A new hybrid clustering algorithm based on K-means and ant colony algorithm

Jue Lu
School of Information
Wuhan University of Technology
Wuhan, P. R. China
luj2006@126.com

Rongqiang Hu
School of Automation
Wuhan University of Technology
Wuhan, P. R. China
hurq@whut.edu.cn

***Abstract*—K-means algorithm and ant clustering algorithm are all traditional algorithms. The two algorithms can complement each other. The combination of two algorithms will improve clustering's accuracy and speed up algorithm' convergence. Tests prove hybrid clustering algorithm is more effective than each above-mentioned algorithm. Especially, the new algorithm has good results in image segmentation.**

Keywords - similarity, data mining, clustering

I. INTRODUCTION

Clustering analysis plays an important role in data mining field. Data can be grouped into different classes or clusters by clustering analysis. There exists better similarity among the objects in the same class and poorer similarity among the objects in different classes. According to the theory of machine learning, clustering is a kind of un-supervised learning because it has no priori knowledge of classification. Clustering analysis is widely applied in image processing, model recognition, document retrieval, medical diagnosis, web analysis etc. At present, it has become a popular research subject in the field of data mining. Many researchers are deeply absorbed in this work.

Clustering analysis dated from 1960s[1][2][3][4]. Traditional classification methods include hierarchical method, partitioning method, density-based method, grid-based method and model-based method etc. In recent years, a lot of new clustering algorithms have been proposed after deep study on clustering. Among them is clustering algorithm based on ant system. Up to now, combination of two or three different clustering algorithms is new to clustering analysis. This paper introduces a novel way that ant system is combined with k-means algorithm to improve clustering performance. The aim is to optimize clustering. Tests prove its has good result.

II. ANT CLUSTERING ALGORITHM

Ant colony algorithm[5][6] is a novel optimization method by simulating evolution. Ant clustering algorithm is its branch, which is a kind of clustering algorithm based on swarm intelligence[7][8] and is often applied in image segmentation[9][10]. Researchers observed the fact that ants can classify their ant babies automatically and orderly. Researchers were enlightened from this, ant clustering algorithm was proposed. Its basic idea is that all un-clustering objects are put on a two-dimensional grid. Each object takes an initial place randomly. Ants are also on the grid and can move randomly. When an ant runs into an object, it will measure colony similarity within its local range. This data will be transformed into the probability that the ant moves the object by. The ant will pick up or drop the object according to the value of this probability. The transportation of objects by ants is a process of self-organization clustering. In the end, different clusters on the flat will take shape.

A. Definitions of colony similarity

Colony similarity is the similarity between an un-clustering object and other objects within its local range. The expression of that is:

$$f(O_i)=\begin{cases}\dfrac{1}{\pi r^2}\sum\limits_{O_i\in Neigh(r)}\left(1-\dfrac{d(o_i,o_j)}{\alpha}\right) & \\ 0 & \text{otherwise}\end{cases} \tag{1}$$

Here: Neigh(r) denotes ant's local range. It refers to a circular area that r is its radius on the grid. $d(o_i,o_j)$ presents the distance between object O_i and O_j in the attribute space. Euclidean distance and cosine distance are often used to express the distance between objects. α is defined as a co-efficient of colony similarity.

B. Probability transformation function

The task of probability transformation function is to transform colony similarity into the probability of ants transporting un-clustering object. Colony similarity is its variable. The co-domain of probability transformation function is from 0 to 1. It usually can be drawn into two corresponding curves on the co-ordinate. One expresses the probability of pick-up, another expresses the probability of drop. The regulation of probability transformation function is as followed: The more the colony similarity is, the less the probability of pick-up is; the less the colony similarity is, the more the probability of pick-up is. To the probability of drop, visa versa. According to this principle, we can choose sigmoid function as probability transformation function. The probability of pick-up and the probability of drop are individually defined as followed:

$$P_p=1-Sigmoid(f(O_i)) \tag{2}$$

$$P_d=Sigmoid(f(O_i)) \tag{3}$$

Here: $Sigmoid(x)=\dfrac{1-e^{-cx}}{1+e^{-cx}}$

c is a constant. The more it is, the faster convergence of algorithm is.

