Clustering Analysis Based on Improved Fuzzy C - Means Algorithm

Qiongqiong Hu,
*Department of Computer Technology and Applications,
Qinghai University 810016 China*
Yiyang Li,
*Department of Computer Technology and Applications,
Qinghai University, Xining 810016 China*
Yong Ge,
*School of Mechanical Engineering,
Qinghai University, Xining 810016 China*
Yu-an Zhang*,
*Department of Computer Technology and Applications,
Qinghai University 810016 China*
Qinglian Ma,
*Interdisciplinary Graduate School of Agriculture and Engineering,
University of Miyazaki, Miyazaki City 889-2192 Japan*
Makoto Sakamoto,
*Faculty of Engineering,
University of Miyazaki, Miyazaki City 810016 Japan*

Abstract

Cluster analysis is one of the most important technologies in the field of data mining, and it is also a hot topic in academic research. So far, it has made great achievements in theory and method, and plays an important role in data analysis in various fields . The fuzzy C-means algorithm (FCM) maintains the simplicity of its thinking, the time complexity is close to linear, and it is efficient and scalable for large-scale data mining. Particle Swarm Optimization (PSO) is an optimization tool with global optimization ability, and the group intelligence is formed by the interaction between particles in the group, and the optimal result is found by the intelligence . However, with the deepening of algorithm research, its research prospects and application areas are increasingly bright and compelling .

Keywords—particle swarm optimization algorithm, clustering analysis, fuzzy C - means algorithm.

I. INTRODUCTION

Clustering analysis is one of the most important technologies in the field of data mining and plays an important role in data analysis in all fields [1]. Clustering is an unsupervised and efficient learning method, and its data is similar and unique to each molecule, so it can be classified according to its different characteristics. At present, clustering analysis has been widely used in many aspects, and with the development of science technology has been more and more attention. In essence, clustering belongs to the local search algorithm, which uses an iterative climbing technique to find the optimal solution [5]. Therefore, the traditional clustering algorithm has two fatal problems: First, with large-scale data processing, the original algorithm fails or takes a lot of time, the second is easy to fall into the local minimum.

Among them, the fuzzy C-means algorithm (FCM) keeps its thinking simple and easy. The time complexity is close to linear, and it has high efficiency and scalability for large-scale data mining [2].After years of development of particle swarm optimization algorithm [3],its optimization speed, quality and algorithmic robustness has been greatly improved. However, most of the current research focused on the realization, improvement and application of algorithms [6]. At present, there are still many unsolved problems in the study of particle swarm algorithm:(1) stochastic convergence analysis; (2) determination of algorithm parameters; (3) discrete / binary particle swarm optimization; (4) the use of different problems to design a corresponding effective algorithm;(5) Should be committed to complement and extend the algorithm with other algorithms or techniques to solve the problem of PSO easy to fall into the local optimum [4].

*Corresponding author: Yu-an Zhang, e-mail:2011990029@qhu.edu.cn

II. THE PSO ALGORITHM

The basic idea of the particle swarm algorithm is that the first step is to initialize a group of random solutions to obtain the particles in the particle swarm. The second step is to find the optimal value in the algorithm by repeating the result of repeated feedback. At each repetition the two optimal values of pbest and gbest are obtained according to Equation 1 and Equation 2 in the feedback. Fig.1 shows the PSO algorithm flow chart.

$$V_i = v_i * w + \text{rand}(1) * c_1 *(\text{rand}(1) * c_2 *(\text{gbest[i]}+ \text{pbest[i]}-\text{present[i]})) \quad (1)$$

$$\text{Present[i]} = \text{present[i]} + \text{v[i]} \quad (2)$$

V_i is the velocity value of particle ***i***, ***present*** **[i]** is the position of the current particle ***i***, ***w*** is the inertia weight, c_1, c_2 is the learning factor, ***rand*** **(1)** is between (0,1) random number.

