clustering datasets compared with the comparative algorithms consistent with the all evaluation measures. A proper balance between exploitation (intensification) and exploration (diversification) search improves the performance of the proposed hybrid KH algorithms. It balances these basic components in H-KHA owing to the combination that added the serving of the HS algorithm after the updating of the positions of the KHA in each iteration.

Table 8: Algorithm performance based on clusters quality

| | | Clustering techniques | | | | | Optimization algorithms | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Measure | K-mean | Kmean++ | Spectral | Agglomerative | DBSCAN | GA | HS | PSO | CS | ABC | KHA | H-GA | H-PSO | H-CS | H-ABC | H-KHA |
| DS1 | Accuracy | 0.234 | 0.254 | 0.245 | **0.299** | 0.244 | 0.221 | 0.244 | 0.248 | 0.235 | 0.222 | 0.215 | 0.227 | 0.280 | 0.278 | 0.279 | 0.292 |
| | Precision | 0.234 | 0.239 | 0.255 | **0.288** | 0.255 | 0.214 | 0.249 | 0.236 | 0.233 | 0.225 | 0.205 | 0.234 | 0.255 | 0.256 | 0.272 | 0.286 |
| | Recall | 0.229 | 0.233 | 0.260 | 0.280 | 0.259 | 0.225 | 0.238 | 0.231 | 0.228 | 0.229 | 0.218 | 0.233 | 0.247 | 0.261 | 0.265 | **0.284** |
| | F-measure | 0.233 | 0.235 | 0.256 | 0.284 | 0.258 | 0.221 | 0.242 | 0.233 | 0.230 | 0.222 | 0.216 | 0.231 | 0.253 | 0.251 | 0.270 | **0.285** |
| | Rank | 10 | 9 | 5 | 2 | 4 | 15 | 8 | 10 | 13 | 14 | 16 | 12 | 6 | 7 | 3 | 1 |
| DS2 | Accuracy | 0.335 | 0.335 | 0.325 | 0.345 | 0.326 | 0.332 | 0.335 | 0.321 | 0.339 | 0.322 | 0.335 | 0.334 | 0.345 | 0.339 | **0.349** | 0.347 |
| | Precision | 0.325 | 0.334 | 0.322 | 0.341 | 0.342 | 0.323 | 0.334 | 0.312 | 0.334 | 0.326 | 0.342 | 0.331 | 0.333 | 0.341 | 0.334 | **0.349** |
| | Recall | 0.329 | 0.332 | 0.326 | 0.349 | 0.335 | 0.331 | 0.345 | 0.322 | 0.329 | 0.329 | 0.338 | 0.233 | 0.336 | 0.339 | 0.332 | **0.354** |
| | F-measure | 0.328 | 0.333 | 0.324 | 0.346 | 0.339 | 0.329 | 0.339 | 0.320 | 0.331 | 0.325 | 0.340 | 0.331 | 0.331 | 0.343 | 0.333 | **0.348** |
| | Rank | 13 | 8 | 15 | 2 | 5 | 7 | 5 | 16 | 10 | 14 | 4 | 10 | 10 | 3 | 8 | 1 |
| DS3 | Accuracy | 0.450 | 0.462 | 0.415 | **0.495** | 0.465 | 0.476 | 0.466 | 0.461 | 0.465 | 0.456 | 0.473 | 0.455 | 0.469 | 0.473 | 0.475 | 0.492 |
| | Precision | 0.452 | 0.466 | 0.422 | **0.487** | 0.460 | 0.468 | 0.464 | 0.459 | 0.461 | 0.460 | 0.468 | 0.451 | 0.465 | 0.467 | 0.470 | 0.488 |
| | Recall | 0.450 | 0.462 | 0.421 | **0.495** | 0.465 | 0.458 | 0.466 | 0.456 | 0.466 | 0.450 | 0.464 | 0.439 | 0.468 | 0.479 | 0.468 | 0.485 |
