Proceedings of the Federated Conference on Computer Science and Information Systems pp. 79–87
DOI: 10.15439/2016F295
ACSIS, Vol. 8. ISSN 2300-5963

Clustering based on the Krill Herd Algorithm with Selected Validity Measures

Piotr A. Kowalski[1,2], Szymon Łukasik[1,2], Małgorzata Charytanowicz [2,3] and Piotr Kulczycki[1,2]

[1] Faculty of Physics and Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Cracow, Poland
Email: {pkowal,slukasik,kulczycki}@agh.edu.pl

[2] Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {pakowal,slukasik,mchmat,kulczycki}@ibspan.waw.pl

[3] Institute of Mathematics and Computer Science
The John Paul II Catholic University of Lublin
Konstantynów 1 H, 20-708 Lublin, Poland
Email: mchmat@kul.lublin.pl

***Abstract*—This paper describes a new approach to metaheuristic-based data clustering by means of Krill Herd Algorithm (KHA). In this work, KHA is used to find centres of the cluster groups. Moreover, the number of clusters is set up at the beginning of the procedure, and during the subsequent iterations of the optimization algorithm, particular solutions are evaluated by selected validity criteria. The proposed clustering algorithm has been numerically verified using twelve data sets taken from the UCI Machine Learning Repository. Additionally, all cases of clustering were compared with the most popular method of k-means, through the Rand Index being applied as a validity measure.**

I. Introduction

EXPLORATORY Data Analysis is essentially centred upon tasks of clustering, classification, data reduction and outliers detection. The procedure of clustering consists of dividing a large data set into smaller subsets called 'clusters'. This partition is achieved through developing a function which assigns individual elements of the data collection into each subset. This technique has been applied to a wide range of problems, including various technical tasks [1], robotics [2] and control approaches [3], to aspects of economics [4], as well as to many agricultural issues [5].

This procedure is considered to be an unsupervised method, therefore, the division of the data is based on information directly discovered (derived) from the data itself. Hence, the separation into clusters is made in such a way that the elements within the clusters are very similar to each other, but show a difference to that held in other clusters. [6].

In data clustering procedures, a few main groupings of algorithms can be distinguished. The first of these are hierarchical [7]. In this case, the process consists of phases in which available set of clusters are merged or divided. An example of an algorithm implementing the aforementioned task, is the "bottom up" approach of Agglomerative Clustering [8]. It starts from a division in which every object is a separate cluster. In each subsequent iteration, the various groups are combined on the basis of the adopted criteria. Finally, all tested elements are placed within one cluster. A further example, albeit an opposite, is the "top down" approach of Divisive Clustering Algorithm [9]. Here, all data items start in one cluster, and this splits recursively, as one element represents one cluster.

A second algorithm group is that called 'centroid-based clustering'. This is based on minimizing variance within the clusters. Here, the best known and most commonly employed method is 'k-means procedure' [10].

The application of fuzzy-logic-based techniques [11] are a still further way of completing a clustering task. In so doing, the individual elements of a considerate data set are assigned to more than one cluster. This feature imparts a significant difference to this category of algorithms, when compared to the other procedures. The most popular algorithm of this group is 'C-fuzzy-means' [12].

Density based methods are included within another group of clustering procedures. One of the more recently introduced algorithms is that referred to as the Complete Gradient Algorithm [13]. It based on the nonparametric methodology of statistical kernel estimators as used for the recognition of data set density. This information provides the number, as well as the shape of the proposed clusters. An interesting feature of this algorithm is that it possibilities of adjustment to the authentic structure of data, and, consequently, the achieved

978-83-60810-90-3/$25.00©2016, IEEE

results are more justifiable with regard to natural point of view.

A further, but similar group of procedures of clustering tasks are those of algorithms based on grid technique. These methods are based on the assumption that data space can be partitioned into a finite number of cells - the grid structure. Subsequently, each cell density is calculated, and, after sorting the cells according to their densities, the clusters centres are determined. What is interesting herein is that this group of algorithms allows for the traversal of neighbour cells. The first algorithm in this group was introduced by Warnekar and Krishna [14]. Nowadays, the most well-known algorithms in this group are CLIQUE, MAFIA, ENCLUS, OptiGrid, O-cluster and CBF. It should be noted that such algorithms can be used for high-dimensional tasks [15].

Yet one more group of clustering algorithms is that based on an optimization algorithm inspired by Nature [16], [17]. In this approach, some metaheuristics are applied for the optimization of adopted division criteria. This action enables the coming about of great similarity of items inside the clusters, and, simultaneously, vast diversity between clusters. The mentioned criteria can be expressed as a specific mathematical formula, using a variety of statistical measures. These criteria are called 'clustering indexes', and their properties are used to assess the quality of the assignment of individual elements of the test set to the appropriate clusters.

Because the task of clustering is a NP-hard problem of combinatorial optimization [18], here – in natural manner – we apply KHA [19] as an optimization technique. This is so as to find the best location for placement of the centre point of cluster. Based on these position of centres, the individual elements of the data set are then assigned to defined groups. Completion of the thus defined clustering method is achieved using the selected three indices separately, and the obtained results are compared with the outcomes of k-means method application, taking into account the Rand Index [20] as a common evaluation criterion.

In next section, the reader will familiarize with some general information concerning optimization tasks and KHA. In Section III, the details of the application of the clustering approach, as well as selected clustering validity measures are being covered. The experimental results of our work are discussed in Section IV. Finally, in the last section of this paper, the reader will find some conclusions regarding the application of the proposed clustering algorithm, as well as intended further research and studies.