The execution flow of the PSO can be described as follows:

1.Determine the population size of the particle N, determine the maximum number of iterations, randomize the velocity and position of the particles;
2.Calculate the fitness value of each particle according to the fitness function;
3.The current fitness value of each particle is taken as the initial optimal value of the particle, and the whole optimal value is obtained by comparing all the initial optimal values.
4.Update the particle's position and speed according to the update formula;
5.Calculate the fitness value of the particles in the new population;
6.Compare the new fitness value of each particle with the previous optimal fitness value of the particle. If it is good, update the individual optimal value of the particle, otherwise it will not be updated.
7.The new fitness value of each particle and the overall fitness value of the comparison, if the good, then update the overall fitness value, or not updated;
8.Repeat steps ***5*** through ***7*** until a certain number of iterations is satisfied.

gbest and ***pbest***, that is, the individual optimal value and the population optimal value, in each iteration process, the two optimal values will be repeated according to the *Formula 1* and *Formula 2*. The advantage of PSO is that the concept is relatively simple, relatively easy to achieve, while the need to adjust the parameters less, and the convergence rate. The basic features of the PSO algorithm can be summarized as follows: Intelligent search; progressive optimization; black box structure; implicit parallelism; flexibility; global optimization; robustness; self-organization.

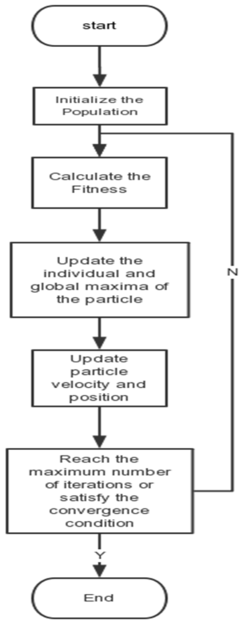

Fig. 1. PSO optimization flow chart

III. CLUSTER ANALYSIS

A. The basic concept of cluster analysis

Clustering is a process of sorting according to data characteristics. If a class as a group, you can cluster as a data compression form. Clustering can be more specific image is defined as follows: Data space in the data set D in the number of data points is n, each data points have d attributes, then each data point can be expressed as $X_i = (x_{i1}, x_{i2}, ..., x_{id})$, the entire data set can be seen as a matrix of ND dimensions. By clustering the data set, we can get k different clusters C_i (*i* = 1,2, ..., k). In the process of clustering, it is possible that some data points do not belong to any clusters. These data points are called outliers or noise points.

B. FCM (Fuzzy C-means Algorithm) clustering algorithm

The fuzzy C-means algorithm (FCM) is an improvement of the general C-means algorithm. The ordinary C-means algorithm is hard for the data partitioning, and FCM is a flexible fuzzy partition.

First, the concept of membership function is described. The membership function is a function that represents the degree of an object x belonging to the set A, usually denoted by A (x), whose range of arguments is all the objects that may belong to the set A (ie, all the points in the space in which the set A is located) The value range is [0,1], that is, $0 \le \mu A(x) \le 1$. $\mu A(x) = 1$ means that x is completely subordinate to set A, which is equivalent to $x \in A$ on the traditional set concept. A membership function defined on space X = {x} defines a fuzzy set A, or a fuzzy subset defined on the universe X = {x}. For the finite objects $X_1, X_2, X_3, ..., X_n$, the fuzzy set can be expressed as:

$$A = \{\mu_A(x_i), \ x_i | x_i \in X\} \tag{3}$$

Through the concept of fuzzy set, in the problem of cluster analysis, we can treat clusters clustered by clustering as a fuzzy set, and all membership points belong to the group [0,1]. In the fuzzy C-means clustering algorithm, it divides n data into c data sets and finds the center on the basis of this, so that the value function of the non-similarity index is minimized. However, for each data set membership, its sum is equal to 1:

$$\sum_{i=1}^{c} \mu_{ij} = 1, \quad \forall j = 1, ..., \ n \tag{4}$$

then, at this time FCM value function (or objective function) is:

$$J(U, c1, ..., \ c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} \mu_{ij}^m d_{ij}^2 \tag{5}$$

