An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence

Article · June 2009

DOI: 10.4304/jsw.4.4.299-306 · Source: DOAJ

Weihui Dai
Fudan University

An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence

Weihui Dai
School of Management, Fudan University, Shanghai 200433, P.R.China
Email: whdai@fudan.edu.cn

Shouji Liu and Shuyi Liang
School of Management, Fudan University, Shanghai 200433, P.R.China
Email: {052025052, 062025047}@fudan.edu.cn

***Abstract*—This paper proposes an improved ant colony optimization cluster algorithm based on a classics algorithm - LF algorithm. By the introduction of a new formula and the probability of similarity metric conversion function, as well as the new formula of distance, this algorithm can deal with the category data easily. It also introduces a new adjustment process, which adjusts the cluster generated by the carry process iteratively. We approve that that the algorithm can improve the efficiency and the convergence of the cluster theoretically. Data experiments show that the improved ant colony algorithm can form more accurate and stability clusters than the K-Modes algorithm, Information Entropy-Based Cluster Algorithm, and LF Algorithm. Scalability experiments show that the running time has an obvious linear relationship with the size of data set. Furthermore, we describe the process and idea of the algorithm usage by a mobile customer classification case and analyze the cluster results. This algorithm can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time.**

***Index Terms*—swarm intelligence, cluster analysis, optimized ant colony algorithm, data mining, category data**

I. INTRODUCTION

Swarm intelligence comes of the scientists' research and observation on the social insect. The so-called swarm intelligence is that a great many of simple, unintelligent units unite into a group and express intellectual behaviors through mutual cooperating with each other. Swarm intelligence exhibits a number of interesting properties such as flexibility, robustness, decentralization and self-organization. It is widely used in portfolio optimization problem, knowledge discovery, communication networks, data-mining and etc [1]-[4]. Cluster analysis plays an important role in the application of data mining. It plots out different sorts according to the size of similitude degree, that is, when the similitude degree is high we consider them the same sort.

Manuscript received Feb 19, 2009; revised March 1, 2009; accepted March 20, 2009. This research was supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

Cluster analysis is widely applied to various fields [5], such as marketing [6], pattern recognition [7], space data analysis, gene classification [8], and so on. In the recent years, some scholars study cluster problem according to the idea of swarm intelligence [9]. The inspiration of ant colony cluster comes of the accumulation of ant bodies and classification of ant larvae. The classic ant colony cluster algorithm takes use of the characteristics of positive feedback of the ant colony. Such algorithm is robust, good convergence, and parallel. However, it is also with the disadvantage of long time convergence, easy stagnation and local optimization.

This paper proposes an improved ant colony cluster algorithm based on the basic model of Denueubourg and LF algorithm. It enables the ants to consult historical information when conveying objects by importing adjusting process and short period memory, and it also does iterative regulating to the cluster formed by the ants. Thus, it advances the convergence speed of the algorithm and the efficiency of the cluster.

The second part of the paper introduces the mathematic model of the algorithm, describes the concept definition, similitude degree and the probability conversion formula, and discusses the characteristic of the optimized solution and the convergence character. The third part is the experimentation, contrast optimized ant colony algorithm with K-modes cluster algorithm and LF algorithm based on current data collect, and discusses the experiment conclusion. The fourth part introduces a practice case, which provides a blue print to solve the problem. The fifth part summarizes the paper, and gives the perspective of future research.

II. THE OPTIMIZED ANT COLONY ALGORITHM BASED ON SWARM INTELLIGENCE

A. Mathematics Model

1) Definitions:

a) Attribute Probability: Suppose data collect D contains n objects, $A_1, A_2, \ldots, A_m$ are the m attributes in D. If attribute A_j of the objects $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ is x_{ij} ,

x_{ij} appears q_j times in D, thus $q_j = sum\left\{x_{kj} = x_{ij} \mid X_k \in D\right\}$, then the attribute probability p_{ij} of object X_i on attribute A_j is:

$$p_{ij} = \frac{q_j}{n} \tag{1}$$

b) Similitude Degree: Suppose D contains n objects, the similitude degree of X_i indicates the algorithm average value of all attribute probability of each attribute, we define $f(X_i)$ as the similitude degree of X_i:

$$f(X_i) = \frac{1}{m}\sum_{j=1}^{m} p_{ij} \tag{2}$$

c) Probability Conversion Function: It is the function that switches the similitude degree to the cluster object probability of the unit. The independent variable is the similitude degree, and the function field is [0, 1]. The larger the similitude degree is, the smaller the pick-up conversion probability is, while the put-down conversion probability follows the contrary rule. In this algorithm, we define probability conversion function as follows:

$$p_p = 1 - \frac{1-e^{-cf(X_i)}}{1+e^{-cf(X_i)}} \tag{3}$$

$$p_d = \frac{1-e^{-cf(X_i)}}{1+e^{-cf(X_i)}} \tag{4}$$

Here, p_p is the probability pick-up function, p_d is the probability put-down function. p_d is an up-convex function, and the convergence speed varies from different c. The larger c is, the nearer the function value is to 1, that is, the corresponding put-down function value is large. So we can expedite the convergence speed by augmenting c. In this paper, let c=5.

