ICIC Express Letters
Volume **11**, Number **2**, February **2017**

 ISSN 1881-803X
pp. **503–508**

HYBRID CLUSTERING BASED ON ANT COLONY OPTIMIZATION AND AGGLOMERATIVE HIERARCHICAL CLUSTERING

Yongqing Wei[1], Lin Lv[2,4], Min Ren[3,4] and Xiao Pan[2,4]

[1]Basic Education Department
Shandong Police College
No. 54, Wenhua East Road, Lixia District, Jinan 250000, P. R. China

[2]School of Information Science and Engineering
Shandong Normal University
No. 88, Wenhua East Road, Jinan 250014, P. R. China
1069315857@qq.com

[3]School of Mathematics and Quantity Economy
Shandong University of Finance and Economics
No. 40, Shungeng Road, Central District, Jinan 250014, P. R. China

[4]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology
Jinan 250014, P. R. China

Received July 2016; accepted October 2016

Abstract. *In order to obtain the global optimal agglomerative hierarchy clustering result, a hybrid clustering algorithm based on ant colony optimization and agglomerative hierarchy clustering is proposed. The algorithm used the state transition rule and pheromone update rule of ant colony optimization algorithm to optimize the agglomerative hierarchical clustering. In this method, the state transition rule was used to determine the next merged points in the hierarchical clustering, and the pheromone update rule was used to find the optimal clustering path. In this way, we can receive high quality hierarchical clustering results. It made experiments on the artificial data set and UCI data set. The results show that the clustering performance of the proposed method is better than that of the traditional clustering algorithm, which improves the accuracy of clustering.*
Keywords: Ant colony optimization, Agglomerative hierarchy clustering, State transition rule, Pheromone update rule

1. **Introduction.** Data clustering is one of the common and effective methods in data mining. It is an important means and method of data partition or grouping. Clustering analysis has been widely used in the fields of statistics, pattern recognition, image processing, marketing and so on [1].

So far, the clustering algorithms include hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering algorithm [2]. In the agglomerative hierarchical clustering algorithm, the Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) algorithm and Clustering Using Representatives (CURE) algorithms are two typical algorithms. The BIRCH algorithm proposed by Zhang et al. in 1996 realized the hierarchical clustering of a large number of data for the first time [3]. The algorithm used the agglomerative hierarchical clustering algorithm to cluster the data objects into each cluster, and then selected another clustering method to cluster each cluster. Its accuracy is lower than other methods. Guha et al. proposed CURE algorithm which is based on representative points [4,5]. The algorithm used the representative points of the cluster to represent the data distribution, and reduce the influence of the noise for the clustering results by shrinkage factor. The accuracy of the algorithm is improved. However, because of the shrinkage factor, the ability of the algorithm to find the data of any shape is reduced. For this problem, [6] proposed an Approximate Binary Hierarchical

Clustering Using Representatives (ABHCURE) algorithm. This algorithm is combined with single-layer multi-clusters merge mode, pseudo-noise mechanism and dynamic minimum number of clusters on the CURE algorithm. It improved the accuracy of clustering data which also has different shapes and noises.

From the above, we can find that if the hierarchical clustering does not make good decision when splitting or merging clusters, it may get low quality clustering results. From the above description and research, this paper considers it as a combinatorial optimization problem.

2. **Ant Colony Optimization Algorithm.** Ant colony optimization algorithm [7,8] is a kind of efficient heuristic global search technique for solving combinatorial optimization problems.

When ants move, they can release a certain concentration of pheromone on the path [9]. In order to understand the basic principle of ant colony optimization algorithm, the following explains the model of ant colony optimization algorithm and its implementation process using TSP problem as an example [10].

