Clustering Analysis of Fusion Similarity Calculation and Improved Genetic Algorithm

Yuan Yao[a], **and Jin Feng**[b,*]

School of Information, Beijing Institute of Technology, Zhuhai, 519088, China

[a]80269741@qq.com, [b]vonfengjing@qq.com

*Corresponding author

Keywords: fuzzy C-means clustering analysis; ST-DBSCAN; genetic simulated annealing algorithm; multi-population genetic simulated annealing algorithm; number of clusters.

Abstract: Aiming at the shortcomings of fuzzy C-means clustering analysis (FCM), which is easy to fall into the local minimum, and the sensitivity to the initial clustering center, this paper firstly used a density-based DBSCAN algorithm (ST-DBSCAN) to determine the number of clusters by calculating the distance and density between data. At the same time, based on the genetic simulated annealing algorithm (SAGA), this paper proposed a clustering analysis based on multi-population genetic simulated annealing algorithm. Firstly, it analyzed and evaluated FCM, and proposed the shortcomings of FCM in determining the number of clusters and clustering process. Then, it determined the number of clusters the ST-DBSCAN algorithm in the FCM. At the same time, it studied the genetic simulated annealing algorithm, and optimized the genetic simulated annealing algorithm by adding multiple groups of parallel genetic ideas. Finally, it combined FCM with a variety of genetic simulated annealing algorithms to optimize the clustering process. The experimental results show that the algorithm has better global search ability and convergence ability, and has certain advantages over traditional clustering algorithms in clustering effect and stability.

1. Introduction

Fuzzy C-means clustering (FCM) is a clustering method that uses the concept of Euclidean space to determine the geometric closeness of samples. It is a clustering method that automatically divides the sample into several different clusters and makes the samples within the same cluster similar, and determines the distance between these clusters. The FCM algorithm is an unsupervised learning technique for searching sample training sets in the field of big data and artificial intelligence. At the same time, since fuzzy C-means clustering is a local search method, its clustering results are easily affected by the initial clustering center. Therefore, if there is an abnormal point in the data set, there is a possibility of falling into a local minimum.

In view of the shortcomings of FCM, the literature [1] selected the point with the largest field density as the first clustering center by calculating the density of each point field in the data set. They designed an experiment that used five data sets to validate, and its results showed that the method had the effect of speeding up the operation. However, because genetic algorithms have the problem of easy premature convergence and backward fitness and low search efficiency, traditional genetic algorithms tend to stagnate in local optimal solutions and it is difficult to obtain global optimal solutions. Therefore, some studies have introduced simulated annealing algorithms into genetic algorithms [3], which have shown good performance.

These fuzzy C-means clustering optimization algorithms are studied separately from the perspective of optimizing the initial clustering center or optimizing the clustering process. However, the determination of the number of clusters in FCM is not clearly stated, and the clustering process can also be further improved. This paper first introduced the method of calculating the distance and density of data points into the FCM to determine the number of clusters of FCM. At the same time, this paper combined multiple groups of parallel genetic ideas with simulated annealing algorithms, and designed genetic coding methods and fitness functions according to the specific conditions of clustering problems. Finally, it introduced the combined algorithm into FCM to improve global search and convergence, optimized the clustering process of FCM and achieved better clustering effect.

2. Fuzzy C-means clustering analysis

Set the sample data to $X=\{x_1,x_2,\cdots,x_n\}$, and divide sample data into classes $C(2\le C\le n)$. At the same time, its cluster centers were represented as $\{v_1,v_2,\cdots,v_c\}$, the corresponding categories were represented as $\{A_1,A_2,\cdots,A_n\}$, and U was a similar classification matrix. The objective function is

$$J_b(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^b(d_{ik})^2 \tag{1}$$

Where μ_k is the membership of class A_k, and fuzzy C-means clustering algorithm requires

$$\sum_{j=1}^{c}\mu_j(x_i)=1\text{, } i=1,2,\cdots n \tag{2}$$

That is to say, the sum of the various degrees of membership is 1, through the following formula

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}}{d_{lk}})^{\frac{2}{b-1}}} \tag{3}$$

$$v_{ij}=\frac{\sum_{k=1}^{n}(\mu_{ik})^b x_{kj}}{\sum_{k=1}^{n}(\mu_{ik})^b} \tag{4}$$

By repeatedly modifying the cluster center and membership degree, the objective function was minimized in the process of repeated iterations, and achieved the purpose of fuzzy clustering of data.

