

Optimal k-means clustering using artificial bee colony algorithm with variable food sources length

Sabreen Fawzi Raheem, Maytham Alabbas

Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq

Article Info

Article history:

Received Aug 17, 2021

Revised May 25, 2022

Accepted Jun 18, 2022

Keywords:

Artificial bee colony algorithm

K-means algorithm

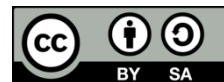
Optimize k-means clustering

Variable-length representation

ABSTRACT

Clustering is a robust machine learning task that involves dividing data points into a set of groups with similar traits. One of the widely used methods in this regard is the k-means clustering algorithm due to its simplicity and effectiveness. However, this algorithm suffers from the problem of predicting the number and coordinates of the initial clustering centers. In this paper, a method based on the first artificial bee colony algorithm with variable-length individuals is proposed to overcome the limitations of the k-means algorithm. Therefore, the proposed technique will automatically predict the clusters number (the value of k) and determine the most suitable coordinates for the initial centers of clustering instead of manually presetting them. The results were encouraging compared with the traditional k-means algorithm on three real-life clustering datasets. The proposed algorithm outperforms the traditional k-means algorithm for all tested real-life datasets.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Maytham Alabbas

Department of Computer Science, College of Computer Science and Information Technology,

University of Basrah

Basrah, Iraq

Email: ma@uobasrah.edu.iq

1. INTRODUCTION

It is important to utilize various data mining techniques, including cluster analysis, to identify, analyze, and categorize data attributes. Researching in data clustering is still active. It is used extensively in various fields, such as medical sciences, image analysis, machine learning, web cluster engines, classification, knowledge discovery, and software engineering. Clustering is splitting the area or population into groups to make data points in the same group more comparable than data points in other groups. It is one of the most often used strategies for unsupervised classification. Several exist a variety of unsupervised clustering algorithms available including k-means [1], cobweb [2], farthest-first [3], expectation-maximization (EM) [4], density-based [5], and hierarchical clustering [6]. On the whole, though, the most widely used is the k-means algorithm.

K-means clustering is a well-known partition algorithm. It was widely utilized in scientific research and industrial applications due to its simplicity, rapid convergence, and suitability for massive data sets processing, among others. The traditional k-means clustering method allocated random beginning points during clustering center initialization and typically found a local optimum clustering result. As a result, the lack of stability affected categorization accuracy, and a globally optimized method is required to overcome the limitations of this algorithm.

Numerous researches have been conducted to address this issue. For instance, Maulik and Bandyopadhyay [7] suggest using a genetic algorithm (GA) to search in the feature space for cluster centers

to optimize a similarity measure. The particle swarm optimization (PSO) algorithm was used to locate the centroids of several clusters that the user specifies [8]–[10]. Kao *et al.* [11] propose a hybrid method, namely combining k-means, Nelder-Mead simplex search, and PSO (K–NM–PSO). Niknam and Amiri [12] suggested the hybrid evolutionary algorithm, fuzzy adaptive particle swarm optimization-ant colony optimization- k-means algorithms (FAPSO–ACO–K), which has a greater chance of clustering. Laszlo and Mukherjee [13] proposed a genetic algorithm (GA) approach for seeding the k-means clustering method with centers using a unique crossover operator that swaps adjacent centers. Nguyen and Cios [14] developed a clustering method, called genetic algorithm k-means logarithmic regression expectation maximization (GAKREM), capable of performing clustering without requiring the predetermined number of groups. It is a hybrid of genetic algorithm, K-means, logarithmic regression, and expectation-maximization. Armano and Farmani [15] suggested an approach the combination of k-means and artificial bee colony algorithms (kABC), that used the artificial bee colony (ABC) algorithm to enhance k-means' capacity to identify global optimal clusters in nonlinear partitioned clustering situations. Karaboga and Ozturk [16] applied the ABC algorithm as data clustering on classification benchmark problems. Zhang *et al.* [17] presented how to divide N items into k clusters using the ABC algorithm. Zou *et al.* [18] introduced the cooperative ABC (CABC), an expanded ABC method that outperforms the original ABC in addressing complex optimization problems. It was used to solve issues with clustering. Bonab *et al.* [19] utilized the k-means algorithm with the ABC algorithm and the differential evolutionary algorithm to get the optimal solution of objects in datasets and images. Wang and Wang [20] employed a hybrid algorithm from the k-means and ABC algorithms to do the clustering procedure. The present work is a step forward in this respect. We will address the defects of the k-means algorithm using the proposed first ABC algorithm with a variable-length food source to predict the optimal number of clusters (the value of k) and the initial centroids for the k-means method.

