Data Mining Model Based on Improved Ant Colony Algorithm

Haoning Wu1*
School of Computer Science, Wuhan University, Wuhan, 430079, China
*Corresponding author’s e-mail: Haoning.Wu@whu.edu.cn

Abstract. Data mining technology is an important means and way of data analysis in the Internet era, and it is a process of extracting potentially useful information from large databases. From the perspective of theory and technology, data mining is a process of discovering the relationship between data and models or between data from a huge database. From the application level, data mining can provide the government and enterprises with valuable and different levels of knowledge, and provide strong technical support for social development. The existing data methods mainly include classification method, association analysis method, cluster analysis method and anomaly detection method, but these methods have some shortcomings. Aiming at the problems of big randomness, sensitivity to parameters and slow convergence speed of existing ant colony clustering algorithm, this paper proposes a data mining model based on improved ant colony algorithm, which will be applied to cluster analysis of experimental data sets. The experimental results show that the improved algorithm has higher accuracy and faster convergence speed, and can effectively realize data information mining.

1. Introduction

With the rapid development of Internet technology, the speed of information growth is faster and faster, which makes a lot of information hidden behind the massive data. Then, how to obtain useful information from these hidden data has become a hot issue at present. In order to solve such problems, data mining technology came into being. The knowledge generated by data mining can be applied in many fields, such as decision support, biopharmaceutical, information management. Obviously, data mining technology has more and more extensive and good application prospects. Ant colony algorithm is mainly used to solve combinatorial optimization problems[1-2]. Through the foraging principle of ants in nature, it can find the shortest or optimal combination path, which can reduce resource consumption and improve efficiency [3-4].

At present, ant colony algorithm has been applied in many industry problems, such as scheduling problem, vehicle routing problem, assignment problem, setup problem, system identification, biological protein folding and other fields, and has achieved good results. This paper uses the improved ant colony algorithm to build a data mining model, using the advantages of the improved algorithm to improve the efficiency and quality of data mining, has good practical significance[5].
2. Ant Colony Clustering Algorithm (ACCA)

2.1. Basic ant colony clustering model

The core idea of the basic ant colony clustering model is that when an ant without load meets an object, it will judge the number of objects of the same kind around it. The fewer objects of the same kind, the more different the object and other objects belong to, the more easily the ant can pick up the object; On the contrary, the load ant can not always carry the object, it will move arbitrarily and try to put down the object. In the process of moving, it will constantly judge the similarity between the surrounding objects and the objects loaded by ants. The more the same objects, the more similar the object is to the surrounding objects, the greater the possibility of the ant to put down the object [6-7].

The probability that the object will be picked up is

\[ P_p = \left( \frac{c_1}{s + c_1} \right)^2 \]  

(1)

The probability of the object being dropped is

\[ P_d = \left( \frac{s_2}{s + c_2} \right)^2 \]  

(2)

Where \( s \) is the number of similar objects around ants, \( c_1 \) and \( c_2 \) are threshold constants. When \( s < c_1 \), the probability of the object being picked up is close to 1, and the probability of the ant choosing to pick up the object will be very large; on the contrary, when \( s \geq c_1 \), the probability of the object being picked up will be close to 0, and the probability of the ant choosing to pick up the object will be very small. The probability of ants dropping is opposite to that of picking them up. When \( s < c_2 \), the probability of the object being put down is close to 1, which means that when the ant moves to the position, there are a lot of data objects and similar objects around it, which can be considered as belonging to the same category. Therefore, the probability of the object being put down is very high. When \( s \geq c_2 \), the probability of the object being put down is close to 0, and the probability of the ant's load object being put down is very small.

2.2. Standard ant colony clustering algorithm (SACCA)

The basic ant colony clustering model is extended and applied to the field of data analysis. The famous standard ant colony clustering algorithm is proposed by using real value elements to cluster data vectors. The main idea of the algorithm is: firstly, the data to be processed is randomly projected into a two-dimensional grid of \( Z \times Z \). At the same time, \( N \) ants randomly move on the two-dimensional grid, constantly picking up or dropping data objects. These ants have a certain observation range and can only observe the data objects in the surrounding \( S \times S \) area. Each ant randomly selects a data object and calculates the similarity between the data object and the observed data object. If the similarity is low, it means that the object is not similar to the same category. Therefore, the probability of the object being put down is very high. When \( s \geq c_2 \), the probability of the object being put down is close to 0, and the probability of the ant's load object being put down is very small.
If an ant encounters the data object \( o_i \) at position \( L \), the local similarity density between \( O_i \) and the surrounding objects is calculated as follows:

\[
D(O_i) = \begin{cases} 
\frac{1}{n} \sum_{O_j \in A_{s>L}} \left[ 1 - \frac{\text{dist}(O_i, O_j)}{\alpha} \right] & \text{if } D(O_i) > 0 \\
0 & \text{else}
\end{cases}
\] (3)

Among them, \( A_{s>L}(L) \) is the range that the ant can observe at position \( L \), that is, the neighborhood area, is a circular area centered on \( L \). \( \text{dist}(O_i, O_j) \) represents the distance between two data objects \( O_i \) and \( O_j \), which is generally Euclidean distance; \( n \) is all the data objects that the ant can observe at position \( L \). \( \alpha \) is the dissimilarity factor and the group similarity coefficient. Its value has an important influence on the number of clusters, the results of clustering and the convergence speed of the algorithm.