II. Optimisation based on Krill Herd Algorithm

KHA is an iterative heuristic procedure inspired by the natural phenomena of krill heard behaviour. This technique is mainly used for solving optimization problems in continuous space. Here, the solution of this problem comes about by finding such an argument x° of space under consideration $S \subseteq R^N$, which satisfies the following formula

$$f(x^\circ) = \min_{x \in S} f(x) \tag{1}$$

where $f(x)$ describes value of cost function.

The KHA originally proposed by Amir Hossein Gandomi and Amir Hossein Alavi in the paper [19], imitates the behaviour of the individual krill moving together as a herd. Such herds, move accordingly to environmental factors such as proximity to neighbours (herd density), dispersion of swarm, food position and any other biological and environmental phenomena.

In order to solve the optimization problem, we apply KHA metaheuristic. Herein, particular elements $x_i = x_i^1, \ldots, x_i^N$ of N dimensional solutions space in the form of P herd's individuals are represented. In the kth iteration, the best solution of the optimization problem as represented by the pth individual is given alternatively by these two equations:

$$x^\circ(k) = arg \max_{p=1,\ldots,P} f(x_p(k)) \quad \text{/for maximalization task/} \tag{2}$$

or

$$x^\circ(k) = arg \min_{p=1,\ldots,P} f(x_p(k)). \quad \text{/for minimalization task/} \tag{3}$$

The above best solution are corresponding with extremal value of cost function $f^\circ = f(x^\circ)$ given as (2) or (3).

The full KHA procedure in flow chart form is shown as Figure 1. This algorithm starts from an initialization of all its parameters, and positions of all P individuals are generated randomly ❶. In next step ❷, the cost function values are calculated for all initial P individuals using (2) or (3). The subsequent stage ❸ is of great importance and is characterized by KHA technique. It consists of formulas describing the movement of particular individuals. Such motion viv-a-vis each individual krill is determined by three main components. They are:

- movement induced by other krill individuals,
- foraging activity,
- random diffusion.

In subsequent time units, vector of movement of ith krill in KHA technique is based on the by Lagrangian formula:

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

where N_i is the motion induced by other krill individuals, F_i denotes the foraging motion and D_i is the physical diffusion of the krill individuals, respectively.

The first factor ❹ is a reflection of the social inspiration of the swarm's individual members. In the herd, individuals are maintained at a high density. Hence, the velocity of each individual is influenced by the movement of others. Thus, the direction of movement by the α_i parameter is induced by the presence of other herd members. This parameter is determined on the basis of the following components: local effect and target effect. The fraction of motion is formulated as:

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

Here N^{max} represents the maximum possible speed that can be induced, ω_n in the range $[0, 1]$ is the inertia weight of a particular krill and N_i^{old} is the motion induced in the previous turn. The α_i parameter is defined as:

$$\alpha_i = \alpha_i^{local} + \alpha_i^{target}, \tag{6}$$

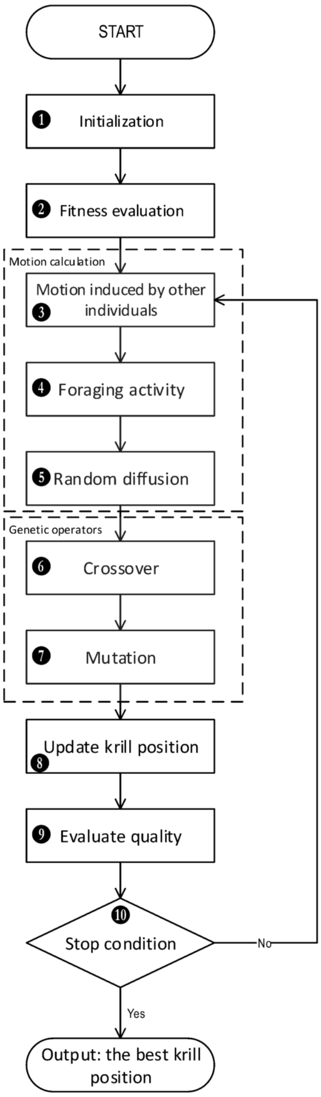


Fig. 1: Flowchart of KHA

where α_i^{local} is the local influence of the neighbours of any particular krill, whereas α_i^{target} is the target direction. The latter is determined by the position and movement of the best individual in a herd.

The α_i^{local} parameters are calculated according to the following formula:

$$\alpha_i^{local} = \sum_{j=1}^{NN} \hat{f}_{ij}\hat{X}_{ij}, \tag{7}$$

where

$$\hat{X}_{ij} = \frac{x_j - x_i}{\|x_j - x_i\| + \epsilon}, \tag{8}$$

and

$$\hat{f}_{ij} = \frac{f_i - f_j}{f^{worst} - f^{best}}. \tag{9}$$

In equation (9), f in describes the fitness value (1) of any investigated krill. Therefore f^{worst} and f^{best} represent, respectively, the worst and the best fitness of individuals in swarm. Additionally, NN provides the identification of the number of reachable krill neighbours, and ϵ is a positive number introduced to avoid singularities in the formula (8).

For determination of distance between particular krills and their neighbours, a parameter designated as being the sensing distance d_s, is introduced. This parameter may be formulated as:

$$d_{s,i} = \frac{1}{5P}\sum_{j=1}^{P} \|x_i - x_j\|. \tag{10}$$

Each individual incorporates its own target vector. This is determined as follows:

$$\alpha_i^{target} = C^{best}\hat{f}_{i,best}\hat{x}_{i,best}, \tag{11}$$

where

$$C^{best} = 2\Big(rand + \frac{k}{K^{max}}\Big). \tag{12}$$

Herein, k, K^{max} designate, respectively, the current iteration number and the maximum number of iterations. Moreover, $rand$ is a random value between 0 and 1, whereas $\hat{f}_{i,best}$ describes the best value of fitness function, while $\hat{x}_{i,best}$ provides the position of the best ith krill individual form the previous iterations.

