

A Hybrid Data Clustering Approach using K-Means and Simplex Method-based Bacterial Colony Optimization

S. Suresh Babu¹ and K. Jayasudha²

^{1,2} Department of Computer Science, Thiruvalluvar Government Arts College, Rasipuram, Tamilnadu, India

Abstract - Clustering is a common data mining and data analysis tool. K-means is a popular clustering approach in which the data is partitioned into K clusters. The k-means method, on the other hand, is highly dependent on the initial state and eventually converges to a local optimum solution. A new hybrid algorithm is proposed in this paper using a k-means algorithm combined with simplex method-based bacterial colony optimization (SMBCO+KM) for finding more efficient groups. The main aim of the hybrid approach is to enhance the clustering quality by utilizing the benefits of both algorithms. The suggested approach outperforms other algorithms according to simulation findings.

Key Words: Bacterial colony optimization, simplex method, k-means, convergence rate, data clustering

1.INTRODUCTION

Data clustering is the task of gathering information into clusters (classes) so that the data in each group has a high similarity while being considerably different from data in other clusters [1]. The three-decade-old k-means technique is among the most widely used partitioning clustering algorithms in a range of areas. Over continuous data, the k-means algorithm is defined only when the initial partitions were near to the final solution. To put it another way, the outcomes of k-means are greatly dependent on the initial state and the time it takes to obtain a locally optimal solution. Many studies in clustering have been conducted to try to overcome this problem.

For example, Y.-T. Kao et al. (2008) have presented a new hybrid method that is combined with simplex search, K-means, and particle swarm optimization [2]. X. Geng et al. (2019) have developed a hybrid method based on k-means and Agglomerative nesting (AGNES) for topic detection [3]. M. A. El-Shorbagy et al. (2021) have developed an algorithm that combines the discovery proficiency of LS and the exploitation proficiency of GOA and incorporates the advantages of both LS and GOA [4]. P. Padmavathi et al. (2018) [5] presented a fuzzy clustering method based on social spider optimization (FSSO) and a hybrid method for fuzzy clustering (2021) [6]. K. Vijayakumari et al (2021) presented a hybrid method based on FBCO and fuzzy c-means algorithm for fuzzy clustering [7].

The BCO is the most widely used well-known algorithm that is applied to several fields of real-time applications. The most significant benefit of the BCO is the ability to share information with others via a communication process. Individual communication and group exchanges are different kinds of communication mechanisms that have started in the BCO algorithm. The communication process is utilized to improve the efficacy of the solutions that have been provided. When tackling a data clustering problem, however, standard BCO has a slow convergence rate and a long calculation time [8]. Because the BCO clustering technique uses some internal iteration to achieve a significant clustering result, it takes longer to compute.

However, traditional SI methods necessitate a high level of development aptitude, which might lead to premature convergence and increased calculation time[9]. Individual algorithms have various advantages and disadvantages on their own [10]. Hence, combining any two methods provides an alternative method of overcoming the limitations of individual algorithms [11]. The various SI techniques are integrated with k-means to improve clustering quality and overcome the shortcomings of a single algorithm. Recently, Revathi et. al. (2021) developed a hybrid method based on BCO and k-means to enhance the performance [12]. However, the conventional BCO has many drawbacks including a slow convergence rate, failure to achieve local optimum values.

In this research, propose a hybrid technique for handling clustering problems that combine an SMBCO and k-means. SMBCO has used simplex method for enhancing the performance of BCO. The main aim of the proposed SMBCO+KM is to enhance the searching ability of both local and global solutions, enhance the convergence rate and avoid the local optima problem.

The suggested hybrid algorithm's goals are to improve the accuracy of the clustering problem while eliminating the flaws of both algorithms. The BCO is utilized to search the complete space for the global optimum in the suggested hybrid algorithm. When the BCO method obtains a solution close to the optimum solution, the process is switched to the k-means algorithm to generate more precise similar groupings. The paper's contribution is as follows:

- The SMBCO+KM algorithm is proposed to solve the data clustering problem
- The k-means is integrated with BCO and BCO is enhanced by the simplex method to produce more similar data partitions
- The proposed hybrid SMBCO+KM utilizes the benefits of both algorithms to overwrite their shortcomings of its.
- The strength of the SMBCO+KM is evaluated on six famous UCI datasets.
- To analyse the strength of SMBCO+KM, objective function and computation time is used
- The performance of the SMBCO+KM method compared with some benchmark algorithms.