| | F-measure | 0.450 | 0.463 | 0.421 | **0.493** | 0.463 | 0.466 | 0.465 | 0.467 | 0.465 | 0.456 | 0.467 | 0.449 | 0.465 | 0.467 | 0.465 | 0.487 |
| | Rank | 13 | 11 | 16 | 1 | 11 | 5 | 7 | 3 | 7 | 15 | 3 | 14 | 7 | 3 | 7 | 2 |
| DS4 | Accuracy | 0.366 | 0.367 | 0.398 | 0.355 | 0.381 | 0.365 | 0.366 | 0.361 | 0.370 | 0.367 | 0.373 | 0.356 | 0.381 | 0.381 | 0.387 | 0.396 |
| | Precision | 0.361 | 0.358 | 0.395 | 0.354 | 0.382 | 0.363 | 0.369 | 0.366 | 0.371 | 0.362 | 0.362 | 0.359 | 0.373 | 0.379 | 0.386 | 0.389 |
| | Recall | 0.366 | 0.356 | 0.380 | 0.352 | 0.379 | 0.367 | 0.351 | 0.368 | 0.365 | 0.365 | 0.366 | 0.356 | **0.381** | 0.377 | 0.380 | 0.380 |
| | F-measure | 0.365 | 0.356 | 0.386 | 0.353 | 0.381 | 0.365 | 0.355 | 0.367 | 0.366 | 0.360 | 0.363 | 0.355 | 0.374 | 0.378 | 0.383 | 0.384 |
| | Rank | 9 | 13 | 1 | 16 | 4 | 9 | 14 | 7 | 8 | 12 | 11 | 14 | 6 | 5 | 3 | 2 |
| DS5 | Accuracy | 0.370 | 0.362 | **0.482** | 0.476 | 0.461 | 0.366 | 0.367 | 0.369 | 0.365 | 0.356 | 0.373 | 0.368 | 0.372 | 0.373 | 0.374 | 0.481 |
| | Precision | 0.375 | 0.366 | 0.475 | 0.479 | 0.466 | 0.368 | 0.366 | 0.370 | 0.363 | 0.364 | 0.369 | 0.367 | 0.373 | 0.374 | 0.372 | **0.482** |
| | Recall | 0.369 | 0.361 | 0.471 | **0.485** | 0.461 | 0.368 | 0.365 | 0.369 | 0.366 | 0.353 | 0.364 | 0.366 | 0.369 | 0.379 | 0.368 | 0.479 |
| | F-measure | 0.372 | 0.363 | 0.473 | **0.483** | 0.462 | 0.366 | 0.364 | 0.370 | 0.365 | 0.356 | 0.366 | 0.365 | 0.371 | 0.368 | 0.369 | 0.480 |
| | Rank | 4 | 14 | 3 | 1 | 15 | 9 | 13 | 6 | 11 | 16 | 9 | 11 | 5 | 8 | 7 | 2 |
| DS6 | Accuracy | 0.320 | 0.331 | 0.354 | 0.389 | 0.346 | 0.343 | 0.336 | 0.341 | 0.325 | 0.346 | 0.362 | 0.355 | 0.368 | 0.371 | 0.373 | **0.394** |
| | Precision | 0.322 | 0.329 | 0.355 | **0.398** | 0.341 | 0.341 | 0.334 | 0.348 | 0.321 | 0.340 | 0.366 | 0.352 | 0.367 | 0.369 | 0.370 | 0.391 |
| | Recall | 0.326 | 0.330 | 0.359 | 0.389 | 0.342 | 0.339 | 0.335 | 0.346 | 0.326 | 0.341 | 0.364 | 0.354 | 0.362 | 0.371 | 0.368 | **0.394** |
| | F-measure | 0.323 | 0.329 | 0.357 | **0.395** | 0.341 | 0.341 | 0.334 | 0.347 | 0.325 | 0.339 | 0.364 | 0.353 | 0.363 | 0.369 | 0.370 | 0.393 |
| | Rank | 16 | 14 | 7 | 1 | 10 | 10 | 13 | 9 | 15 | 12 | 5 | 8 | 6 | 4 | 3 | 2 |
| **Mean rank** | | 10.83 | 11.50 | 7.83 | 3.83 | 8.16 | 9.16 | 10.00 | 8.50 | 10.66 | 13.83 | 8.00 | 11.50 | 6.66 | 5.00 | 5.16 | 1.66 |
| **Final rank** | | 13 | 14 | 6 | 2 | 8 | 10 | 11 | 9 | 12 | 15 | 7 | 14 | 5 | 3 | 4 | 1 |

Note: The lowest ranked algorithm is the best one.