(1) When all ants have visited n city, the ants update their information on their paths. The update formula is as follows:

$$\tau_{ij}(t+n) = \rho.\tau_{ij}(t) + \Delta\tau_{ij} \tag{1}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{2}$$

$$\Delta\tau_{ij}^{k} = \frac{Q}{L_k} \tag{3}$$

m represents the total number of ants, Q is a constant, and ρ is a coefficient less than 1. $\tau_{ij}(t)$ represents edge (i,j)'s concentration of pheromone at time t. $\Delta\tau_{ij}^{k}$ represents each unit length's pheromone is placed by the ant k at time t and $t+n$ on edge (i,j).

(2) Before the ants have visited all the cities, it is not allowed to visit the city that has been visited.

(3) The transition probability of ants from a certain city to the next city is:

$$p_{ij}^{k}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum [\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}} \; j \in T_k \tag{4}$$

T_k represents the cities that the ants can choose. α represents the relative importance of pheromone evaporation factors. β represents the relative importance of heuristic information.

3. **Agglomerative Hierarchical Clustering Based on Ant Colony Optimization.**

Definition 3.1. ***Distance standard.*** *The distance between the two data points is calculated by the Euclidean distance.*

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2} \tag{5}$$

$$x_i = (x_{i1}, \ldots, x_{im}) \tag{6}$$

$$x_j = (x_{j1}, \ldots, x_{jm}) \tag{7}$$

x_i and x_j are two m dimensional data points.

The distance of the two clusters uses the minimum distance to measure similarity. It is the most common measure method of cohesive hierarchical clustering algorithm.

$$d_{\min}(c_i, c_j) = \min \|p_i - p_j\| \tag{8}$$

$p_i \in c_i$, $p_j \in c_j$, p_i, p_j are two data points, c_i, c_j are two clusters.

Definition 3.2. ***Object function.*** *It is set as the square sum of the clustering error [11] (Suppose there are C cluster centers after the clustering is completed):*

$$E = \sum_{j=1}^{c} \sum_{x_i \in c_j} |x_i - c_j|^2 \tag{9}$$

$$C_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i \tag{10}$$

C_j represents a centroid. m_j represents the number of data contained in a cluster. Getting the minimum object function means getting the optimal hierarchical clustering.

3.1. **Hybrid clustering based on ant colony optimization and agglomerative hierarchical clustering.** The aim of the agglomerative hierarchical clustering based on ant colony optimization algorithm is to build the shortest clustering path after traveling all the data. Its main method is to refer to the foraging behavior of ants [12,13].

The basic flow chart of the algorithm is shown in Figure 1.

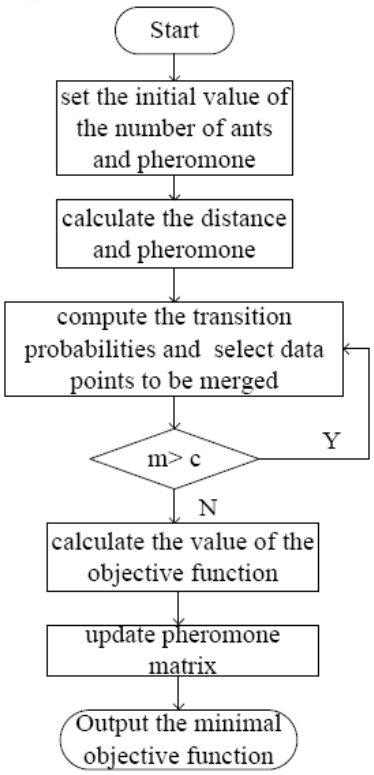

FIGURE 1. Flow chart of the algorithm

Its basic idea is as follows.

Step 1: Initialize the algorithm's parameters. It contains the number of ants, weight parameter α, β, evaporation factor ρ, and so on.

Step 2: For each ant m, each instance is used as a cluster in the clustering process. When the ant m is in the data point x_i, we calculate its distance from other clusters. And then it uses the pheromone τ_{ij} to calculate the probability of the next point to be merged (merge probability).

$$p_{ij}^{m} = \frac{(\tau_{ij})^{\alpha} (\eta_{ij})^{\beta}}{\sum_{l \in N_i^m} (\tau_{il})^{\alpha} (\eta_{il})^{\beta}} \; j \in N_i^m \tag{11}$$

$$\eta_{ij} = \frac{1}{d_{ij}} \tag{12}$$

d_{ij} represents the distance between the two data points or two clusters. η_{ij} represents heuristic information based on distance. α and β are two weight parameters. They determine the relative influence of pheromone and heuristic information. If $p_{ij}^m \geq p_0$ (p_0 is obtained by experiments), x_j and x_i are merged. If not, they are not merged.