In the remaining portion of this paper, the k-means algorithm and ABC algorithm are briefly reviewed in section 2. Section 3 describes the proposed method. Results are explained in section 4. Section 5 discusses the current findings. Section 6 presents the conclusion.

2. BACKGROUND

2.1. K-means clustering algorithm

K-means clustering algorithm is simple, an iterative, numerical, unsupervised, and non-deterministic technique [21]. It is commonly used in data mining for grouping huge sets of datasets. It is a partitioning clustering algorithm that divides the supplied datasets into k distinct clusters over an iterative operation that meets a local minimum.

K-means algorithm starts with k, a user-specified parameter, initial cluster centers randomly chosen from the dataset. In each iteration, each point at a given dataset is assigned to the closest cluster center. After categorizing all data points into clusters, the new centroid of each cluster is re-calculated as the mean of all cluster points. The procedure is continued until the centroids are still the same.

2.2. Artificial bee colony algorithm

The ABC algorithm is based on actual honey bee behavior [22]. It comprises three different groups of bees: bees that are working, watching, and scouting. There are observers in the first half of the colony, which are then employed artificial bees, and in the second half, which are the onlookers, there are employed artificial bees. Only one bee is used in the collection of each type of food. This statement can be summarized as saying that the number of employed bees equals the number of food sources. When the bees have depleted the food supply, the hired bee takes on the role of a scout.

The ABC algorithm is divided into four stages: initialization, employed bees, spectator bees, and scout bees, as presented in the following lines.

- Initialization: a random number of solutions is placed in the search space using (1):

$$X_{ij} = Lb_j + rand(0,1) * (Ub_j - Lb_j) \quad (1)$$

where $X_{i=1, \dots, SN} = \{x_{i1}, x_{i2}, \dots, x_{iD}\}$, SN is the set of all possible solutions, D is the number of optimization parameters that need to be determined, and the lower and upper limits of the solution site in dimension j are defined as Lb_j and Ub_j , respectively.

- Employed bees phase: employed bees and candidate solutions have one-to-one communication. Employed bees visit each of their solution candidates and change one dimension of the visited solution until they find the solution using (2):

$$V_{ij} = X_{ij} + \varphi_{ij} * (X_{ij} - X_{kj}) \quad (2)$$

where j is a uniformly chosen dimension at random from $\{1, 2, \dots, D\}$, φ is a numeric value generated at random $[-1, 1]$, X_k solution is chosen at random from the population with the condition ($i \neq k$). In a manner similar to how the solutions in the bee's memory replace each other, when V_{ij} is larger than the previous position, the new solution replaces the old one; otherwise, the previous solution's location is retained. Once all worker bees have completed their search, they perform a waggle dance to alert spectator bees to their food sources.

- Onlooker bees phase: employed bees assess the fitness, $fitness_i$, of a solution X_i using (3):

$$fitness_i = \begin{cases} \frac{1}{1+f(X_i)} & \text{if } f(X_i) \geq 0, \\ 1 + abs(X_i) & \text{if } f(X_i) < 0. \end{cases} \quad (3)$$

Each onlooker bee visits a solution based on the selection probability of a candidate solution X_i using (4).

$$P_i = \frac{fitness_i}{\sum_{i=1}^{SN} fitness_i} \quad (4)$$

When an onlooker bee chooses a solution, it looks for a new, better option. As a result, the employed bees step controls the ABC algorithm's diversification behavior, whereas the observer bees step controls the algorithm's intensification behavior.

- Scout bees step: in this phase, every food source that does not improve beyond a specific 'limit' of trials is abandoned and replaced with a new location, which serves as a scout for the bee employed, using (1).

3. THE PROPOSED METHOD

To overcome the previously mentioned limitations of the k-means algorithm, this section discusses the first mechanism to enhance the basic ABC algorithm to deal with variable-length representation (ABCVL). Different lengths, i.e., numbers of features, individuals may focus on various search space areas. Based on this ability, we first develop a new initialization technique to generate variable-length individuals to provide a suitable diversity level for the whole. In addition, a special search equation is designed to deal with variable-length individuals. The ABCVL algorithm was adopted to select the best number of clusters (K) and the initial centroids for the k-means algorithm, as discussed below.