2) The characteristic of the optimized solution:

The algorithm uses probability function as the basis of pick-up or put-down. Thus, the algorithm aims to increase the similitude degree of the whole system, maximizing the system's average similitude degree. We define the objective function in this algorithm as follows:

$$F(D) = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} f(X_i) \tag{5}$$

$$0 \le f(X_i) \le 1$$

Here, $\sum_{i=1}^{k} n_l = n$

$f(X_i)$ is the similitude degree of X_i in cluster k.

This is a complicated optimization problem. It aims to find out clustering fashion to maximize the average similitude degree when giving certain cluster.

According to (1) and (2), (5) can be rewrite as follows:

$$\begin{aligned} F(D) &= \frac{1}{n}\sum_{l=1}^{k}\sum_{j=1}^{n_l}\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{A_i}\frac{q_{it}}{n_l} \\ &= \frac{1}{nm}\sum_{l=1}^{k}\frac{1}{n_l}\sum_{i=1}^{m}\sum_{t=1}^{A_i} q_{it}^2 \end{aligned} \tag{6}$$

In (6), q_{it} is the times of A_t which is the i[th] attribute in cluster l. Thus, $\sum_{t=1}^{A_i} q_{it} = n_l$ .That is, the sum of the times of the attribute's value is total objects in the cluster. So we get:

$$F(D_l) = \frac{1}{mn_l}\sum_{i=1}^{m}\sum_{t=1}^{A_i} q_{it}^2 \tag{7}$$

If the attributes are independent, then we can calculate them independently:

$$F(D_{l-A_i}) = \frac{1}{n_l}\sum_{t=1}^{A_i} q_t^2 \tag{8}$$

As discussed previously $\sum_{t=1}^{A_i} q_{it} = n_l$, we can get:

$$F(D_{l-A_i}) = \frac{1}{n_l}\sum_{t=1}^{A_i} q_t^2 = n_l\sum_{t=1}^{A_i}\frac{q_t^2}{n_l^2} \le n_l\left(\sum_{t=1}^{A_i}\frac{q_t}{n_l}\right)^2 \le n_l \tag{9}$$

Only when $q_s = n_l, q_r = 0 (r \ne s)$, the equal mark comes into existence. That is, only when all the objects in the cluster get the same value, the similitude degree gets the maximum. Furthermore, the more concentrative the value of each attribute is (the attribute of the objects is similar), the larger the average similitude degree is.

In (5), the more similar the objects in the cluster are, the larger the average similitude degree is. Contrarily, the larger the similitude degree of the sub-cluster data collect is, the larger the whole similitude degree is.

So, the characteristic of the optimized solution is that all the attributes are most concentrative. As each value of the objects' attribute is random, the optimized solution may be more than one.

3) The algorithm's convergence:

To data collect D, when we divide the larger cluster into smaller ones, the similitude degree will get smaller. In the beginning phase of the ant cluster, we consider the whole object space as a whole data collect. As the mathematic model discussed above, in order to maximize the average similitude degree, we must divide the larger data collect into smaller ones. So the whole object space will be divided in to some small clusters gradually. Thus we know the cluster algorithm can divide the objects effectively.

When the data collect was divided to a certain degree, if the remained division keeps the attribute proportion constant, then the data collect will not be divided again. This characteristic ensures the data collect will not be divided to a cluster which only contains one object.

Suppose there are two clusters: D_1 and D_2. D_1 contains the object X_i. The similitude degree of X_i in D_1 is $f(X_i^{(1)})$, and the similitude degree in D_2 is $f(X_i^{(2)})$. If we convert X_i from D_1 to D_2, according to the definition of the probability conversion function, then the mathematic expectation of the change to similitude degree is:

According to (1) and (2), (5) can be rewrite as follows:

$$\Delta(X_i) = p_p(f(X_i^{(1)})) \times p_d(f(X_i^{(2)})) \times (f(X_i^{(2)}) - f(X_i^{(1)})$$

$$p_p \geq 0, p_d \geq 0 \tag{10}$$

So the direction of the change to similitude degree is decided by the similitude degree of the objects in two clusters. When $f(X_i^{(1)}) < f(X_i^{(2)})$, $\Delta(X_i) > 0$. We derive the probability put-down function p_d:

$$p_d{}'(x) = \frac{2ce^{-cx}}{\left(1+e^{-cx}\right)^2} \tag{11}$$

So the probability put-down function p_d is an increasing function. p_d increases with the increasing of $f(X_i^{(2)})$. And we can also know the probability pick-up function p_p is a decreasing function. p_p increases with the decreasing of $f(X_i^{(1)})$. We get the conclusion: The larger the change to the similitude degree of the object X_i in D_1 and D_2, the more probably it will be converted from D_1 to D_2, the whole average similitude degree will get larger.

For our algorithm, similitude degree is the estimation basis of whether pick-up or put-down for the ant conveying. Thus, this estimation basis ensures the increasing of the average similitude degree during the ant conveying process. The maximum of the average similitude degree is 1, so the average similitude degree has an upper bound.

So we get the conclusion: The average similitude degree will converge to a certain value, so the algorithm has a convergence trait.

B. The Algorithm Process

The main process of the ant cluster algorithm is the ant conveying process. Ant decides whether pick-up the current object by object's probability conversion function. Similarly, when ant conveys the object to the destination, it also considers the similitude degree between the current object and the surrounding objects to decide whether to put-down or not. In this process, the ant doesn't know the other ants' location distributing and load status, neither the other objects' distributing status outside its observing scope. So we can say the ant conveying process is an easy and absolute individual behavior. Yet, it is this easy individual behavior makes the objects divided into various clusters during long-time and concurrent process.

Besides the similitude degree and the probability conversion function, the main factor that influences the ant conveying process is the ant's observing radius. The smaller the observing radius is, the more efficient the cluster is [17]. Yet, if the observing radius is too small, it will cause many isolated points which influence the combination of the clusters, and cause the inefficiency of the cluster capability. If the observing radius is large, though the cluster result is coarse, it expedites the algorithm's convergence speed in return. Fig. 1 is the cluster effect figure with two different observing radiuses.

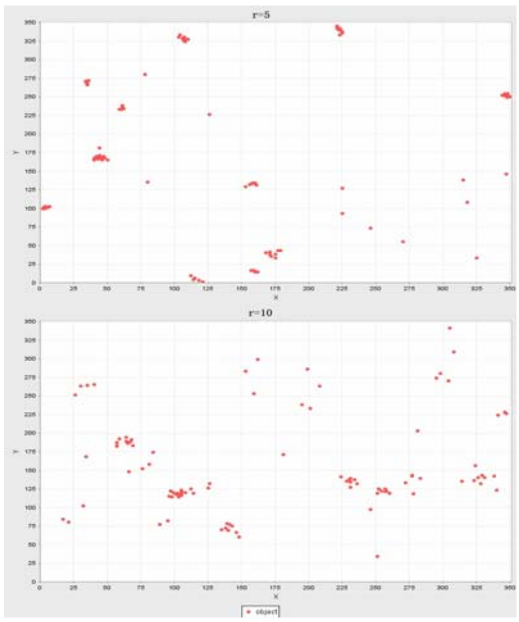

Figure 1. Cluster effect figure with different observing radius

The observing radius in the nether figure is twice of the upper figure. We can see that the cluster of the upper figure is close, but it has 11 isolated points, and it only has 3 in the nether figure in contrast. In order to improve the cluster process, our algorithm betakes gradual-changing observing radius. In the beginning phase, we use larger observing radius to expedite the conveying process; in the anaphase, we use smaller ones to deal with the formative clusters more accurately. The change of the observing radius is linear, which is propitious to adjust the accuracy and convergence speed of the cluster.

Through this kind of examination checking up, the inducting fashion makes the conveying process more efficiently, and avoids the inefficiency caused by the iterative pick-up and put-down behaviors. The adjusting of the observing radius and the ant's "memory" function is one of the advantages that differ from traditional ant cluster algorithm, also is the magnitude ameliorating aspect.

Compared with traditional ant cluster algorithm, the improvement of the optimized ant algorithm in this paper is:

- The optimized algorithm has an adjusting process which improves the efficiency of the algorithm, and avoids the local optimality and stagnancy as well.
- Dynamic adjusting of observing radius.
- Adopt a new similitude degree formula.
- Short-term memory. Endow the ant with a short-term memory to reduce the repeated behavior of the ant.

The subsequent experiment proves these improvements in our algorithm are excellent in both accuracy and efficiency.