### *6.5.4. Convergence analysis*

One of the strong criteria for evaluating metaheuristic algorithms is the convergence rate to an optimal solution. The criterion for evaluating the clustering algorithm is their convergence rate to find the optimal solution accourding to the ASDC measure. Fig. 3 shows the convergence behaviors of H-KHA and the comparative optimization algorithms (i.e., GA, HS, PSO, CS, ABC, KHA, H-GA, H-PSO, H-CS, H-ABC and H-KHA) on the text document dataset. The researchers conducted 20 independent runs for each dataset with randomly generated initializations. Thereafter, the average value is calculated based on the convergence behavior of each algorithm.

Fig. 3 clearly shows that the convergence of the original KH algorithm compared with H-KHA is faster because the KH algorithm may be stuck in the local optima and got premature convergence. However, H-KHA is more efficient than the original KH algorithm with regard to the algorithm performance and execution time and generates a much better clustering quality than do the popular algorithms. The results show that the proposed hybrid algorithm (H-KHA) outperforms the component algorithms with regard to cluster quality.

Moreover, the convergence of original KH algorithm, which is faster than the H-KHA versions and comparative algorithms, shows constant progress during the execution. H-KHA obtain the best solution among the competitive algorithms. Thus, the proposed hybrid H-KHA show competitive performance. Although the original KH algorithm converge faster at the beginning, they soon become trapped at the local optimum and sometime got premature convergence. The objective function (ASDC) values in the H-KHA follow a smooth curve from the initial values to the final optimum solution with no acute moves. Another unusual point is that the original KH algorithm obtained premature convergence regarding their ASDC.

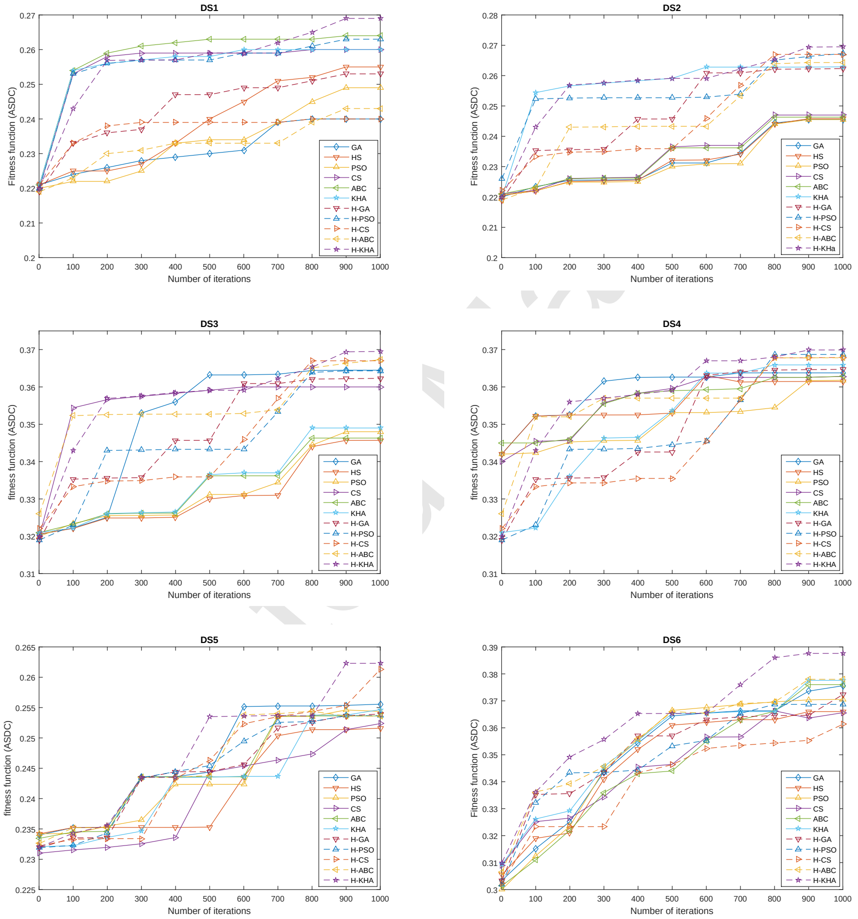


Figure 3: Convergence behaviour of text clustering optimization algorithms.