3.1. Initial population

Each food source is represented by a variable length from one source to another. We follow the principle of variable-length chromosomes [23], [24] to represent individuals in the population. In this phase, the population of SN individuals is randomly initialized with a different value of optimization parameters (k). Each individual, ind , in the population is represented in the following:

$$ind_j = \{s_1, s_2, s_3, \dots, s_{k_j}\}, \quad 1 \leq j \leq SN$$

where SN is the number of food sources (solutions) in the population, the value of s is selected randomly from the dataset, k_j indicates that the individual j has k centers of clustering whose value varies from one individual to another depending on the minimum, nC_{min} , and maximum, nC_{max} , number of centers, where

$$nC_{min} \leq k \leq nC_{max}$$

3.2. Objective function

To measure the overall k-means clustering quality of food sources, we used different objective functions (the nectar) based on the most important metrics to evaluate the goodness of a clustering, like daviess-bouldin index (DBI) measure [25], Silhouette coefficient, homogeneity measure, completeness measure, v-measure, and Inertia. In addition to hybrid metrics, such as DBI-homogeneity-completeness measure, DBI-Silhouette measure, and DBI-V-measure, are applied. Below is an explanation of the concept of each of these metrics.

- The DBI measure is calculated by comparing each cluster's average similarity to the most similar. The smaller the average similarity, the more distinct the groups are and the more accurate the clustering result.

- The Silhouette coefficient, aka the Silhouette score, is used to calculate the quality of a clustering method. The result of this coefficient is in the range [-1,1], where: i) 1: means that clusters are well-defined and distinct; ii) 0: means that clusters are indifferent to one another or that the distance between them is insignificant; and iii) -1: means that clusters were misallocated.
- A homogeneity measure is a statistic that indicates the proportion of samples from a particular class that belong to a single cluster. The fewer distinct classifications that are included inside a cluster are, the better. The lower and upper limits of this measure should be 0.0 and 1.0, respectively, where higher is better.
- Completeness measure: a perfectly complete clustering is one where all data points belonging to the same class are clustered into the same cluster. Completeness and homogeneity are symmetrical.
- V-measure is equal to the harmonic mean of completeness and homogeneity.
- Inertia is a simple criterion derived from the sum of square error, which is used as the criterion clustering function. It is the sum of distance squares between all samples and their centers of clustering.
- Hybrid scale 1: In this scale, the DBI measure is combined with homogeneity and completeness measures using (5).

$$\text{Hybrid scale}_1 = 0.5 * DBI_{score} + 0.25 * \text{homogeneity}_{score} + 0.25 * \text{completeness}_{score} \quad (5)$$

- Hybrid scale 2: In this scale, the DBI measure is combined with Silhouette coefficient using (6).

$$\text{Hybrid scale}_2 = 0.6 * DBI_{score} + 0.4 * (1 - \text{Silhouette}_{score}) \quad (6)$$

- Hybrid scale 3: In this scale, the DBI measure is combined with V-measure using (7).

$$\text{Hybrid scale}_3 = 0.6 * DBI_{score} + 0.4 * (1 - V_{measure}) \quad (7)$$

3.3. Modified search equation

The basic ABC algorithm was improved to process variable-length solutions. We have modified the search equation, namely $MSEq$, by following various modification directions, as discussed below. Figure 1 shows the possible cases to produce a new individual V_i from two individuals X_i and X_k , with the lengths L_1 and L_2 , respectively.

3.3.1. MSEq₁

In the first case, $MSEq_1$, the modification position j is chosen within the range of the smaller individual length that is shared between the two individuals, and (2) is applied. The length of the new individual is the same as the length of the first individual. Figure 1(a) shows the schematic diagram of the $MSEq_1$ example, assuming that the length of the selected food sources (solutions) is 7 and 5, respectively. The position j value, therefore, must be randomly chosen within the range [1,5]. In Figure 1(a), for instance, the value of j is selected as 3.

3.3.2. MSEq₂

In the second case, $MSEq_2$, if L_1 is less than L_2 and the modification position j is chosen within the range of L_2 but outside L_1 , then the values of X_i are saved to V_i and the value of $X_{k,j}$ is saved to $V_{i,j+1}$. Therefore, the length of the new individual is L_1+1 . Figure 1(b) shows the schematic diagram of the $MSEq_2$ example, assuming that the length of the selected food sources (solutions) is 3 and 5, respectively. The position j value, therefore, is randomly chosen within the range [1,5]. If j falls outside 3, here is 4, then the value of $X_{k,4}$ will be saved to $V_{i,4}$.