Where μ_{ij} is between 0 and 1; c_i is the clustering center of fuzzy group I, $d_{ij} = \|c_i - x_j\|$ is the Euclidean distance between the i_{th} cluster center and the j_{th} data point; Is a weighted index. By constructing the *Formula 6*, we obtain the necessary conditions for the *Formula 5* to reach the minimum value:

$$\bar{J}(U, c_1, ..., c_c, \lambda_1, ..., \lambda_n) = J(U, c_1, ..., c_c) + \sum_{j=1}^{n} \lambda_j (\sum_{i=1}^{c} \mu_{ij} - 1) \tag{6}$$

$$= \sum_{i=1}^{c} \sum_{j}^{n} \mu_{ij}^m d_{ij}^2 + \sum_{j=1}^{n} \lambda_j (\sum_{i=1}^{c} \mu_{ij} - 1)$$

here, λ_j, j = *1* to *n*, is the *Formula 4* of the n constraints of the Lagrangian multiplier. Generate guidance for all input parameters. It can be seen that the necessary condition for the *Formula 5* to be minimized is:

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^n x_j}{\sum_{j=1}^{n} \mu_{ij}^m} \tag{7}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{2/(m-1)}} \tag{8}$$

that the fuzzy C-means clustering algorithm is essentially a repetitive feedback process. When running in a batch mode, the FCM algorithm uses the following steps to determine the clustering center c_i and the membership matrix ***U[1]***:

Step 1: Initialize the membership matrix U This value requires a random number in the range 0 to 1 to satisfy the constraint in *Equation 4*.

Step 2: Calculate c clustering centers ci, i = 1, ..., c using *Formula 7*

Step 3: Calculate the value function according to *Formula 5*. According to the value of the function to determine whether the algorithm to stop.

Step 4: Use the *Formula 8* to compute the new $\boldsymbol{U}$ matrix. Return to step 2. The algorithm can also initialize the cluster center and then repeat it.

1) The Achievement of FCM Algorithm

Where $\boldsymbol{x}$ is the 400 points from 0 to 1, that is, 400 sets of data, using the Rand () function in excel. By running the algorithm, the data of the four types of data and the five types of data are realized, and the value function (objective function) Jb of the corresponding non-similarity index is obtained. The following related images are generated using MATLAB 7.0, and the different shapes in the picture represent different clusters. Each population center has a triangle that is used to represent the cluster center. In ***Fig.2***, the four small circles It is shown that the number of groups is 4 clusters, and the five small circles in ***Fig.3*** show that the number of groups is 5 clusters, and the Jb values are compared to prove the enhancement of the improved algorithm.

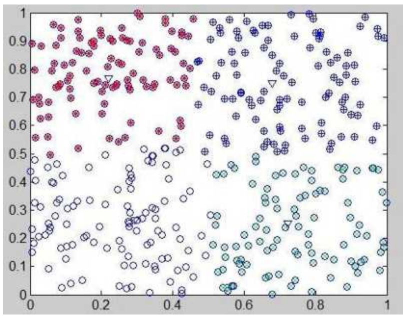

Fig. 2. When cn = 4, Jb = 3.3191 (result of a random experiment)

C. Apply the particle swarm algorithm to fuzzy C-means clustering

Through the above we can see that the FCM clustering in the processing of 400 sets of data, the implementation process is in the gradient method of decline, very rigid, step by step the implementation process will not record each of the optimal solution steps will be achieved on the line Error stops, and outputs the objective function. Based on the limitations of FCM algorithm, it is easy to fall into the local optimal solution. Therefore, it is considered that PSO will replace the gradient descent method in FCM and change it to iterative calculation. This advantage is that the whole algorithm is advancing in the optimal solution, which is more conducive to finding the optimal solution. Although a large number of calculations are added, the similarity in the group is improved, the non-similarity is reduced, and the jb value is optimized.