2. DATA CLUSTERING

The data samples (patterns) $(X = \{x_1, x_2, \dots, x_n\})$ are found out the group of the N data patterns into the K groups $(C = C_1, C_2, \dots, C_K)$. The data clustering problem must meet the following requirements.

$$\begin{cases} \bigcup_{i=1}^K C_i = X; \\ C_i \cap C_j = 0, \quad i, j = 1, 2, \dots, K; i \neq j; \\ C_i = 0, \quad i = 1, 2, \dots, K; \end{cases} \quad (1)$$

Hence, the clustering problem could be determined as follows;

$$\begin{cases} C_i = \{x_j \mid \|x_j - z_j\| \leq \|x_j - x_p\|, x_j \in X\}, \\ \quad p \neq i, p = 1, 2, \dots, K, \\ z_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \quad i = 1, 2, \dots, K, \end{cases} \quad (2)$$

Where, $\| \cdot \|$ - represents the distance value which is the calculated difference between two given data samples. z_j - represent the center of the cluster. Hence, the major goal of clustering approaches is to reduce the sum of squared errors (SSE) [13] which can be defined as follows,

$$SSE = \sum_{k=1}^K \sum_{x_j \in C_i} \|x_j - z_i\|^2 \quad (3)$$

3. K-MEANS ALGORITHM

The unsupervised k-means methodology is a distinguished method for handling the clustering problem. It's a partitioned clustering procedure that's basic,

Algorithm 1: K-means algorithm

Step 1: Choose k cluster centroid values at random.

Step 2: Calculate distance (Equation (3))

Step 3: Values for the cluster centroid is updated (Equation (5))

Step 4: Execute the termination procedure. If yes, proceed to step 2; then, the process is terminated.

straightforward, and low-cost to compute. [14]. This algorithm starts with cluster centre values that are chosen at random. Based on similarity, each data point is assigned to the group's closest point. Distance values are used to calculate the similarity of data points. The Euclidean distance is an extensively used measure of similarity. The following is the definition of the distance function:

$$D(x_p, z_j) = \sqrt{\sum_{i=1}^d (x_{pi} - z_{ji})^2} \quad (4)$$

Here, x_p represents the p^{th} data sample, z_j is

representing the j^{th} cluster center value, and d is the number of data features. The cluster center is updated by using the mean value of the related data samples belonging to appreciate class.

$$z_j = \frac{1}{n_j} \sum_{x_p \in C_j} x_p \quad (5)$$

Here, n_j is denotes the data objects associated with j . C_j

is denotes a subset group from the cluster C . When one of the following conditions is met, the k-means clustering method expires: when the maximum iteration is reached or when no cluster membership changes. The k-means clustering method is depicted in algorithm 1.

4. SIMPLEX METHOD

Spendley et al. discovered the simplex approach (1962). It is defined by a set of points that is one greater than the dimensions of the search space. The simplex approach provides a number of advantages, including a rapid search speed, a small calculation area, and a high ability to search locally [15, 16]. The SM method's detailed procedure is outlined below,

Step 1: Calculate all of the solutions in the population (bacteria). Elect the best global X_g and the second best

X_b , assuming X_s is the spider that has to be changed and $f(X_g)$, $f(X_b)$ and $f(X_s)$ are associated fitness value.

Step 2: Using the formula below, get the middle point X_c of X_g and X_b :

$$X_c = \frac{X_g + X_b}{2} \quad (6)$$

Step 2: Using the formula below, find the reflection point X_r . Typically, the reflection coefficient α is set to 1.

$$X_r = X_c + \alpha(X_c - X_s) \quad (7)$$

Step 4: Compare the fitness values of $f(X_r)$ and $f(X_g)$. If $f(X_r) < f(X_g)$, the following equation was used to execute the extension operation:

$$X_e = X_c + \gamma(X_r - X_c) \quad (8)$$

Where γ - denotes the extension coefficient, which is usually set at 2. After that, compare the fitness of the extension point X_e with the global best X_g . If $(X_e < X_g)$ then X_s is replaced by X_e . Else X_r will be used instead of X_s .

Step 5: Match the fitness values of X_s and X_r . If $f(X_r)$ is greater than $f(X_s)$, the compression operation is conducted using the following formula:

$$X_t = X_c + \beta(X_s - X_c) \quad (9)$$

Where, the condense coefficient β s commonly set at 0.5. then, compared the condense point X_t and the point X_s . If $f(X_t) < f(X_s)$, then X_s should be exchanged for X_t . Else, X_r will be used instead of X_s .

