

A New Cluster Analysis Based on Combinatorial Particle Swarm Optimization Algorithm

Jin Jin*

College of Electrical Engineering
Northwest University for Nationalities
Lanzhou, China
e-mail: jinjin_2000@163.com
*Corresponding Author

Ma Zhong

College of Electrical Engineering
Northwest University for Nationalities
Lanzhou, China
e-mail: 18093180439@126.com

Xue Lin

School of Foreign Studies
University of Science and Technology Beijing
Beijing, China
e-mail: ustbxuelin@163.com

Tian Changhui

School of Foreign Studies
University of Science and Technology Beijing
Beijing, China
e-mail: tian_changhui@aliyun.com

Abstract—Inspired by the swarm intelligence in self-organizing behavior of real Particle Swarm Optimization various Particle Swarm Optimization algorithms were proposed recently for many research fields in data mining such as clustering Compared with the previous clustering approaches such as K-means the main advantage of Particle Swarm Optimization based clustering algorithms is that no additional information is needed such as the initial partitioning of the data or the number of clusters In this paper, we discuss the clustering analysis way by a combination of advantages of particle swarm optimization in the clustering, since Particle Particle Swarm Optimization has the good global searching quickly. Firstly, the center and number of clustering are determined by using the Particle Swarm Optimization, and then the above clustering results are optimized by the K-means algorithm combining with the optimization algorithm. The simulated experiments show that the combining algorithm is obviously superior to some common clustering algorithms since it has obvious advantage in optimization capacity, more efficient and more robust than previous research such as the classical K-means clustering algorithm.

Keywords-Particle swarm algorithm; cluster analysis; combinatorial optimization; K-means

I. INTRODUCTION

Bio-cluster analysis is based on the difference between different objects, pattern classification based on specific guidelines, its application is quite extensive. Comparison of cluster analysis methods, mainly of hierarchical clustering methods BIRCH, CURE and ROCK, density clustering OPTICS, DBSCAN algorithm and DENCLUE, STING and grid clustering algorithm, CLIQUE and Wave Cluster algorithms, and so on. Based on K-means algorithm is Mac Queen classical

algorithm for clustering problem, widely used in the field of data mining and knowledge discovery [1]. Due to the K-means algorithm for clustering likely to be initially selected the center of premature convergence in suboptimal solution, many scholars have suggested that the clustering method based on genetic algorithm [2-3], in 2002 by Strehl and Ghosh use a combination of hypergraph model clustering method [8]. Literature describes a document based on a combination of two different clustering techniques topic to discover new methods[9].

Particle Swarm Optimization algorithms (Particle Swarm Optimization, PSO is an evolutionary computation technique Eberhart and Kenney in 1995 [11], from the predatory behavior of the birds, is based on the iteration of the optimization tool. Widely used for solving complex optimization problems [12], the algorithm with fast convergence, less need to adjust parameter settings, and so on. In recent years a large number of applications to engineering practice, its search model of velocity and displacement, simple operation, low complexity, both with great probability to ensure optimal solutions, and improve the speed of convergence of the local area.

Using Particle Swarm algorithms to solve for K-means clustering problem, PSACO-KMEAMS clustering algorithm is proposed (k-Means Cluster Optimization Based on Combination of Particle Swarm Algorithm and Ant Colony Optimization), the basic idea is: Ant Colony algorithm based fast global search capability and rapidity of convergence of particle swarm optimization algorithm and Ant Colony algorithm generates a data object used in the initial the initial cluster centers, using Particle Swarm Optimization algorithm in the late segmentation cluster and cluster division between, improving the clustering structure,

reaching algorithm complement each other. The algorithm and K-means algorithm, Ant Colony Optimization, Particle Swarm Optimization algorithm for contrast, achieve better results of cluster analysis.