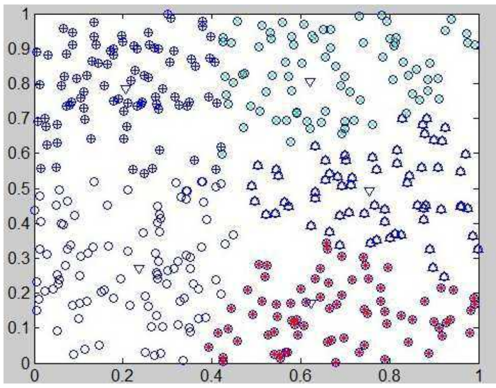

Fig. 3. When cn = 5, Jb = 2.0944 (a random result)

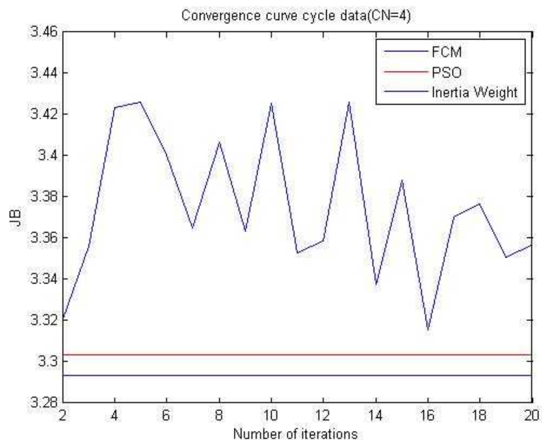

Fig. 4. Under the condition of CN = 4, the JB value curve under different strategies ((FCB represents the fuzzy C-means algorithm,

PSO represents the idea of integrating the particle swarm algorithm in FCM, and the Inertia Weight represents the linearly decreasing inertia weight))

D. Apply linear inertia weighting strategy

Taking into account the optimization in this hope that the early have a higher global search capabilities, and later have a higher open capacity to speed up the convergence rate, so the value of inertia weight should be diminishing, and therefore on the basis of the initial improvement plus the decreasing linear weight, the inertia weight value is no longer used with a fixed value of 1, and it is set to a weight drop between 0.9 and 0.4.

Note: In order to describe the convenience in the form, the improvement refers to: only the use of FCM, the initial improvement refers to: FCM in the integration of the idea of particle swarm algorithm; re-improvement refers to: the initial improvement on the basis of the use of Linearly decreasing inertia weight values. The following description has the same meaning.

The above two graphs show that the improved Jb value is stable and the value is small, which converges to 3.2934 and 2.0811, respectively, when the program is run for 2 to 20 times and the average of the Jb values is obtained.

TABLE I
CONVERGENCE CURVE CYCLE DATA(CN=4)

Cycle-index	Before improvement	First improvement	Secondary improvement
2	3.319930289	3.30345825	3.293458078
3	3.355932514	3.303458248	3.293458177
4	3.42322518	3.303458172	3.293458117
5	3.426067281	3.303458416	3.293458145
6	3.400340325	3.303458182	3.293459015
7	3.364774648	3.303458318	3.293459013
8	3.406441959	3.30345819	3.293458177
9	3.363434596	3.303458244	3.293458196
10	3.425287692	3.303458303	3.293458745
11	3.352790831	3.303458203	3.293458676
12	3.358763164	3.303458242	3.29345866
13	3.426013215	3.303458326	3.29345836
14	3.33766005	3.303458286	3.293458688
15	3.388138626	3.303458241	3.293458702
16	3.315347435	3.303458217	3.293458521
17	3.370346834	3.30345827	3.293458323
18	3.376711892	3.303458259	3.293458509
19	3.350576588	3.303458256	3.293458746
20	3.3567382	3.303458276	3.293458488

TABLE II
CONVERGENCE CURVE CYCLE DATA(CN=5)