C. The Algorithm Pseudocode

For every item O_i do
    Place O_i randomly on grid
End For
For all ants do
    Place ant at randomly selected site
End For
/* main loop */
For t=1 to $t_{\max}$ do
    For all ants do
      If ((ant unladed) and (site S_j occupied by item O_i))then
        Compute the similarity $f(O_i)$ of O_i in R×R area
        Calculate the p_p and generate a random number Q
        If p_p > Q then /*pick-up rule*/
          Pick up item O_i
          Remember the $f(O_i)$ and current position
          Move the ant with the item O_i to a random site
        Else
          Move the empty ant to a random site
        End If
      Else If ((agent carrying item O_i) and (site empty)) then
        Compute the similarity $f(O_i)$ of O_i in this place
        Calculate the p_d and generate a random number Q
        If p_d > Q then /* put-down rule*/
          Drop item
        End If
        Move to a randomly selected neighboring site
      End If
    End For
    If ((t > $0.5t_{\max}$) and (t meet the radius change condition)) then
        Reduce the radius
        Generate the clusters iteratively and calculate the cluster center
        Unite the clusters with the same cluster center
        Relocate the items with poor similarity
    End If
End For
Print location of items /* export cluster result*/

III. Clustering Data Experimentation

A. Data Collect

In order to check up the validity of the algorithm and the veracity of the clustering result, we use several groups of data to put up the experiment. The data all come from UCI machine learning data-base [18]. This paper chooses Car, Soybean, Voting, Zoo and Nursery as the five main data collect. The former four data collect test the clustering result, data collect Nursery tests the algorithm's expansion character. We use K-modes, ECA (Entropy-based clustering algorithm), LF algorithm and OACA (Optimized ant clustering algorithm presented by this paper) to test the five data collect, and analyze the experiment result.

Data collect Car: Contains 6 classified attribute, 340 groups of data. Each group of data involves the purchase price and the technology information of a certain car.

Data collect Soybean: Contains 35 classified attribute, 160 groups of data. These data have 7 kinds which represents one kind of soybean disease, that is: Diaporthe Stem Canker, Brown Spot, Bacterial Blight, Bacterial Pustule, Purple Seed Stain, Alternarialeaf Spot, and Frog Eye Leaf Spot.

Data collect Voting: Contains the voting notes of 1984 American Congress. Each group of data involves the attitude (agree or disagree) of one councilor to 16 topics for discussion (such as beat crime, tax-free, education consume, and so on).

Data collect Zoo: Contains 16 classified attribute, 101 groups of data. Each group of data represents a certain animal. These animals have 7 sorts: mammal, bird, creeping animal, amphibian, fish, carapace, and insect.

Data collect Nursery: Contains 8 classified attribute, 12960 groups of data. These data can be divided to 5 sorts: Not_recom, Recommend, Very_recom, Priority, and Spec_prior.

The table below lists the attribute number, recoding number of every data collect:

TABLE I. DATA COLLECT LIST

Collect	Recording	Attribute	Sorts	Missing
Car	340	6	4	No
Soybean	160	35	7	No
Voting	435	16	2	Yes
Zoo	101	16	7	No
Nursery	12960	8	5	No

B. The measurement of Clustering Effect

We can consider two aspects to measure the clustering effect: validity and veracity. Validity: Whether the algorithm can find out all internal kinds in the data collect. Veracity: Whether the algorithm can put the same kind of data to same cluster, different kinds of data to different clusters. We define Clustering shrinking rate (C_r) and Veracity rate (V_r) to measure the clustering effect:

$$C_r = \frac{m_{best}}{m_{result}} \times 100\% \tag{12}$$

m_{best} is the best clustering number, and m_{result} is the actual clustering number after clustering. Clustering shrinking rate measures the degree of objects put into clusters. To the algorithm that fix the number of clusters (K-modes, ECA), the rate does no sense. But to ant clustering algorithm, as the number of cluster after clustering is unfixed, the rate can reflect the shrinking effect of the clustering.

$$V_r = \frac{m_{right}}{m_{all}} \times 100\% \tag{13}$$

m_{right} is the number of objects that putted to right cluster, and m_{all} is the total number of objects.

C. Result Analysis

According to the experiment, we can see that the validity and veracity of the OACA algorithm satisfy the request. The algorithm can divide the different data collect efficiently. Fig. 2 shows the validity of every algorithm.

We get the conclusion from Fig. 2 that OACA algorithm works most efficiently except on data collect Zoo. For this data collect, the validity of OACA is only lower than k-modes algorithm. This shows the clustering effect of OACA algorithm is excellent compared with the other three algorithms. Furthermore, the validity of the four algorithms is different to different data collect which indicates the validity is related to the data collect itself to a certain extent. If the differentiation degree of data (from different kind) in the data collect is low, this will cause low validity.