Step 6: the shrink operations are accomplished to identify the condense point X_w $f(X_g) < f(X_r) < f(X_s)$. This is defined as follows,

$$X_w = X_c - \beta(X_s - X_c) \quad (10)$$

Here, β is the shrink coefficient. If $f(X_w) < f(X_s)$, X_s must be swapped by X_w , else X_r will be used instead of X_s .

5. BACTERIAL COLONY OPTIMIZATION

BCO is the newest optimization algorithm developed by Niu et al. (2012) [17]. In comparison to other bacteria algorithms like BFO [18] and BC [19], the BCO algorithm looks for nutrients by communicating information between individuals, a process known as communication. Chemotaxis, communication, elimination, reproduction, and migration are the five primary phases of the BCO. Algorithm 2 shows the BCO algorithm with simplex method.

The chemotaxis process is carried out in two different ways: running and tumbling. The goal of the running process is to improve the efficiency of convergence. The purpose of the tumbling procedure is to avoid problems

with local optima. The most important task in BCO communication is the communication phase. Two types of processes in the communication process are used: dynamic neighbor-oriented (randomly oriented study) and group-oriented. The communication procedure is adopted to improve searching capabilities while also lowering computational costs and preventing premature convergence. The tumbling process can be determined as follows,

$$\begin{aligned} Position_i(T) &= Position_i(T-1) + C(i) \\ & * [f_i \cdot (G_{best} - Position_i(T-1)) + (1 - f_i)] \quad (11) \\ & * (P_{best_i} - Position_i(T-1)) + turb_i] \end{aligned}$$

The running process performed as follows,

$$\begin{aligned} Position_i(T) &= Position_i(T-1) + C(i) \\ & * [f_i \cdot (G_{best} - Position_i(T-1)) + (1 - f_i)] \quad (12) \\ & * (P_{best_i} - Position_i(T-1)) \end{aligned}$$

$$C(i) = C_{min} + \left(\frac{Iter_{max} - Iter_j}{Iter_{max}} \right) (C_{max} - C_{min}) \quad (13)$$

Where, $turb_i$ - turbulent direction variance, $C(i)$ - chemotaxis step size, $f_i \in (0, 1)$, G_{best} - global best value and P_{best} - personal best or local best, $Iter_{max}$ - maximum iteration and $Iter_j$ - current iteration.

6. PROPOSED SMBCO+KM

The proposed SMBCO+KM is combined with three algorithms k-means, BCO, and simplex method. The k-

Algorithm 2: BCO for clustering

- Step 1: Each bacterial colony
- Step 2: Chemotaxis and communication process
- Step 3: Calculate each colony's cluster centre
- Step 4: Calculate the distance between the data samples and the cluster center
- Step 5: Reproduction and elimination process
- Step 6: Migration process
- Step 7: Bacterial colony updating using **simplex method**
- Step 8: If the terminating state is not met, then go to step 2.1.

means is a well-known fast clustering algorithm. But, it

produced low accuracy and fell into local optima. On the other hand, BCO is a well-known swarm intelligence global optimization algorithm and also it produced high accuracy. But, its convergence rate is low and fails to find a local optimum. The simplex method is a well-known local searching method that is used to enhance the performance of various optimization algorithms. Hence, the proposed method uses the merits of the above three algorithms for obtaining a more efficient solution and to dismiss the shortcomings. In the proposed SMBCO+KM, First, the searching ability of BCO is enhanced by the simplex method. Then, the results of the SMBCO are used as the initial condition of the k-means algorithm. The step by step algorithm for hybrid SMBCO+KM is mentioned in

Algorithm 3: proposed hybrid SMBCO+KM

Begin Hybrid method

Step 1: Initialize population required parameters

Step 2: Begin BCO

Step 2.1: Each bacterial colony

Step 2.2: Chemotaxis and communication process

Step 2.3: Calculate each colony's cluster centre

Step 2.4: Calculate the distance between the data samples and the cluster center

Step 2.5: Reproduction and elimination process

Step 2.6: Migration process

Step 2.7: Bacterial colony updating using **simplex method**

Step 2.8: If the terminating state is not met, then go to step 2.

Step 2: End BCO

Step 3: Begin K-means

Step 3.1: Assign the best solution of SM-BCO as the initial cluster center

Step 3.2: Calculate distance (Equation (3))

Step 3.3: Values for the cluster centroid is updated (Equation (5))

Step 3.4: Execute the termination procedure. If yes, proceed to step 2; then, the process is terminated.

