Research and Application on Fuzzy Clustering Based on Genetic Algorithm

ZHANG YuHong[1, a]

[1]Department of Basic Education，City Institute，Dalian University of Technology，Dalian 116000，People's Republic of China

[a]518568@qq.com

Keywords: fuzzy clustering; genetic algorithm; information entropy; FCM

Abstract. As for shortages such as wiring and expansion in traditional fire alarm system, under condition of lacking of wireless fire early-warning product, this paper applies wireless sensor network technology and multi-sensor fusion technology to design one wireless fire probability classified early-earning system. This system can adopt ZigBee technology to quickly establish star early-warning network, as well as gives early warning of fire occurrence tendency according to probability etc, so that it can adapt to fire early warning and alarm requirement in various of complicated environment.

1. Introduction

Fuzzy clustering method has advantages of simple, easy and better clustering effect, it has universal application in plenty of actual problems, fuzzy C-average and clustering algorithm have more universal applications in fuzzy clustering algorithm, and however, FCM clustering algorithm has some problems, such as confirmation of clustering number, option of initial clustering center, easy to fall into local minimum etc. It is especially proper for non-linear function optimization problem, as for problems of clustering number confirmation, this paper puts forward one kind of method based on average information entropy, it adopts density functional method to confirm initial clustering center, it has advantage such as quick convergence speed and precise classification etc compared with the traditional fuzzy clustering method[1].

2. Fuzzy clustering method

Set finite limit X={x_1，x_2，…，x_n}，of which, $x_j \in R^p$，j=1，2…，n. FCM adopts square error and function as clustering criterion function:

$$J = \sum_{i=1}^{m} \sum_{j=1}^{n} (d_{ij})^h \left\| x_j - v_i \right\|^2 \quad (1)$$

In the formula，n is sample number，m is the given category number, and 1<m<n，h is weighted power exponent, vi is the clustering center of the i category, d_{ij} is the i category degree of sample j，which meets $\sum_{i=1}^{m} d_{ij} = 1, j = 1,2,...n$ .FCM algorithm requires is the clustering result of making J reach the minimum，makes respectively corresponds to J d_{ij} ，v_i is derivation，command their differential coefficient as 0 and substitute condition $\sum_{i=1}^{m} d_{ij} = 1$， the answer is as follows：

$$\_ij = \frac{\left[\frac{1}{\left\|x_j - v_i\right\|^2}\right]^{\frac{1}{h-1}}}{\sum_{k=1}^{m}\left[\frac{1}{\left\|x_j - v_k\right\|^2}\right]^{\frac{1}{h-1}}}, i = 1,2,..., \ m; j = 1,2..., \ n \quad (2)$$

$$V_i = \frac{\sum_{j=1}^{n}(d_{ij})^h x_j}{\sum_{j=1}^{n}(d_{ij})^h}, i = 1,2,..., m \quad (3)$$

FCM algorithm is completed through iteration of formula (2) and (3), when J converges to the minimum, it gets the final clustering center and so that it gets the final clustering result.

3. Fuzzy clustering analysis based on genetic algorithm

A. Confirmation of A clustering number, entropy is used to describe disorder degree of atomic distribution, while in the information theory, information entropy is the information amount measurement contained by certain information sent out by information source, when information sent out by certain information source becomes more confirmed, information entropy of this information source is smaller. Distribution of data point is similar to atomic distribution, according to information entropy theory when clustering division becomes much more reasonable and belongingness of data point on certain clustering becomes more confirmed, information entropy of this clustering is much smaller[2]. Based on the above-mentioned information entropy theory, it uses average information entropy as standard to measure clustering number. Firstly it should confirm expected clustering number range[m_{min}， m_{max}].

$$H(k) = -\sum_{i=1}^{k}\sum_{j=1}^{n}\left[\frac{d_{ij}\log_2(d_{ij}) + (1 - d_{ij})\log_2(1 - d_{ij})}{n}\right] \quad (4)$$

In the formula，d_{ij} is the i clustering degree of sample j，$d_{ij} \in [0, 1]$，i，j=0，1，…when k increases from m_{min} to m_{max}， it will produce m_{max}-m_{min}+1 H(k)， it chooses clustering umber k corresponds to one minimum H(k)as the final clustering number m. When clustering number is gradually changing, the membership degree of every data is also gradually changing, according to information entropy theory, when average information entropy becomes smaller, which indicates membership degree of every data belongingness category. According to information entropy theory, when average information entropy becomes smaller, which indicates the membership degree of data belongs to this category, meanwhile, it indicates clustering number division is more reasonable. The basic steps of using average information entropy to confirm optimal clustering number are as follows: ①set the maximum clustering number m_{max} and the minimum clustering number m_{min}，threshold value X，and set k=m_{min}-1；②random initial membership matrix $U^{(t)}$，t=0，k=k+1；③update membership matrix $U^{(t)}$ and clustering center $V^{(t)}$，t=t+1；④when $|J^{(t)}-J^{(t-1)}|>X$，back to ③；⑤calculate H(k)，write down clustering number k at this time. If H(m)=0，then H(m)=H(k)；if H(m)>H(k)，then $H_m(x)=H_k(x)$，it uses current k value to update m value；if k>m_{max}，then m is the final clustering number, otherwise back to②。