The next main factor F_i of equation (4), is connected with the food foraging task. This F_i is defined as:

$$F_i = V_f\beta_i + \omega_f F_i^{old}, \tag{13}$$

where V_f is the food foraging speed and ω_f, denotes the inertia of the movement. In this previous equation (13), the food fitness of the ith individual is determined as follows:

$$\beta_i = \beta_i^{food} + \beta_i^{best}. \tag{14}$$

The aforementioned food aspect is determined by way of its location. Therefore, the virtual centre of food concentration is defined via KHA. This conception by the "centre of mass" approach is inspired. Hence, the food concentration in each iteration is calculated according to following formula:

$$X^{food} = \frac{\sum_{i=1}^{P}\frac{1}{f_i}x_i}{\sum_{i=1}^{P}\frac{1}{f_i}}. \tag{15}$$

Moreover, the food attraction for the ith krill individual is described via:

$$\beta_i^{food} = C^{food}\hat{f}_{i,food}\hat{X}_{i,food}. \tag{16}$$

The food coefficient in (16), expresses the global attraction of the food centre (15), and may be calculated as:

$$C^{food} = 2\Big(1 - \frac{k}{K^{max}}\Big). \tag{17}$$

The second part of equation (14) is as follows:

$$\beta_i^{best} = \hat{f}_{i,best}\hat{x}_{i,best}. \tag{18}$$

In this equation, $f_{i,best}$ is the best fit achieved by a given ith krill individual so far. This is characterised by its position $\hat{x}_{i,best}$.

The last element of the Lagrangian equation (4) is related to random physical diffusion ❺, notated as D_i. In essence, this component is of fully random character. This sub-part of movement is focused upon the diversity of population. In addition, it allows the individual krill to escape krill swarm in a situation of local optimum. Moreover, this part of equation (4) represents a trade-off between exploration and exploitation. The following formula describes this aspects of a random diffusion:

$$D_i \;=\; D^{max}\Big(1 - \frac{k}{K^{max}}\Big)\delta, \tag{19}$$

where, D^{max} is the maximum diffusion factor and δ describes the random directional vector.

Finally, the motion process can be formally summarized. This employs all the above effective parameters. The position of ith krill during the interval t to $t + \Delta t$ is, thus, determined by the following formula:

$$x_i(t + \Delta t) \;=\; x_i(t) + \Delta t \frac{dx_i}{dt}. \tag{20}$$

Here, it must be emphasized that parameter Δt is very sensitive to the speed and accuracy of optimisation task. In this respect, the Δt may be interpreted as being a scale factor of krill movement. This parameter can be obtained by way of the following equation:

$$\Delta t = C_t \sum_{j=1}^{N} (UB_j - LB_j). \tag{21}$$

In this equation, C_t is an empirically found constant number from the interval $[0, 2]$. What is more, UB_j and LB_j are, respectively, the upper and lower bounds of the jth feature $(j = 1, \ldots, N)$ of data set $X = x_1, \ldots, x_P$.

In the next stage of the KHA, the implementation of two basic evolutionary operators is applied. Firstly, in step ❻ the crossover function is considered. This operator is controlled by the crossover probability of the Cr parameter. In this approach, this operator is defined randomly. The crossover results in a change of the mth coordinate of ith krill as shown below by the formula:

$$x_{i,m} = \left\{ \begin{array}{ll} x_{r,m} & \text{for} \quad \gamma \leq Cr \\ x_{i,m} & \text{for} \quad \gamma > Cr \end{array} \right. , \tag{22}$$

where $Cr = 0.2\hat{K}_{i,best}$; $r \in \{1, 2, ..., i-1, i+1, ..., P\}$ denotes a random index, and γ is a random number drawn from the interval $[0, 1)$ generated according to the uniform distribution. In this approach the crossover operator is calculated upon a single individual.

The last part of the main loop of the KHA employs the mutation operator ❼. This modifies the mth coordinate of the ith krill, doing so via the following formula:

$$x_{i,m} = \left\{ \begin{array}{ll} x_{gbest,m} + \mu(x_{p,m} - x_{q,m}) & \text{for} \quad \gamma \leq Mu \\ x_{i,m} & \text{for} \quad \gamma > Mu \end{array} \right. , \tag{23}$$

wherein $Mu = 0.05/\hat{K}_{i,best}$; $p, q \in \{1, 2, ..., i{-}1, i{+}1, ..., P\}$ and $\mu \in [0, 1)$.

This operation completes the evolutionary procedures. Subsequently, we can now obtain individuals that are readily utilizable within the next iteration. In so-doing, in the last stage ❽ of the main loop, we should calculate the cost function for all the swarm members. Herein, the algorithm's stop condition ❿ decides whether the next iteration or the optimization algorithm is to be completed. The form of stop condition could be that of a time limit, or the reaching of a desired fitness level or a combination of these two.

More information about KHA can be found in [19]. Regarding KHA parameters, the tuning of the KHA is described in publications: [21], [22] and [23]. Notably articles [24] and [25] include other proposed modifications of the algorithm. The KHA procedure has been verified positively in discrete optimization tasks [26]. Furthermore, a parallel version of this algorithm can be found in [27]. It should also be underlined that this heuristic procedure has been applied in data base domains [28], medical tasks [29], in mechanism and machine theory [30], and also in neural learning process [31], e.t.c.

III. Clustering and Selected Clustering Indices

In this section a fusion of KHA with a variety of clustering task assessment methods is to be presented. The stated validation methods are based on characterisation indexes.