Ants usually pick up or put down a data object according to the probability conversion function, which is a function with local similarity density as a parameter, which can be divided into pick-up probability \( P_p \) and drop probability \( P_d \), with the value range of \([0,1]\)

In each cycle, the empty ant randomly selects a data object and calculates the pick-up probability \( P_p \)

\[
P_p(O_i) = \left( \frac{c_i}{s(O_i) + c_i} \right)^2
\] (4)

A random number \( R \) is generated. If the value of \( P_p \) is greater than \( R \), the ant picks up the data object. Otherwise, the ant continues to move until it encounters the next unpicked data object. If the ant has no load, the probability \( P_d \) of the ant is calculated

\[
P_d(O_i) = \begin{cases} 
2s(O_i) & \text{if } s(O_i) < c_2 \\
1 & \text{else}
\end{cases}
\] (5)

It also generates a random number \( R \). if the value of the drop probability \( P_d \) is greater than \( R \), the ant drops the data object. Otherwise, the ant continues to carry the data object and move to the next position that can be dropped.

### 2.3. Shortcomings of standard ant colony clustering algorithm

Through the analysis of the ant colony clustering algorithm, it is found that although the ant colony clustering algorithm has many advantages, such as no need to set the cluster center and the number of clusters in advance, has good scalability, can find arbitrary shape clusters and so on. Through the research of many scholars on the SACCA, it is found that the algorithm has the following shortcomings [10]:

1. The randomness of ant movement is large, and the movement has no direction, which will lead to a large number of invalid movements of ants, which may cause ants to pick up and put down the same object, which will reduce the efficiency of the algorithm.

2. The algorithm is sensitive to the parameters, and the parameter setting is often based on the user's experience. The quality of the parameter setting will directly affect the final clustering effect, leading to the reduction of the robustness of the algorithm.

3. When dealing with a large number of data, a large number of iterations and the process of ants putting down and picking up objects are needed. The convergence speed of the algorithm is very slow.
3. Improvement of Algorithm

3.1. Improvement of moving strategies

In the traditional standard ant colony algorithm, the movement of ants is random. In the whole clustering process, some isolated points have not been picked up, or the current ants will repeatedly pick up the data points that other ants have picked up, and the same data is constantly picked up and put down. In this paper, a record matrix is introduced to record the position of the data object. If the position of the data object does not change, it indicates that it has not been picked up. When the empty ant moves, it gives priority to those data objects that have not been picked up, and uses a two-dimensional data object load state matrix to record the unique identification and load status of each data object. The algorithm needs to first determine whether the data object has been loaded. If the answer is yes, then the ant will give up the load of the data object and continue to move. If there are multiple data objects that have not been picked up, one data object is randomly selected as the moving target, and then the pick-up probability is calculated to decide whether to pick up the object and modify the load state. The criterion for judging the end of an iteration is that all data objects have been loaded. Then reinitialize the load state matrix, and then proceed to the next round of iteration until the maximum number of iterations is reached. The algorithm ends and the clustering result is obtained.

This can prevent some data object coordinates not change or can not be put down and the problem of circulation, but also conducive to picking up isolated points, can improve the efficiency of the algorithm.

3.2. Improvement of observation radius

The observation radius $s$ of traditional standard ant colony algorithm is fixed. In the existing improved algorithms, some increase the observation radius according to the number of iterations, and some reduce the observation radius according to the number of iterations. In this paper, the two methods are combined. Firstly, the observation radius is gradually reduced. After reaching a certain number of iterations, the observation radius is gradually increased.

The advantages of this method are: (1) in the first half of the algorithm, the observation range of ants is changed from large to small, and the clustering results are gradually accurate. (2) In the second half of the algorithm, the observation range of ants is changed from small to large, so as to accelerate the convergence speed.

3.3. Improvement of probability transformation function

In the definition of similar density, the Euclidean distance or cosine distance is generally used for two data objects in the grid. In this paper, the linear combination of two distances is used. Not only the Euclidean distance is used to calculate the distance between two objects, but also the cosine function is used to calculate whether the trend of two objects is the same. The linear combination of these two distances can complement each other well.

$$d(i,j) = \sqrt{\sum_{j=1}^{m} \left[ \zeta (x_i - x_j) \right]^2} + (1 - \zeta) \cdot (1 - \sin(x_i - x_j))$$  \hspace{1cm} (6)

3.4. Process of improvement

Step 1: initialize the values of each parameter.
Step 2: randomly distribute the virtual ants and data objects to two-dimensional plane.
Step 3: the ant moves to the position $r$ of any object to be clustered.
Step 4: the empty ants arrive at the random position, and calculate the similarity within the observation radius. Random number $R$ generated in $[0,1]$ obey uniform distribution. If the $P_r > R$ and the load state of $O_i$ is $N$, the data object is picked up and moved to the new position, otherwise the empty ant randomly selects the next data object.
Step 5: the load ant arrives at an idle position, calculates the similarity within the observation radius of the location, and generates a random number $R$ with uniform distribution in $[0,1]$. If $P_d > R$, put down $O_i$, or the load ant moves to a new location and tries to drop the data object.