3.3.3. MSEq₃

In the third case, $MSEq_3$, if L_1 is bigger than L_2 and the modification position j is chosen outside L_2 , (2) is applied on the value of $X_{i,j}$ with any a random position at $X_{k,j}$. The rest positions at X_i are neglected. Therefore, the length of the new individual is j .

Figure 1(c) shows the schematic diagram of the $MSEq_3$ example, assuming that the length of the selected food sources (solutions) is 5 and 3, respectively. The position j value, therefore, is randomly chosen within the range [1,5]. If j falls outside 3, here is 4, then (2) is applied between $X_{i,4}$ and a random position at X_k , here 2, to represent $V_{i,4}$. The length of the new individual is 4. As can be seen above, the $MSEq_1$ differs from $MSEq_2$ and $MSEq_3$ in the length of the new solution, which may vary from the original solutions at the $MSEq_2$ and $MSEq_3$ but at $MSEq_1$ still is the same as the first individual. The pseudocode of the ABCVL algorithm is given in algorithm 1.

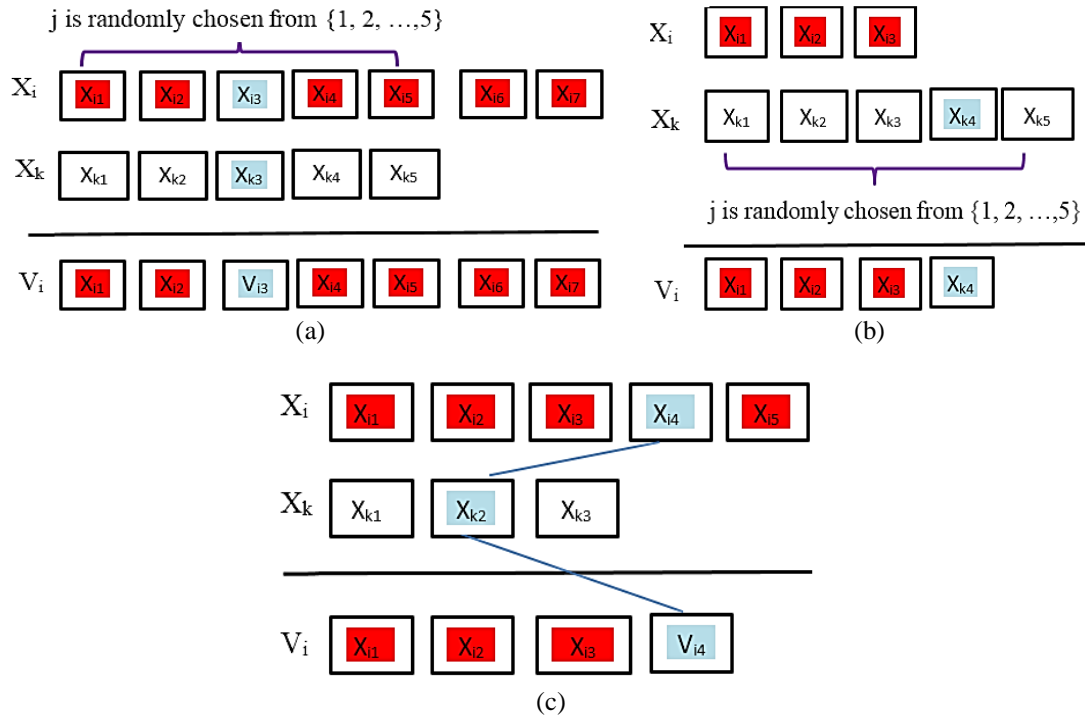


Figure 1. The schematic diagram of (a) $MSEq_1$, (b) $MSEq_2$, and (c) $MSEq_3$

Algorithm 1. The pseudocode of the ABCVL algorithm

L_j is the length of X_j

L_2 is the length of X_k

- Initialization phase using variable-length individuals (see Section 3.1);
- Repeat
- Employed bees phase using the $MSEq$ technique;
- Onlooker bees phase using the $MSEq$ technique;
- Scout bees phase;
- Memorize the best solution achieved so far;
- Until (termination conditions are met).

4. RESULTS

4.1. Tested datasets

To evaluate the efficiency of the ABCVL algorithm, three real-life datasets are used. The summary details of these datasets are: i) *Mall_Customers* dataset [26] consists of 200 samples with five features: *Customer_ID* (Unique ID for each customer), gender (Customer's gender, male or female), age (Customer's age), annual income (Customer's income per year, in \$), and spending score (Customer's spending score, from 1-100). This dataset is to be tested for two features (annual income and spending score). The optimum number of clusters is 5; ii) digits dataset [27] consists of 1797 8×8 grayscale images of handwritten digits. Each feature is an integer in the range 0...16. The optimum number of classes is 10 (code 0...9); and iii) breast cancer dataset [28] consists of 569 samples with two features (real or positive). It is divided into two classes: 212 malignant and 357 benign samples with a dimensionality of 30. The optimum number of classes is 2.