Consider a Y as being a data set matrix with dimensions D and M, respectively

$$Y = [y_1, \ldots, y_M]. \tag{24}$$

Herein, each data set element is represented by one column of this matrix. Moreover, the D feature describes each data item. The goal of the clustering task is to devise the particular division of the data set (24) into the individual C subsets, including the assignment of individual elements $y_1, \ldots, y_M$ to clusters $CL_1, \ldots, CL_C$,. In such process, as a rule, the number of clusters C is considerably smaller than the cardinality of set Y, i.e. $C \ll M$.

Individual clusters, along with their associated elements of the set Y, are characterized by points deemed the centroid of clusters $O = O_1, \ldots, O_C$. Each of these is calculated as:

$$O_c = \frac{1}{\#CL_c} \sum_{y_i \in CL_c} y_i, \tag{25}$$

where $\#CL_c$ denotes the number of elements assigned to the cth cluster. In a similar way, the center of gravity for all the investigated elements (24) is defined:

$$O_Y = \frac{1}{M} \sum_{i=1}^{M} y_i. \tag{26}$$

In this paper, the assignment of individual elements of the data set Y (24), to the clusters, is made through employing the KHA procedure. In undertaking this, krills are encoded as vectors that contain the centroid of clusters O_c. In this case, the number of clusters is established in advance, and the grouping of the individual elements of the data set is made on the basis of the rules of the nearest centroid. Thus, for each point y_i (for $i = 1, \ldots, M$) the distance to each cluster centre O_c is calculated. In so-doing, the ith element belongs to cluster CL_c if the distance $dist(y_i, O_c)$ is the smallest of the tested distances.

Furthermore, the division of elements of the set Y, is evaluated in such a way as to minimize the cost function (1). This aspect is individually determined for each of the clustering index. The formula pertaining to individual functions will be described later in this work.

A. *Rand Index*

The Rand Index is the first in the sequence that will be presented. This is considered to be a so-called supervised method for validating clustering procedures. To use this index, it is assumed that the reference distribution of membership of individual elements of the data set Y with regard to the pertinent cluster, is known to be similar as that in the case of handling the data for the classification task. This index is expressed as:

$$I_R = \frac{a+d}{a+b+c+d}, \tag{27}$$

where a is the number of elements placed in the same reference group that in the cluster grouping, b denotes the number of elements that are placed in the reference group and in the different cluster sets, c defines the number of elements placed in other reference groups and in the same cluster, and, finally, d indicates the number of elements placed in the different reference groups and in the different cluster's groups.

Building upon above definition, it can be observed that I_R can yields value between 0 and 1. Furthermore, its maximum value points out the degree of full compliance of the clustering division result, with a reference set. In the reported studies, this index is employed for comparing the division by way of applying the clustering procedure that is based on KHA, with the division arising from the structure of the reference data (i.e. the label of classes). With this index, it is possible to compare the obtained results with the reference data, as well as with other clustering indices applied in the optimization cost function.

More information about the Rand Index can be found at [20], [32].

B. *Calinski-Harabasz Index*

The following indexes are designated as being unsupervised methods for validating clustering procedures. In such, the assessment of the quality of the division stems from the properties of the dataset and the individual clusters. Consequently, such induces, in terms of measuring ability, can be utilized within the evaluation function (1) at the KHA stage.

The Celinski-Harabasz criterion has its foundation within the concept of data set variance. This index is defined as:

$$I_{CH} = \frac{V_B}{V_W}\frac{M-C}{C-1}, \tag{28}$$

where V_B and V_W denote overall between-cluster and within-cluster variance respectively. These are calculated according to the following formulas:

$$V_B = \sum_{c=1}^{C} \#CL_c \|O_c - O_Y\|^2, \tag{29}$$

and

$$V_W = \sum_{c=1}^{C} \sum_{y_i \in CL_c} \|y_i - O_c\|^2, \tag{30}$$

here, $\|\cdot\|$ is the L^2 norm (Euclidean distance) between the two vectors.

It must be underlined that high values of Celinski-Harabasz Index designate well-defined partitions. More information about this index can be found at [33].

C. *Davies-Bouldin Index*

The Davies-Bouldin Index is one of the more commonly utilized unsupervised evaluations of clustering results criteria. This function consists of a ratio of within-clustering and between-clustering distances. This index is described via:

$$I_{DB} = \frac{1}{C}\sum_{c=1}^{C} \max_{c \neq p}\{D_{c,p}\}, \tag{31}$$

where $D_{c,p}$ denotes within-to-between cluster distance for the cth and pth cluster

$$D_{c,p} = \frac{\overline{d_c} + \overline{d_p}}{d_{c,p}}. \tag{32}$$

In (32) notation $\overline{d_p}$ designates the average distance between each element of the pth cluster and centre point of this group. Moreover $d_{c,p}$ is the distance between the centres of the cth and pth clusters. In this case,the smallest value of the Davies-Bouldin Index delineates a well-defined clustering solution. More information about this measure is obtainable in [34].