II. OPTIMIZATION CLUSTERING ALGORITHM BASED ON PARTICLE SWARM OPTIMIZATION

A. Mathematical description of the K-means clustering algorithm

Based population samples $X = \{X_i, i = 1, 2, \dots, n\}$, where X_i is a M dimensional pattern vector, according to the similarity of the sample will divide it into $C = \{C_1, C_2, \dots, C_m\}$, meet:

$$X = \bigcup_{i=1}^m C_i$$

$$C_i \neq \Phi (i = 1, 2, \dots, m) \quad (1)$$

$$C_i \cap C_j = \Phi (i, j = 1, 2, \dots, m; i \neq j)$$

Always between-class scatter and are:

$$J_c = \sum_{k=1}^m \sum_{X_i \in C_k} d(X_i, Z_k) \quad (2)$$

Z_k is the center of k -th clustering, criterion J_c poly category clustering for samples to the sum of the distances from the Center. After the cluster Center, clustering Division by nearest neighbor rule. The sample X_i , if the cluster Center Z_j category-meet (2), then X_i belongs to the class of j . $d(X_i, Z_j)$ is the distance between samples and corresponding cluster centers using Euclidean distance:

$$d(X_i, Z_j) = \|X_i - Z_j\| \quad (3)$$

B. Particle Clustering Algorithm

Search model of particle swarm optimization algorithm with speed and position. Particle Swarm consists of many particles, each $x_i(t)$ represents the position of the particle problem in the search space of candidate solutions, pros and cons of solutions degree determined by the fitness function $f(Z_j)$.

Every iteration, the particles by track two extremal is dynamics to update its speed $v_i(t)$ and $x_i(t)$. A particle from the initial to the current iteration of the search of optimal solutions: the individual extreme values of $p_i(t)$, the other is the global extremum for $g_i(t)$ PSO is currently the best solution. Of which:

$$p_i(t+1) = \begin{cases} p_i(t) & \text{if } f(x_i(t+1)) < f(p_i(t)) \\ x_i(t+1) & \text{if } f(x_i(t+1)) \geq f(p_i(t)) \end{cases} \quad (4)$$

$$g(t+1) = \min\{p_i(t+1), \text{ for } i\} \quad (5)$$

Each particle according to equation (6) and (7) updated its speed and location:

$$v_i(t+1) = c_0 v_i(t) + c_1 r_1 [p_i(t) - x_i(t)] + c_2 r_2 [p_g(t) - x_i(t)] \quad (6)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (7)$$

Style, $i, g=1, 2, \dots, m$; v_i is the velocity vector of the particle i , x_i is the current location of particle i and t for the current iteration number, c_0 is the inertia weights, it keeps the particle motion of the inertia, and enable them to explore new areas, and generally the (0,1) random number between. c_1 , c_2 Coefficient of Group Cognition, take (0,2) a random number between, r_1 , r_2 is between (0,1) on the random number.

Particle code using real coding based on cluster centers, each code corresponds to a feasible solution. If k represents the number category, d for sample vector of dimension, then the position of the particle is a $k \times d$ size matrix, the velocity of the particle is also a $k \times d$ dimension variable.

In this way, $2(k \times d) + 1$ the length of the coding structure of the particles can be used as follows:

$Z_{11}Z_{12} \dots Z_{1d} \dots Z_{k1} \dots Z_{kd}$	$V_1V_2 \dots V_{k \times d}$	$f(X)$
---	-------------------------------	--------

Evaluate the fitness of each particle. Fitness calculation according to the type of particles (8):

$$f(Z_j) = \frac{\varepsilon}{J_c + J_0} = \frac{\varepsilon}{\sum_{k=1}^m \sum_{X_i \in C_k} d(X_i, Z_k) + J_0} \quad (8)$$

Better if clustering, J_c is smaller, the greater fitness $f(Z_j)$. Type ε is a normal number, J_0 is a small positive number, said in order to avoid the J_c tends to 0 zero overflow situation, J_c clustering results in all types of distance and, it's meaning and the symbol (2). Iteration algorithm of particle swarm is to search fitness function $f(Z_j)$ maximum.

C. Combinatorial Particle Swarm Algorithm

Step1: first initializes the population, when the particle is initialized, each randomly assigned to a certain category of first, as the initial clustering, and various types of cluster centers, the initial position of the particle codes, calculate the fitness of particles, and initialize the particle's velocity.

Step2: for each particle, compared with its fitness and has been the best fitness value of $p_i(t)$, if better, updated $p_i(t)$;

Step3: according to (4) and (5) to adjust the speed and the position of the particle;

Step4: according to your current location, the sample principle of distribution according to the minimum distance is assigned to k cluster centers;

Step5: according to (8) calculates the fitness $f(Z_j)$;

Step6: If after all the particles traverse the output $p_i(t)$ to local best location;

Step7: according to output $p_i(t)$ to local the best location, the best location to find out what groups experienced $g_i(t)$ output global fitness $f(Z_j)$;

Step8: k optimized for new individuals.