C. Description of the algorithm

The process of ant clustering is described as following:

a) Initialization of the algorithm parameters: including *ant-number*, square of the flat, *cycle-counter*, α *r* etc.

b) Data are randomly put on the two dimensional flat and each data object is set as a coordinate (*x*, *y*).

c) An object is assigned to an ant. Assuming all ants are unloaded.

d) Cycle counter++

e) A group of ants begin clustering circle.

f) The colony similarity will be calculated according to formula (1), which the object loaded by an ant is centered with the radius of *r*.

g) If an ant is unloaded, P_p is calculated according to formula (2).

h) Comparing P_p to a random probability P_r, if $P_p < P_r$, the ant doesn't pick up the object and another object will be assigned to this ant. Otherwise the ant will pick up the object. The status of ant also be changed into the loaded and a new coordinate will be randomly assigned to it.

i) If the ant is loaded, P_d is calculated according to formula (3).

j) Comparing P_d to a random probability P_r, if $P_d > P_r$, the ant drops the object and the coordinate of ant will be assigned to this object. The status of the ant will be changed into the unloaded. At this time the algorithm randomly assigns another data object to the ant. Otherwise the ant continues moving with object loaded and will be randomly assigned a new coordinate.

k) If all the ants end up moving, the algorithm will execute step *l*, otherwise the program will choose the next ant and jump to step *e*.

l) If *cycle numbe*r<MAXCYCLE_NUMBER, jump to step *d*, otherwise the program will output the clustering result.

III. K-MEANS ALGORITHM

K-means algorithm is an algorithm based on partition, the algorithm assumes that there is a database consisting of *n* objects and *k* is known as the number of clustering. We can make use of the partition method to build *k* partitions ($k \le n$). Each partition denotes a cluster. Clustering is also based on the similarity between objects. Usually the distance such as Euclidean distance and cosine distance is measured as the similarity.

The result of clustering analysis is that objects in the same class are as similar as possible and objects belonging to different classes are as un-similar as possible. The often-used partition-methods are *k*-means algorithm and *k*-center algorithm.

A. The principle of k-means algorithm

K objects are randomly chosen from *n* objects as initial clustering centers. Then the algorithm calculates the distance from each object to *k* clustering centers and judge which distance is nearest to the clustering center. If one object is that, it will be assigned to this cluster. When all the work of computation is done, it will form *k* new clusters. Next, the algorithm re-computes a mean value of each new cluster as its new clustering center. According to above-procedure, the algorithm will repeat calculating the distance and iterating till criterion function converges. The sum of square error is often-used as the criterion function. It's defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \tag{4}$$

E is the sum of square error to all objects in the database. *P* is a point in the space that expresses a given object. m_i is the mean of clustering C_i .According to this criterion, data belonging to the same class is as more similar as possible and data from different classes are as more different as possible.

B. Step of K-means algorithm

We can describe the process of K-means algorithm as followed:

Input: cluster number *k*, a database of *n* objects

Output: *k* clusters to ensure minimize the sum of square error.

a) *K* objects are randomly chosen as initial clustering centers.

b) The distance between each object and correspondent clustering center is calculated. According to the criterion of the shortest distance, the algorithm assigns them to the correspondent cluster.

IV. THE COMBINATION OF ANT CLUSTERING ALGORITHM AND K-MEANS ALGORITHM

One advantage of ant clustering algorithm is, without any information about classification, this algorithm can finish a primary clustering. However two problems exist. One is, when the algorithm ends up, there are some data not to be assigned to a stack. We call them free data, including the data ants loaded and the data located on the grid individually. Another problem is, if data is placed in a data stack incorrectly, it will cost long time to be moved to the correct data stack by ant. K-means algorithm can make the square error of *k* clustering minimize. In case of many data, it is effective and smart. However you must assign *k* value before clustering. It is difficult to some inexperienced users. On the other hand, the choosing of initial clustering centers will make influence on the final clustering result.

A. Hybrid clustering algorithm

Hybrid clustering algorithm is based on the combination of ant clustering algorithm and K-means algorithm. According to random ergodic theory, ant clustering algorithm doesn't need priori knowledge to do clustering analysis. This algorithm can avoid local optimization but will cost long time to compute. K-means algorithm needs an initial clustering to analyze by means of certainty inspiring principle. However it can converge fast. Two algorithms have complementary advantages. K-means algorithm is

convergent in theory. K-means algorithm is able to remove the incorrect point quickly and assign free data object to the corresponding cluster. Hybrid algorithm doesn't need input initial cluster information and can avoid bad results because of incorrect initial information. Obviously, the combination will accelerate the running of algorithm and improve clustering.

B. The description of hybrid clustering algorithm

Hybrid clustering algorithm is described as following:

a) Each data object is made to do clustering with ant clustering algorithm.

b) From *a*, k clusters and their clustering centers are chosen.

c) Input above-information and cluster with K-means algorithm.

d) Output the result.

V. ALGORITHM ANALYSIS AND CONCULSION

This paper uses Iris plant samples to do individual clustering test on three clustering algorithms. In this database, there are 150 samples of three plants. Each sample has 4 attributes. We have done three tests for every algorithm to measure their accuracy and elapsed time. In this test, parameters are set as followed: *Ant number*=2, *cycle number*=100000, $\alpha = 1$ *r*=2. We use run-time and F-measure to evaluate the performance of three algorithms. The result is seen as table 1. From table 1, we can find ant clustering algorithm costs much time compared with two other algorithms and its accuracy (F-measure) is lower. K-means algorithm is fastest. But its F-measure ranks second. Although hybrid clustering algorithm take more time than K-means algorithm, but its F-measure is highest. This algorithm makes full use of initial result from ant clustering algorithm and provides convenience for K-means algorithm. So it is more stable.