B.Option of initial clustering center. In the initialization of FCM algorithm, it dose not only include initialized clustering number, but also include confirmation of initial clustering center. The above has already used information entropy to confirm optimal clustering number, the following uses

one kind of probability density to choose initial clustering center. It defines density function of sample point as follows:

$$D_i^{(0)} = \sum_{j=1}^{N} \frac{1}{1 + f_d \left\| x_i - x_j \right\|^2} \qquad (5)$$

In the formula，fd=4/r^{2}d，rd is the adjacent density effective radius, its option should be related to distribution characteristic of data set. Take rd as 1.2 times of N sample root-mean-square distance, it is as follows:

$$r_d = \frac{1}{2} \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{i=1}^{N} \left\| x_i - x_j \right\|^2 \qquad (6)$$

From formula (6)it can be known that sample point become more dense around x_i, then $D_i^{(0)}$ value becomes bigger，so it is used to reprsent density degree of sample point in sample space. Command D_1^* =max{$D_i^{(0)}$，i=1，2，…，N}，corresponding x_1^* is the first initial clustering center, then density function adjustment relation formula of the k iteration is as follows:

$$D_i^{(k)} = D_i^{(k-1)} - D_k^* \frac{1}{1 + f_d \left\| x_i - x_k^* \right\|^2}, k = 1,2,..., m-1 \qquad (7)$$

In the formula，m is clustering number, command D_k^* =max{$D_i^{(k-1)}$，i=1，…，N}，the corresponding sample point x_k^* takes central position of the k initial clustering. Formula (5),(6),(7)determines central initialized method.

C. Fuzzy clustering analysis based on genetic algorithm

Genetic algorithm, its short term is GA, which is based on evolution principle of survival of the fittest, survival of the fittest by Darwin it repeatedly uses basic operation of genetics on the population including possible solution, gradually generates new population and make population gradually evolve, meanwhile it uses overall parallel search technology to search the optimal individual in the optimized population, so that it gets optimal solution meets requirement. The basic operation of genetic algorithm mainly includes the following: option, intersection and variation. But except for these 3 basic operation, it also includes affiliated operation, such as coding, adaptability evaluation etc. The iteration step of fuzzy clustering analysis based on genetic algorithm is as follows: ① initial parameter h=1.1; ②initial population, parameter implements binary coding; ③calculate the real number corresponds to every individual in population ④ randomly generates classification population；⑤ it use FCM method to calculate category center and new classification matrix, of which, the smaller of X, result is much more precise ⑥it calculates fitness value corresponds to population individual according to target evaluation function f(U，V)=T_S(m)+D(m) ⑦it adopts rotating disk method to implement option operation of genetic algorithm ⑧it adopts single point intersected operation to implement intersected operation of genetic algorithm ⑨it implements variation operation to judge whether reaches termination condition or not, if it reaches termination condition, implement step 10, otherwise jump to step ③. Triangle membership function is one usual membership function in fuzzy modeling and control, the shape of this membership function is only related to straight line slope, which belongs to linear membership function. When using software to design fuzzy controller, linear membership function is easy to be realized, operation time is short, and it is superior to curve membership function on real-time, it is used more in self-adaptation fuzzy identification and control of membership function online adjustment. Figure 1 is triangle membership function on fuzzy division c=5, here it adopts diagonal triangle membership function[3].

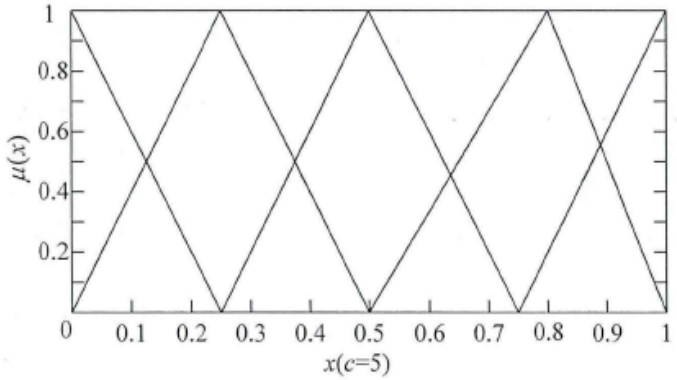

Figure 1 Triangle membership function

4 Algorithm analysis and design

Fuzzy clustering has outstanding outcome on sample grouping; genetic algorithm has mature application in aspect of partial solution. Firstly it should make fuzzy grouping on all the samples, in order to avoid genetic sequence excessively complicated, it set membership degree independent point, as for every group and independent point, every group member respectively makes genetic algorithm calculation, and integrates result into one complete route.

