The New Method Definite Initial Cluster Center for Fuzzy Risk Clustering Neural Networks

Kaiqi Zou Jie Cui

College of Information Engineering, Dalian University, Dalian 116622, P. R. China

Abstract

Neural network has a powerful parallel processing capability, along with its rise in the fuzzy risk cluster analysis which has occupied an important position, however, the quality of fuzzy risk clustering results is influenced by the initial value of options. The initial cluster center method for fuzzy risk clustering neural network, based on density and grid method, automatically determine the number of clusters and the initial cluster centers. Compared with the classical method simulation FCM, we can see a valid and effective method that can effectively speed up the convergence.

Keywords: Fuzzy clustering, Initial cluster center, Fuzzy risk clustering neural networks

1. Introduction

Fuzzy cluster analysis is applied in a good many domains and obtains satisfactory effects and objective benefits[1]-[4]. At present, there are three prime implement approaches, which are alternate optimization, evolutionary algorithm and neural network. In the above methods, neural network play a crucial part. However, like most non-linear optimization problem, fuzzy risk clustering results become bad when we choose an inappropriate initial values, also it maybe lead to a convergence of the results not expected minimums or clustering process at a very slow speed of convergence. Fuzzy risk clustering optimization of the process, the objective function of each local minima have an absorption surrounding jurisdictions, If the choice of the initial value of the absorption jurisdictions distance attractor close location, optimization process quickly converge to the extreme point; Instead, the very slow pace of convergence. If the initial value of the domain of attraction, the fuzzy risk optimization process is likely to converge to the other local minimum points.

Aimed to solve the above problem, such many techniques and means are applied, currently, the comparatively appropriate centers of clustering appear localities which it bears the comparatively centralize denseness. Based on the foundational standpoint, Yager and Filev bring a method of ascertaining cluster center which is called “the mountain method”. However, the compute quantity of this method will be exponential increase as the samples’ dimensions increasing [5]. Chiu improves on the mountain method, and computes the influence of every sample instead of the grid points, hence this improved method eliminates the relation between compute quantity and dimensions of samples. Nevertheless due to several parameters chosen reasonable, it appends a certainty difficulties to application [6]. In recent years, based on density and grid method is widely used to determine the cluster neural network initialization. The author, in paper [7], using the method which bases on grid and density, optimizes the clustering neural networks, and he considers the values of each dimension as factitious input parameters when plotting the grid, consequently the problem fells into the new optimization, then affects the function in practice.

Based on the above groundwork, this paper proposes a new initial method of fuzzy risk clustering neural network, which can automatically acquire the number of clustering and the initial the center of clustering, according to the data of risk sample, based on grid and density method. By the emulator on computer, the method is available and correct, and can effectively optimize the fuzzy risk clustering neural network.

2. Problem description and interrelated conceptions

Suppose that $S=(X_1,X_2,\cdots,X_n)$ is N -dimension space, therein samples set about risk data $D=\{v^1,v^2,\cdots,v^N\}$ is the n -samples set, and $v^i=(v_1^i,v_2^i,\cdots,v_n^i)$ denotes the i -th sample, $(\forall i,j)(v_j^i \in X_j)$, therein $1\le i\le N, 1\le j\le n$. In this paper, it becomes the first resolved question how

to automatically acquire the number of interval in each dimension, then gain the number of categories (or the number of clustering and the initial center of clustering, according to the given data of samples. We suppose that the change region of variable on each dimension is $[l_j, h_j]$, $j = 1, \cdots, n$ .The number of intervals on each dimension is denoted n_j , and $subset\_n_j^h = [left\_l_j^h, right\_h_j^h]$ denotes the h-th sub-interval on j-th dimension, where $left\_l_j^h$ and $right\_h_j^h$ stand for the values of left and right ends.

Definition 1 Based on density method, acquire the number of intervals on each dimension in space S and the opposite interval change range which stands for $subset\_n_j^h$ .Dividing the whole space S into *part*, where the value of *part* is $n_1 \times n_2 \times \cdots \times n_n$, let $U_{(x_1, x_2, \cdots, x_n)}$ stand for a grid cell constructed by *n* intervals, each which is the x_j -th interval on j -th dimension, therein for every integer j with condition $1 \le j \le n, x_j \in \{x | x \in Z, 0 \le x \le n_j\}$.