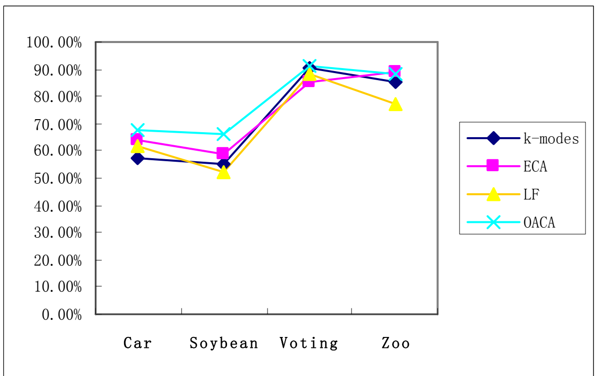

Figure 2. Validity of the Four Algorithms

As to LF algorithm and OACA algorithm, clustering shrinking rate is also a remarkable index. Fig. 3 shows this index of the two algorithms.

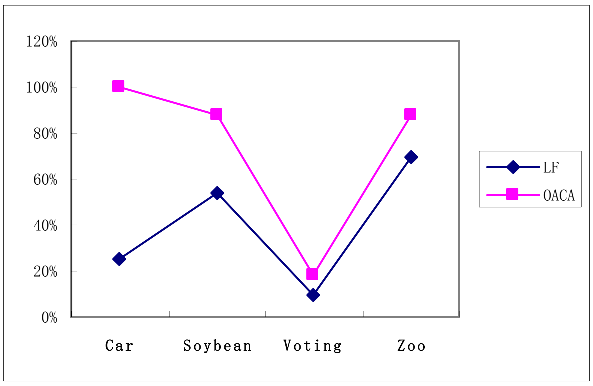

Figure 3. Clustering Shrinking Rate of the two algorithms

The clustering shrinking rate varies to different data collect of both algorithms. This shows the clustering shrinking rate is related to the data collect mostly. And to the four data collect, the shrinking rate of OACA is higher than LF algorithm, this shows OACA does better on the clustering effect than LF algorithm.

For parameter setting, the parameter number of OACA is fewer than the other algorithm. Table II lists the parameter settings of every algorithm:

TABLE II. PARAMETER SETTINGS

Algorithm	Parameter	Iterative times	Cluster number	Other Parameters
k-modes	2	Y	Y	None
ECA	3	Y	Y	Adjusting percentage
LF	6	Y	N	K_1, k_2, α, ant number
OACA	4	Y	N	R_1, R_S, ant number

The most important parameter of all the parameters is the cluster number. This parameter decides whether to appoint the cluster number beforehand. To the actual data collect, estimating the cluster number beforehand is very difficult. This parameter shows the effect and validity of the clustering directly. OACA algorithm doesn't need to appoint the cluster number beforehand. This is the advantage of this algorithm. Compared with LF algorithm, OACA needs less parameter. Furthermore, k_1, k_2, α of the LF algorithm affect the algorithm's convergence speed and clustering effect, and OACA is independent of these parameters. So it is more stable than LF algorithm, and more practical.

D. Expansibility Experiment

The circulating time of the CACA is related to iterative times, ant number and the size of the data collect. The complexity of the algorithm's circulating time is $O(mNn)$, m is the ant number, N is the iterative times, n is the size of the data collect.

In order to test the expansibility of OACA, we use data collect Nursery to carry through the experiment. We have two steps: In part one, we set the same ant number, different size of data collect. We take out the experiment data from data collect Nursery, m=200, n is from 1000 to 10000, increasing extent is 1000; In part two, we fix the size of data collect, n=1000, m is from 50 to 500, increasing extent is 50. Iterative times are 1000 in both experiment, repeat the test for 10 times, the result is its average value. See Fig. 4 and Fig. 5:

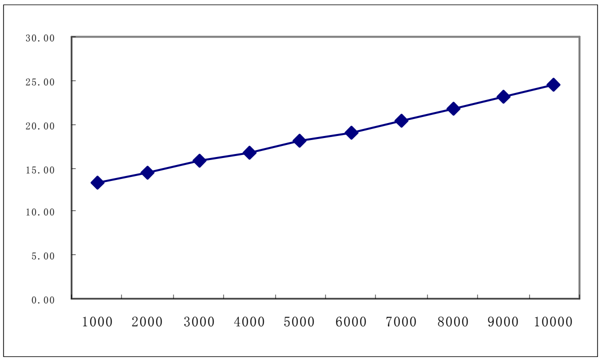

Figure 4. Circulating time varies with data collect size

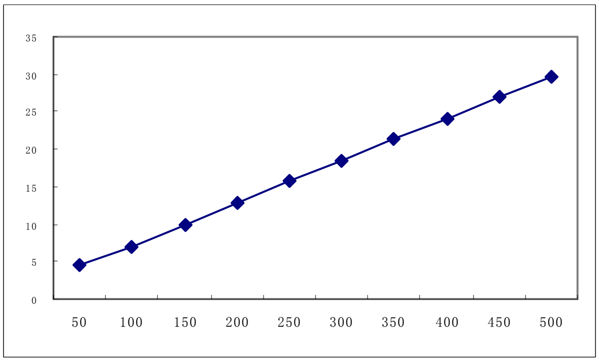

Figure 5. Circulating time varies with ant number

When fixing the ant number, the circulating time increase with the increasing of data collect size. Fixing the data collect, the increasing of ant number will also increase the circulating time. According to the curve slope in both figures, we know that the increasing of ant number makes a more strong impact on the circulating time than the data collect size. But to the same data collect, if the ant number is too small, it will cause the increasing of time to get convergence. So we must choose a proper ant number. In this paper, we set the ant number as $0.2*n$ through many experiments, n is the object number of the data collect.

We can also know from the two figures that the relation between the data collect size or the ant number and circulating time is linear, this indicates we can do clustering analysis of larger data collect. From Fig.6 we know, if we increase the ant number with the increasing of data collect size in-phase, the circulating time also increases linearly. So we get the conclusion: OACA has a good expansibility.

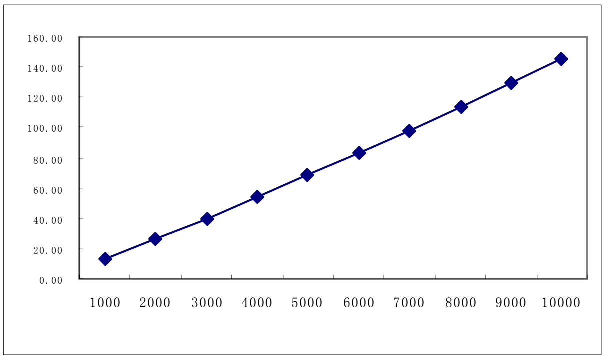

Figure 6. Circulating time varies with both data collect size and ant number

IV. THE APPLICATION OF THE ANT CLUSTERING ALGORITHM

A. Case Background and Objective

In order to improve the customer service quality, a mobile communication company investigates the client who use a certain brand of the company. The investigate data involves clients' background information, utilizing habit, and income. To do the analysis efficiently, they need to classify the client data according to the background and income. There exists a certain rate of missing value in the investigating data which caused by the unwillingness to answer the question or writing errors. So, before the clustering analysis, we should do the pretreatment to the data.

It is obviously that the client sorts are unknown before clustering analysis, and the number of client group sorts. Thus we cannot use the algorithm that fit the cluster number like k-modes algorithm and ECA, and that is the advantage of OACA. Besides, the data collect involve numerical value data and classified data, which cannot be disposed by traditional ant clustering algorithm such as LF algorithm. The clustering analysis is just an elementary classifying. More elaborate classifying should consider client behavior.

B. Data Pretreatment

The data from investigating can be sorted to three kinds: clients' discriminating data, background data, and clients' behavior data. As discussed formerly, the clients' behavior data is the basic of the subsequent research after classifying. In order to estimate the clustering results efficiently, we should distill part of them as the discriminating variable. Thus the clustering analysis use clients' background data as the clustering variable, 5 attributes in total; and clients' discriminating data as the discriminating variable, 4 attributes in total.

The investigating data have 400 observing sample, there exists missing value in these sample. We have two ways to deal with the missing value, the first way is to delete the observing sample, and the second way is to put special tags on the missing value. We use the second way in this paper: set -1 for all the missing value.

There are two types of clients' background: value data and classified data. OACA can deal with the classified data, so we should disperse the value data. The dispersing function is:

$$x = \begin{cases} \dfrac{v}{l} + 1(v \bmod l \neq 0) \\ \dfrac{v}{l}(v \bmod l = 0 \end{cases} \tag{14}$$

For classified data, some classified segment is thin, as the sample data is too small, it is disadvantage to the clustering analysis. So we should combine part of the classified segment.