TABLE I. TEST RESULT OF THREE ALGORITHMS

	ant clustering algorithm		K-means algorithm		hybrid clustering algorithm	
	time	accuracy	time	accuracy	time	accuracy
1	7.3s	56	0.018s	83	6.5s	88
2	7.1s	58	0.017s	75	6.6s	79
3	7.2s	55	0.018s	77	6.5s	85

In addition, this new algorithm has good application in image segmentation. In order to test its segmentation effect, we use K-means method, ant clustering algorithm and hybrid algorithm to get segmentation for two images individually. The segmentation objects are the image 'Lena' and 'peppers'. 256×256, the gray-level is 256. The test is under Windows XP environment.

Comparing two segmented images, to segmented images with K-means method, we can find the object was not segmented completely from the background and some details in the image such as hair of Lena also lost. The information about edge in images is less than that new method. So is ant clustering algorithm. Obviously, the images segmented by hybrid method is better. Their contours are more clear. The results prove the new hybrid algorithm is more effective in image segmentation and we believe it will have good application in other fields.

Fig.1 Original images

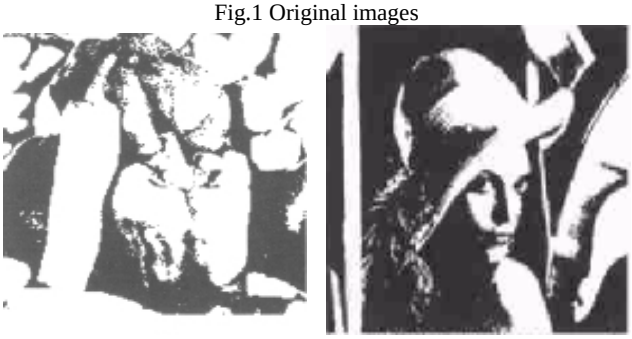

Fig.2 Results by K-means method

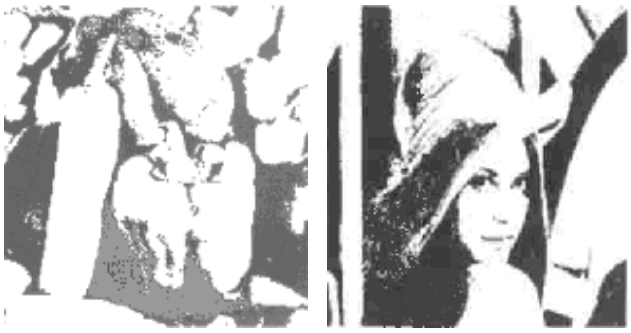

Fig.3 Results by ant clustering algorithm

Fig.4 Results by hybrid algorithm

ACKNOWLEDGMENT

The authors are grateful to ZuoDai, professor of WHUT for his remarks and good suggestions.

REFERENCES

[1] Al-Sultan K S. A Tabu Search Approach to the Clustering Problem[J]. Pattern Recogn, 1995, 28: 1443-1451

[2] Maulik U, Bandyopadhyay S. Genetic Algorithm-based Clustering Technique[J]. Pattern Recognition, 2003, 33(9): 1455-1465.

[3] Zhang R, Peng H. A Faster Simulated Annealing Algorithm for the Data Clustering and Its Application [J]. ComputerEngineering and Application , 2001, 15(1): 85-87.

[4] Kao Y, Cheng K. An ACO-Based Clustering Algorithm[C]//ANTS 2006, LNCS 4150·Berlin: Springer, 2006: 340-347

[5] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization", *Artificial Life*, vol. 5, no. 2, pp. 137-172, 1999

[6] M Dorigo,V Maniezzo and A Colorni. The Ant System: Optimization by a colony of cooperating agents[J]. *IEEE Transactions on Systems, Man and Cybemetics*-part B, 1996,26(1):1-13

[7] Shelokar P S, Jayaraman V K, Kulkarni B D. An Ant Colony Approach for Clustering [J]. Analytica Chimica Acta, 2004, 509: 187-195

[8] Wu B,Shi Z.A clustering algorithm based on swarm intelligence[A]. Proceedings IEEE international conference on info-tech&info-net proceeding [C].Beijing,2001.58-66

[9] Zheng H, Wong A, Nahavandi S. Hybrid ant colony algorithm for texture classification. Proceedings of the 2003 congress on evolutionary computation, 2003,4: 2648-2652

[10] Zhang Y J. *Image Engineering (I): Image Processing and Analysis*, pp. 179-215, Tsinghua University Press, Beijing, 1999