Int. J. Nonlinear Anal. Appl. In Press, 1–10
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2024.33516.4995

A biogeography-based optimization algorithm for data clustering

Mohammadreza Shahriari[a,*], Arash Zaretalab[b]

[a]Faculty of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran
[b]Department of Business Management, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran

(Communicated by Seyed Hossein Siadati)

Abstract

Data clustering is a pivotal technique in data mining, essential for organizing data into meaningful groups across diverse domains such as engineering, medicine, and biology. This study introduces a Biogeography-based Optimization (BBO) algorithm to optimize data partitioning by effectively navigating the solution space towards optimal cluster configurations. The algorithm leverages migration and mutation mechanisms inspired by natural biogeography to enhance clustering accuracy. The proposed method is evaluated using various datasets of different scales and complexities, and its performance is benchmarked against conventional clustering algorithms, including K-means, Genetic Algorithm (GA), Simulated Annealing (SA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO). Comprehensive comparative analyses demonstrate that BBO not only achieves superior clustering accuracy but also exhibits robustness in handling diverse data distributions, underscoring its potential as a valuable tool in data clustering applications.

Keywords: data clustering, biogeography-based optimization, k-means, ACO, PSO, GA
2020 MSC: 91C20

1 Introduction

Clustering is an important problem that must often be solved as a part of the more complicated tasks in image processing, anomaly detection, medicine, construction management, marketing, data retrieval, reliability, portfolio optimization, selecting a supplier, and data envelopment analysis. Clustering is partitioning a set of objects into clusters, where the objects in the same cluster are more similar to each other. Hence, the clustering method is also known as hierarchy, mixture model, learning network, and objective function-based and partition-based clustering.

Shahriari M. [23] in a paper proposes a cultural algorithm for data clustering. The study introduces a novel approach to clustering data using cultural algorithms inspired by cultural evolution principles. This research contributes to the field of industrial mathematics by exploring innovative techniques for data clustering and offering potential advancements in clustering methodologies [19].

The k-means clustering algorithm is one of the most popular and classic clustering algorithms [8]. This method is simple, efficient, and fast with linear time complexity. However, the results of k-means highly depend on the initial

*Corresponding author
Email addresses: shahriari.mr@gmail.com (Mohammadreza Shahriari), arash_zaretalab@yahoo.com (Arash Zaretalab)

Received: March 2024 *Accepted:* April 2024

state in order for them to reach the local optimal solution. There are a large number of researchers who have applied different optimization techniques to eliminate this problem. For example, a Genetic Algorithm-based method to solve the clustering problem was proposed by Cowgill et al. [6] and Maulik and Bandyopadhyay [16]. A tabu search-based heuristic for clustering was developed by Sung and Jin [38]. Shelokar et al. [36] have proposed an ant colony optimization-based approach for optimal clustering N objects into K clusters.

Shahriari M. [26] presented a soft computing approach based on a modified Multiple Criteria Decision Making (MCDM) methodology using intuitionistic fuzzy sets. The study explores the application of soft computing techniques to address decision-making problems characterized by uncertainty and imprecision, offering insights into the integration of intuitionistic fuzzy sets.

In the study conducted by Sharifi et al. [35], the research investigates the impact of technical and organizational activities on the redundancy allocation problem, considering the choice of selecting redundancy strategies. The Memetic Algorithm is employed to address this problem, aiming to optimize redundancy allocation while considering both technical and organizational factors. Additionally In the study of Sharifi et al. [32] concentrates on optimizing reliability and cost in a system employing a k-out-of-n configuration by using a hybrid heuristic approach.

Shahriari M. [30] focused on redundancy allocation optimization within series-parallel systems, employing the fuzzy universal generating function approach. The study introduces an innovative methodology to optimize redundancy allocation, leveraging fuzzy logic to handle uncertainty and imprecision inherent in such systems.

An HBMO algorithm, inspired by the marriage process in the real honey-bee world, was used to solve the clustering problem by Fathian et al. [9]. Kao et al. [15] have introduced a hybrid technique that combines the PSO algorithm, Nelder–Mead simplex search, and K-means algorithm. Cao and Cios [4] have proposed a hybrid algorithm (GAKREM) based on the genetic algorithm, K-means, and logarithmic regression expectation maximization. GAKREM has three main advantages, namely, there is no need to specify the number of clusters a priori, it avoids being trapped in a local optimum, and it requires no lengthy computations. A study focusing on utilizing genetic algorithms to optimize systems featuring repairable components and multi-vacations for repairmen was conducted by Shahriari M. [28] and Shahriari [27]. The research explores the application of genetic algorithms as a tool for optimizing maintenance strategies in complex systems. This study contributes to advancing optimization techniques for systems with repairable components, aiming to enhance system reliability and efficiency.