## 7. Conclusion

Metaheuristic optimization algorithms have been a prominent focus of research in solving complex optimization problems. Krill herd (KH) is a new optimization algorithm for solving many difficult global optimization problems. In this paper, a hybrid krill herd algorithm is proposed to solve the clustering problems. The original krill herd was quickly saturated and subsequently trapped in the local optimum. An enhanced krill herd was invented by introducing the global exploration operator of harmony search to relieve the premature convergence of krill herd. Using this hybridization, H-KHA quickly converges to optimal solutions under the harmony control. The main contributions of this paper are hybrid krill herd with serving harmony search algorithm based on a new probability value ($Def$) in order to control the harmony search operator to explore the exploration search effectively, as well as an analysis of the proposed method. An improved version of H-KHA aims to deal with the global search problems.

To evaluate the proposed H-KHA, five evaluation measures are adopted to text clustering technique such as ASDC, precision, recall, accuracy and F-measure, and two evaluation measures are adopted to data clustering technique such as Error rate and objective function. These measures are the most popular evaluation criteria in data and text mining domain to evaluate the newly proposed clustering method. The proposed H-KHA can produce the best-recorded results for all benchmark datasets used compared with other versions along with the several successful clustering methods and techniques from the literature. Thus, KH algorithm with serving of the harmony search is an effective method for clustering techniques and expects a huge number of upcoming successful stories in the domain of data and text clustering. The results show that the proposed hybridization is active and efficient for tackling the clustering problems. The experimental results are compared with the other comparative algorithms, which showed that the proposed hybridization of KHA (H-KHA) is suitable for solving the clustering problems as regards data and text.

H-KHA is a suitable addition for clustering domain. Other clustering problem can be used in the future to ensure the capability of this algorithm in this domain. Moreover, other powerful local search can be hybridized to further improve the exploitation capability of KH algorithm. Future research can investigate the proposed algorithm on benchmark function datasets.

## References

[1] L. M. Abualigah, A. T. Khader, M. A. AI-Betar, I.-J. AI-Huson, Multi-objectives-based text clustering technique using k-mean algorithm.

[2] P. Manikandan, S. Selvarajan, Data clustering using cuckoo search algorithm (csa), in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer, 2014, pp. 1275–1283.

[3] I. B. Saida, K. Nadjet, B. Omar, A new algorithm for data clustering based on cuckoo search optimization, in: Genetic and Evolutionary Computing, Springer, 2014, pp. 55–64.

[4] A. Hatamlou, Black hole: A new heuristic optimization approach for data clustering, Information sciences 222 (2013) 175–184.

[5] S. J. Nanda, G. Panda, A survey on nature inspired metaheuristic algorithms for partitional clustering, Swarm and Evolutionary computation 16 (2014) 1–18.

[6] R. K. Mishra, K. Saini, S. Bagri, Text document clustering on the basis of inter passage approach by using k-means, in: Computing, Communication & Automation (ICCCA), 2015 International Conference on, IEEE, 2015, pp. 110–113.

[7] M. M. Zaw, E. E. Mon, Web document clustering by using pso-based cuckoo search clustering algorithm, in: Recent Advances in Swarm Intelligence and Evolutionary Computation, Springer, 2015, pp. 263–281.

[8] S. Karol, V. Mangat, Evaluation of text document clustering approach based on particle swarm optimization, Open Computer Science 3 (2) (2013) 69–90.

[9] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, Unsupervised feature selection technique based on genetic algorithm for improving the text clustering, in: Computer Science and Information Technology (CSIT), 2016 7th International Conference on, IEEE, 2016, pp. 1–6.

[10] R. Forsati, M. Mahdavi, M. Shamsfard, M. R. Meybodi, Efficient stochastic algorithms for document clustering, Information Sciences 220 (2013) 269–291.

[11] L. M. Q. Abualigah, E. S. Hanandeh, Applying genetic algorithms to information retrieval using vector space model, International Journal of Computer Science, Engineering and Applications 5 (1) (2015) 19.