D. *Silhoutte Value Index*

The Silhouette Value Index (SH) is the last clustering index to be dealt within this part of this paper. Herein, for each ith point of data set Y, the distance between all points in the same cluster and the separation distance presented by the nearest neighbours, are calculated. This criteria is defined as follow:

$$I_{SV} = \frac{1}{M}\sum_{c=1}^{C}\sum_{y_i \in CL_c} \frac{b(i,c) - a(i,c)}{max(a(i,c), b(i,c))}. \tag{33}$$

Here, $a(i,c)$ describes the mean distance of the ith point to other points in the same cluster CL_c, while $b(i,c)$ represents a minimum of average distance from the ith point in cluster pth to points in other clusters. These values are obtained through the following formulas:

$$a(i,c) = \frac{1}{\#CL_c}\sum_{y_j \in CL_c \& j \neq i} dist(y_i, y_j), \tag{34}$$

and

$$b(i,c) = \min_{CL_l \in C \setminus CL_c} \frac{1}{\#CL_l} \sum_{y_j \in CL_l} dist(y_i, y_j). \tag{35}$$

For a single ith data point, a high value of the component of this criteria denotes that this element y_i is well-matched to its group, and, simultaneously is weakly-match to other clusters. What is interesting, is that a low value of I_{SV} index reveals that the number of clusters is overestimated.

By way of formulas (33)-(35), one can observe that this criteria yields a value between -1 and $+1$. Of note: a well-defined clustering solution is represented by a value close to 1.

More information concerning the SH Clustering Index can be found in [35], [36]. With regard to clustering quality measures as a whole, more information is obtainable in [37], [38].

IV. Numerical results

This section is intended to inform the reader of several numerical verification procedures that are useful in assessing the quality of the proposed clustering methods. In order to verify the quality of the clustering algorithm, 12 sets of data obtained from the UCI Machine Learning Repository were taken into consideration [39]. With regard to these, Table I provides a characterization of all the data sets that were applied in generating a numerical verification within this paper. Evident in this table is that it includes names, abbreviations, numbers of items, dimensionality, number of classes and references to the description of the presented data sets. Herein, synthetic data collection is placed within the first four rows. These data sets are two-dimensional, and, therefore, they serve as being very good explanatory examples upon which Figure 2 is outlined.

In the presented approach, the vector of cluster centre represents the solution in state space for KHA. Thus, the product value $D{\cdot}C$ expresses the dimensionality of a particular optimization task.

In this work, a quite difficult task that the researcher must undertake is to determine a suitable set of KHA parameters. Thus, for several data sets, pilot-tests are calculated. In each test, one parameter of the KHA optimisation procedure is made variable. In addition, in these studies, the CH Index is applied as a validation parameter. As a result of this research, it is found that for almost all data set cases, the same suboptimal sets with best parameters values are calculated. Indeed, it has been discovered that it is only in the case of the Sonar and Ionosphere data sets that the achieved parameters differ. The reason for this is thought to be the higher dimension of these datasets. The following parameters of KHA were established after pilot-tests:

- $P = 20$,
- $K^{max} = 200$,
- $N^{max} = 0.01$,
- $\omega_n = 0.5$,
- $V_f = 0.02$,
- $D^{max} = 0.01$,
- $C_t = 0.5$.

Each clustering test is made of only 200 iterations of the KHA optimization procedure. For this task, three clustering indexes CH, DB and SV are employed, and these validity measures are applied in assessing the value of the cost function (1) for KHA. Because of their different properties (described in Section III), for the indices used here, the following forms of cost functions are formulated

$$f_{CH} = \frac{1}{I_{CH}} + \#CL_{\text{empty}}, \tag{36}$$

$$f_{DB} = 2I_{DB} + \#CL_{empty}, \tag{37}$$

and, finally,

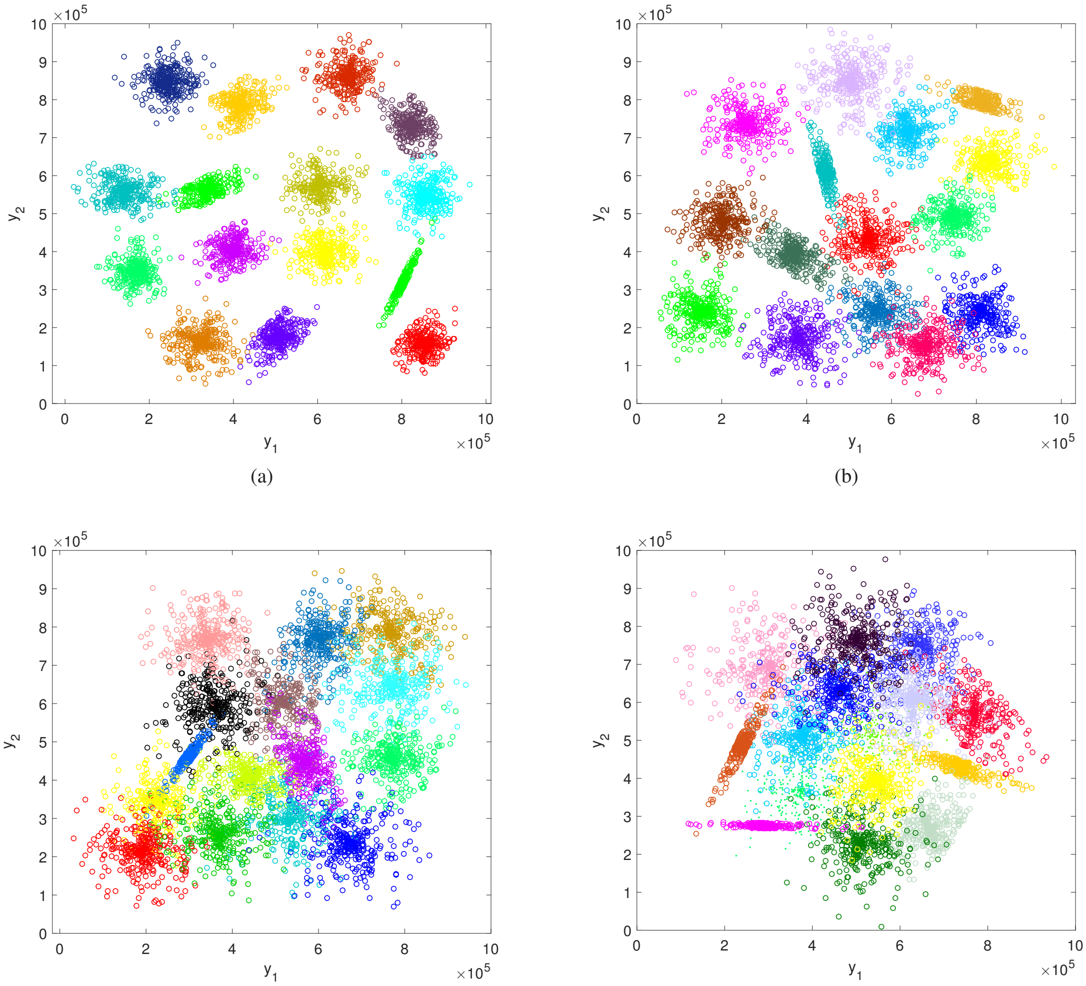
$$f_{SV} = \frac{1}{I_{SH} + 1.01} + \#CL_{empty}. \tag{38}$$