3. Determination of cluster number based on similarity calculation

Most of the current optimizations for the FCM algorithm were based on the case where the number of clusters was known, or simply determined the number of clusters based on the product of density and distance. The fault tolerance of these methods was quite low, and the dependence on algorithm parameters was relatively high.

This paper used a density-based algorithm (ST-DBSCAN) to calculate the density of each point in the data set [4]. The algorithm has ability to find cluster centers based on the spatial, non-spatial, and temporal values of the object. First define two parameters of DBSCAN: ε, the field of the given radius of each point of the data set. $MinPts$, the minimum number of objects included in the sample field that could be the center of the cluster. Set sample data $X=\{x_1,x_2,\cdots,x_n\}$ to any $x_i\in U$. The distance between samples is

$$d_{ik}=d(x_k-v_i)=[\sum_{j=1}^{m}(\mathrm{x}_{kj}-v_{ij})^2]^{\frac{1}{2}} \tag{5}$$

The density of object A is

$$\rho_i=\sum_{j=1}^{n}\varphi(\varepsilon-d_{ij}) \tag{6}$$

where $\varphi(x)=\begin{cases}1,x\ge 0\\0,x<0\end{cases}$.

At the same time, based on the hypothesis of the literature [5]: The cluster center is a high-density data point compared to the surrounding neighbors and is far from the center of the other clusters. This can better calculate the distance of points in the cluster. Set the minimum distance for a larger density to δ_i, and the expression is:

$$\delta_i=\underset{\forall P_j\in U,\rho_j>\rho_i}{Min}d_{ij} \tag{7}$$

If data points have the largest local density, then equation (7) becomes:

$$\delta_i = \underset{\forall P_j \in U}{Max}\, d_{ij} \tag{8}$$

Through the above calculation of data density and distance, the similarity matrix between objects in the data set can be obtained. After setting the appropriate threshold and similarity parameters, analyze the statistical results of the data set could determine the number of clusters in the data set. After determining the number of FCM clusters, this paper will optimize the clustering process of FCM.

4. Multi-group parallel genetic algorithm based on simulated annealing

4.1 Multi-group parallel genetic ideas

Genetic algorithm (GA) originates from Darwin's Biological Evolution and Mendel's Genetics. It draws on the natural selection thought and natural genetic mechanism in evolution and genetics. It is a computational model of natural evolutionary system and a general global random search algorithm. Genetic algorithm has a good effect on improving the global search ability of FCM. Therefore, this paper introduces the algorithm to optimize the clustering process of FCM.

However, traditional genetic algorithms often stagnate in local optimal solutions and it is difficult to obtain global optimal solutions. This paper will improve the performance of genetic algorithms by combining other intelligent algorithms, and then optimize the clustering process of FCM. Among them, multi-group parallel inheritance [6] is a better method to improve the performance of genetic algorithms. It breaks through the traditional method of genetic evolution of genetic algorithm by a single group, introducing multiple populations to search simultaneously, and different populations are controlled by different control parameters (For example, each sub-population has different crossover probability P_c , mutation probability P_{m} , annealing cooling coefficient, etc.). Multiple sub-populations evolve independently. When the number of evolutions reaches S , the current optimal individuals are assigned to all sub-populations to achieve multi-group co-evolution.

In the multi-population parallel inheritance, the immigration operator is added to contact each independent sub-population. Introduce the best individuals that appear in the evolution of various groups into other populations to achieve information exchange between various groups. The comprehensive result of co-evolution of multiple populations is the optimal solution.

4.2 Combination of simulated annealing genetic algorithm and multi-group genetics

The simulated annealing algorithm (SA) originates from the simulation of solid annealing process by statistical physics [7]. It compares the actual problem optimization solution process with the solidification of the heated solidified solid in the heat balance to a regular crystal process. The algorithm accepts new solutions mainly through the Boltzmann standard rules. This rule adds the cooling coefficient as the termination condition of the algorithm, which finally gives an approximate optimal solution. The combination of simulated annealing algorithm and multi-population genetic algorithm can better avoid the premature phenomenon in genetic algorithm [8], and the two can complement each other to obtain better results.