[12] K. Mizooji, A. Haghighat, R. Forsati, Data clustering using bee colony optimization, in: 7th International Multi-Conference on Computing in the Global IT, 2012, pp. 189–194.

[13] A. H. Gandomi, A. H. Alavi, Krill herd: a new bio-inspired optimization algorithm, Communications in Nonlinear Science and Numerical Simulation 17 (12) (2012) 4831–4845.

[14] R. Jensi, G. W. 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 46 (2016) 230–245.

[15] A. L. Bolaji, M. A. Al-Betar, M. A. Awadallah, A. T. Khader, L. M. Abualigah, A comprehensive review: Krill herd algorithm (kh) and its applications, Applied Soft Computing 49 (2016) 437–446.

[16] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, M. A. Awadallah, A krill herd algorithm for efficient text documents clustering, in: Computer

Applications & Industrial Electronics (ISCAIE), 2016 IEEE Symposium on, IEEE, 2016, pp. 67–72.

[17] E. Amiri, S. Mahmoudi, Efficient protocol for data clustering by fuzzy cuckoo optimization algorithm, Applied Soft Computing 41 (2016) 15–21.

[18] L. M. Abualigah, A. T. Khader, Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering, The Journal of Supercomputing (2017) 1–23.

[19] D. Karaboga, C. Ozturk, A novel clustering approach: Artificial bee colony (abc) algorithm, Applied soft computing 11 (1) (2011) 652–657.

[20] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, E. S. Hanandeh, A new hybridization strategy for krill herd algorithm and harmony search algorithm applied to improve the data clustering, management 9 11.

[21] O. Mohd Alia, M. A. Al-Betar, R. Mandava, A. T. Khader, Data clustering using harmony search algorithm, in: International Conference on Swarm, Evolutionary, and Memetic Computing, Springer, 2011, pp. 79–88.

[22] F. Yang, T. Sun, C. 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.

[23] G.-G. Wang, A. H. Gandomi, A. H. Alavi, G.-S. Hao, Hybrid krill herd algorithm with differential evolution for global numerical optimization, Neural Computing and Applications 25 (2) (2014) 297–308.

[24] G.-G. Wang, A. H. Gandomi, X.-S. Yang, A. H. Alavi, A new hybrid method based on krill herd and cuckoo search for global optimization tasks, International Journal of Bio-Inspired Computation.

[25] K. K. Bharti, P. K. Singh, Chaotic gradient artificial bee colony for text clustering, Soft Computing 20 (3) (2016) 1113–1126.

[26] G.-G. Wang, S. Deb, A. H. Gandomi, A. H. Alavi, Opposition-based krill herd algorithm with cauchy mutation and position clamping, Neurocomputing 177 (2016) 147–157.

[27] G. Wang, L. Guo, H. Wang, H. Duan, L. Liu, J. Li, Incorporating mutation scheme into krill herd algorithm for global numerical optimization, Neural Computing and Applications 24 (3-4) (2014) 853–871.

[28] G.-G. Wang, A. H. Gandomi, A. H. Alavi, S. Deb, A hybrid method based on krill herd and quantum-behaved particle swarm optimization, Neural Computing and Applications 27 (4) (2016) 989–1006.

[29] G.-G. Wang, S. Deb, S. M. Thampi, A discrete krill herd method with multilayer coding strategy for flexible job-shop scheduling problem, in: Intelligent systems technologies and applications, Springer, 2016, pp. 201–215.

[30] Z. W. Geem, J. H. Kim, G. Loganathan, A new heuristic optimization algorithm: harmony search, Simulation 76 (2) (2001) 60–68.

[31] M. Mahdavi, H. Abolhassani, Harmony k-means algorithm for document clustering, Data Mining and Knowledge Discovery 18 (3) (2009) 370–391.

[32] O. Mohd Alia, M. A. Al-Betar, R. Mandava, A. T. Khader, Data clustering using harmony search algorithm, in: International Conference on Swarm, Evolutionary, and Memetic Computing, Springer, 2011, pp. 79–88.

[33] Y.-T. Kao, E. Zahara, I.-W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 34 (3) (2008) 1754–1762.

[34] H. Nikbakht, H. Mirvaziri, A new clustering approach based on k-means and krill herd algorithm, in: Electrical Engineering (ICEE), 2015 23rd Iranian Conference on, IEEE, 2015, pp. 662–667.