Sharifi et al. [34] present a study on the availability optimization of a system with k-out-of-n sub-systems, taking into account various types of component failures. The proposed BBQ (Biogeography-Based Quantum-behaved) algorithm is employed for this optimization task. This research contributes to enhancing the reliability and availability of systems by effectively considering different types of component failures and utilizing a novel optimization approach [31].

Shahriari M. [29] employs a Hybrid NSGA-II algorithm to address the redundancy allocation model for series-parallel systems. The study focuses on optimizing redundancy allocation in such systems to enhance reliability and performance, utilizing a hybrid approach combining the strengths of NSGA-II (Non-dominated Sorting Genetic Algorithm II) for multi-objective optimization. Also, Shahriari M. [24] studied a bi-objective redundancy allocation model formulated to optimize the reliability and cost of series-parallel systems using the NSGA-II algorithm.

Sharifi et al. [33] presented, the NSGA-II algorithm is utilized to address a three-objective redundancy allocation problem involving k-out-of-n sub-systems. The research aims to optimize redundancy allocation while considering multiple objectives, with a focus on enhancing system reliability and performance.

Mohagheghi et al. [17] proposed a novel interval type-2 fuzzy optimization approach for evaluating R&D projects and selecting project portfolios. The study introduces a sophisticated methodology that leverages fuzzy logic to handle uncertainty and imprecision inherent in project evaluation and selection processes. The research contributes to advancing decision-making techniques in R&D project management, offering a valuable tool for organizations to optimize their project portfolios effectively [1, 2, 20, 21].

Niknam and Amiri [18] have presented a hybrid evolutionary optimization algorithm according to a fuzzy adaptive PSO, ACO, and K-means, called FAPSO-ACO-K, to solve the clustering problem. Using the advantages of the K-means algorithm and also the output of the hybrid FAPSO–ACO algorithm is considered as an initial state of K-means. Chuang et al. [5] showed the outstanding application of PSO in multi-dimensional space clustering performance. However, the rate of convergence when searching for global optima is still not sufficient [15]. For this reason, they combined Chaotic-map Particle Swarm Optimization (CPSO) with an accelerated convergence rate strategy. This technique allows the ACPSO algorithm to cluster arbitrary data better than previous algorithms. Results of the conducted experimental trials on a variety of data sets taken from several real-life situations demonstrate that ACPSO

was superior to the K-means, PSO, NM-PSO, CPSO, K-PSO and K-NM-PSO algorithms [15]. Cura [7] has proposed a particle swarm optimization approach to clustering. Apart from many of the previously proposed approaches, the PSO algorithm is applicable when the number of clusters is either known or unknown.

In order to solve this model using the Multi-Objective Particle Swarm Optimization (MOPSO) algorithm, Shafiei et al. [22] proposed a Multi-Objective Mathematical Model for the Time-Cost trade-off Problem, considering the Time Value of Money. A Non-Dominated Sorting Genetic Algorithm as a tool for addressing the multi-objective optimization of discrete time-cost tradeoff problems within project networks employed by Shahriari [25] and Hosseinzadeh Lotfi, et al. [11, 12, 13, 14].

In this study, the BBO algorithm is extended to solve clustering problems. The algorithm's performance has been tested on different-scale datasets and compared with several other proposed clustering algorithms.

The data clustering analysis is discussed in Section 2. In Section 3, the Biogeography-based Optimization algorithm is presented. Section 4 illustrates the implementation of the BBO algorithm in clustering. In Section 4, the performance of the proposed algorithm is demonstrated and compared with that of the original GA, SA, PSO, and K-means for different datasets. Finally, conclusions are presented in Section 5.