Step 3: End K-means

Step 4: Perform the termination condition if yes then go to step 2, otherwise terminate the process and go to step 5.

Step 5: store the final cluster center as the best solution

Step 6: End

algorithm 3.

7. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results are conducted by using MATLAB. The strength of the proposed SMBCO+KM is evaluated on six different prominent UCI datasets. The proposed SMBCO+KM technique's performance is evaluated using the objective function. The strength of the proposed SMBCO+KM matched with some benchmark algorithms such as k-means [20], PSO [21], BFO [22], BCO[23], SMBCO

7.1 Datasets collections

The developed and compared methods are applied to six different datasets to obtain experimental results. The datasets were retrieved from the UCI machine learning database, the details of which are shown in Table 1 and discussed as follows,

- From 214 data samples, the Glass is divided into six groups, each with nine attributes.
- The 303 data samples in the Heart are divided into two unique classes based on six attributes
- Fisher's iris is a collection of 150 samples divided into three groups with four attributes each.
- There are 871 data samples in Vowel, which are divided into six types based on three features.
- There are 178 data samples in the wine, which are divided into three classes with thirteen attributes each.
- Wisconsin breast cancer (WBC) has 683 samples that are divided into two classes based on nine features.

7.2 Parameter settings

The best parameter settings can produce more efficient outcomes for the given solutions. The following parameter settings are considered in this present research paper. High computation time is taken when choosing the high value of the chemotaxis step. Hence, this present research work selects as $N_c = 100$. The selecting the swim step is $N_s = 4$, and the reproduction value is

Table 1: Description of datasets

Datasets	Instances	Features	Clusters
Glass	214	9	6
Heart	303	76	2
Iris	150	4	3
Vowel	871	3	6
WBC	683	9	2
Wine	178	13	3

$N_{re} = 4$. The lowest step length value C_{min} is 0.01. The

higher step length C_{max} is 0.2. The probability value in the elimination and dispersal step is 0.25.

7.3 Performance indicators

The performances of the developed algorithms are investigated with the help of a performance analyzer such as an objective function. The objective function's goal is to close the distance between two data samples, such as data object values and the center of the cluster. The value of the objective function is divided into three categories: maximum, worst, and medium. As a result, the lowest value is seen as the best.

7.4 Discussions

Table 2 shows the performance comparisons of the developed data clustering algorithms. To analyze the performance of developed clustering algorithms, in this research work, considering six different benchmark datasets such as glass, heart, Iris, WBC, wine, and vowel. The qualities of clustering algorithms are analyzed using the objective function, standard deviation, and computational time.

Table 2 shows the performance of developed algorithms based on the objective function. The low value of the objective function is considered as best performance. For example, the low objective value is produced by the proposed SMBCO+KM algorithm for iris datasets such as 85.80, and low computation time is taken to convergence

Table 2: Comparative analysis results of the objective values

Datasets	Techniques	Objective values				Time (s)
		Best	Mean	Worst	SD	
Glass	K-Means	244.85	253.53	260.55	4.32	0.0396
	PSO	227.31	234.35	240.63	3.96	15.084
	BFO	219.31	225.31	231.48	3.65	12.169
	BCO	208.34	214.64	219.78	3.14	10.823
	SMBCO	197.04	204.13	207.17	2.81	9.858
	SMBCO+KM	187.98	188.71	190.10	1.17	8.926
Heart	K-Means	4530.98	5089.41	5406.96	230.50	0.0802
	PSO	4485.06	4757.32	5049.54	172.89	19.437
	BFO	4321.44	4607.83	4772.54	118.68	18.028
	BCO	4213.12	4393.84	4497.06	72.24	16.792
	SMBCO	4130.36	4245.05	4347.92	61.27	14.819
	SMBCO+KM	4097.41	4158.93	4197.50	26.59	12.951
Iris	K-Means	145.58	197.83	227.86	21.55	0.0628
	PSO	91.37	100.11	109.77	5.61	5.6598
	BFO	88.45	96.61	104.73	4.63	5.1481
	BCO	80.83	88.76	95.66	4.09	4.5180
	SMBCO	76.78	84.30	90.71	3.75	4.1381
	SMBCO+KM	73.77	80.93	85.80	3.29	3.8712
WBC	K-Means	2516.24	2926.13	3279.84	212.01	0.0926
	PSO	2331.89	2679.86	2984.67	191.81	16.184
	BFO	2232.43	2620.73	2884.57	184.39	15.118
	BCO	2139.69	2342.74	2499.02	93.69	13.902
	SMBCO	2055.70	2119.54	2193.51	42.66	11.362
	SMBCO+KM	1970.26	2036.99	2095.03	36.15	10.752
Wine	K-Means	18791.69	19824.04	20695.52	529.96	0.08726
	PSO	17989.09	18520.85	18999.75	313.10	14.3623
	BFO	16848.34	17610.17	17890.36	254.06	11.7212
	BCO	16536.52	16957.24	17181.56	176.14	7.2810
	SMBCO	16383.72	16561.36	16891.29	154.77	6.6482
	SMBCO+KM	16268.65	16411.67	16573.67	92.22	5.9261
Vowel	K-Means	134227.58	148569.91	161084.74	7610.08	0.0820
	PSO	129623.26	140510.24	147086.88	4582.54	17.871
	BFO	126995.69	133261.81	139473.93	3450.50	16.108
	BCO	121231.89	125020.95	129747.96	2718.64	13.581
	SMBCO	115778.54	119875.35	121849.42	1555.52	12.984
	SMBCO+KM	114714.65	116077.57	117810.54	924.57	11.821