1) Set the initial parameters of the algorithm: initial temperature T_i , cooling coefficient C_i , number of subpopulations M , evolution number S , crossover probability P_{ic} , mutation probability P_{im} .

2) Generate M initial population; initialize evolutionary algebra $\mathrm{loop1} = 0$, count variable $\mathrm{loop2} = 0$.

3) Repeat the following operations for sub-population i , until the number of sub-populations reaches M .

a) Evaluate individual fitness function within sub-populations $f(x_j), j = 1, 2, \cdots, 2N+1$.

b) Individuals x_j and x_k are randomly selected from the i population to perform cross operations, and calculate the fitness function values $f(x'_j)$ and $f(x'_k)$. If $\min\{1, \exp-(f(x'_j) - f(x_j)/T_i)\} > \mathrm{random}$, then receive x'_j as a new individual, and if $\min\{1, \exp(-(f(x'_j) - f(x_j))/T_i)\} > \text{random}$, receive x'_k as a new individual, where random is a random number over the $[0,1]$ interval.

c) After the crossover, the individual performs the mutation operation, and it is determined by step *b)* whether to accept the new individual. Repeat steps b) and *c)*.

4) $\text{loop1} \leftarrow \text{loop1}+1$, if $\text{loop2} < S$, then $\text{loop2} \leftarrow \text{loop2}+1$, go to step *3)*, otherwise proceed to the next step.

5) Obtain the best individual among all current sub-populations, and then select the current optimal individual and assign it to all the populations.

If the current total optimal individual satisfies the convergence condition, then return to the global optimal solution. If loop1 still does not reach the maximum number of evolutions, modify the annealing temperature, let $T_i \leftarrow C_i T_i (i = 1 \sim M)$ and return to step *3)*.

5. Cluster analysis based on multi-group genetic simulated annealing algorithm

Introduce the simulated annealing multi-group parallel genetic algorithm into fuzzy C-means clustering analysis. According to the requirements of FCM, design the chromosome coding method, fitness function, cross mutation operation and cooling operation in the clustering process to improve the global search ability and convergence of FCM and optimize the clustering process.

1) Chromosome coding method. At present, the research on chromosome coding mode in genetic algorithm [10] includes binary coding and floating-point number coding. This paper adopted the clustering-based floating-point coding method in coding mode. Divide each chromosome into c cluster centers by the number of cluster centers calculated based on the density and distance.

2) Fitness function. After selecting the cluster center, divide the sample vector $x_l = [x_1, x_2, \cdots, x_m]^T$ into classes of center c_i by calculating the Euclidean distance: $x_l = [x_1, x_2, \cdots, x_m]^T$. Define the objective function

$$J = \sum_{k=1}^{m} \sum_{x_k \in C_i} | x_k - c_i | \tag{9}$$

and fitness function

$$f = \frac{1}{1+J} \tag{10}$$

When the sum of the internal dispersions is small, the objective function value J is also smaller, and the value of the fitness function is also larger.

3) Consistent with the multi-group genetic algorithm in the crossover and mutation operations, the two parent parts are replaced by the crossover probability P_{ic} to generate new individuals. At the same time, the floating point number of each gene position is mutated by the mutation probability P_{im}, and replace the mutated gene by a random number.

4) Repeat steps 3), 4) in the multi-population genetic algorithm to obtain optimal individuals and assign them to various groups.

5) Temperature T_i decreases the control parameter of its value as the algorithm progresses. If $T_i < T_{end}$, the algorithm ends successfully and returns the global optimal solution; otherwise, $T_{i+1} = kT_i$ returns.

In summary, the algorithm performs a large cycle with the final performance index as the requirement, and controls the process of the multi-group genetic algorithm by the termination temperature of the simulated annealing algorithm. In the algorithm, calculate the density and distance to obtain the clustering number. At the same time, optimize the fuzzy C-means clustering by multi-group genetic algorithm to obtain the optimal result.