Cycle-index	Before improvement	First improvement	Secondary improvement
2	2.116946479	2.091145987	2.081188193
3	2.127751484	2.091163385	2.081261095
4	2.114908503	2.091140444	2.081247198
5	2.123793236	2.091163526	2.081226471
6	2.118510332	2.091142244	2.0812091
7	2.116134805	2.091163586	2.081188193
8	2.108827784	2.091180206	2.081193212
9	2.107951404	2.091192388	2.081185936
10	2.110614058	2.091165532	2.081183326
11	2.11809571	2.091161191	2.081178895
12	2.127541954	2.091169919	2.081189653
13	2.115151641	2.091148045	2.08119233
14	2.113399393	2.091161057	2.081160813
15	2.109794086	2.09116062	2.081184913
16	2.109062738	2.09116664	2.081197084
17	2.129499201	2.091152719	2.081190877
18	2.120138722	2.091167363	2.081191144
19	2.110598647	2.091155037	2.081188943
20	2.115272508	2.091170088	2.081193307

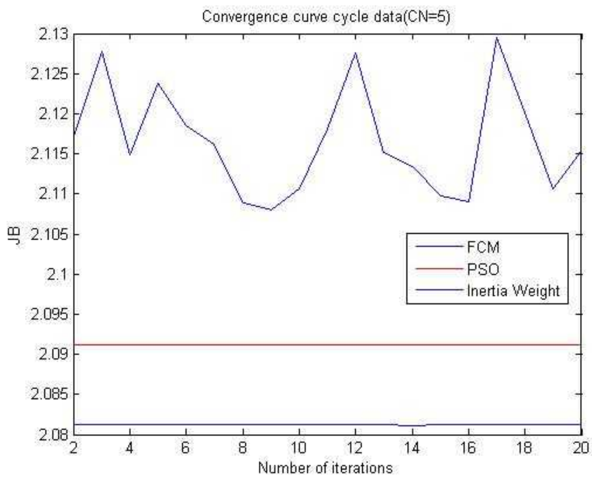

Fig. 5. Under the condition of CN = 5, the JB value curve under different strategies ((FCB represents the fuzzy C-means algorithm, PSO represents the idea of integrating the particle swarm algorithm in FCM, and the Inertia Weight represents the linearly decreasing inertia weight))

IV. CONCLUSION AND PROSPECT

In this paper, the main goal is to achieve the combination of particle swarm optimization and clustering analysis, so as to optimize the clustering analysis algorithm and optimize it by introducing PSO algorithm. Finally, the improved algorithm is proved on the convergence curve is superior to the original algorithm. With the advent of large data age, large data analysis in the future will be more and more people's attention. The intelligent, diversity, good self-service checkability and adaptability of the group intelligence theory will bring great prospects for the common development of particle swarm optimization and clustering algorithm.

Acknowledgements

This work is supported by Natural Science Foundation of China(No.61640206); Chunhui Research Program from Ministry of Education (Z2016082,Z2015050)； China Scholarship Council (CSC:201608635010)；Youth Foundation of Qinghai University(No.2014-QGY-30).

References

1. Lect.Shital P.Bora.Data Mining and Ware Housing[C]//*3rd International Conference on Electronics Computer Technology(ICECT)*,Kanyakumari,2011,1:15.
2. M.Kholghi, H.Hassanzadeh, M.R.Keyvanpour, Classification and Evaluation of Data Mining Techniques for Data Stream Requirements[C]//*International Symposium on Computer Communication Control and Automation (3CA)*, Tainan, 2010, Vol.1: 474-478.
3. Song Z, Chen J, Zheng B. A Multi-Objective Optimization Evolutionary Algorithm with Multiple Indicators Enhanced[J]. *Journal of South-Central University for Nationalities* (Natural Science Edition), 201130389-93.
4. Jianguo ChenZhongshan Song Bojin Zhenget al. A Multi-objective Optimization Evolutionary Algorithm With Better Performances on Multiple Indicators *International Symposium on Intelligent Computation and Application ISICA* 2010201047-56.
5. Vanden Bergh F. An analysis of particle swarm optimizer[C]. *Proceedings of the 1998 conference of Evolutionary computation* Piscataway IEEE Press199867-63.
6. Luo Huilan, WEI Hui. Clustering Algorithm for Mixed Data Based on Clustering Ensemble Technique[J]. *Computer Science*, 2010(11):234-238.