2 Data clustering analysis

The K-means algorithm [8] searches for the cluster centers, $C_1, C_2, ..., C_K$, in such a way that the sum of the squared distances (i.e., objective function) of each data point (X_i) to its nearest cluster center (C_k) is minimized, as shown in Equation (2.1), where d is a distance function. Typically, d is chosen as the Euclidean distance, which is derived from the Minkowski metric and can be defined as Equation (2.2).

$$f(X, C) = \sum_{i=1}^{N} \left(\min_{k=1,2,...,K} d(X_i - C_k) \right)^2 \tag{2.1}$$

$$d(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r} \Rightarrow d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{2.2}$$

The main procedures of the K-means algorithm are as follows.

1. To randomly select k points as initial centroids.
2. To assign each point to the nearest centroid.
3. To update the locations of each centroid by calculating the mean value of the objects assigned to it.
4. To stop, if the termination criterion is met, or to go to step 2, if otherwise.

This means that either the iterations reach the maximum number, or the location of the centroids does not change.

3 Biogeography-based optimization

The Biogeography-based Optimization (BBO) algorithm is an evolutionary optimization technique that aims to find optimal solutions for a given problem. Here are the key points about BBO:

- BBO optimizes a function by stochastically and iteratively improving candidate solutions based on a given measure of quality (fitness function). It belongs to the class of metaheuristics, making it applicable to a wide range of problems without specific assumptions about the problem structure. Unlike classic optimization methods (such as gradient descent), BBO does not require the function to be differentiable.
- BBO draws inspiration from biogeography, which studies the distribution of biological species across time and space. Mathematical models of biogeography describe processes like speciation, migration, and extinction of species on islands. Islands with high habitat suitability (HSI) can support many species, while those with low HSI can support only a few. BBO treats the objective function as a black box, relying solely on the quality measure provided by the function for candidate solutions.
- BBO maintains a population of candidate solutions. New solutions are created by combining existing ones using a simple formula. The function's gradient is not needed, making it suitable for discontinuous functions.

- BBO is typically used to optimize multidimensional real-valued functions. It can be applied to various domains without relying on specific problem characteristics. BBO was introduced by Dan Simon [37]. It leverages principles from biogeography to guide its search process.

In the genetic algorithm, each chromosome was considered as an individual member and had its own fitness. Similarly, in Biogeography-based Optimization (BBO), each biogeographical region is considered as an individual member and has its own Habitat Suitability Index (HSI). In this algorithm, similar to the genetic algorithm (where higher fitness indicated a better solution), a solution (biogeographical region) with higher HSI represents a good solution. In BBO, properties from regions (solutions) with higher HSI migrate to regions with lower HSI. In other words, by acquiring properties from regions with higher HSI, regions with lower HSI become like them and improve. This migration pattern involves two types of migration operators: output migration and input migration. Output migration is for a solution with a higher HSI that shares its properties, while input migration is for a solution with a lower HSI that accepts properties. Now, the BBO algorithm seeks solutions (biogeographical regions) that maximize HSI using these two operators. It is worth mentioning that each of these two types of migration has its own rate, known as the input rate and the output rate.

3.1 Selection strategy

This pivotal stage stands out as one of the defining features that sets Biogeography-Based Optimization (BBO) apart from its counterparts. In BBO, the selection process encompasses two distinct strategies: one tailored for migration operators (input and output), and another designated for the mutation operator. What distinguishes BBO further, as previously mentioned, is its departure from conventional algorithms like Genetic Algorithms (GA), where population members undergo complete replacement. Instead, in BBO, these members are subject to modification throughout various iterations, ensuring continuity and evolution within the population.

The primary objective of the selection strategy within BBO is twofold: first, to discern whether a specific region warrants modification; second, to determine the source from which this region should acquire its new attributes. This modification concept is dichotomized into migration and mutation strategies, each demanding careful consideration during every iteration. Decisions regarding executing these strategies for each solution are pivotal in shaping the evolutionary trajectory of the algorithm.

In subsequent sections, we delve deeper into the intricacies of these two types of strategies, offering comprehensive insights into their roles and mechanisms within the BBO framework.

The two sub-sections below will present the details of these two types of strategies.

1. **Selection Strategy for Migration Operators**
   In this section, we face 2 decisions. The first is whether a specific region wants to change or not? We compare a randomly generated number with the input rate to make this decision. The second decision is to determine which region the susceptible region to change wants to accept the property from. For this purpose, we use the roulette wheel on the output rates.
2. **Selection Strategy for Mutation**
   To determine this strategy, we compare a randomly generated number with the mutation rate. The output of this section determines whether the region in question should mutate or not.