4.2. Parameters setting

In the artificial bee colony with variable-length (ABCVL) algorithm, the colony size (NP) was 40, with the number of food sources (SN) equal to NP/2. The minimum (nC_{min}) and maximum (nC_{max}) number of optimization parameters (D) were 2 and 15, respectively. The value of abandonment limit counter "limit" was selected as SN*D. The maximum number of cycles was taken as 100.

4.3. Results

We carried a test using the ABCVL algorithm on two types of experiments. The first experiment, Experiment I, was performed to evaluate the effectiveness of the objective function on the performance of the

ABCVL algorithm. On the other hand, the second experiment, Experiment II, was performed to evaluate the variable-length individuals' effectiveness.

4.3.1. Experiment I

To select a suitable metric as an objective function, we tested eight metrics (Section 3.2) to evaluate the ABCVL algorithm for clustering problems on three datasets with taking into consideration an evaluation of performance. We executed the ABCVL algorithm ten times with each metric and chose the best one among them. Table 1 shows the results of this experiment in terms of the best value, mean, optimal k, and the mean of ten values of k, where the best results are highlighted in bold.

Table 1. The results of Experiment I

Method	Datasets											
	Mall_Customers				Digits				Breast_cancer			
	Best	Mean	Optimal k	Mean (k)	Best	Mean	Optimal k	Mean (k)	Best	Mean	Optimal k	Mean (k)
DBI	0.566707	0.570686	5	5.8	1.690358	1.704200	10	10.3	0.504404	0.506289	2	2.2
Silhouette	0.446068	0.446068	5	5.0	0.809094	0.809863	12	11.9	0.302735	0.302735	2	2.0
Homogeneity	-	-	-	-	0.154497	0.1611149	14	13.8	0.334320	0.336008	9	8.8
Completeness	-	-	-	-	0.152885	0.158477	2	2.8	0.483191	0.483191	2	2.0
V_measure	-	-	-	-	0.193189	0.201859	12	12.1	0.529086	0.529086	3	3.0
Inertia	12738.31	13067.55	14	14.0	1043333.01	1052333	14	13.8	9432896.69	9601898	9	9.0
Hybrid-scale1	-	-	-	-	0.975439	0.986869	10	11.5	0.517427	0.517427	2	2.0
Hybrid-scale2	0.521398	0.521398	5	5.0	1.337068	1.355614	10	10.0	0.423736	0.423736	2	2.0
Hybrid-scale3	-	-	-	-	1.164177	1.176221	10	11.0	1.120295	1.132735	2	2.0

4.3.2. Experiment II

To test the performance of the ABCVL algorithm on clustering problems, we compared its results with the basic k-means algorithm on three clustering datasets. The results of the current experiment, in terms of the best value, mean and standard deviation produced by each algorithm throughout 30 runs, are shown in Table 2, where the best results are highlighted in bold. In addition, the ABCVL algorithm with the hybrid-scale2 function evolution curve is shown in Figure 2(a) to (c) for the three tested datasets. The coordinates of the cluster center produced are given in Tables 3-5 for three tested datasets.

Table 2. The results of Experiment II

Dataset	The basic K-means algorithm			The ABCVL algorithm		
	Best	Mean	SD	Best	Mean	SD
Mall_Customers	0.5213980	0.5895242	0.1056726	0.52139798	0.5209177	0.002586571
Digits	1.382326	1.503292	0.06446432	1.336862	1.345410	0.008708451
Breast_cancer	0.4237363	0.4237363	1.110×10 ⁻¹⁶	0.42373629	0.4237363	1.110×10 ⁻¹⁶

Table 3. Coordinates for cluster centers obtained by the ABCVL algorithm for the *Mall_Customers* dataset

Dimensionality=2	Clustering Centers for Mall_Customers dataset				
	1	2	3	4	5
Annual Income	73.0	54.0	30.0	99.80159093758257	54.436110937843466
Spending Score	88.0	48.0	73.0	81.04616983940875	99.0

Table 4. Coordinates for cluster centers obtained by the ABCVL algorithm for digits dataset

