A Novel Spatial Clustering Analysis Method Using Bat Algorithm

Guanghui LIU, Heyan HUANG, Shumei WANG, Zhaoxiong CHEN

1 School of Computer Science and Technology, Nanjing University of Science & Technology, Nanjing, 210094, China, lgh025@163.com
2 School of Computer Science and Technology, Beijing Institute of Technology, Beijing, 100081, China, hhy63@bit.edu.cn
3 School of Information Engineering, Nanjing University of Finance & Economics, Nanjing, 210046, China

Abstract

In data mining, clustering can be regarded as the process of finding some optimal centers. Bat algorithm is a powerful method for solving many multi-objective optimization problems. It has fast convergence and global optimization ability of the search features. In this paper, we present a bat clustering algorithm (BCA) based on similarity, and improve the corresponding fitness function. Contrary to Particle Swarm Optimization (PSO) clustering algorithm using Euclidean distance, BCA performs a more correct search in the entire solution space. Spatial data sources are from IRIS dataset and wine dataset. These data are clustered by clustering analysis based on BCA-similarity algorithm. Experimental results show that BCA-similarity clustering method can achieve fast and accurate spatial data clustering.

Key words: Clustering; Bat algorithm; similarity; Euclidean distance; PSO

1. Introduction

Spatial clustering analysis can be regarded as the process of dividing the concrete or abstract data into several groups or classes. Each set of data of clustering is called a cluster, and each cluster data is called an object. Purpose of clustering is to make characteristics of objects in the same cluster as similar as possible, characteristic differences between objects in the different cluster as large as possible. Similar or dissimilar measure is based on the values of the data object’s description property to determine. Object distance or similarity function is usually used to describe. Clustering analysis can be used as an unsupervised learning method to mine association rules from the characteristic data of study objects. In many fields, clustering analysis has a wide range of applications, and achieves satisfactory results. It is a powerful information processing method.

Particle swarm optimization (PSO) was originally developed by Eberhart and Kennedy in 1995[1]. PSO algorithm was inspired by the swarm behavior such as fish and bird schooling in nature. In the PSO algorithm, particles can be considered as simple agents “flying” through a problem space. A particle’s location in the multi-dimensional problem space represents one solution for the problem. When a particle moves to a new location, a different problem solution is generated[2,3]. This solution is evaluated by a fitness function that provides a quantitative value of the solution’s utility[4,5].

By idealizing some of the echolocation characteristics of micro-bats[6,7,8,9], Yang proposed a new optimization algorithm, namely, Bat Algorithm (BA), in 2010. Although the original BA presents
superior results in the experiments than PSO, we notice that the performance and the accuracy of the original BA still have the capacity to present better.

In this paper, based on improved bat algorithm and similarity, we give a novel acquisition method to clustering rules. This classification method can guarantee the global convergence. According to the experimental results, our proposed algorithm is better than PSO algorithm using Euclidean distance.

2. Bat Algorithm [BA]

The bat algorithm [10,11] is based on the echolocation behavior of bats. Micro-bats emit a very loud sound pulse (echolocation) and listens for the echo that bounces back from the surrounding objects. The \textit{i}th bat flies randomly with velocity \( v_i \) at position \( x_i \) with a fixed frequency \( f_{\text{min}} \). The bats search for food through varying their wavelength \( \lambda \) and loudness \( A_0 \). In Yang’s method, the movement of the virtual bat is given by Eq. (1) – Eq.(3).

\[
\begin{align*}
  f_i &= f_{\text{min}} + (f_{\text{max}} - f_{\text{min}}) \cdot \beta \\
  v_i^t &= w \cdot v_i^{t-1} + (x_i^t - x_{\text{best}}) \cdot f_i^t \\
  x_i^t &= x_i^{t-1} + v_i^t
\end{align*}
\]

Where \( f \) is the frequency used by the bat searching for its prey, while the suffixes, \( \text{min} \) and \( \text{max} \), represent the minimum and maximum value, respectively, \( x_i \) denotes the location of the \( i \)th bat in the solution space, \( v_i \) represents the velocity of the bat, \( t \) indicates the current iteration, \( \beta \) is a random vector, which is drawn from a uniform distribution, and \( \beta \in [0,1] \), and \( x_{\text{best}} \) indicates the global near best solution.

