Hybrid data clustering approaches using bacterial colony optimization and k-means

J Revathi¹, V P Eswaramurthy² and P Padmavathi³

¹,³Department of Computer Science, Periyar University, Salem, Tamilnadu, India.
²Department of Computer Science, Government Arts and Science College, Komarapalayam, Namakkal, Tamilnadu, India

E-Mail: ¹revathigacslm@gmail.com, ²eswar_murthy@yahoo.com, ³padmaperiyasamy99@gmail.com

Abstract. Data clustering is a fashionable data analysis technique in the data mining. K-means is a popular clustering technique for solving a clustering problem. However, the k-means clustering technique extremely depends on the initial position and converges to a local optimum. On the other hand, the bacterial colony optimization (BCO) is a well-known recently proposed data clustering algorithm. However, it is a high computational cost to complete a given solution. Hence, this research paper proposes a new hybrid data clustering method for solving data clustering problem. The proposed hybrid data clustering algorithm is a combination of the BCO and K-means called BCO+KM clustering algorithm. The experimental result shows that the proposed hybrid BCO+KM data clustering algorithm reveal better cluster partitions.

Keywords. Data clustering, k-means clustering algorithm, bacterial colony optimization, hybrid clustering algorithm

1. Introduction

Data clustering is the most significant unsupervised learning problem that deals with finding a structure in a collection of unlabeled data. In the data clustering, the similar data samples are falling into same groups and dissimilar data samples are falling into different groups. The data clustering is an essential tool for examining data analysis, because prior information about the data set is not required during the learning process. K-means clustering is a well-known partition data clustering algorithm. The k-means clustering techniques produce better results only when the initial values were close to the expected solution. The outcomes of k-means techniques extremely depend on the initial state of solution and achieve to local optimal solution. To solve the shortcomings of k-means clustering algorithm, the swarm intelligence (SI) [1] algorithm is applied including artificial bee colony optimization (ABC) [2], particle swarm optimization (PSO) [3] [4], social spider optimization (SSO) [5], bacterial foraging optimization algorithm (BFO) [6] and bacterial colony optimization (BCO) [7].

Recently, we present a BCO algorithm for solving data clustering problem [7]. The BCO clustering algorithm achieved a higher clustering efficiency compared with k-means and PSO. The most important gain of BCO algorithm is an exchanging the information with others by using a communication method. Two types of communication mechanism have commenced in BCO algorithm
including individual interaction and group exchanges. The communication process is used to improve the efficiency of given solutions. However, conventional BCO is a low convergence speed and high computation time when solving data clustering problem. Since, the BCO clustering algorithm obtains more computation time due to the method of some internal iteration in order to attain significant clustering outcome.

However, the conventional SI algorithms are required to large development aptitude that may just to fall into premature convergence and takes more computation time. The conventional individual data clustering algorithm has many advantages and many disadvantages by own [8]. Hence, the hybridization of any two algorithms is an alternative way to achieve the challenges of individual data clustering algorithm [9]. The various SI approaches are combined with k-means clustering in order to enhance the performance clustering quality and overwrite demerits of an individual data clustering algorithm.

In this paper, we proposed a new hybrid data clustering algorithm is a combination of BCO and k-means for solving data clustering problem. The objectives of proposed hybrid data clustering algorithm is used to enhance the performance of the data clustering problem and overwrite the demerits of both data clustering algorithms. In the proposed hybrid data clustering algorithm, the BCO is used to search the entire space for the global optimum solution. Then, the cluster process is change to k-means clustering algorithm for obtaining more accurate similar groups when the BCO algorithm achieves a solution near to the optimum solution. The contribution of the paper is,

- The k-means clustering algorithm is combined with BCO in order to produce more similar data partitions
- The proposed hybrid data clustering algorithm is utilize the benefits to overwrite the shortcomings of both algorithms

2. Data clustering

Given a data samples \( X = \{x_1, x_2, \ldots, x_n\} \) find out a partition of the data samples \( N \) into \( K \) cluster \( C_1, C_2, \ldots, C_K \). The objective of the clustering algorithm is to reduce the sum squared error (SSE) and it can be defined as follows

\[
SSE = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2
\]  

Where, \( \| x_i - c_k \|^2 \) - denotes distance between data points \( x_i \) and center point of \( c_k \), in this study, Euclidean distance measure is used for measuring the sum of squared error.