Dimensionality=64	Clustering centers for Digits dataset									
	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
1	3.01693	0	0.22672	0	0	0	0	0	0	0
2	8.35953	0	3.3012	13.1017	1	6	0	0	12	0
3	16.000	13.215	14.478	14.369	8.984	15	8	8	16	4
...
60	12.508	14.633	16.000	0.000	9.853	5	7	14	0	16
61	16.000	16.000	7.505	0	3.303	0	0	16	0	6
62	1.031	11.529	1.6	0	0	0	0	4	0	0
63	0	0	0	0	0	0	0	0	0	0

Table 5. Coordinates for cluster centers obtained by the ABCVL algorithm for *Breast_cancer* dataset

Dimensionality=30	Clustering centers for Breast_cancer dataset	
	1	2
1	14.95	17.20
2	18.77	24.52
3	97.84	114.20
...	689.5	929.4
26
27	0.2500	0.6566
28	0.08405	0.18990
29	0.2852	0.3313
0	0.09218	0.13390

5. DISCUSSION

The results in Table 2 indicate that the best metric for evaluating the clustering process is the hybrid-scale2, i.e., the combination of DBI measure with the Silhouette coefficient. This metric gives the optimal number of k for all datasets and all experiment runs. Therefore, this metric will be used in the current system to evaluate the performance of the clustering method. The results have shown that applying the current work produces better performance than setting the parameters using trial-and-error. In addition, using the current work produces better results than the basic k-means algorithm relating to the best value, mean, standard deviation, and speed of finding the best value for the same tested datasets. As shown in Figure 2, the ABCVL gives a faster convergence speed for the *Mall_Customers* dataset in Figure 2(a) than it is with the *Breast_cancer* dataset in Figure 2(c), which in turn is faster than it is with the digits dataset in Figure 2(b).

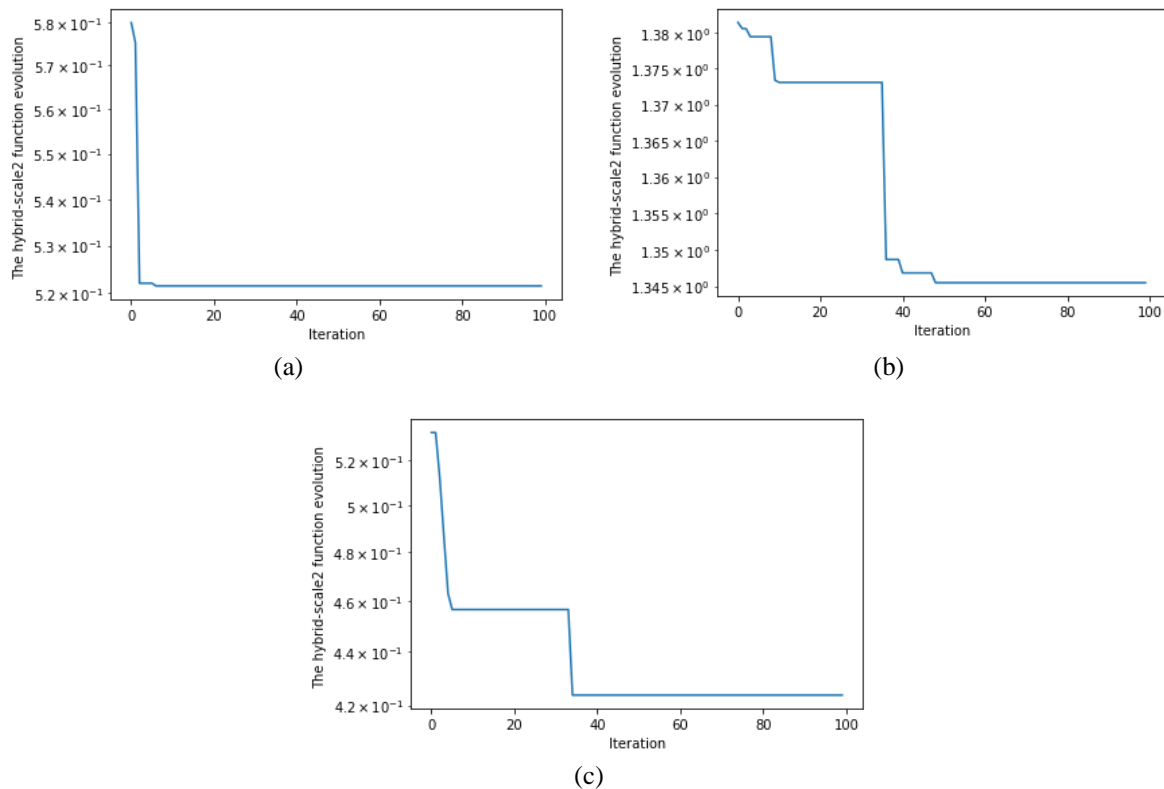


Figure 2. The curve of the hybrid scale2 evolution for three tested datasets (a) *Mall_Customers* dataset, (b) digits dataset, and (c) *Breast_cancer* dataset