C. Clustering Analysis

In this case, we use optimized clustering algorithm to do the clustering analysis. Through many times of tests, we set a 350×350 two dimension plane, the ant number is 80, iterative times is 100000, observing radius is [2,10]. We repeat ten times clustering analysis to the data collect, and adopt the best effect as the analysis result. Fig.7 shows the analysis result. We see that OACA can divide the data collect efficiently. Except fewer isolated dots, most data are assembled to 18 clusters. These 18 clusters represent the initial client sorts. Table IV shows the detailed information.

Adjacent field X and adjacent field Y indicate the distributing section of the object on horizontal coordinate and vertical coordinate. The average distance indicates the distance between the dots of every clusters and cluster center. Except these 18 clusters, there exist 3 isolated dots, which locates at (266,122), and its isolated object is (1,5,3,1,2). Fig.8 shows the distance between clusters and cluster center.

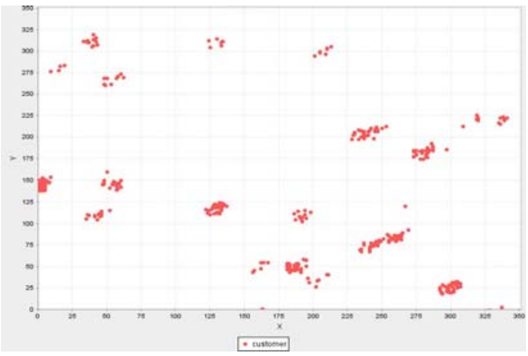

Figure 7. Clustering Effect

TABLE IV. CLUSTERING RESULTS LISTS

Cluster	Clustering Center	Size	Average Distance
cluster1	(1,2,4,3,1)	57	0.9912
cluster2	(3,0,1,2,4)	4	1.1250
cluster3	(0,-1,4,3,-1)	15	0.4333
cluster4	(0,3,4,1,1)	13	0.8077
cluster5	(1,1,2,1,0)	15	0.7000
cluster6	(1,1,1,3,2)	10	0.8000
cluster7	(1,2,4,1,0)	30	0.7333
cluster8	(1,-1,-1,0,-1)	7	0.5000
cluster9	(0,3,-1,1,0)	46	1.2500
cluster10	(0,2,1,1,0)	10	0.9000
cluster11	(2,2,4,2,1)	11	0.7727
cluster12	(0,-1,-1,0,-1)	7	0.8571
cluster13	(2,2,3,2,3)	32	1.3438
cluster14	(1,-1,2,2,-1)	47	1.4348
cluster15	(1,3,3,3,0)	31	0.8548
cluster16	(1,2,5,2,1)	36	0.9028
cluster17	(2,2,2,3,1,)	13	1.1923
cluster18	(2,1,0,3,2)	15	1.4286

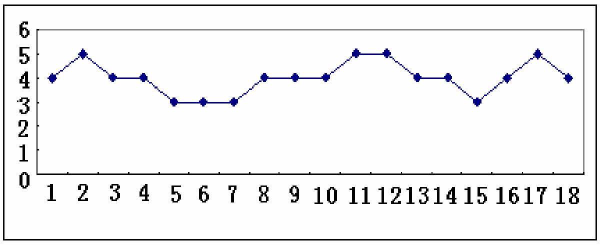

Figure 8. Distance between Isolated Dots and Cluster Center

The distance between isolated dots and each cluster center is larger than the average distance between clusters. This shows isolated dots are different from the exist clusters.

D. Results Discussion

Clustering can differentiate the clients that have different character. Of the 18 clusters from clustering, we choose cluster7 and cluster15 as the examples to analyze the clusters that can mark the clients group.

Cluster7 contains 30 observing sample, the cluster center is (1,2,4,1,0), the average monthly mobile cost is ￥51-150, expending percentage on daily life is 31%-50%, cash paying percentage of all the daily cost is 71%-90%, education level is high school, monthly income is less than 2000. The clients are mainly female, their age is from 31-35, Easy-Own is their frequent using brand. From the occupation distribution, we know the occupation distributing centralized degree is not distinct. So we can ignore the occupation distribution when marking the clients' sorts.

Cluster15 contains 31 observing sample, its cluster center is (1,3,3,3,0), the average monthly mobile cost is ￥51-150, expending percentage on daily cost is 51%-70%, cash paying percentage of all the daily cost is 51%-70%, education level is bachelor, monthly income is less than 2000. We see that the gender character is not distinct. The clients' age is from 21-25, their occupation is mainly common employees, Easy-Own and M-Zone are their frequent using brands. The users of this sort are mainly young men, their career paths are at starting phases, and they are mostly common employees. Expending percentage on daily cost is high, and monthly mobile cost is low comparatively.

Through combining cluster center and discriminating variables, we can compartmentalize different clients' type.