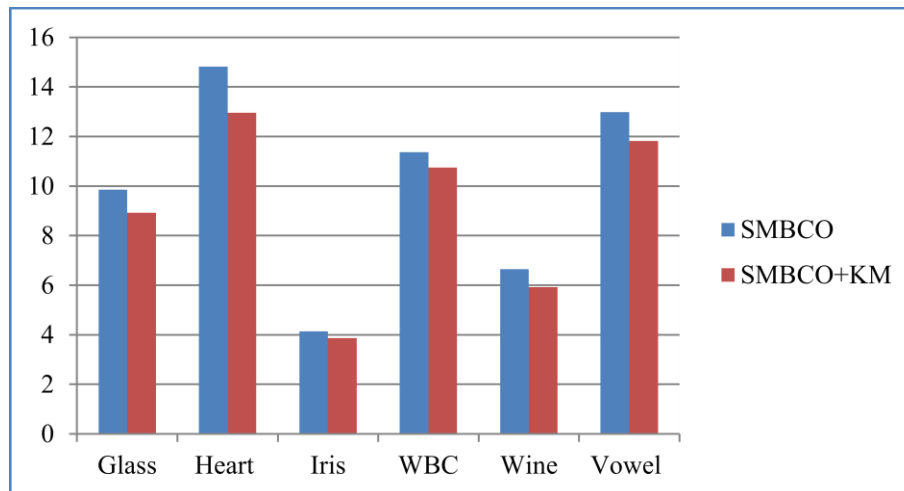


Figure 1 : Performance comparisons based on computational time

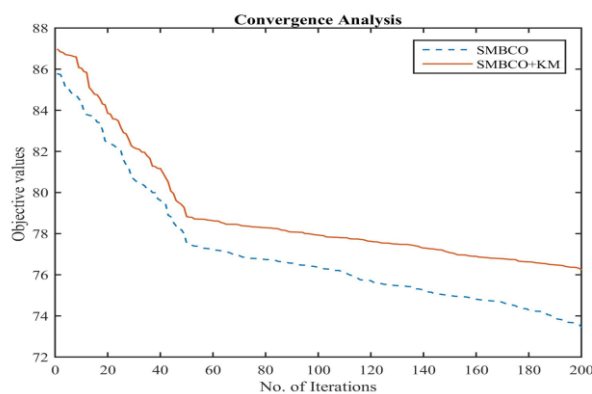


Figure 2 : convergence rate of SMBCO and SMBCO+KM

such as 3.8712 when compared with other algorithms. According to Table 2, the proposed SMBCO+KM achieves higher performances compared with other compared algorithms. Figure 1 shows the performance of the computational algorithm for SMBCO and proposed SMBCO+KM. Figure 2 shows the convergence rate of the SMBCO and SMBCO+KM algorithms.

8. CONCLUSIONS

The well-known clustering methods k-means and BCO have their benefits and drawbacks. For the data clustering problem, the present article developed a new hybrid method based on k-means and simplex BCO. The performance analysis is carried out on six datasets using five well-known techniques. When using the proposed hybrid clustering technique to solve clustering problems, the proposed method obtains the best solution by determining the best clustering center vector for each bacterial individual. When compared to existing methods, the experimental findings showed that the proposed SMBCO+KM method offered the best solution.