For local search part, a new solution for each bat is generated locally using random walk based on the best selected current solution.

\[
x_{\text{new}} = x_{\text{old}} + \varepsilon A^t
\]

Where \( \varepsilon \in [-1,1] \) is a random number, while \( A^t \) is the average loudness of all bats at this time step. The loudness decreases as a bat tends closer to its food and pulse emissions rate increases.

\[
\begin{align*}
  A_i^{t+1} &= \alpha \cdot A_i^t \\
  r_i^{t+1} &= r_i^0 \left[ 1 - e^{-\gamma t} \right]
\end{align*}
\]

Where \( \alpha \) and \( \gamma \) are constants.

3. Modified Bat Algorithm

In this paper, we use a modified Bat Algorithm (MBA). Updating velocities and locations can be reformulated into Eq.(7) and Eq.(8)
\[ v^t_i = w_1 \cdot v^{t-1}_i + (x_{\text{best}} - x^t_i) \cdot f^t_i \] (7)

\[ x^t_i = w_2 \cdot x^{t-1}_i + v^t_i \] (8)

Where \( x_{\text{best}} \) is the global best of all the bats, while \( w_1 \) and \( w_2 \) are inertia weights. The parameters \((w_1,w_2)\) can balance the proportional relationship between the global convergence ability and the local convergence ability, and improve the performance of the algorithm. In general, \((w_1,w_2) \in [0.7,1]\).

4. Similarity metric between the variables - similarity coefficient

Similarity has been an important research topic in several fields such linguistic, Artificial intelligence. When evaluating similarity in a taxonomy, the most natural way to access similarity is to evaluate the Distances between the two concepts being compared. Therefore the shorter is the path from one to another means that they are more similar\(^{12}\). This approach has been used as measure of similarity. However, one of the main drawbacks is that it relies on the notion that links in a taxonomy represent uniform distances (Resnik 1995; 1998), Our own view is that similar entities are assumed to have common features, Our similarity algorithm assess concept similarity and relation similarity. When \( p \) target variables cluster, the similarity coefficient could measure the degree of similarity between the variables (or relevance). In general, \( C_{\alpha\beta} \) represents the similarity coefficient between variables \( x_\alpha \) and \( x_\beta \), should be met:

1) \( |C_{\alpha\beta}| \leq 1 \) and \( C_{\alpha\alpha} = 1 \);
2) \( C_{\alpha\beta} = \pm 1 \iff x_\alpha = c x_\beta (c \neq 0) \);
3) \( C_{\alpha\beta} = C_{\beta\alpha} \);

The absolute value of \( C \) is closer to 1, the correlation between variables \( A \) and \( B \) is greater.

Assume sets of variables \( X= (x_{\alpha_1}, x_{\alpha_2}, ..., x_{\alpha_n})^T \) and \( Y= (x_{\beta_1}, x_{\beta_2}, ..., x_{\beta_n})^T \) are \( n \)-dimensional vectors, \( r \) define the similarity coefficient of \( X, Y \) :

\[ r = \frac{\sum_{i=1}^{n} x_{\alpha_i} \cdot x_{\beta_i}}{\sqrt{ (\sum_{i=1}^{n} x_{\alpha_i}^2) \cdot (\sum_{i=1}^{n} x_{\beta_i}^2) } } \] (9)

According to the above definition, the similarity coefficient of input sample \( X^k \) and sample the cluster center \( P_j \) is modified as:

\[ r^k_j = \frac{\sum_{i=1}^{n} x_{\alpha_i} \cdot p_{\beta_i}}{\sqrt{ (\sum_{i=1}^{n} x_{\alpha_i}^2) \cdot (\sum_{i=1}^{n} p_{\beta_i}^2) } } \] (10)

Assume the cluster center node \( P^* \) with the maximum similarity coefficient to win the competition, \( P^*_j \) should be met:

\[ r^k_j = \max_{p\in[1,2,...,m]} \{ r^k_p \}. \]
In order to prevent premature convergence to find the maximum value in the PSO algorithm, here we use the reciprocal of the maximum similarity coefficient: 

\[ r_i^{**} = \min_{j \in [2, \ldots, m]} \left\{ \frac{1}{r_j} \right\} \]

change to find the minimum.

5. Spatial clustering based on bat algorithm

In spatial clustering analysis based on bat algorithm, each bat represents the \( C \) centroids of \( K \) classes. Bat consists of many candidate classification schemes. The evaluation of the pros and cons of the classification scheme is the key to the use of optimization algorithms for spatial clustering[13].

The objective of the clustering algorithm is to discover the proper centroids of clusters for minimizing the intra-cluster distance as well as maximizing the distance between clusters[14,15]. The clustering algorithm performs a globalized searching for solutions whereas most other partitional clustering procedures perform a localized searching.