[35] P. A. Kowalski, S. Łukasik, M. Charytanowicz, P. Kulczycki, Clustering based on the krill herd algorithm with selected validity measures, in: Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on, IEEE, 2016, pp. 79–87.

[36] D. Van der Merwe, A. P. Engelbrecht, Data clustering using particle swarm optimization, in: Evolutionary Computation, 2003. CEC'03. The 2003 Congress on, Vol. 1, IEEE, 2003, pp. 215–220.

[37] R. Jensi, G. W. Jiji, Mba-lf: A new data clustering method using modified bat algorithm and levy flight., ICTACT Journal on Soft Computing 6 (1).

[38] S. Zeng, X. Tong, N. Sang, Study on multi-center fuzzy c-means algorithm based on transitive closure and spectral clustering, Applied Soft Computing 16 (2014) 89–101.

[39] I. Davidson, S. Ravi, Agglomerative hierarchical clustering with constraints: Theoretical and empirical results, in: European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2005, pp. 59–70.

[40] D. Ienco, G. Bordogna, Fuzzy extensions of the dbscan clustering algorithm, Soft Computing (2016) 1–12.

[41] U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern recognition 33 (9) (2000) 1455–1465.

[42] S. Rana, S. Jasola, R. Kumar, A review on particle swarm optimization algorithms and their applications to data clustering, Artificial Intelligence Review 35 (3) (2011) 211–222.

[43] P. Berkhin, A survey of clustering data mining techniques, in: Grouping multidimensional data, Springer, 2006, pp. 25–71.

[44] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[45] Q. Li, B. Liu, Clustering using an improved krill herd algorithm, Algorithms 10 (2) (2017) 56.

[46] W. Song, Y. Qiao, S. C. Park, X. Qian, A hybrid evolutionary computation approach with its application for optimizing text document clustering, Expert Systems with Applications 42 (5) (2015) 2517–2524.

[47] X. Cui, T. E. Potok, P. Palathingal, Document clustering using particle swarm optimization, in: Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005., IEEE, 2005, pp. 185–191.

[48] K. A. Prabha, N. K. Visalakshi, Improved particle swarm optimization based k-means clustering, in: Intelligent Computing Applications (ICICA), 2014 International Conference on, IEEE, 2014, pp. 59–63.

[49] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, O. A. Alomari, Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering, Expert Systems with Applications.

[50] T. Basu, C. Murthy, A similarity assessment technique for effective grouping of documents, Information Sciences 311 (2015) 149–162.

[51] L. M. Abualigah, A. T. Khader, M. A. AlBetar, E. S. Hanandeh, Unsupervised text feature selection technique based on particle swarm optimization algorithm for improving the text clustering (2017).

[52] C. Cobos, H. Muñoz-Collazos, R. Urbano-Muñoz, M. Mendoza, E. León, E. Herrera-Viedma, Clustering of web search results based on the cuckoo search algorithm and balanced bayesian information criterion, Information Sciences 281 (2014) 248–264.

[53] Z. Fan, S. Chen, L. Zha, J. Yang, A text clustering approach of chinese news based on neural network language model, International Journal of Parallel Programming 44 (1) (2016) 198–206.

[54] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, A survey of multiobjective evolutionary clustering, ACM Computing Surveys (CSUR) 47 (4) (2015) 61.

[55] P. Jaganathan, S. Jaiganesh, An improved k-means algorithm combined with particle swarm optimization approach for efficient web document clustering, in: Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on, IEEE, 2013, pp. 772–776.

[56] M. Deepa, P. Revathy, P. Student, Validation of document clustering based on purity and entropy measures, International Journal of Advanced Research in Computer and Communication Engineering 1 (3) (2012) 147–152.

[57] R. Jensi, D. G. W. Jiji, A survey on optimization approaches to text document clustering, arXiv preprint arXiv:1401.2229.

- A novel hybrid of krill herd algorithm with harmony search algorithm
- The enhancement includes adding the operator of the harmony search algorithm to the krill herd algorithm
- A new probability value *(Def*) is proposed to control the harmony search operator to explore the search space effectively
- Investigate the proposed algorithm for the text and data clustering problems