REFERENCES

- [1] S. S. Babu and K. Jayasudha, "A survey of nature-inspired algorithm for partitional data clustering," in *Journal of Physics: Conference Series*, 2020, vol. 1706, no. 1: IOP Publishing, p. 012163.
- [2] 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, 2008.
- [3] X. Geng, Y. Zhang, Y. Jiao, and Y. Mei, "A novel hybrid clustering algorithm for topic detection on chinese microblogging," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 289-300, 2019.
- [4] M. A. El-Shorbagy and A. Ayoub, "Integrating grasshopper optimization algorithm with local search for solving data clustering problems," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 783-793, 2021.

- [5] P. Padmavathi, V. Eswaramurthy, and J. Revathi, "Fuzzy social spider optimization algorithm for fuzzy clustering analysis," in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018: IEEE, pp. 1-6.
- [6] P. Padmavathi, V. Eswaramurthy, and J. Revathi, "Hybridization of Fuzzy C-Means and Fuzzy Social Spider Optimization for Clustering," in Advances in Electrical and Computer Technologies: Springer, 2021, pp. 179-187.
- [7] K. Vijayakumari and V. Baby Deepa, "Fuzzy C-Means Hybrid with Fuzzy Bacterial Colony Optimization," in Advances in Electrical and Computer Technologies: Springer, 2021, pp. 75-87.
- [8] K. Tamilarasi, M. Gogulkumar, and K. Velusamy, "Data clustering using bacterial colony optimization with particle swarm optimization," in 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021: IEEE, pp. 1-5.
- [9] K. Tamilarasi, M. Gogulkumar, and K. Velusamy, "Enhancing the performance of social spider optimization with neighbourhood attraction algorithm," in Journal of Physics: Conference Series, 2021, vol. 1767, no. 1: IOP Publishing, p. 012017.
- [10] 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.
- [11] J. Ji, H. Xiao, and C. Yang, "HFADE-FMD: a hybrid approach of fireworks algorithm and differential evolution strategies for functional module detection in protein-protein interaction networks," Applied Intelligence, pp. 1-15, 2020.
- [12] J. Revathi, V. Eswaramurthy, and P. Padmavathi, "Hybrid data clustering approaches using bacterial colony optimization and k-means," in IOP Conference Series: Materials Science and Engineering, 2021, vol. 1070, no. 1: IOP Publishing, p. 012064.
- [13] A. Abraham, S. Das, and S. Roy, "Swarm intelligence algorithms for data clustering," in Soft computing for knowledge discovery and data mining: Springer, 2008, pp. 279-313.
- [14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 881-892, 2002.
- [15] J. A. Nelder and R. Mead, "A simplex method for function minimization," The computer journal, vol. 7, no. 4, pp. 308-313, 1965.
- [16] Y. Zhou, Y. Zhou, Q. Luo, and M. Abdel-Basset, "A simplex method-based social spider optimization algorithm for clustering analysis," Engineering Applications of Artificial Intelligence, vol. 64, pp. 67-82, 2017.
- [17] B. Niu and H. Wang, "Bacterial colony optimization," Discrete Dynamics in Nature and Society, vol. 2012, 2012.
- [18] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," IEEE control systems magazine, vol. 22, no. 3, pp. 52-67, 2002.
- [19] S. D. Muller, J. Marchetto, S. Airaghi, and P. Kournoutsakos, "Optimization based on bacterial chemotaxis," IEEE Transactions on Evolutionary Computation, vol. 6, no. 1, pp. 16-29, 2002, doi: 10.1109/4235.985689.
- [20] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern recognition, vol. 36, no. 2, pp. 451-461, 2003.
- [21] I. De Falco, A. Della Cioppa, and E. Tarantino, "Facing classification problems with particle swarm optimization," Applied Soft Computing, vol. 7, no. 3, pp. 652-658, 2007.
- [22] M. Wan, L. Li, J. Xiao, C. Wang, and Y. Yang, "Data clustering using bacterial foraging optimization," Journal of Intelligent Information Systems, vol. 38, no. 2, pp. 321-341, 2012.
- [23] J. Revathi, V. Eswaramurthy, and P. Padmavathi, "Bacterial colony optimization for data clustering," in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019: IEEE, pp. 1-4.