6. CONCLUSION

Choosing a suitable combination of parameters for the k-means algorithm, like the number of clusters k and the initial centroids, is a challenge and becomes more complicated when trading with complex problems. Selecting reasonable parameters depends on the nature of problems and applications. One potential solution to overcome this challenge using trial and error based on an expert valuation, but it may not be

performed efficiently in areas where the level of expertise is low. Moreover, even if expertise level is relevant, specifying a fair value for each parameter can become a difficult task and time-consuming. Another solution to determine the optimal number of clusters is semi-automatic approaches such as elbow, silhouette, and gap statistics. Unfortunately, these approaches are subjective. An alternative solution is to assign each parameter automatically. To achieve this goal, we proposed the first ABC algorithm with variable-length individuals, ABCVL; in fact, we have found no such attempt so far for the ABC algorithm. In the ABCVL algorithm, we first develop a new initialization technique to generate variable-length food sources to provide a suitable diversity level for the whole. Then, a special search equation is designed to deal with variable-length individuals. Overall, the current work achieved more encouraging findings than the traditional k-means algorithm on three real-life clustering datasets. They seem to be consistent with those of other studies and suggest that further work in this direction would be very worthwhile. In future work, we will take the order of the data into account since it impacts the final results.





REFERENCES

- [1] I. S. A. AL-Forati, A. Rashid, and A. Al-Ibadi, "IR sensors array for robots localization using k-means clustering algorithm," *International Journal of Simulation: Systems, Science and Technology*, Mar. 2019, doi: 10.5013/IJSSST.a.20.S1.12.
- [2] K. Lee, H. W. Kim, C. Moon, and Y. Nam, "Analysis of vocal disorders using Cobweb clustering," in *1st International Conference on Artificial Intelligence in Information and Communication*, Feb. 2019, pp. 120–123, doi: 10.1109/ICAIIIC.2019.8669011.
- [3] L. H. Trang, N. Le Hoang, and T. K. Dang, "A farthest first traversal based sampling algorithm for k-clustering," in *Proceedings of the 2020 14th International Conference on Ubiquitous Information Management and Communication*, Jan. 2020, doi: 10.1109/IMCOM48794.2020.9001738.
- [4] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, "Relational neural expectation maximization: unsupervised discovery of objects and their interactions," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, Feb. 2018.
- [5] A. Jabbar, "Local and global outlier detection algorithms in unsupervised approach: a review," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 17, no. 1, pp. 1–12, Mar. 2021, doi: 10.37917/ijeee.17.1.9.
- [6] S. Al-Dabooni and D. Wunsch, "Model order reduction based on agglomerative hierarchical clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1881–1895, Jun. 2019, doi: 10.1109/TNNLS.2018.2873196.
- [7] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition*, vol. 33, no. 9, pp. 1455–1465, Sep. 2000, doi: 10.1016/S0031-3203(99)00137-5.
- [8] D. W. van der Merwe and A. P. Engelbrecht, "Data clustering using particle swarm optimization," in *The 2003 Congress on Evolutionary Computation, 2003*, 2003, vol. 1, pp. 215–220, doi: 10.1109/CEC.2003.1299577.
- [9] M. Omran, A. P. Engelbrecht, and A. Salman, "Particle swarm optimization method for image clustering," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 3, pp. 297–321, May 2005, doi: 10.1142/S0218001405004083.
- [10] M. Zhao, H. Tang, J. Guo, and Y. Sun, "Data clustering using particle swarm optimization," in *Lecture Notes in Electrical Engineering*, vol. 309, Springer Berlin Heidelberg, 2014, pp. 607–612, doi: 10.1007/978-3-642-55038-6_95.
- [11] Y.-T. Kao, E. Zahara, and I.-W. Kao, "A hybridized approach to data clustering," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1754–1762, Apr. 2008, doi: 10.1016/j.eswa.2007.01.028.
- [12] 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, Jan. 2010, doi: 10.1016/j.asoc.2009.07.001.