Commonly used fitness function is the Euclidean distance function, although the between-class distance and within-class distance have a certain relationship, However, due to the uncertainty of this relationship, leads to incomplete evaluation strategy. This method isn’t used in this paper, instead, the similarity as fitness function.

Each bat \( O_i = (o_{i1}, o_{i2}, \ldots, o_{in}) \) represents the centroids of the \( k \) clusters, where \( O_{ij} \) denotes the centroids spatial coordinate vector corresponds to the \( j \) class of the \( i \) particle. The fitness function \( f(x) \) is the reciprocal of the sum that the similarity coefficients of input sample and the group cluster centroids plus.

That is:

\[ f(x) = \frac{1}{\sum_{i=1}^k r_i} \]

(11)

Where

\[ r_i = \frac{\sum_{j=1}^n x_i \cdot o_{ij}^T}{\left( \sum_{j=1}^n x_i \cdot o_{ij}^T \right) \ast \left( \sum_{j=1}^n o_{ij} \cdot o_{ij}^T \right)} \]

(12)

\( r_i \) denotes the similarity coefficient between a spatial data vector \( x_i \) and a cluster centroids \( O_{ij} \).

Assume the cluster centroids \( O'_i \) with the maximum similarity coefficient to win the competition, here we use the reciprocal of the maximum similarity coefficient, so fitness function is changed to find the minimum. \( O'_i \) should be met:

\[ f' = \min_{j \in [2, \ldots, m]} \{ f' \} \]

Algorithm steps:

Step1: Bat Swarm initialization; Create a swarm with \( n \) bats. Initialize dimensions \( D \) and inertia weight \( w \). Initialize the bat population \( x_i \) \((i = 1, 2, \ldots, n)\) and \( v_i \). Define Pulse frequency \( f_i \) at \( x_i \). Initialize the
rates $r_i$ and the loudness $A_i$. Initialize current-global-best and current-local-best for the swarm.

**Step2:** Load and standardize test dataset.

**Step3:** Randomly initialize each bat's position and flight velocity vector, position and flight velocity vectors are composed of a spatial vector $x = (x_1, ..., x_d)^T$.

**Step4:** Determine each bat's spatial clustering program; according to the Eq.(11) and Eq. (12). Calculate similarity function of sets $(x_1, x_2, ..., x_n)$ that will be classified and K cluster centers corresponds to bats. According to sorting the size of the fitness, classify $(x_1, x_2, ..., x_n)$.

**Step5:** According to the current position and velocity, search out the optimal position ($x_{best}$) of each bat so far.

**Step6:** Update bats' velocity and position as following:

While $(t < \text{Max number of iterations})$

Generate new solutions by adjusting frequency and updating velocities and locations/solutions [Eq.(1), (7) and (8)],

If $(\text{rand} > r_i)$

Select a solution among the best solutions [Eq.(4)]

Generate a local solution around the selected best solution

End if

Generate a new solution by flying randomly

If $(\text{rand} < A_i \& F(x_j) < F(x_{best}))$

Accept the new solutions [Eq.(8)]

Increase $r_i$ and reduce $A_i$ [Eq.(5) and Eq.(6)]

End if

Calculate objective function value $F(x)$

Rank the bats and find the current best $x_{best}$

End while

**Step7:** If the current optimum bats does not meet the convergence conditions, re-clustering divide the current bat swarm. Repeat steps (2)-(6) until one of the following termination conditions is satisfied. (a) The maximum number of iterations is exceeded or (b) The average change in centroid vectors is less than a predefined value.

**Step8:** Record global optimal solution.

---

6. Simulation and Comparison

To verify the validity of the proposed BA clustering algorithm, several simulations were done, and were compared with the PSO clustering algorithm using Euclidean distance or similarity. Various studies show that PSO algorithms can outperform genetic algorithms (GA) and other conventional algorithms for solving many clustering problems[16,17,18,19]. Bat algorithm, as a improved clustering algorithm which is better than PSO, must be validated from experiments.

The experiments are taken on the personal computer with an Intel Core-i7 2600K 3.4GHz CPU,

6.1. Simulations of IRIS dataset

In this study, the well-known IRIS dataset is used. The dataset contains three classes of four-dimensional sample: Setosa, Versicolor, Virginica. The number of samples per class is 50. With 2, 3, 4-dimensional data as example, spatial distribution of the three types of samples are shown in figure 1. For comparison, the same parameters are used as far as possible in the experiment. PSO parameters are shown in Table 1, and the parameter settings for our proposed method MBA are listed in Table 2. The experimental results are shown in figure 2, figure 3 and Table 3.