At this foundation, we can analyze different clients' behavior character. Thus we can design the service and product that match the clients.

V. Conclusion and Expectation

The optimized ant colony clustering algorithm expands the traditional ant algorithm such as LF algorithm, which can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time.

The current clustering efficiency of OACA is expected to be improved. There exists repeated vibrancy in the conveying process, so the algorithm needs further improvement that can assure the algorithm single convergence in the every conveying process. The algorithm mainly aims at classified data, but we should disperse the value data in the dealing process. So, we should do further improvement to the algorithm to adapt the mixed data directly.

Acknowledgment

This research was supported by the National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

References

[1] Kennedy J., Eberhart R. C., *Swarm Intelligence*, San Francisco: Morgan Kaufmann, 2001.

[2] Qi Kang, *Particle Swarm Optimization and Scheduling Algorithms*. Shanghai: Tongji University, 2005.

[3] Qidi Wu and Lei Wang, *Intelligent Ant Colony Algorithm and Application*. Shanghai: Technology Press, 2004.

[4] Qidi Wu and Lei Wang, *Intelligent Particle Swarm Optimization Algorithm Research and Application*. NanJing: Jiangsu Education Press, 2005.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys,* vol.31(3), pp. 264-323, March 1999.

[6] Huimin Ma, Yiqing Lu, and Hanbin Yin, "Customer segmentation based on the share of the customer," *Journal of Wuhan University of Technology Information & Management Engineering*, vol. 25(3), pp.184-187, March 2003.

[7] Yumin Tian, Lina Meng, and Lixia Mei, "Region-based image retrieval system using integrative feature description," *Computer Engineering*, vol. 31(19), pp.167-169, Oct. 2005.

[8] Xiangyang Ku, Wenxiang Peng, Huifeng Xue, and Jijun Li, "The research on the spatial clustering algorithm based on GIS". *Computer Engineering and Applications*, vol.41(29), pp.24-26, Oct. 2005.

[9] Ng M. K., and Huang Z., "Data-mining massive time series astronomical data: challenges, problems and solutions," *Information and Software Technology*, vol.41(9), pp.545-556, Sept. 1999.

[10] Deneubourg J. L., Goss S., and Franks N., "The dynamics of collective sorting: robot-like ant and ant-like robot," *Proceedings First Conference on Simulation of Adaptive Behavior: From Animals to Animates*. Cambridge, MA: MIT Press, 1991.

[11] Lumer. E. and Faieta. B., "Diversity and adaptation in populations of clustering ants,".in *Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats,* vol.3, J. A. Meyer and S.W. Wilson, Eds. Cambridge, MA: MIT Press/Bradford Books,1994, pp.501-508.

[12] Yang Y., and Kamel M., *Clustering Ensemble Using Swarm Intelligence Symposium. Piscataway,* vol.51. NJ: IEEE service center, 2003.

[13] Everitt and Brian, *Cluster analysis*, New York: Halsted Press, 1993.

[14] Dorigo M. and Gambardella L. M., "Ant colony system: a cooperative learning approach to the traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, Vol.(1), Jan. 1997.

[15] Bonabeau, E. Dorigo M., and Theraulaz G, *Swarm Intelligence: From Natural to Artificial Systems*. New York: Oxford University Press, August 1999.

[16] Chunhung Cheng, Ada Waichee Fu, and Yi Zhang, "Entropy-based subspace clustering for mining numerical data," In *Proceedings of The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp.84-93.

[17] Weili Zhao, Yanxin Sun, and Zhiguo Zhang, "An improved ant colony clustering algorithm based on information entropy," *Journal of Shenyang Institute of Chemical Technology*, vol.29(4), pp.296-300, Dec. 2005.

[18] Machine Learning Repository, http://archive.ics.uci.edu/ml/, Feb. 6, 2009.

Weihui Dai received his BS degree in Automation Engineering in 1987, his MS degree in Automobile Electronics in 1992, and his PhD in Biomedical Engineering in1996, all from Zhejiang University, China. He is currently an Associate Professor at the Department of Information Management and Information Systems, School of Management, Fudan University, China. He has published more than 100 journal papers and conference articles in Software Engineering, Information Systems, Service Management, Net-ecology, e-Business, etc. Dr. Dai became a member of IEEE in 2003, a senior member of China Computer Society in 2004, and a senior member of China Society of Technology Economics in 2004.

Shouji Liu received his BS degree in Information Management and Information Systems in 2005 from Fudan University, China.

Shuyi Liang received her BS degree in Applied Mathematics in 2006 from Fudan University, China. She is current a master student at School of Management, Fudan University, China.