Step9: If you reach the end good enough conditions of location or number of iterations, the end, else go to Step2.

For a new generation of particles, following the K-means algorithm for optimization:

Rule number one: according to the particle cluster Center coding, in accordance with the nearest neighbor rule to determine the particle clustering;

Rule two: according to the clustering, calculate the new cluster centers, fitness update particles, instead of encoded values.

If empty clusters were randomly removed from some other non-empty clustered farthest from the cluster Center pattern vector and the vector into an empty cluster, repeat this process until no clustering available in the Division so far.

D. Improved Combinatorial Particle Swarm Algorithm

By the literature that, improvement of Particle Swarm Optimization and K-mean several paper [13] improved method makes the following algorithm:

Method 1: the improvement is first mixed with the K-means algorithm, the specific method is by using the K-means algorithm for classification, the result is as a result of one of the particles, and later in Step7 using particle swarm of Ant Colony clustering as a result of initial value assignment to Ant Colony clustering search, this method is called Kmean+PSO+ACO.

Method 2: using K-means algorithm idea, later in Step4, recalculated on the new classification based on the new cluster, and then update the current position, and later in Step7 using particle swarm of Ant Colony clustering as a result of initial value assignment to Ant Colony clustering search, this method is called PSO+Kmean+ACO.

Method 3: improved method 1 and method 2, this is called Kmean+PSO+Kmean+ACO.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Experimentation platform for Windows XP, the machines up to Inter the Core2 Duo CPU P8700 (2.53 GHz), by compiling all software running under MATLAB algorithm.

Particle Swarm parameter setting: $r_1, r_2(0,1)$ random number between, c_1 and c_2 are called factors. In General, $c_1 = c_2 = 1.5$. The c_0 is the weighting coefficient, value 0.9 to 0.1 decrease from high to low. Ant Colony optimization algorithm for parameter setting: $\alpha = 1$, $\beta = 4$, $\rho = 0.8$, $Q = 1000$, $\tau_0 = 1$, maximum number of iterations $\max N = 100$.

Experimentation data set from the UCI machine learning Library [14] about for Iris,Wine,zoo and Glass. Data for each sample data set size, number properties, and classification are shown in table I.

TABLE I DATABASE DESCRIPTION

Neighborhood	Iris	Wine	Zoo	Glass
Data size	150	178	101	214
The number of attributes	4	13	17	10
Number of categories	3	3	7	7

Although the results on the above data using the new algorithm PSACO-KMEANS, K-means, ACO and the PSO and the GA test, each algorithm of 15 experimental, the results are shown in table II.

TABLE II THE RESULTS OF CLUSTERING ALGORITHM OF SUM OF MEAN SQUARE ERROR

Algorithm	Iris	Wine	Zoo	Glass
K-means	104.7	348.7	258.9	7986
ACO	158.9	248.9	288.9	7854
PSO	150.6	216.7	278.5	7736
Kmean+PSO	148.9	102.6	232.7	7856
PSO+Kmean	146.5	109.8	198.6	7248
Kmean+PSO+ACO	116.3	107.5	187.6	7432
PSO+Kmean+ACO	119.6	121.5	178.3	7197
Kmean+PSO+Kmean+ACO	103.5	98.7	180.9	7158

It is evident results from the table, means K-means algorithm for Iris data set test dataset other than good when some poor test results, algorithm, PSO and GA on data collection of Wine test results, test results of the data collection of Zoo and Glass, various algorithm closer. Taken as a whole the results, the Kmean+PSO+Kmean+ACO algorithm results is more stable, showing a high degree of optimization of advantage.