6. Emulation Experiment

This paper used the three data sets in Table 1 to test the proposed algorithm. These three data sets have classification tables, which can be used for performance evaluation.

Table 1. Description of the tested data sets.

Name of data sets	Sample quantity	Attribute quantity	Classification quantity
Iris	150	4	3
Wine	178	13	3
Ecoli	327	7	5

For the effect of the optimization process of this algorithm, this paper compared the traditional FCM algorithm with the algorithm of this paper. Taking the Iris data set as an example, the iteration number, cluster center and final objective function values of the two algorithms are shown in the following table.

Table 2. Comparison of FCM algorithm and the algorithm in this paper

Name algorithm	**Number of iterations**	**Cluster center**		**Objective function value**
FCM algorithm	3	5.5378,3.1190	3.0114,0.8850	61.1495
		6.2305,2.9381	4.7627,1.6106	
		5.3595,3.2098	2.5072,0.6774	
	6	5.0875,3.3501	1.7328,0.3513	45.7663
		5.3021,3.1248	2.5254,0.6735	
		6.3926,2.9304	5.1129,1.7934	
	9	5.0015,3.3902	1.4935,0.2516	29.1403
		5.8896,2.7848	4.3500,1.3823	
		6.6669,3.0281	5.5185,2.0194	
	12	5.0001,3.3891	1.4945,0.2520	29.1119
		5.9175,2.7940	3.3871,1.4002	
		6.6925,3.0368	5.5493,2.0348	
The algorithm in this paper	3	5.0013,3.3902	1.4852,0.2528	29.1182
		5.9183,2.7948	3.3876,1.4011	
		6.6933,3.0374	5.5498,2.0356	
	6	5.0007,3.3989	1.4851,0.2525	29.1178
		5.9180,2.7938	3.3873,1.4009	
		6.6927,3.0369	5.5495,2.0353	
	9	5.0002,3.3898	1.4949,0.2524	29.1152
		5.9159,2.7947	3.3870,1.4005	
		3.3927,3.0364	5.5493,2.0351	

It can be seen from the data in the table that as the number of iterations increases, the clustering center and objective function values of the FCM algorithm gradually become stable. From the beginning of 3 iterations, the objective function value was 61.1495 and it was reduced to 22.1119 after 12 iterations. The algorithm of this paper is close to the effect of 12 iterations of FCM algorithm in 3 iterations. Although the variation of the cluster center after 9 iterations was not very large, the algorithm was more stable and more effective than the FCM algorithm from a global perspective, and had a high fast convergence and better global convergence ability.

At the same time, in the aspect of the final clustering results, this paper adopted an external evaluation method F-measure [11]. This evaluation method defines the precision of cluster j and classification i .

$$\text{precision}(i, j) = \frac{N_{ij}}{N_i} \tag{11}$$

and recall rate

$$\text{recall}(i, j) = \frac{N_{ij}}{N_j} \tag{12}$$

The classified F-measure expression is

$$F(i)=\frac{2\times \text{precision}(i,j)\times \text{recall}(i,j)}{\text{precision}(i,j)+\text{recall}(i,j)} \tag{13}$$

and the weighted arithmetic mean is

$$F_p=\frac{\sum(|i|\times F(\mathrm{i}))}{\sum|i|} \tag{14}$$

Where $|i|$ is the number of individuals classified. This paper tested the three data sets for 100 times and got the mean of the F-measure, whose results were shown in table 3.

Table 3. Means of F-measure.

Name of algorithm	Iris	Wine	Ecoli
FCM	0.8930	0.4718	0.5845
GASA-FCM	0.9001	0.6406	0.6254
The algorithm in this paper	0.9134	0.6913	0.6544

When acting to the second data set, it can be seen that when the number of classification is consistent, the bigger the sample quantity and attributes are, the more obvious the superiority of the algorithm is. The main reason is that the clustering algorithm has benefited from the combination of the multi-population parallel genetic idea and the simulated annealing algorithm, which improves the global search capability of the clusters and solves the problem of converging to local optimum efficiently, and finally, it improves the clustering effects.