In the investigation presented here, it is assumed that, firstly, a clustering procedure based on KHA is performed by way of one selected index at a time. The result of this experiment is the clustering of the explored data set. In the next step of the test, the Rand Index calculated versus class labels is employed, as this is a commonly used evaluator of clustering performance. Thus, the obtained KHA optimization procedure solution is compared with the reference label of the class (cluster) which came from the data set. Additionally, for comparison purposes, outcomes from utilizing the k-means algorithm are also reported (with corresponding Rand Index values). Results generated by means of aforementioned steps can be seen in Table II. Throughout the testing runs, both KHA-based clustering procedures, as well as the k-means clustering algorithm were performed 30 times.

Table II consists of two parts. The first incorporates the 2nd and 3rd columns, and it contains the mean values $\overline{R}$ and the standard deviations σ_R of the Rand Index that was obtained while using the k-means clustering function. The second part of the table lists the Rand Index results (as in the first part). However, these were obtained by the way of following the KHA-clustering procedure. Here, each of three subparts provides the application results for Celinski-Harabasz ($\overline{R_{CH}}$ and $\sigma_{R_{CH}}$), Davies-Bouldin ($\overline{R_{DB}}$ and $\sigma_{R_{DB}}$) and Silhoutte Value ($\overline{R_{SV}}$ and $\sigma_{R_{SV}}$) Indexes, respectively.

While comparing all the obtained results, it can be seen that it is only in the case of the ION data set when Rand Index of clustering that was performed with k-means procedure achieves better quality then the one attained by the application of the KHA-clustering procedure. In all other cases, the results obtained via the KHA clustering method yield much better evaluation notes. These cases in Table II are emphasized with a bold font.

Based on presented results, one can observe that the Celinski-Harabasz Index clustering validation measure proved to be the best evaluation index applicable in metaheuristic procedures used in clustering. However, with regard to the other indexes, the results generated by way of the Davies-Bouldin Index are better than that obtained via the k-means



(a)

(b)

(c)

(d)

Fig. 2: Plots of 2 dimensional *s1* (a), *s2* (b), *s3* (c) and *s4* (d) datasets

TABLE I: Data sets used for experimental verification

Name of Data set	Abreviation	Number of			Bibliographical
	in paper	elements (M)	features (D)	classes (C)	reference
Synthetic 1	S1	5000	2	15	[40]
Synthetic 2	S2	5000	2	6	[40]
Synthetic 3	S3	5000	2	3	[40]
Synthetic 4	S4	5000	2	6	[40]
Ionosphere	ION	351	34	2	[41]
Iris	Iris	150	4	3	[42]
Seeds	Seeds	210	7	3	[43]
Sonar	SON	208	60	2	[44]
Thyroid	TH	7200	21	3	[45], [46]
Vehicle	VH	846	18	4	[47]
Wisconsin Breast Cancer	WBC	683	10	2	[48]
Wine	Wine	178	13	3	[49]

TABLE II: Results summary

	k–means clustering		KHA clustering					
Data set	$\overline{R}$	σ_R	$\overline{R_{CH}}$	$\sigma_{R_{CH}}$	$\overline{R_{DB}}$	$\sigma_{R_{DB}}$	$\overline{R_{SV}}$	$\sigma_{R_{SV}}$
S1	0.9748	0.0093	**0.9782**	0.0078	0.9200	0.0138	**0.9775**	0.0090
S2	0.9760	0.0072	**0.9839**	0.0053	0.9610	0.0128	0.9664	0.0096
S3	0.9522	0.0072	**0.9548**	0.0053	0.9138	0.0177	0.9436	0.0093
S4	0.9454	0.0056	**0.9484**	0.0048	0.8942	0.0229	0.9388	0.0080
Iris	0.8458	0.0614	**0.8872**	0.0145	0.7846	0.0099	0.8321	0.0563
Ionosphere	**0.5945**	0.0004	0.5573	0.0124	0.5393	0.0239	0.5682	0.0090
Seeds	0.8573	0.0572	**0.8709**	0.0156	0.6234	0.0952	0.8341	0.0586
Sonar	0.5116	0.0016	**0.5145**	0.0078	**0.5196**	0.0015	**0.5151**	0.0022
Vehicle	0.5843	0.0359	**0.6076**	0.0194	0.4854	0.0342	0.5192	0.0157
WBC	0.5448	0.0040	**0.5456**	0.0000	**0.5465**	0.0002	**0.5456**	0.0000
Wine	0.7167	0.0135	**0.7257**	0.0073	0.3979	0.0378	0.6708	0.0084
Thyroid	0.5844	0.0982	0.4535	0.0339	**0.8148**	0.1007	**0.8423**	0.0757

algorithm in only three of the applications, and that of the Silhouette Value Index, four.