3.2 Migration operator

Migration is an operator that is used to modify a solution using other solutions. The main idea of this operator is the same as migration in biogeography, which indicates the movement of species and biological properties among different biomes. In this process, each solution, according to its input rate, is selected to receive properties (and species) and in this regard, the solution, according to its output rate, is selected to share properties (and species).

Properties and species migrate from solutions with high HSI (good solutions) to solutions with low HSI (weak solutions). This interaction causes that with the increase in the number of species in a biome and its desirability, the output migration rate (sharing properties) in it increases.

As mentioned, in BBO, it must first be determined whether a specific solution needs to be modified in each iteration or not? And then if there is a need for modification, it should be seen which solution the relevant solution should get the property from. The explanation of these two decisions has been presented in the selection strategies section. The general structure of the BBO algorithm is as follows:

Parameter Setting (number of iterations, Pop Size, m_{max})
Best solution = [ ]
for $i = 1$ to number of Pop Size **do**
  habitat(i)=Randomly
  fitness habitat (i)=evaluate (habitat (i))
end for
for $it = 1$ to number of iterations **do**
  calculate (λ_i, μ, p, m) according to habitats rank
  for $i = 1$ to number of Pop Size **do**
    for $siv = 1$ to number of nvar **do**
      if $rand \leq \lambda_i$ **then**
        x = Roulette wheel Selection (μ)
        habitat(i,siv) = x(siv)
        fitness of habitat (i) =evaluate (habitat (i))
      end if
      if $rand \leq m_i$ **then**
        habitat (i,siv)= Randomly
        fitness of habitat (i)= evaluate (habitat (i))
      end if
    end for
  end for
  Update (Best solution)
end for

4 Results and discussion

We experimented with the BBO on five different scale datasets and compared it with other well-known algorithms. All algorithms are implemented in MATLAB software and executed on a 2 GHz laptop with 6GB of RAM. Two of the datasets are artificial, taken from Kao et al. [15], and the three of them are well-known iris, thyroid, and Beverage $N1$ datasets taken from the Machine Learning Laboratory [3]. Many authors have considered them to study and evaluate the performance of their algorithms, and can be described as follows:

Dataset 1: Artificial data set one (n = 600, d = 2, k = 4). This is a two-feature problem with four unique classes. A total of 600 patterns were drawn from four independent bivariate normal distributions, where classes were distributed according to

$$M_2 = \left(\mu = \begin{pmatrix}\omega_i \\ \omega_i\end{pmatrix}, \sum \left[\begin{bmatrix}0.5 & 0.05 \\ 0.05 & 0.5\end{bmatrix}\right]\right) \qquad i = 1, 2, ..., 4, \quad \omega_1 = -3, \quad \omega_1 = 0, \quad \omega_1 = 3, \quad \omega_1 = 6$$

μ and $\sum$ being mean vector and covariance matrix, respectively. The data set is illustrated in Figure 1.

Dataset 2: Artificial data set two (n = 250, d = 3, k = 5). This is a three-feature problem with five classes, where every feature of the classes was distributed according to Class 1-Uniform (85, 100), Class 2-Uniform (70, 85), Class 3-Uniform (55, 70), Class 4-Uniform (40, 55), Class 5-Uniform (25, 40). The data set is illustrated in Figure 1.

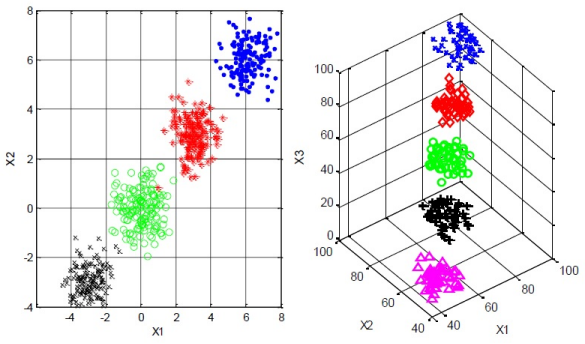

Figure 1: Two artificial data sets

Dataset 3: The Iris flower dataset, also known as Fisher's Iris dataset, is a multivariate dataset introduced by Sir Ronald Fisher in 1936 as an illustrative example of discriminant analysis. It comprises three categories, each containing 50 objects, representing different types of iris plants.