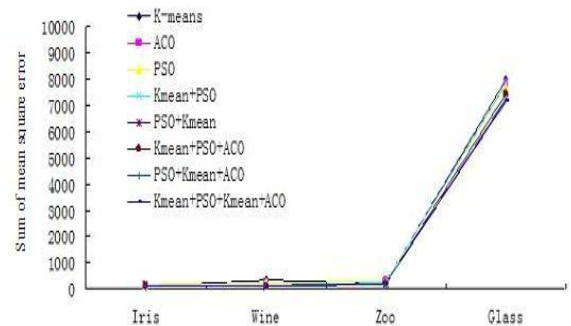


Chart I Comparison of clustering combined algorithm

IV. CONCLUSION

It proposes a particles group algorithm and Ant group algorithm fusion of algorithm, will particles group algorithm into Ant group algorithm of specific process in the, using particles group algorithm find of results with into Ant group algorithm among, using particles group algorithm fast convergence of advantages to speed up Ant group algorithm of convergence speed, and avoid Ant group algorithm in local optimal of, expand has Ant

group algorithm of solution range. Firstly, K-means hybrid algorithm applied to solving clustering problem solved two ways to find system data under the conditions. Secondly, designing a fusion algorithm based on adaptive function and correlation clustering the data set reflects more accurately.

ACKNOWLEDGMENT

This work is supported by the Fundamental Research Funds for the Central Universities of China for Northwest University for Nationalities (Grant No. 31920140087), and the Introduction of Talent Project for Northwest University for Nationalities (Grant No. xbmuyjrc201312).

The authors would like to express thanks to Prof. M. J. Khan with the School of PN Engineering, National University of Sciences and Technology, Islamabad, Pakistan and Prof. J. Cao with the Research Institute of Information Technology, Tsinghua University, Beijing, China for their beneficial discussions about this interesting topic.

REFERENCES

- [1] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principle*, Addison Wesley, Reading, 1974.
- [2] S. K. Pal and P. P. Wang, *Genetic Algorithms for Pattern Recognition*, Boca Raton: CRC Press, 1996.
- [3] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of Clusters and Application to Image Classification", *Pattern Recognition*, vol. 35, pp. 1197-1208, 2002.
- [4] N. Cai, J. Cao, H. Ma, C. Wang, "Swarm stability analysis of nonlinear dynamical multi-agent systems via relative Lyapunov function", *Arab. J. Sci. Eng.*, vol. 39, pp. 2427-2434, 2014.
- [5] N. Cai, J. Cao, M. J. Khan, "A controllability synthesis problem for dynamic multi-agent systems with linear high-order protocol", *Int. J. Control Automat. Syst.*, vol. 12, pp. 1366-1371, 2014.
- [6] N. Cai, M. J. Khan, "On swarm stability of linear time-invariant descriptor compartmental networks", *IET Control Theory Appl.*, vol. 9, pp. 793-800, 2015.
- [7] N. Cai, J. Cao, M. J. Khan, "Almost decouplability of any directed weighted network topology", *Physica A*, vol. 436, pp. 637-645, 2015.
- [8] A. Strehl and J. Ghosh, *Cluster ensembles a knowledge reuse frame work for combining partitionings*. Edmonton : AAAI/MIT Press , 2002.
- [9] H. Ayad and M. Kamel, *Topic discovery from text using aggregation of different clustering methods*. Calgary, 2002.
- [10] M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem", *IEEE Trans on Evolutionary Computation (S1089-778X)*, vol. 1, pp. 53-66, 1997.
- [11] J. Kennedy and R. C. Eberhart, "Particle Swarm Optimization", *Proceedings of IEEE International Conference on Neural Networks*, vol. 1, pp. 1942-1948, 1995.
- [12] R. C. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources". *Proc. Congress on Evolutionary Computation 2001*, vol. 1, pp. 81-86, 2001.
- [13] S. Gao and J. Y. Yang, "New Clustering Method Based on Particle Swarm Algorithm", *Journal of Nanjing University of Aeronautics and Astronautics*, vol. 38, pp. 62-64, 2006. (In Chinese)
- [14] C. L. Blake and C. J. Merz, *UCI Machine Learning repository of machine learning databases*. 1998. <http://www.ics.uci.edu/mlearn/MLSummary.html>.
- [15] P. R. Patnalk, "An integrated hybrid neural system for noise filtering, simulation and control of a fed-batch recombinant fermentation", *Biochemical Engineering Journal*, vol. 15, pp. 165-175, 2003.
- [16] Y. Yang and S. K. Mohamed, "An aggregated clustering approach using multi-ant colonies algorithms", *Pattern Recognition*, vol. 39, pp. 1278-1289, 2006.