Definition 2 The density of grid cell $U_{(x_1, x_2, \cdots, x_n)}$ is that:

$$D\left(U_{(x_1, x_2, \cdots, x_n)}\right) = C_{(x_1, x_2, \cdots, x_n)} / V_{(x_1, x_2, \cdots, x_n)} \qquad (1)$$

Where $C_{(x_1, x_2, \cdots, x_n)}$ denotes the number of samples in grid cell $U_{(x_1, x_2, \cdots, x_n)}$, and $V_{(x_1, x_2, \cdots, x_n)}$ means the volume of grid cell $U_{(x_1, x_2, \cdots, x_n)}$. The under equations (2) and (3) are the different pattens.

$$V_{(x_1, x_2, \cdots, x_n)} = \prod_{j=1}^{n} subset\_n_i^{x_i} \qquad (2)$$

equally:

$$V_{(x_1, x_2, \cdots, x_n)} = \prod_{j=1}^{n} \left(right\_h_i^{x_i} - left\_l_i^{x_i}\right) \qquad (3)$$

Definition 3 grid cell $U_{(x_1, x_2, \cdots, x_n)}$ is dense one if and only if $D\left(U_{(x_1, x_2, \cdots, x_n)}\right) \ge \tau$, where value of density threshold τ is decided by provided situation.

3. The initial method of clustering neural networks

3.1 Method of initial cluster center

There are many methods of initial cluster center, comparatively common one is that choosing random C vectors as initial center to cluster in data samples characteristic space R^s or selecting the C , what is the number of cluster, samples as initial centers. But the above method makes cluster results fall into probability of global optimum less, meanwhile requires to confirm the cluster number C , however in practice, the cluster number can not be aware in advance, it is always directly decided by the distributing of samples, and also the distributing of samples are mutative.

Aimed to the above problems, this paper proposes a method to affirm the cluster centers and cluster number for fuzzy risk clustering neural network, according to the fact distributing of data samples. There is a method called CLIQUE (Clustering In QUEst) in data mining domain[8]. In this method a property is that if k-dimension is dense, then it's projection on $k-1$-dimension must be dense. In other words, given the k -dimension dense candidate cell, if it's result of rejection on $k-1$ -dimension is not dense, then we surely induce that it must be not dense. According to this theory, we can use the density method to gain n_j , which is the number of intervals on each dimension and $subset\_n_j^h$, which stands for interval ranges on each dimension. Suppose the number of dense grid cell is C , we can say that the available number of cluster is C , then search relevant cluster centers in every dense grid cell as initial cluster centers for fuzzy risk clustering neural networks, by means of fuzzy k -means clustering algorithm.

3.2 Design the algorithm of initial centers

DBSCAN (Density-based Spatial Clustering of Application with Noise) is a clustering algorithm based on density. Merging condition to approach to arbitrary-shaped clusters is designed carefully so that a single threshold can distinguish correctly all clusters in a large spatial dataset though some density-skewed clusters live in it. In this method, a group of points set, which has the linked dense, is called a cluster.

For the sample set D in n -dimension space S , we find out the cluster centers and interval ranges on each dimension by applying the DBSCAN. On each dimension, the primary algorithms are like the below, where we take the j -th dimension for example.

Step 1. Make the every sample standard which is data v_j on j -th dimension, in other words, let all data reject into given interval which is from 0 to 1, including 0 and 1.

Step 2. Sort data which is denoted v_j on j -th dimension.

Step 3. Select the neighborhood radius which is denoted ε and Min_num which stands for the number of least points in neighborhood. Generally speaking, value of Min_num can follow the below formula (4):

$$1 < Min\_num = \mathrm{int}(\varepsilon N) < N, (0 < \varepsilon < 0.5) \quad (4),$$

Here, $\mathrm{int}(\varepsilon N)$ is regarded the integer part of εN as the value.

Step 4. Select the element v_j^i in turn, then check up whether the number of elements in ε 's neighborhood, ultimately seek the biggest set that the density is conjoint. Repeat Step 4 until it is the last element.

Step 5. After Step 4, we can get n_j sub-intervals $subset\_n_j^h$ with conditions $1 \le h \le n_j, 1 \le j \le n$ on j -th dimension.

From the above research, we can induce that the method of initial cluster centers mainly includes the below five steps:

First, adopting method of DBSCAN, we can get n_j, the number of cluster, and $subset\_n_j^h$, the intervals on each dimension.

Second, according to equation (1) in Def. 3, calculate the density of each gird cell $D(U_{(x_1,x_2,\cdots,x_n)})$.

Third, judge every grid cell whether a dense one or not, according to Def. 3, in other words, judge the grid cell whether a available cluster or not.

Fourth, after executing third, we can receive value of C, which is the number of available cluster.

Fifth, for C available cluster subset, we adopt k -means clustering method to find C cluster centers.