3. K-means clustering algorithm

K-means clustering technique is a well-known unsupervised method for solving the data clustering problem. It is partitioned data clustering algorithm, simple, easy to write, straightforward and low computation cost [10]. This algorithm is commencing with cluster center values which are selected randomly. The each data points are allocated to the closest center of the groups based on similarity. Similarity of data points is calculated by using distance values. Euclidean distance measures a well-known and widely used similarity measures. The distance function is defined as follows,

\[
D(x, z) = \sqrt{\sum_{j} (x_j - z_j)^2}
\]

Here, \( x \) is represents the \( p^{th} \) data sample, \( z \) is representing the centroid values of the cluster \( j \) and

**Algorithm 1: K-means clustering algorithm**

1: Randomly select k cluster centroid values
2: Find distance vector using the equation (2)
3: Update the cluster centroid values using equation (3)
4: Perform the termination condition if yes then go to step 2, Otherwise, terminates the process
\(d\) is the number of data features. The center of groups is restructured by using the mean value of the associated data point belongs to respected class.

\[
z_j = \frac{1}{n_j} \sum_{p \in c_j} x_p\]  

(3)

Here, \(n_j\) is represents the data samples associated with \(j\). \(c_j\) is represents a subset group from the cluster \(C\). The process of k-means clustering algorithm expired when any one of the following condition is satisfied: when exceed the maximum number of iterations or when there are no cluster membership change. The algorithm 1 shows the process of the k-means clustering algorithm.

4. Bacterial colony optimization (BCO)

BCO is a latest SI algorithm designed by Niu and Wang (2012) that mimics the behavior of artificial bacteria [11]. The major difference between BCO and other bacteria inspired heuristic algorithms such as BFO [12] and BC [13] is that the bacteria in BCO searches for nutrients by exchanging information between individuals instead of just swimming randomly. The BCO has five steps including chemotaxis and communication, elimination and reproduction, migration process. In which, the chemotaxis and communication is major and newly added scheme in order to enhance searching ability. The communication mechanism is established based on an interactive communication scheme with each bacterium in order achieves global optimum values. Hence, this communication mechanism can reduce the computation cost and avoid premature convergence. The algorithm 2 shows the process of the BCO clustering algorithm. Therefore, the updating the position of the each bacterial colony as follows,

\[
\text{Position}_i(T) = \text{Position}_i(T-1) + C(i) \cdot [f_i \cdot (G_{best} - \text{Position}_i(T-1)) + (1-f_i) \cdot (P_{best} - \text{Position}_i(T-1)) + turb]
\]  

(4)

Algorithm 2: BCO clustering algorithm

Begin

Step 1: Initialize population \(s\)
Step 2: Each bacterial colony
Step 3: Chemotaxis and communication process
Step 4: Compute the cluster center for each colony
Step 5: compute the distance between cluster centroid and data samples
Step 6: An interactive exchange process
  Step 6.1: If Individual Exchange
    If Dynamic neighbour oriented
    If found poorer fitness bacterial colony, then change with random values
    Else Random oriented
    Change poorer colony after finding poor fitness value of bacterial colony
  from the given data sample selected randomly
  Step 6.2: Else if Group exchange
    Change the poorer after finding the best bacterial colony dependent upon the
    fitness values
Step 7: Reproduction and elimination process
  Step 7.1: Find the strong bacterial colony
  Step 7.2: Execute the reproductions and elimination
Step 8: Migration process
Step 9: Bacterial colony updating
Step 10: Search the position of minimum distance class label
Step 11: Assigning the each data sample into associated cluster
Step 12: If terminating state is not met, then go to step 2.

End
Position_i(T) = Position_i(T-1) + C(i) × \left[ f_i × (G_{best} - Position_j(T-1)) + (1 - f_i) \times (P_{best} - Position_j(T-1)) \right] 
\tag{5}

C(i) = C_{min} + \left( \frac{\text{Iter}_{\max} - \text{Iter}_j}{\text{Iter}_{\max}} \right) \left( C_{\max} - C_{\min} \right) 
\tag{6}

Where, turbi is a turbulent direction variance, C(i) is the chemotaxis step size, f_i ∈ (0,1), G_{best} is the global best and P_{best} is the personal best (local best). Iter_{\max} and Iter_j are the maximum iteration and current iteration respectively.