Step 3: Determine whether the number of ants k reached the total number of M; if not, $k = k + 1$. Go to Step 2.

Step 4: All the ants continue to cluster the object, and the clustering center and pheromone are updated.

Clustering center:

$$C_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i \tag{13}$$

m_j is the total number of the data points which belongs to the c_j cluster.

Pheromone updating: the pheromone will have different strength if the path length of the ant is different. The pheromone which the ant released has cumulative effect in the path. It is assumed that the initial pheromone on each route is equal. When the clustering begins, the pheromone is updated. The first is the pheromone evaporation:

$$\tau_{ij} = (1 - \rho)\tau_{ij} \tag{14}$$

ρ is the evaporation rate of the pheromone, $0< \rho \leq 1$. Its effects are to avoid unlimited information accumulation, and make the algorithm "forget" the poor path which is selected before. After this, all ants release the pheromone:

$$\tau_{ij} = \tau_{ij} + \sum_{k=1}^{n} \Delta\tau_{ij}^m \tag{15}$$

$\Delta\tau_{ij}^m$ represents the information of the ant m's path which is from data x_i to the cluster c_j.

$$\Delta\tau_{ij}^m = \frac{1}{d\left(x_i, c_j\right)} \tag{16}$$

Step 5: Calculate the minimum objective function.

Step 6: Then it outputs the optimal solution.

4. Experiment and Result Analysis.

4.1. Experimental data preprocessing.

The data sets are standardized and normalized.

Standardized process (There is mean standardization)

$$x_i^* = \frac{x_i - \overline{x_i}}{s_i} \tag{17}$$

$\overline{x_i}$ represents mean data. s_i represents the standard deviation of the data.

4.2. Artificial data set.

D1, D2 are two-dimensional data point sets which are two artificial structures. They all have two types. D1 contains 150 data. D2 contains 100 data. The distribution of D1 is relatively scattered. The distance between the two clusters is relatively close. However, the distribution of D2 is relatively concentrated, and the distance between the two clusters is relatively far.

Table 1 is the initial parameters of the algorithm which are used to do experiments on artificial data sets. The parameter value of the algorithm in this paper is arranged as follows.

Table 1. Experimental parameters setting

parameter	parameter value
the number of ants	$m = 6$
pheromone volatile factor	$\rho = 0.7$
weight parameter	$\alpha = \beta = 0.6$
merge probability	$p_0 = 0.7$

The algorithm of this paper was compared with traditional agglomerative hierarchical clustering algorithm and CURE algorithm. CURE algorithm is typical in the hierarchical clustering. They mainly compared the accuracy of all the algorithms. Accuracy results of the three algorithms are given in Table 2. Among them, the algorithm proposed in this paper has the highest accuracy rate. The accuracy rate of the traditional agglomerative hierarchical clustering algorithm is the lowest. And the accuracy of all the algorithms in the data set D2 is higher than that in D1.

Table 2. Comparison of accuracy of algorithms

data set	sample number	agglomerative hierarchical clustering	CURE	algorithm in this paper
D1	150	75.3%	80.4%	84.5%
D2	100	78.9%	81.5%	85.8%

4.3. **UCI dataset.** This experiment uses Iris data set, wine data set, and thyroid data set.

Because these three data sets are different from the artificial data sets, we need to reset the initial parameters. Parameter settings are shown in Table 3.

Table 3. Experimental parameters setting

parameter	parameter value
the number of ants	$m = 7$
pheromone volatile factor	$\rho = 0.6$
weight parameter	$\alpha = \beta = 0.5$
merge probability	$p_0 = 0.6$

Input the Iris data set, wine data set, and thyroid data set. The results are shown in Table 4.