After executing the above five steps, we have been automatically got the number of cluster and relevant cluster centers, according to data of samples, meanwhile, we get ready for initial cluster neural network.

3.3 The optimized algorithm of fuzzy risk cluster neural network

When designing the competitive neural network, we set the number of nerve cell on output layer stand for the number of cluster. Unfortunately, the cluster number is always uncertain. Hence the tentative method is commonly adopted to design the structure of neural network, and it becomes necessary to store several redundant nerve cells, then it affects the convergent speed and the validity of cluster. However, we can be able to optimize the structure of neural network after the pretreatment for samples, but also enormously short training time through decreasing training degrees because of the initial power values. The idiographic processes can be divided the four below steps.

Step 1. Adopt the cluster algorithm based on grid and density to acquire cluster centers and number of cluster which is denoted C.

Step 2. According to cluster number C, design the structure of neural network and initial the power values of neural network.

Step 3. Input the set composed by samples, then cluster it.

Step 4. Output the result of cluster.

4. Experiments and emulators

In order to affirm the algorithm available, we will take a group samples on 2 -dimension space in Tab. 1 for example, with conditions $N = 100$, $\tau = 0.1$ and $Min\_num = 10$.

Sample $v^i (1 \le i \le 100)$	X of v^i	Y of v^i
v^1	200	150
v^2	135	140
⋮	⋮	⋮
v^{100}	120	130

Tab. 1: Part of data of risk parameters

After making all data in Tab.1 standard, using the method proposed in this paper, we can get the result like Figure 1. In Fig. 1, in order to prove the method available, we let star ($*$) stand for initial cluster center given by new method and plus sign (+) created by classical method – FCM (Fuzzy C -Means). Through observing the below picture, we can clearly and confidently make the conclusion that the algorithm is available.

5. Conclusions

We can induce the below conclusion from the experiment and emulator, it is feasible and available that the initial cluster centers for fuzzy risk cluster neural network. In data analysis domain, propose a disposal method, aiming to data needed disposed to cluster objectively, which effectively shorts the

learning time of neural network. As the threshold needed to artificially set, so the density of the wide disparity between different cluster may not receive satisfying result, even losing some cluster, then the next step is to study the use of the relative density, to eliminate a given threshold value caused by the impact.

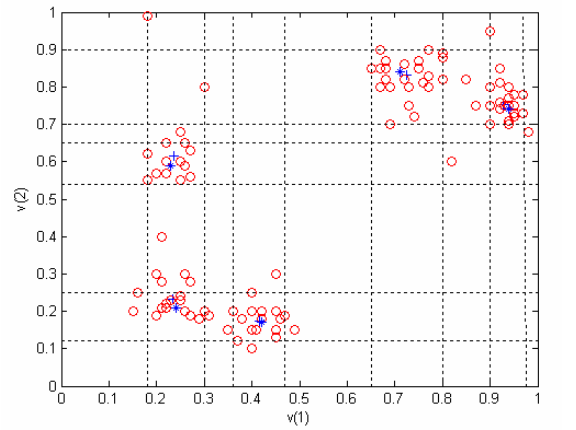

Fig. 1: Fuzzy risk clustering result compare to FCM.

Acknowledgement

This work is partially supported by National Nature Science Foundation of China (Grant No. 60573072).

References

[1] S. Chen and G. J. Gibson, Adaptive channel equalisation using a polynomial-perceptron structure. Proc. Inst. Elec. Eng., 137(5): 257~264, 1990.

[2] T.A. Runkler and R.H. Palm, Identification of nonlinear systems using regular fuzzy c-elliptotype clustering. FUZZ-IEEE'96, 1026~1030, 1996.

[3] J.C. Bezdek, Castelaz P F. Prototype classification and feature selection with fuzzy sets. IEEE Trans. SMC, 2(7): 87~92, 1977

[4] J.Z. Liu, A Fuzzy Clustering Method for Image Segmation Based on Two-Dimensional Histogram. Acta Electronica Sinica ,20(9) : 40~46, 1992.

[5] R.R. Yager and D.P. Filev, Approximate clustering via the mountain method. IEEE Trans. On SMC, SMC-24(8): 1279~1284, 1994.

[6] S.L. Chiu, Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy System, 2(3): 267~278, 1994.

[7] G.N. Deng, K.Q. Zou,Optimizing the structure of clustering neural networks based on grid and density. Journal of Harbin Engineering University, Vol. 27 Supp1, 2006.

[8] M. Zhu, Data Mining. The Publish of University of Science and Technology of China, 2002.