Looking closely at all the results obtained by way of an application of the KHA procedure, it can be stated that for data collections S1, S2, S3, S4, Iris, Seeds, VH and Wine, the employment of the Celinski-Harabasz Index as a part of the cost function in KHA-clustering procedure gives the best results. Similarly, for the data set SON, applying the Davies-Bouldin Index, and, for the TH data set, using the Silhouette Value Index, yield the best result. Furthermore, in the situation of tests with use the WBC data collection, clusterings incorporating all three indexes provide the same result.

V. Summary

This paper is a presentation of research describing various clustering methods based on metaheuristic procedures and several validation measures. Here, in optimizing the cluster centroid locations, the biologically-inspired KHA procedure was employed. For the evaluation of particular KHA generated solutions, the paper assessed the quality of using Celinski-Harabasz, Davies-Bouldin and Silhouette Value Indexes as three clustering variants. Moreover, the Rand Index was calculated so as to evaluate the quality of the derived solutions of the analyzed clustering procedures. The proposed algorithm, in its three versions, was also confronted via the application of the well-known and commonly enrolled k-means method.

As a result of the study, it was established that the results obtained via the KHA-clustering method are much better than for that which were generated via k-means clustering procedure. What is more, the Celinski-Harabasz Index, as well as the KHA-clustering method, qualify for being considered superior for clustering tasks.

Future research will be targeted on deeper analysis of new clustering quality validation methods, as well as on applying the new procedures of swarm intelligence to the task of clustering.

References

[1] P. Kulczycki, M. Charytanowicz, P. A. Kowalski, and S. Lukasik, "The complete gradient clustering algorithm: properties in practical applications," 2012.

[2] P. A. Kowalski, S. Łukasik, M. Charytanowicz, and P. Kulczycki, "Data-driven fuzzy modeling and control with kernel density based clustering technique," *Polish Journal of Environmental Studies*, vol. 17, pp. 83–87, 2008.

[3] S. Łukasik, P. Kowalski, M. Charytanowicz, and P. Kulczycki, "Fuzzy models synthesis with kernel-density-based clustering algorithm," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, vol. 3, Oct 2008. doi: 10.1109/FSKD.2008.139 pp. 449–453.

[4] S. Breschi and F. Malerba, "The geography of innovation and economic clustering: some introductory notes," *Industrial and corporate change*, vol. 10, no. 4, pp. 817–833, 2001.

[5] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, "Complete gradient clustering algorithm for features analysis of x-ray images," in *Information Technologies in Biomedicine*, ser. Advances in Intelligent and Soft Computing, E. Piętka and J. Kawa, Eds. Springer Berlin Heidelberg, 2010, vol. 69, pp. 15–24. ISBN 978-3-642-13104-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13105-9_2

[6] L. Rokach and O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 321–352. ISBN 978-0-387-24435-8. [Online]. Available: http://dx.doi.org/10.1007/0-387-25465-X_15

[7] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[8] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *Knowledge Discovery in Databases: PKDD 2005*. Springer, 2005, pp. 59–70.

[9] S. M. Savaresi, D. L. Boley, S. Bittanti, and G. Gazzaniga, "Cluster selection in divisive clustering algorithms." in *SDM*. SIAM, 2002, pp. 299–314.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66*, 1967, pp. 281–297.

[11] M.-S. Yang, "A survey of fuzzy clustering," *Mathematical and Computer modelling*, vol. 18, no. 11, pp. 1–16, 1993.

[12] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[13] P. Kulczycki and M. Charytanowicz, *A Complete Gradient Clustering Algorithm*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 497–504. ISBN 978-3-642-23896-3. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-23896-3_61

[14] C. Warnekar and G. Krishna, "A heuristic clustering algorithm using union of overlapping pattern-cells," *Pattern Recognition*, vol. 11, no. 2, pp. 85 – 93, 1979. doi: http://dx.doi.org/10.1016/0031-3203(79)90054-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320379900542

[15] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*. CRC Press, 2013.

[16] C.-W. Tsai, W.-C. Huang, and M.-C. Chiang, "Recent development of metaheuristics for clustering," in *Mobile, Ubiquitous, and Intelligent Computing*, ser. Lecture Notes in Electrical Engineering, J. J. J. H. Park, H. Adeli, N. Park, and I. Woungang, Eds. Springer Berlin Heidelberg, 2014, vol. 274, pp. 629–636. ISBN 978-3-642-40674-4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40675-1_93

[17] T. Niknam and B. Amiri, "An efficient hybrid approach based on pso, {ACO} and k-means for cluster analysis," *Applied Soft Computing*, vol. 10, no. 1, pp. 183 – 197, 2010. doi: http://dx.doi.org/10.1016/j.asoc.2009.07.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494609000854

[18] W. J. Welch, "Algorithmic complexity: three np- hard problems in computational statistics," *Journal of Statistical Computation and Simulation*, vol. 15, no. 1, pp. 17–25, 1982. doi: 10.1080/00949658208810560. [Online]. Available: http://dx.doi.org/10.1080/00949658208810560

[19] A. H. Gandomi and A. H. Alavi, "Krill herd: A new bio-inspired optimization algorithm," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 12, pp. 4831–4845, 2012. doi: 10.1016/j.cnsns.2012.05.010. [Online]. Available: http://dx.doi.org/10.1016/j.cnsns.2012.05.010