7. Conclusion

The emulation experiment results of the data set show that the algorithm in this paper has an advantage over the traditional FCM algorithm in terms of solving clustering problems. This algorithm shows better clustering effects and stability.

Genetic algorithm is an optimization algorithm which is applicable to those complicated problems with difficult object models construction and large search space, whose openness determines its great breakthrough in computational capability when combined with other algorithms. This paper combines the multi-population parallel genetic idea with the simulated annealing algorithm and then with the FCM algorithm, taking an advantage of the global search capability of the simulated annealing algorithm and the rapid convergence of the FCM algorithm, which brings this algorithm a better global search capability and much more rapid convergence.

Acknowledgments

In terms of funding, this work was supported by the Provincial Key Platforms and Major Scientific Research Projects in Guangdong Universities (2017KQNCX247), and National College Student Innovation Training Program Funding Project (201813675001).

In terms of article, the translation and proofreading of this paper was supported by Ms. Lin Jinqi.

References

[1] Y. Liu, B. Kang and T. Hou, Optimized Initial Value C-means Algorithm, Journal of Jilin University (Engineering Edition), vol. 48, no. 6, pp:306-311, 2018.

[2] Z. H. Wu, Research on Clustering Method Based on Genetic Algorithm, Shandong Normal University, 2006.

[3] X. M. Wang and Y. H. Wang, Combination of Simulated Annealing Algorithm and Genetic Algorithm, Chinese Journal of Computers, pp. 381–384, 1997.

[4] D. Birant and A. Kut, ST-DBSCAN: An algorithm for clustering spatial–temporal data, Data &

Knowledge Engineering, vol. 60, no. 01, 2006.

[5] R. F. Bie, R. Mehmood, S. S. Ruan, Y. C. Sun and H. Dawood. Adaptive fuzzy clustering by fast search and find of density peaks. Personal and Ubiquitous Computing, vol.20, no. 05, pp. 785-793, 2016.

[6] Y. H. Zhou, Y. C. Lu and C. Y. Shi, AdaptiveParallel Genetic Algorithm Based on Overcoming Premature Convergence, Journal of Tsinghua University (Natural Science Edition), pp. 785-793, 1998.

[7] L. S. Kang, Non-Numerical Parallel Algorithm——Simulated Annealing Algorithm, Beijing: Science Press, pp. 22-25, 1997.

[8] L. J. Yin, L. J. Yang, M. M. Hu and Y. C. Deng, Fuzzy C-means Clustering Algorithm based on Hybrid Genetic Simulated Annealing, Journal of Hubei Automotive Industry Institute, vol. 29, no. 03, pp. 62-65, 2015.

[9] H. Y. Wu, B. G. Chang, C. C. Zhu and J. H. Liu, Multi-group parallel genetic algorithm based on simulated annealing mechanism, Journal of Software, no. 3, pp. 416-420, 2000.

[10] H. Q. Mo, Research on Genetic Algorithm Search Ability and Coding Mode, South China University of Technology, 2001.

[11] Y. Yang and F. Jin, K. Mohamed, Summary of Clustering Effectiveness Evaluation, Application Research of Computers, pp. 1630-1632+1638, 2008.

[12] Z. H Wu, G. J. Zhang and X. Y. Liu, Cluster Analysis Based on Simulated Annealing Genetic Algorithm, Application Research of Computers, pp. 30-32, 2005.

[13] Z. H. Wang, Z. J. Liu and D. H. Chen, Research on Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization, Computer Science, vol. 39, no. 09, pp. 166-169, 2012.

[14] Z. M. Zhang, New Data Clustering Algorithm Combined of Ant Colony Algorithm and Improved Fuzzy C-Means Algorithm, Proceedings of 2016 International Conference on Communications, Information Management and Network Security (CIMNS2016), 2016.

[15] H. E. Assaad, A. Samé, G. Govaert and P. Aknin, A variational Expectation–Maximization algorithm for temporal data clustering, Computational Statistics and Data Analysis, 2016.

[16] L. C. C. Heredia and A. R. Mor, Density-based clustering methods for unsupervised separation of partial discharge sources, International Journal of Electrical Power and Energy Systems, 2019.