- [13] M. Laszlo and S. Mukherjee, "A genetic algorithm that exchanges neighboring centers for k-means clustering," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2359–2366, Dec. 2007, doi: 10.1016/j.patrec.2007.08.006.
- [14] C. D. Nguyen and K. J. Cios, "GAKREM: a novel hybrid clustering algorithm," *Information Sciences*, vol. 178, no. 22, pp. 4205–4227, Nov. 2008, doi: 10.1016/j.ins.2008.07.016.
- [15] G. Armano and M. R. Farmani, "Clustering analysis with combination of artificial bee colony algorithm and k-means technique," *International Journal of Computer Theory and Engineering*, vol. 6, no. 2, pp. 141–145, 2014, doi: 10.7763/IJCTE.2014.V6.852.
- [16] D. Karaboga and C. Ozturk, "A novel clustering approach: artificial bee colony (ABC) algorithm," *Applied Soft Computing*, vol. 11, no. 1, pp. 652–657, Jan. 2011, doi: 10.1016/j.asoc.2009.12.025.
- [17] C. Zhang, D. Ouyang, and J. Ning, "An artificial bee colony approach for clustering," *Expert Systems with Applications*, vol. 37, no. 7, pp. 4761–4767, Jul. 2010, doi: 10.1016/j.eswa.2009.11.003.
- [18] W. Zou, Y. Zhu, H. Chen, and X. Sui, "A clustering approach using cooperative artificial bee colony algorithm," *Discrete Dynamics in Nature and Society*, vol. 2010, pp. 1–16, 2010, doi: 10.1155/2010/459796.
- [19] M. B. Bonab, S. Z. M. Hashim, A. K. Z. Alsaedi, and U. R. Hashim, "Modified k-means combined with artificial bee colony algorithm and differential evolution for color image segmentation," in *Advances in Intelligent Systems and Computing*, vol. 331, Springer International Publishing, 2015, pp. 221–231, doi: 10.1007/978-3-319-13153-5_22.
- [20] X. M. Wang and J. B. Wang, "Improved artificial bee colony clustering algorithm based on k-means," *Applied Mechanics and Materials*, vol. 556–562, pp. 3852–3855, May 2014, doi: 10.4028/www.scientific.net/AMM.556-562.3852.
- [21] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [22] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Tech. Rep., 2005.
- [23] U. Maulik and S. Bandyopadhyay, "Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 5, pp. 1075–1081, May 2003, doi: 10.1109/TGRS.2003.810924.
- [24] Z. Qiongbing and D. Lixin, "A new crossover mechanism for genetic algorithms with variable-length chromosomes for path optimization problems," *Expert Systems with Applications*, vol. 60, pp. 183–189, Oct. 2016, doi: 10.1016/j.eswa.2016.04.005.
- [25] S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling IDS Clusters," in *11th Nordic Workshop on Secure IT-systems*, 2006, pp. 53–64.
- [26] Shwetabh123, "Mall_Customers," *Kaggle*. <https://www.kaggle.com/shwetabh123/mall-customers> (accessed Jul. 15, 2021).





- [27] "sklearn.datasets.load_digits," *Scikit learn*. Accessed: Jul. 15, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits
- [28] "sklearn.datasets.load_breast_cancer," *Scikit learn*. Accessed: Jul. 15, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

BIOGRAPHIES OF AUTHORS



Sabreen Fawzi Raheem     received the B.Sc. degree in Computer Science from College of Science, University of Basrah, Iraq, in 2016. Currently, she is a master student at the Department of Computer Science, College of Computer Science and Information Technology, University of Basrah. She has authored or coauthored more than three refereed journal and conference papers. Her research interests include artificial intelligence, machine learning, and computational linguistic. She can be contacted at email: sabreen.fawzi@uobasrah.edu.iq.



Maytham Alabbas     is currently a Professor in the Department of Computer Science at the University of Basrah where he has been a faculty member since 2003. He received his Ph.D. degree (2013) in Computer Science from the University of Manchester, UK. His Ph.D. thesis has been awarded the 2014 Best Thesis Prize of the School of Computer Science at the University of Manchester. He received his M.Sc. (2002) and B.Sc. (1999) in Computer Science from the University of Basrah, Iraq. His current research concerns are artificial intelligence, NLP, machine learning, computational linguistic, and language engineering. He has served on different conference and workshop program committees such as IntelliSys 2021-2019, AMLTA 2019, ACLing 2018, and AISI 2018. He published more than 22 journal papers and 16 conference papers. He has ACM professional membership. He can be contacted at email: ma@uobasrah.edu.iq.