A. Algorithm frame, the best planning algorithm frame based on genetic algorithm and fuzzy clustering.

B. Algorithm realization

Fuzzy grouping. It sets membership degree: every point has its belonging membership degree as for every randomly generated central point, set one membership degree, if it exceeds this membership, then distribute it to its corresponding central point and becomes to be group member, if it exceeds this membership degree, it becomes to be isolated point, it dose not belong to any group.

(1) Grouping number. As for test data (30 points), it works out 2 conditions: ①group number must les than or equalto6, because excessive group number will make group member become smaller, so it place relatively complicated genetic algorithm calculation problem to the independent calculation of outer group, the test demonstrates that if group number is bigger than 6, it will cause complication increase in time and calculation, reduction in effect. ②group member is must bigger than 2, if group member is 1, it should become to be isolated point. If grouping result dose not conform to the above 2 conditions, then it needs grouping again until it conform to the conditions.

(2) Membership degree and isolated point. Every point has its membership as for every central point; one has high membership it belongs to this group, all the membership value of every point equals to 1 when added together. This paper sets membership degree as variable variables, it is set as 0.6、0.7、0.8 to implement test. If membership of certain point is less than the set threshold value, this point will become to be isolated point.

B. Genetic algorithm

(1) Gene coding. This paper uses simulated test in table 1 to make gene coding, X_0 indicates the first point, X_1 indicates the second point……, corresponding value of the other points are indicated by the following table 1:

Table 1 Simulated test data

Point	Point position	Point	Point position	Point	Point position
X_0	(0000，0000)	X_{10}	(2233，7796)	X_{20}	(2233，7796)
X_1	(8111，1465)	X_{11}	(9529，3269)	X_{21}	(9625，1993)
X_2	(9709，7792)	X_{12}	(6452，6792)	X_{22}	(3371，0456)
X_3	(9676，1659)	X_{13}	(5183，4073)	X_{23}	(1157，8598)
X_4	(0882，9845)	X_{14}	(5304，4581)	X_{24}	(0235，1155)
X_5	(9728，3648)	X_{15}	(8237，8768)	X_{25}	(5128，8521)

X_6	(1356，0110)	X_{16}	(8695，7320)	X_{26}	(5439，0968)
X_7	(0086，8410)	X_{17}	(6881，1740)	X_{27}	(2619，3025)
X_8	(0713，6392)	X_{18}	(7452，2020)	X_{28}	(2923，9794)
X_9	(7829，6025)	X_{19}	(8306，9441)	X_{29}	(7746，8235)

Suppose the generated initial string format is $\{X_0, X_1, X_2, X_3, X_4\}$, this paper implements genetic algorithm calculation on this string.

(2) Mating system. This paper uses the best string copy to the nest generation, and Fitness function will determine the mating number of every string. Because it has chosen its entrance and retreat point for every group after fuzzy grouping (the first point and the last point in this group), so when it makes genetic algorithm calculation on every group member, it can not move the first point and the last point[4].

(3) Fitness function, definition of fitness function in this paper is as follows:

$$f = \underset{x}{Min}(\sum_{i=1}^{n-1} D(x_i, x_{i+1}) + D(x_n, x_0))$$

Of which，x indicates the input string, n indicates all the points position of x, D indicates distance between every point, x_i indicates point position in string.

(4) Variation condition. Variation is random event; the probability of variation can not be too excessive. The probability of variation defined by this paper is 10%, if the entrance and retreat point of every group (the first point and the last point in group) is confirmed, then it dose not participate in variation, the rest points makes random exchange.

5 Conclusions

This paper establishes one kind of new fuzzy C-average (FCM) clustering algorithm based on overall search ability of genetic algorithm. The main research work includes the following: firstly, considering the past clustering analysis result affected by clustering number and initial clustering center, this paper puts forward method based on average information entropy to confirm clustering number, secondly, it constructs a new target function, thirdly, it applies genetic algorithm to get center and membership value of every category, and then it applies fuzzy C-average to make iteration solution[5].

References

[1]Tang Chenghua，Liu Pengcheng，Tang Shensheng，Xie Yi. Abnormal Invade Behavior Test of Fuzzy Clustering Analysis Based on Characteristics Option[J].Computer Research and Development，2015，03：718-728.

[2]Zhang Yongku，Yin Lingxue，Sun Jinguang. Fuzzy Clustering Algorithm Based on Improved Genetic Algorithm [J]. Journal of Intelligent Systems，2015，04：627-635.

[3]Fan Yu. Fuzzy Clustering of Optimization Genetic Algorithm in Image Segmentation [J].Electronic Testing，2013，05：279-281.

[4] Wu Gang，Xu Feng，Wang Bing，Xu Sizhe. Improved BP Neural Network Wind Speed Prediction Based on Genetic Algorithm[J].Electronic Design Engineering，2016，11：120-123.

[5]Zhu Changjiang，Chai Xiuli. Research and Application on Fuzzy Clustering Based on Improved Genetic Algorithm [J].Science Technology and Engineering，2013，10：2863-2866+2870.