The dataset consists of 150 instances, with each instance having four attributes:

- Sepal length in centimeters
- Sepal width in centimeters
- Petal length in centimeters
- Petal width in centimeters

These attributes provide measurements of various characteristics of iris flowers, enabling researchers to explore patterns and relationships within the dataset for classification and analysis purposes.

Dataset 4: Beverage $N1$ dataset. These data are the results of a chemical analysis of Beverage N1s grown in the same region in Italy, extracted from three different cultivars. This dataset contains 178 instances with 13 continuous numeric attributes. The attributes are alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted Beverage $N1$s, and proline. All attributes are continuous. There is no missing attribute value.

Dataset 5: Contraceptive Method Choice (CMC) dataset. This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples were married women who either were not pregnant or did not know if they were at the time of the interview. The problem is to predict the current contraceptive method choice (no use of long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics. The attributes are wife's age, wife's education, husband's education, number of children ever born, wife's religion, wife's now working, husband's occupation, standard-of-living index, media exposure, and contraceptive method used.

To evaluate the performance of the BBO, we have compared it with the following clustering algorithms: K-means, GA, TS, SA, ACO, and PSO, which are taken from Niknam and Amiri [18]. The comparison of results for each dataset is based on the best solution, obtained after more than 20 different simulations, for each algorithm. The sum of the intra-cluster distances, i.e., the distances between data vectors within a cluster and the centroid of this cluster, as defined in the Equation used to measure the quality of a clustering. Clearly, the smaller the sum of the distances, the higher the clustering quality. MATLAB software is used in a 2 GHz laptop with 6GB of RAM to encode this algorithm.

Table 1: Comparison of different clustering algorithms

Data set	**Criteria**	**Algorithms**					
		K-means	**GA**	**SA**	**ACO**	**PSO**	**BBO**
Artset1	Best	516.04	518.09	518.95	517.87	515.93	**515.85**
	(Std)	295.84	189.86	195.15	2.01	180.24	0.00
	Average	721.57	638.094	684.68	519.88	627.74	515.85
Artset2	Best	1746.9	**1743.20**	**1743.20**	**1743.20**	**1743.20**	**1743.20**
	(Std)	720.66	437.05	429.02	134.06	415.02	2.01
	Average	2762.00	2667.30	2686.84	1948.97	2517.20	1745.40
Iris	Best	97.333	113.98	97.45	97.10	96.89	**96.44**
	(Std)	14.631	14.56	2.01	0.36	0.34	0.01
	Average	106.05	125.19	99.95	97.17	97.23	96.49
Beverage N1	Best	16555.68	16530.53	16473.48	16530.53	16345.96	**16289.19**
	(Std)	793.21	0.00	753.08	0.00	85.49	0.01
	Average	18061.00	16530.53	17521.09	16530.53	16417.47	16290.52
CMC	Best	5842.20	5705.63	5849.03	5701.92	5700.98	**5692.27**
	(Std)	47.16	50.36	50.86	45.63	46.95	0.00
	Average	5893.60	5756.59	5893.48	5819.13	5820.96	5693.82

Niknam and Amiri [18] provide the results of K-means, GA, SA, ACO, and PSO. The highest values are indicated in bold type. Table 1 summarizes the results of the simulations comparing the proposed Biogeography-Based Optimization (BBO) algorithm with different algorithms on Artset1, Artset2, Iris, Beverage N1, and Liver disorder datasets. For Artset1, the BBO algorithm achieved a result of 515.87, significantly outperforming other algorithms. Similarly, for Artset2, BBO converged to the global optimum of 1743.20, matching the performance of GA, SA, ACO, and PSO algorithms. It's worth noting that the standard deviation of solutions obtained by BBO was lower compared

to other algorithms. In the case of the Iris dataset, the proposed BBO algorithm reached the global optimum of 96.54, a result that other algorithms failed to achieve even after more than 20 runs. Additionally, for the Beverage N1 dataset, BBO produced the best result among all algorithms.