Table 4. Experimental results

data set	agglomerative hierarchical clustering	CURE	algorithm in this paper
Iris	76.2%	81.2%	84.1%
wine	41.3%	57.4%	61.1%
thyroid	76.6%	79.9%	84.9%

From Table 4, we can see that compared with those two algorithms, the algorithm of this paper has significantly improved in accuracy. Furthermore, we can see that whether it is the algorithm of this paper or traditional agglomerative hierarchical clustering and CURE algorithm, the accuracy of the experiment on the Iris data set and thyroid data set is higher than that of the corresponding experiment on the wine data set.

5. **Conclusions.** In this paper, the agglomerative hierarchical clustering based on ant colony optimization algorithm improved the accuracy of the agglomerative hierarchical clustering by introducing ant colony optimization, and we can determine more accurate merge points. In future research and practice, we will combine the other aspects of intelligent optimization algorithm to develop hierarchical clustering, so that it can be more efficient to achieve hierarchical clustering.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China (61373148, 61502151), Shandong Province Natural Science Foundation (ZR2012FM038, ZR2014FL010), Shandong Province Outstanding Young Scientist Award Fund (BS2013DX033), Science Foundation of Ministry of Education of China (14YJC860042) and Project of Shandong Province Higher Educational Science and Technology Program (No. J13LN19, No. J15LN02).

REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles et al., Image inpainting, *Proc. of ACM SIGGRAPH Conference on Computer Graphics*, New York, pp.417-424, 2000.
[2] R. Xu and D. Wunsch, Survey of clustering algorithms, *IEEE Trans. Neural Networks*, vol.16, no.3, pp.645-678, 2005.
[3] T. Zhang, R. Ramakrishnan and M. Livny, BIRCH: An efficient data clustering method for very large databases, *ACM SIGMOD International Conference on Management of Data*, vol.25, no.2, pp.103-144, 1996.
[4] S. Guha, R. Rastogi and K. Shim, An efficient clustering algorithm for large databases, *Information Systems*, vol.26, no.1, pp.35-58, 2001.
[5] S. Guha, R. Rastogi and K. Shim, A robust clustering algorithm for categorical attributes, *Information Systems*, vol.25, no.5, pp.345-366, 2000.
[6] Y.-T. Wang, J.-D. Wang and H.-Y. Chen, An algorithm for approximate binary hierarchical clustering using representatives, *Journal of Chinese Computer Systems*, vol.36, no.2, pp.215-218, 2015.
[7] M. Dorigo, C. G. Di and L. M. Gambardella, Ant algorithms for discrete optimization, *Artificial Life*, vol.5, no.2, pp.137-172, 1999.
[8] J. Xiao and L. Li, A hybrid ant colony optimization for continuous domains, *Expert Systems with Applications*, vol.38, no.7, pp.11072-11077, 2011.
[9] H. Azzag, G. Venturini, A. Oliver et al., A hierarchical ant based clustering algorithm and its use in three real-world applications, *European Journal of Operational Research*, vol.179, no.3, pp.906-922, 2007.
[10] Y. Zhou and Z. Huang, Artificial glowworm swarm optimization algorithm for TSP, *Kongzhi Yu Juece/Control & Decision*, vol.27, no.27, pp.1816-1821, 2012.
[11] J. Xie and Y. Wang, K-means algorithm based on minimum deviation initialized clustering centers, *Computer Engineering*, vol.40, no.8, pp.205-210, 2014.
[12] B. Biswal, P. K. Dash and S. Mishra, A hybrid ant colony optimization technique for power signal pattern classification, *Expert Systems with Application*, vol.38, no.5, pp.6368-6375, 2011.
[13] C. Ma, A. Cao and Y. Zhou, Primary research on improved algorithm of ant colony clustering combination, *Journal of Shenyang Jianzhu University (Natural Science)*, vol.27, no.4, pp.798-803, 2011.