Step 6: judge whether the observation radius and the movement speed of ants need to be changed according to the number of iterations.

Step 7: under the iterative control mechanism, the maximum number of iterations is reached, the algorithm ends and the result is output.

4. Analysis of Experimental Results

4.1. Experimental environment and parameter setting
The computer configuration used in this paper is: CPU i9-9900K, memory 64GB, network bandwidth 300m, hard disk 500GB solid state disk + 2TB mechanical hard disk The parameters of the algorithm are as follows: the maximum number of iterations is 1000, the number of ants is 2000, the area of two-dimensional grid is $200 \times 200$, the initial observation radius is 10, the transfer probability parameter is 5, the initial velocity of ants is 10, the difference factor is 0.2, the upper bound of observation radius is 10, the lower bound of observation half diameter is 4, the maximum velocity of ants is 10, and the minimum velocity of ants is 4.

4.2. Experimental result
In this paper, improved ant colony clustering algorithm (IACCA) and standard ant colony clustering algorithm (SACCA) are used to test Iris and Wine, Zoo and Image Segmentation data sets in UCI machine learning library respectively. After 100 tests on four data sets, the calculated average F-measure values are shown in Table 1.

| Data set      | SACCA   | IACCA   |
|---------------|---------|---------|
| Iris          | 0.887   | 0.926   |
| Zoo           | 0.832   | 0.901   |
| Wine          | 0.721   | 0.799   |
| Image Segmentation | 0.561   | 0.612   |

According to the data in the table above and the broken line comparison chart, IACCA has better clustering effect compared with SACCA no matter what type of data is processed. In addition, the results of ant colony clustering algorithm on discrete data are also significantly better than the results of continuous data clustering. From the perspective of algorithm principle, the improved algorithm reduces the influence of artificial parameters on clustering results, and increases the robustness of the algorithm. The observation radius of ants is reduced first and then increased, which not only makes the clustering results more accurate, but also speeds up the convergence speed of the algorithm. By using the mechanism of load state matrix, the state of ants can be adjusted dynamically to reduce the randomness of ants moving and the situation of repeatedly picking up a data object, which can speed up the clustering process and improve the efficiency of the algorithm.

5. Conclusion
This paper first expounds the necessity of data mining, and then briefly introduces the standard ant colony clustering algorithm, and describes the idea and specific process of the algorithm in detail, and
points out its shortcomings, mainly including the sensitivity of the algorithm to parameters, the randomness of ant movement and other issues, and targeted to improve the algorithm, mainly including the improvement of mobile strategy, the optimization of observation radius and then the improved algorithm is applied to the standard data set, and the experimental analysis is carried out. The results show that the improved algorithm has better clustering effect than the standard algorithm. The data mining model established in this paper has good mining performance and can realize clustering analysis of data.

References
[1] Aggarwal Y, Aggarwal P, Sihag P, et al. (2019) Estimation of Punching Shear Capacity of Concrete Slabs Using Data Mining Techniques. International Journal of Engineering Science, 32(7):908-914.
[2] Gharoun H, Keramati A, Nasiri M M, et al. (2019) An integrated approach for aircraft turbofan engine fault detection based on data mining techniques. Expert Systems, 36(2):1-18.
[3] Faradonbeh R S, Taheri A. (2019) Long-term prediction of rockburst hazard in deep underground openings using three robust data mining techniques. Engineering with Computers, 35(2):659-675.
[4] Marozzo F, Talia D, Trunfio P. (2018) A Workflow Management System for Scalable Data Mining on Clouds. IEEE Transactions on Services Computing, 11(3):480-492.
[5] Ghimatgar H, Kazemi K, Helfroush M S, et al. (2019) An improved feature selection algorithm based on graph clustering and ant colony optimization. Knowledge-Based Systems, 159(NOV.1):270-285.
[6] Wu L, Tian X, Wang H, et al. (2019) Improved ant colony optimization algorithm and its application to solve pipe routing design. Assembly Automation, 39(1):45-57.
[7] Ababou M, Bellafikh M, Kouch R E. (2018) Energy Efficient Routing Protocol for Delay Tolerant Network Based on Fuzzy Logic and Ant Colony. International Journal of Intelligent Systems and Applications, 10(1):69-77.
[8] Sitarz P, Powałka B. (2018) Dual ant colony operational modal analysis parameter estimation method. Mechanical Systems & Signal Processing, 98(jan.1):231-267.
[9] Mokhtari N A, Ghezavati V. (2018) Integration of efficient multi-objective ant-colony and a heuristic method to solve a novel multi-objective mixed load school bus routing model. Applied Soft Computing, 2018, 68:92-109.
[10] Rahman M, Ong Z C, Chong W T, et al. (2019) Wind Turbine Tower Modeling and Vibration Control Under Different Types of Loads Using Ant Colony Optimized PID Controller. Arabian Journal for Science & Engineering, 44(2):707-720.