Figure 1. IRIS sample spatial distribution

Table 1. Parameters of PSO clustering algorithm

<table>
<thead>
<tr>
<th>dimension (D)</th>
<th>learning factor1(c1)</th>
<th>learning factor2(c2)</th>
<th>number of particles (N)</th>
<th>Max iterations</th>
<th>Inertia weight(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.5</td>
<td>2</td>
<td>40</td>
<td>100</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 2. The parameter setting for MBA.

<table>
<thead>
<tr>
<th>initial (A_i^0)</th>
<th>Initial (r_i^0)</th>
<th>([f_{min}, f_{max}])</th>
<th>iterations</th>
<th>(\varepsilon)</th>
<th>(\alpha)</th>
<th>(\gamma)</th>
<th>(v_{max})</th>
<th>(w_1)</th>
<th>(w_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>[0,1]</td>
<td>100</td>
<td>[-1,1]</td>
<td>0.9</td>
<td>0.9</td>
<td>10</td>
<td>0.75</td>
<td>0.85</td>
</tr>
</tbody>
</table>
In our implementation, the total population size is set to 40. Each test containing the full iterations is repeated by 25 runs with different random seeds. The final result (correct rate) is obtained by taking the average of the outcomes from all runs.

Table 3. The comparison of average correct rate between PSO with BA in 25 runs

<table>
<thead>
<tr>
<th>dataset</th>
<th>clustering algorithm</th>
<th>correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>first category</td>
</tr>
<tr>
<td>IRIS</td>
<td>PSO &amp; Euclidean distance</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>BA &amp; similarity</td>
<td>100%</td>
</tr>
</tbody>
</table>

From the tests, the correct rate of Bat Algorithm [BA] classification rules based on the similarity in the IRIS dataset was 100%, 96%, 92%; but the correct rate of PSO classification rules based on the Euclidean distance was 100%, 94%, 74%. Experimental results show that BA clustering algorithm based on similarity can successfully classify all instances of the IRIS dataset. Compared with the PSO clustering algorithm based on Euclidean distance, the correct rate has increased.

6.2. Simulations of Wine dataset

In comparison tests, wine dataset is also used. These wine data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Some characteristics are listed below:

- Data size: 178 entries
- 3 classes
- Data distribution: 59, 71, and 48 entries for each class
In the experiments, wine dataset is respectively classed with PSO method and Bat Algorithm. Euclidean distance and similarity are both used as fitness function with above two algorithms. We will use the same population size of n=40 for all algorithms in all simulations. We have carried out extensive simulations and each algorithm has been run at least 25 times so as to carry out meaningful statistical analysis. The parameter settings for PSO and MBA are shown in Table 1 and Table 2. The experimental results are shown in figure 4, figure 5 and figure 6. Table 4 is the comparison of average correct rate between PSO with BA in 25 runs.

First, we will compare the convergence characteristic of the BA and PSO for similarity fitness function, as shown in Figure 4. The fitness function \( f(x) \) is the reciprocal of the sum that the similarity coefficients [Eq.(4)].

![fitness convergence graphs](image)

**Figure 4.** The convergence graphs of the PSO and BA for similarity fitness function.

Studying the convergence graphs in Figure 4, we can see that the BA is noticeably more efficient in finding the global optima. This is no surprising as the aim of developing the improved bat algorithm was to try to use the advantages of existing algorithms and other interesting feature inspired by the fantastic behavior of echolocation of micro-bats.

![Wine Clustering results of PSO algorithm based on Euclidean distance](image)

**Figure 5.** Wine Clustering results of PSO algorithm based on Euclidean distance
From the experiments, PSO and bat algorithm based on similarity have better clustering results. But from the overall experimental results analysis, it can be seen that our proposed bat algorithm makes the clustering fitness function value to converge minimum. This shows that the similarity is larger and the clustering is more compact. From figure 4, figure 5, figure 6 and Table 4, it shows that the correct number and the correct rate of clustering using bat algorithm have been greatly improved. Experiments
show that the proposed bat algorithm produces more accurate classification models than PSO classifier.

7. Conclusions

In this paper, by combining similarity with the bat algorithm, and by reanalyzing the characteristics of the bat and redefining the corresponding operations based on the basic framework of Bat Algorithm (BA), we have successfully proposed a newly improved Bat Clustering Algorithm (BCA). The experimental results indicate that our proposed BCA produces a more accurate outcome. We can see it is potentially more powerful than PSO. However, for noisy datasets, whether this algorithm can produce better clustering results, it could be an important topic for further research.

8. References