Finally, the BBO algorithm achieved an optimum value of 9851.72 for the Liver disorder dataset, significantly surpassing other algorithms' performance. These findings underscore the effectiveness and robustness of the BBO algorithm across diverse datasets. It consistently delivers superior results and demonstrates its potential as a powerful optimization technique. In summary, the results presented above unequivocally demonstrate the superiority of the proposed Biogeography-Based Optimization (BBO) algorithm across all datasets. Notably, the BBO algorithm consistently outperformed other algorithms by delivering high-quality solutions and exhibiting small standard deviations. This indicates the algorithm's robustness and reliability in finding optimal solutions.

Moreover, the BBO algorithm showcased its capability to converge to the global optimum in all runs across different datasets. In contrast, other algorithms may struggle with local optima, highlighting the BBO algorithm's ability to navigate complex solution spaces effectively and avoid getting trapped in suboptimal solutions. Overall, these findings affirm the effectiveness of the BBO algorithm as a powerful optimization technique capable of delivering superior performance and reliable results across diverse datasets and problem domains.

Statistical testing illustrates the significant differences between the results of the proposed BBO algorithm and those of other clustering algorithms. Specifically, we utilize the Friedman and Iman–Davenport tests to ascertain whether significant differences exist in the clustering algorithm results. If statistically significant differences are detected, we proceed with the Holland post hoc test to compare the control method against the remaining algorithms.

Garcia et al. [10] work provides further insights into the classification problem and detailed methodology. It's important to note that a significance level (α) of 0.05 is utilized as the threshold for all analyses, ensuring robust statistical inference. Table 2 depicts the average ranking of clustering algorithms computed through Friedman's test. The proposed BBO algorithm stands alone in the first rank, followed by PSO, ACO, GA, SA and K-means, successively. Table 3 presents the p-value computed by the Friedman test and the Iman–Davenport test, which confirms the existence of significant differences among the performance of all the clustering algorithms.

Table 2: Average ranking of clustering algorithms

Friedman	**K-means**	**GA**	**SA**	**ACO**	**PSO**	**BBO**
Ranking	4.8	4.5	4.6	3.5	2.2	1.4

Table 3: Results of Friedman's and Iman–Davenport's tests

Method	**Statistical value**	**p-Value**	**Hypothesis**
Friedman	14.29	0.014	Rejected
Iman-Davenport	5.3371	0.0028	Rejected

Table 4: Results of Holland's method (BBO is the control algorithm)

i	**Algorithm**	**Z**	**p-Value**	$1-(1-\alpha)^i$	**Hypothesis**
5	K-means	2.4286	0.015158	0.2262	Rejected
4	SA	2.2857	0.022272	0.1855	Rejected
3	GA	2.2143	0.026808	0.1426	Rejected
2	ACO	1.5	0.133614	0.0975	Not rejected
1	PSO	0.5714	0.567728	0.05	Not rejected

Therefore, Holland's method is carried out as a post hoc test to detect effective statistical differences between the control approach, i.e., the one with the lowest Friedman's rank, and the remaining approaches, the results of which are shown in Table 4. The results of Holm's method reveal that the control algorithm (BBO) is statistically better than K-means, GA, and SA. There is no significant difference in the ACO and PSO cases based on Holland's method results. However, the results reported in Table 1 show that the BBO algorithm achieved the best results among all the algorithms.

5 Conclusion

In conclusion, clustering remains a critical technique in data analysis with extensive applications across diverse fields such as engineering, medicine, biology, and social sciences. This study presented the Biogeography-Based Optimization (BBO) algorithm as a novel approach for clustering data vectors across five distinct scale datasets. The BBO

algorithm effectively minimizes the objective function of the clustering problem within an N-dimensional Euclidean space, especially when the number of clusters is predefined and clearly specified. The algorithm leverages migration and mutation processes inspired by natural biogeography, allowing for efficient exploration and exploitation of the solution space. The simulation results unequivocally demonstrated the efficiency, robustness, and computational effectiveness of the proposed BBO algorithm in achieving optimal clustering configurations. Moreover, the findings suggest that integrating BBO with other metaheuristic or machine learning techniques, such as Particle Swarm Optimization (PSO) or Genetic Algorithm (GA), could further enhance its performance in complex and high-dimensional datasets. Moving forward, we anticipate that this algorithm will be widely adopted in numerous fields, contributing to more accurate data segmentation, pattern recognition, and decision-making processes. Additionally, further exploration of adaptive parameter tuning and hybridization strategies could pave the way for even more advanced clustering solutions. Overall, the BBO algorithm holds substantial promise for addressing complex clustering challenges, paving the way for innovative research and practical applications in data science and beyond.