5. Proposed hybrid clustering algorithm

The k-means clustering method tends to converge faster than the BCO due to it need less functions evaluations. However, it produced low clustering accuracy. On the other hand, The BCO algorithm produced high cluster accuracy, but it takes more computation cost when compared with K-means clustering algorithm. Hence, the k-means clustering algorithm hybrid with BCO algorithm is

Algorithm 3: hybrid BCO+KM clustering algorithm

Begin hybrid BCO+KM

Begin BCO

Step 1: Initialize population s
Step 2: Each bacterial colony
Step 3: Chemotaxis and communication process
Step 4: Compute the cluster center for each colony
Step 5: compute the distance between cluster centroid and data samples
Step 6: An interactive exchange process
  Step 6.1: If Individual Exchange
   If Dynamic neighbour oriented
   If found poorer fitness bacterial colony, then change with random values
   Else Random oriented
   Change poorer colony after finding poor fitness value of bacterial colony
   from the given data sample selected by randomly
  Step 6.2: Else if Group exchange
   Change the poorer after finding the best bacterial colony dependent upon the fitness values
Step 7: Reproduction and elimination process
  Step 7.1: Find the strong bacterial colony
  Step 7.2: Execute the reproductions and elimination
Step 8: Migration process
Step 9: Bacterial colony updating
Step 10: Search the position of minimum distance class label
Step 11: Assigning the each data sample into associated cluster
Step 12: If terminating state is not met, then go to step 2.

End BCO

Begin K-means

Step 13: Calculate the cluster center using global best G_{best}
Step 14: Find distance vector by using equation (2)
Step 15: Update the cluster centroid values using (3)
Step 16: Perform the termination condition if yes then go to step 2, otherwise terminate the process

End K-means

End Hybrid BCO+KM
proposed in this paper. The algorithm 3 shows the process of the proposed hybrid BCO+KM clustering algorithm.

6. Experimental results and discussions
The experimental outcomes are developed by using MATLAB 2015b, i5 processor, 1.90 GHz and 4GB RAM computer. The outcomes of every data clustering techniques are achieved from the average values of fifty independent runs. The strength of the proposed hybrid BCO+KM clustering algorithm is compared with some benchmark algorithm such as k-means [14], PSO [3], BFO [15] and BCO [7]. The experimental results confirmed that the proposed hybrid BCO+KM clustering algorithm produced higher clustering accuracy.

Data collections
In this research work, ten well-known machine learning datasets are used for analyzing the performance of the proposed hybrid clustering method. The details of used datasets are shown in Table 1.
- Artificial datasets 1, it contains two features problem, four distinctive classes and six hundred data samples [8]. The classes were scattered according to,
  \[ N(\mu = \begin{pmatrix} m \\ 0 \end{pmatrix}, \sum = \begin{pmatrix} 0.05 & 0.5 \\ 0.5 & 0.05 \end{pmatrix}) \]
  \( i=1,2,3,4 \) \( m = 1, m = 2, 3, 4 \) \( m = 6 \), here, \( \mu \) and \( \sum \) represent the mean vector and covariance matrix respectively. Artificial dataset 2, it consist of three features, five different classes and [8]. Such as 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).
- Balance data set is created based on weight of the balance scale and distance. It contains six hundred and twenty five data samples with 4 features and 3 unique classes.
- Contraceptive Method Choice (CMC) dataset contains 1473 data points with 3 categories and 9 features
- Glass datasets have two hundred and four data points with six classes with 9 features in each type.
- The Heart data set is based on the heart disease diagnosis. It has seventy six attributes for each data sample with 303 data sample and 2 unique classes.
- Fisher’s iris consists of three different species of the iris flower, one fifty samples and four features were collected in each species.
- Wisconsin Breast Cancer (WBC) data set has 683 data samples and two classes with 9 features.
- The wine dataset contains of 178 data samples and 3 types of class with 13 features.
- The vowel dataset has 871 indian vowels and 3 features with 6 classes

| Datasets        | Instances | Features | Clusters |
|-----------------|-----------|----------|----------|
| Artificial dataset 1 | 600       | 2        | 4        |
| Artificial dataset 2 | 250       | 3        | 5        |
| Balance         | 625       | 4        | 3        |
| Iris            | 150       | 4        | 3        |
| Glass           | 214       | 9        | 6        |
| Heart           | 303       | 76       | 2        |
| WBC             | 683       | 9        | 2        |
| Wine            | 178       | 13       | 3        |
| CMC             | 1,473     | 9        | 3        |
| Vowel           | 871       | 3        | 6        |
Parameters settings
The choosing best possible parameters of the clustering algorithms can be improving the performance of cluster superiority. In this paper, the following parameter settings are followed in order to investigate the performance of the proposed hybrid clustering method. In k-means, hundred is set for the maximum iterations. In PSO clustering algorithm, \( P = 100, \ w = 0.9 \rightarrow 0.1, \ c_1 = c_2 = 2 \). In BFO data clustering algorithm, \( S = 100, \ N_C = 100, \ N_C = 4, \ N_s = 4, \ N_{ed} = 2 \) and \( P_{ed} = 0.25 \). In the proposed hybrid clustering method, the convergence rate of BCO is determined by depending upon the its chemotaxis step swim step \( N_s \) and \( N_C \) values. The high value of chemotaxis step of the BCO is taking more computation cost. Hence, chemotaxis step is taking as a \( N_C = 100 \). Similarly, \( N_s = 4 \). The value of reproduction is chosen as \( N_{re} = 4 \) and the value of dispersal step is selected as \( N_{ed} = 2 \).