[20] H. Parvin, H. Alizadeh, and B. Minati, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[21] P. A. Kowalski and S. Łukasik, "Experimental study of selected parameters of the krill herd algorithm," in *Intelligent Systems'2014*. Springer Science Business Media, 2015, pp. 473–485. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11313-5_42

[22] G. P. Singh and A. Singh, "Comparative study of krill herd, firefly and cuckoo search algorithms for unimodal and multimodal optimization," *IJISA*, vol. 6, no. 3, pp. 35–49, 2014. doi: 10.5815/ijisa.2014.03.04. [Online]. Available: http://dx.doi.org/10.5815/ijisa.2014.03.04

[23] P. K. Adhvaryyu, P. K. Chattopadhyay, and A. Bhattacharjya, "Application of bio-inspired krill herd algorithm to combined heat and power economic dispatch," in *2014 IEEE Innovative Smart Grid Technologies - Asia*. IEEE, 2014. doi: 10.1109/isgt-asia.2014.6873814. [Online]. Available: http://dx.doi.org/10.1109/isgt-asia.2014.6873814

[24] L. Guo, G.-G. Wang, A. H. Gandomi, A. H. Alavi, and H. Duan, "A new improved krill herd algorithm for global numerical optimization," *Neurocomputing*, vol. 138, pp. 392–402, 2014. doi: 10.1016/j.neucom.2014.01.023. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2014.01.023

[25] G.-G. Wang, A. H. Gandomi, and A. H. Alavi, "Stud krill herd algorithm," *Neurocomputing*, vol. 128, pp. 363–370, 2014. doi: 10.1016/j.neucom.2013.08.031. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2013.08.031

[26] G.-G. Wang, S. Deb, and S. M. Thampi, *Intelligent Systems Technologies and Applications: Volume 1*. Cham: Springer International Publishing, 2016, ch. A Discrete Krill Herd Method with Multilayer Coding Strategy for Flexible Job-Shop Scheduling Problem, pp. 201–215. ISBN 978-3-319-23036-8. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23036-8_18

[27] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, "Increasing the Speed of the Krill Herd Algorithm through Parallelization," in *Information Technology, Computational and Experimental Physics*. AGH University of Science and Technology Press, 2016, pp. 117–120. ISBN 978-83-7464-838-7

[28] ——, "The column-oriented database partitioning optimization based on the natural computing algorithms," in *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Łódź, Poland, September 13-16, 2015*, 2015. doi: 10.15439/2015F262 pp. 1035–1041. [Online]. Available: http://dx.doi.org/10.15439/2015F262

[29] A. Mohammadi, M. S. Abadeh, and H. Keshavarz, "Breast cancer detection using a multi-objective binary krill herd algorithm," in *Biomedical Engineering (ICBME), 2014 21th Iranian Conference on*, Nov 2014. doi: 10.1109/ICBME.2014.7043907 pp. 128–133.

[30] R. R. Bulatović, G. Miodragović, and M. S. Bošković, "Modified krill herd (mkh) algorithm and its application in dimensional synthesis of a four-bar linkage," *Mechanism and Machine Theory*, vol. 95, pp. 1 – 21, 2016. doi: http://dx.doi.org/10.1016/j.mechmachtheory.2015.08.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094114X15001895

[31] P. Kowalski and S. Łukasik, "Training neural networks with krill herd algorithm," *Neural Processing Letters*, 2015. doi: 10.1007/s11063-015-9463-0

[32] E. Achtert, S. Goldhofer, H. P. Kriegel, E. Schubert, and A. Zimek, "Evaluation of clusterings – metrics and visual support," in *2012 IEEE 28th International Conference on Data Engineering*, April 2012. doi: 10.1109/ICDE.2012.128. ISSN 1063-6382 pp. 1285–1288.

[33] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[34] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, April 1979. doi: 10.1109/TPAMI.1979.4766909

[35] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[36] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[37] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[38] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[39] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[40] P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," *Pattern Recognition*, vol. 39, no. 5, pp. 761 – 775, 2006. doi: http://dx.doi.org/10.1016/j.patcog.2005.09.012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320305003778

[41] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Technical Digest*, vol. 10, no. 3, pp. 262–266, 1989.

[42] P. A. Kowalski and P. Kulczycki, "Interval probabilistic neural network," *Neural Computing and Applications*, pp. 1–18, 2015. doi: 10.1007/s00521-015-2109-3. [Online]. Available: http://dx.doi.org/10.1007/s00521-015-2109-3

[43] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Lukasik, and S. Zak, "Complete gradient clustering algorithm for features analysis of x-ray images," in *Information Technologies in Biomedicine*, ser. Advances in Intelligent and Soft Computing, E. Pietka and J. Kawa, Eds. Springer Berlin Heidelberg, 2010, vol. 69, pp. 15–24. ISBN 978-3-642-13104-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13105-9_2

[44] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural networks*, vol. 1, no. 1, pp. 75–89, 1988.

[45] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[46] J. R. Quinlan, P. J. Compton, K. Horn, and L. Lazarus, "Inductive knowledge acquisition: a case study," in *Proceedings of the Second Australian Conference on Applications of expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1987, pp. 137–156.

[47] R. Setiono and W. Leow, "Vehicle recognition using rule based methods," *Turing Institute Research Memorandum TIRM-87-018*, vol. 121, 1987.

[48] J. Zhang, "Selecting typical instances in instance-based learning," in *Proceedings of the ninth international conference on machine learning*, 1992, pp. 470–479.

[49] S. Aeberhard, D. Coomans, and O. De Vel, "Comparison of classifiers in high dimensional settings," *Dept. Math. Statist., James Cook Univ., North Queensland, Australia, Tech. Rep*, no. 92-02, 1992.