References

[1] E. Ahmady, N. Ahmady, T. Allahviranloo, and M. Shahriari, *A multi-step method to solve bipolar-fuzzy initial value problem*, Comput. Appl. Math. **43** (2024), no. 1, 74.

[2] M. Akram, I. Ullah, T. Allahviranloo, and M. Shahriari, *An integrated weighted multi-criteria decision making method using Z-number and its application in failure modes and effect analysis*, J. Ind. Inf. Integ. **30** (2025), 100805.

[3] C.L. Blake and C.J. Merz, *UCI repository of machine learning databases*, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[4] D. Cao and K.J. Cios, *GAKREM: a novel hybrid clustering algorithm*, Inf. Sci. 178 (2008), 4205–4227.

[5] L.Y. Chuang, C.J. Hsiao, and C.H. Yang, *Chaotic particle swarm optimization for data clustering*, Expert Syst. Appl. **38** (2011), 14555–14563.

[6] M.C. Cowgill, R.J. Harvey, and L.T. Watson, *A genetic algorithm approach to cluster analysis*, Comput. Math. Appl. **37** (1999), 99–108.

[7] T. Cura, *A particle swarm optimization approach to clustering*, Expert Syst. Appl. **39** (2012), 1582–1588.

[8] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

[9] M. Fathian, B. Amiri, and A. Maroosi, *Application of honey-bee mating optimization algorithm on clustering*, Appl. Math. Comput. **190** (2007), 1502–1513.

[10] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power*, Inf. Sci. **180** (2010), 2044–2064.

[11] F. Hosseinzadeh Lotfi, T. Allahviranloo, W. Pedrycz, M. Shahriari, H. Sharafi, and S. Razipour GhalehJough, *Foundations of decision*, In: Fuzzy decision analysis: multi attribute decision making approach, Stud. Comput. Intell. Springer, Cham. **1121** (2023), 1–56.

[12] F. Hosseinzadeh Lotfi, T. Allahviranloo, W. Pedrycz, M. Shahriari, H. Sharafi and S. Razipour GhalehJough, *Weight determination methods in fuzzy environment*, Fuzzy decision analysis: multi attribute decision making approach, Stud. Comput. Intell. Springer, Cham. **1121** (2023), 83–100.

[13] F. Hosseinzadeh Lotfi, T. Allahviranloo, W. Pedrycz, M. Shahriari, H. Sharafi and S. Razipour GhalehJough, *The Criteria importance through inter-criteria correlation (CRITIC) in uncertainty environment*, In: Fuzzy decision analysis: multi attribute decision making approach, Stud. Comput. Intell. Springer, Cham. **1121** (2023), 309–324.

[14] F. Hosseinzadeh Lotfi, T. Allahviranloo, W. Pedrycz, M. Shahriari, H. Sharafi, and S. Razipour GhalehJough, *The multi-objective optimization ratio analysis (MOORA) in uncertainty environment*, Fuzzy decision analysis: multi attribute decision making approach, Stud. Comput. Intell. Springer, Cham. **1121** (2023), 325–344.

[15] Y.T. Kao, E. Zahara, and I.W. Kao, *A hybridized approach to data clustering*, Expert Syst. Appl. **34** (2008), 1754–1762.

[16] U. Maulik and S. Bandyopadhyay, *Genetic algorithm-based clustering technique*, Pattern Recogn. **33** (2000), 1455–1465.

[17] V. Mohagheghi, S.M. Mousavi, B. Vahdani and M.R. Shahriari, *R&D project evaluation and project portfolio selection by a new interval type-2 fuzzy optimization approach*, Neural Comput. Appl. **28** (2017), 3869–3888.

[18] T. Niknam and B. Amiri, *An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis*, Appl. Soft Comput. **10** (2010), 183–197.

[19] R.W. Po, Y.Y. Guh and M.S. Yang, *A new clustering approach using data envelopment analysis*, Eur. J. Oper. Res. **199** (2009), 276–284.

[20] B. Rahmaniperchkolaei, Z. Taeeb, M. Shahriari, F. Hosseinzadeh Lotfi, and S. Saati, *Discrete and combinatorial optimization*, T. Allahviranloo, W. Pedrycz and A. Seyyedabbasi (Eds.), Decision-making models, Elsevier, Academic Press, 2024, pp. 177–208.