Parameter measures
The clustering process can be defined based on the optimization techniques. The major aim of the objective function is try to reduce the sum of squared error (SSE) between the data samples and the center of cluster. The lowest value of objective function is considered as the best values. The objective of the clustering problem is to reduce the (SSE) between data samples. Three kinds of values are calculated from the objective function such as best, average and worst. The low value of

| Table 2: Performance of the proposed model based on objective functions |
|---|---|---|---|
| **Datasets** | **Methods** | **Objective Function Values** |
| | **Best** | **Average** | **Worst** |
| Artificial dataset 1 | K-Means | 585.53 | 653.035 | 720.54 |
| | PSO | 612.60 | 637.255 | 661.91 |
| | BFO | 579.39 | 599.475 | 619.56 |
| | BCO | 566.19 | 541.79 | 517.39 |
| | Hybrid BCO+KM | 556.19 | 562.71 | 569.23 |
| Artificial dataset 1 | K-Means | 2338.92 | 2595.415 | 2851.91 |
| | PSO | 2231.59 | 2365.425 | 2499.26 |
| | BFO | 2194.21 | 2247.07 | 2299.93 |
| | BCO | 2014.63 | 2099.4 | 2184.17 |
| | Hybrid BCO+KM | 1917.46 | 1957.825 | 1998.19 |
| Balance | K-Means | 754.18 | 818.68 | 883.18 |
| | PSO | 679.61 | 694.935 | 710.26 |
| | BFO | 657.19 | 664.39 | 671.59 |
| | BCO | 636.18 | 652.825 | 669.47 |
| | Hybrid BCO+KM | 619.71 | 622.395 | 625.08 |
| | K-Means | 109.87 | 150.755 | 191.64 |
| | PSO | 95.63 | 97.56 | 99.49 |
| Iris | BFO | 92.46 | 93.445 | 94.43 |
| | BCO | 91.90 | 92.175 | 92.45 |
| | Hybrid BCO+KM | 89.62 | 90.325 | 91.03 |
| Glass | K-Means | 214.67 | 235.49 | 256.31 |
| | PSO | 246.16 | 249.55 | 252.94 |
| | BFO | 204.15 | 221.755 | 239.36 |
| | BCO | 192.21 | 199.09 | 205.97 |
| | Hybrid BCO+KM | 191.19 | 194.19 | 197.19 |
the objective function values is called best value which is considered as a best performance. A highest value of objective function is considered as a worst value which is considered as a worst performance. The average is calculated the average values of the both high and low values.

**Discussions**

The present part is discussing the different key factors for performance comparisons of data clustering algorithm in order to evaluate performance of the proposed hybrid clustering algorithm. Table 2 Table 3 shows the performance comparison of the proposed hybrid methods based objective function for all datasets. The graphical representation of performance comparison shows in Figure 1. Compared with objective value, the proposed hybrid clustering method is formed low objective values when compared with other existing clustering techniques.

For instance, the outcome obtained on the iris dataset demonstrates that the proposed hybrid clustering method converges to the global optimum value of 89.62. As well, The BCO, BFO, PSO and k-means algorithm reaches to 91.90, 92.46, 95.63, and 109.87 respectively. Similarly, compared with other datasets, the proposed OBL+BCO are obtained minimum objective values. On the average, the proposed hybrid clustering algorithm produced higher efficient cluster quality. The experimental results confirmed that proposed method is an effective clustering method and it can be applied to real world clustering problem with efficient manner.

**Table 3(Continue...): Performance of the proposed model based on objective functions**

| Datasets | Methods       | Objective Function Values |
|----------|---------------|----------------------------|
| Heart    | K-Means       | 4891.67 5155.285 5418.90 |
|          | PSO           | 4217.26 4543.725 4870.19 |
|          | BFO           | 4181.61 4222.66 4263.71  |
|          | BCO           | 4082.27 4100.81 4119.35  |
|          | Hybrid BCO+KM | 3898.93 3918.6 3938.27  |
|          | K-Means       | 2923.82 3169.99 3416.16  |
|          | PSO           | 2705.82 2836.505 2967.19 |
|          | BFO           | 2663.17 2712.375 2761.58 |
|          | BCO           | 2632.53 2625.07 2617.61  |
|          | Hybrid BCO+KM | 2680.09 2664.895 2649.70 |
|          | K-Means       | 16583.12 17350.135 18117.15 |
|          | PSO           | 16491.94 16542.945 16593.95 |
|          | BFO           | 16219.48 16266.645 16313.81 |
|          | BCO           | 16237.62 16228.45 16219.28 |
|          | Hybrid BCO+KM | 16264.82 16289.38 16313.94 |
|          | K-Means       | 5889.49 5931.83 5974.17  |
|          | PSO           | 5779.34 5835.58 5891.82  |
|          | BFO           | 5629.39 5700.3 5771.21  |
|          | BCO           | 5690.81 5664.225 5637.64  |
|          | Hybrid BCO+KM | 5547.27 5713.105 5878.94  |
|          | K-Means       | 149548.16 154884.94 160221.72 |
|          | PSO           | 148934.42 149457.5 149980.58 |
|          | BFO           | 147014.49 147253.4 147492.31 |
|          | BCO           | 146671.46 146721.37 146771.27 |
|          | Hybrid BCO+KM | 146271.75 146443.07 146614.39 |
Figure 1: Average performance comparisons of proposed hybrid data clustering algorithm

7. Conclusions
This present research paper is investigates the data clustering problem by using a hybrid BCO+KM algorithm. The experimental results are conducted on ten different machine learning datasets. The experimental results show that the proposed a hybrid BCO+KM algorithm is reveal the higher clustering accuracy when compared with some clustering algorithm.

References
[1] Abraham A Das S and Roy S 2008 Swarm intelligence algorithms for data clustering Soft computing for knowledge discovery and data mining Springer.
[2] Zhang C Ouyang D and Ning J 2010 An artificial bee colony approach for clustering Expert Systems with Applications 37 4761-4767.
[3] De Falco I Della Cioppa A and Tarantino E 2007 Facing classification problems with particle swarm optimization Applied Soft Computing 7 652-658.
[4] Cui X Potok T E and Palathingal P 2005 Document clustering using particle swarm optimization Swarm Intelligence Symposium IEEE
[5] Padmavathi P Eswaramurthy V P and Revathi J 2018 Fuzzy Social Spider Optimization Algorithm for Fuzzy Clustering Analysis International Conference on Current Trends towards Converging Technologies (ICCTCT) IEEE.
[6] Wan M Li L Xiao J Wang C and Yang Y 2012 Data clustering using bacterial foraging optimization 38 321-341.
[7] Revathi J Eswaramurthy V P and Padmavathi P 2019 Bacterial Colony Optimization for Data Clustering IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) IEEE.
[8] Niknam T and Amiri B 2010 An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis Applied soft computing 10 183-197.
[9] Ji J Xiaoy H and Yang C 2020 HFADE-FMD: a hybrid approach of fireworks algorithm and differential evolution strategies for functional module detection in protein-protein interaction networks Applied Intelligence, 1-15.
[10] Kanungo T Mount D. M Netanyahu N. S Piatko C. D Silverman R and Wu A Y 2002 An efficient k-means clustering algorithm: Analysis and implementation IEEE transactions on pattern analysis and machine intelligence 24 881-892.
[11] Niu B and Wang H 2012 Bacterial colony optimization Discrete Dynamics in Nature and Society 2012.
[12] Passino K M 2002 Biomimicry of bacterial foraging for distributed optimization and control IEEE control systems magazine 22 52-67.
[13] Muller S D Marchetto J Airaghi S and Kournoutsakos P 2002 Optimization based on bacterial chemotaxis IEEE Transactions on Evolutionary Computation 6, 16-29.
[14] Likas A Vlassis N and Verbeek J J 2003 The global k-means clustering algorithm Pattern recognition 36 451-461.
[15] Wan M Li L Xiao J Wang C and Yang Y 2012 Data clustering using bacterial foraging optimization Journal of Intelligent Information Systems 38 321-341.