[21] B. Rahmaniperchkolaei, Z. Taeeb, M. Shahriari, F. Hosseinzadeh Lotfi, and S. Saati, *Applied optimization problems*, T. Allahviranloo, W. Pedrycz and A. Seyyedabbasi (Eds.), Decision-making models, Elsevier, Academic Press, 2024, pp. 237–299.

[22] M.A. Shafiei, M.R. Shahriari, F. Lotfi Hosseinzadeh, and R. Radfar, *Presenting a multi-objective mathematical model for time-cost trade off problem considering time value of money and solve it by mopso algorithm*, J. New Res. Math. **4** (2018), no. 14, 35–50.

[23] M.R. Shahriari, *A cultural algorithm for data clustering*, Int. J. Ind. Math. **8** (2016), no. 2.

[24] M.R. Shahriari, *Set a bi-objective redundancy allocation model to optimize the reliability and cost of the Series-parallel systems using NSGA II problem*, Int. J. Ind. Math. **8** (2016), no. 3, 171–176.

[25] M. Shahriari, *Multi-objective optimization of discrete time–cost tradeoff problem in project networks using non-dominated sorting genetic algorithm*, J. Ind. Eng. Int. **12** (2016), no. 2, 159–169.

[26] M.R. Shahriari, *Soft computing based on a modified MCDM approach under intuitionistic fuzzy sets*, Iran. J. Fuzzy Syst. **14** (2017), no. 1, 23–41.

[27] M. Shahriari, *Using genetic algorithm to optimize a system with repairable components and multi-vacations for repairmen*, Int. J. Nonlinear Anal. Appl. **13** (2022), no. 2, 3139–3144.

[28] M. Shahriari, *Set a bi-objectives model for suppliers selection with capacity constraint and reducing lead-time with meta-heuristic algorithms*, Int. J. Nonlinear Anal. Appl. **13** (2022), no. 2, 3291–3305.

[29] M. Shahriari, *Using a hybrid NSGA-II to solve the redundancy allocation model of series-parallel systems*, Int. J. Ind. Math. **14** (2022), no. 4, 503–513.

[30] M. Shahriari, *Redundancy allocation optimization based on the fuzzy universal generating function approach in the series-parallel systems*, Int. J. Ind. Math. **15** (2023), no. 1, 69–77.

[31] M. Shahriari, F. Hosseinzadeh Lotfi, B. Rahmaniperchkolaei, Z. Taeeb, and S. Saati, *Data optimization and analysis*, T. Allahviranloo, W. Pedrycz and A. Seyyedabbasi (Eds.), Decision-making models, Elsevier, Academic Press, 2024, pp. 209–236.

[32] M. Sharifi, G. Cheragh, K. Dashti Maljaii, A. Zaretalab, and M. Shahriari, *Reliability and cost optimization of a system with k-out-of-n configuration and choice of decreasing the components failure rates*, Scientia Iranica, **28** (2021), no. 6, 3602–3616.

[33] M. Sharifi, P. Pourkarim Guilani, and M. Shahriari, *Using NSGA II algorithm for a three objectives redundancy allocation problem with k-out-of-n sub-systems*, J. Optimization Industr. Eng. **9** (2016), no. 19, 87–96.

[34] M. Sharifi, M.R. Shahriari, and S. Khoshniat, *Availability optimization of a system with k-out-of-n sub-systems considering different types of components failure using BBQ algorithm*, Int. J. Ind. Math. **11** (2019), no. 4, 239–248.

[35] M. Sharifi, M.R. Shahriari and A. Zaretalab, *The effects of technical and organizational activities on redundancy allocation problem with choice of selecting redundancy strategies using the memetic algorithm*, Int. J. Ind. Math. 11 (2019), no. 3, 165–176.

[36] P.S. Shelokar, V.K. Jayaraman, and B.D. Kulkarni, *An ant colony approach for clustering*, Anal. Chem. Acta **509** (2004), 187–195.

[37] D. Simon, *Biogeography-based optimization*, IEEE Trans. Evolut. Comput. **12** (2008), no. 6, 702–713.

[38] C.S. Sung and H.W. Jin, *A tabu-search-based heuristic for clustering*, Pattern Recogn. **33** (2000), 849–858.