International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-4 Issue-13 December-2013

A Selective Fuzzy Clustering Ensemble Algorithm

Kai Li[1], Peng Li[2]

Abstract

To improve the performance of clustering ensemble method, a selective fuzzy clustering ensemble algorithm is proposed. It mainly includes selection of clustering ensemble members and combination of clustering results. In the process of member selection, measure method is defined to select the better clustering members. Then some selected clustering members are viewed as hyper-graph in order to select the more influential hyper-edges (or features) and to weight the selected features. For processing hyper-edges with fuzzy membership, CSPA and MCLA consensus function are generalized. In the experiments, some UCI data sets are chosen to test the presented algorithm's performance. From the experimental results, it can be seen that the proposed ensemble method can get better clustering ensemble result.

Keywords

Clustering ensemble, Fuzzy membership, Selection of members, Hyper-graph

1. Introduction

Clustering is one of the most important tools for data analysis. It groups data objects into multiple classes or clusters, wherein data have much high similarity in same cluster whereas they have great difference in different cluster. One of the most commonly used methods is k-means clustering. Since Zadeh proposed the concept of fuzzy in 1965, some researchers began to study fuzzy clustering and applied in various fields, for example, image processing pattern recognition, medical, etc. However, fuzzy clustering algorithm has some drawbacks. In order to solve these problems, some researchers proposed the improved fuzzy clustering algorithms by introducing entropy into objective function.

This work is support in part by Natural Science Foundation of China under Grant 61375075 and Nature Science Foundation of Hebei Province under Grant F2012201014 and F2012201020.

Kai Li, School of Mathematics and Computer, Hebei University, Baoding, China.

Peng Li, School of Mathematics and Computer, Hebei University, Baoding, China.

For unifying these fuzzy clustering algorithms, Li et al. [1, 2] proposed a unified objective function based on generalized entropy. They used multi-synapse neural network and joined the augmented Lagrange multipliers method instead of Lagrange multipliers to solve optimization problem with generalized entropy's objective function and presented fuzzy clustering algorithm GEFCMNN which refers to some parameters such as fuzzy index *m*, generalized entropy index *a* and ratio coefficient *b*. It is well known that each clustering algorithm has its own advantages and disadvantages. A single clustering algorithm is impossible to have a good clustering result to all datasets [3] and it is very difficult to choose an appropriate algorithm for a given data set [4]. Thus, clustering ensemble appears to be an effective method to overcome these limitations and it can improve the robustness and stability of the clustering result. The concept of clustering ensemble was first appeared in Strehl and Ghosh's paper[5], and they proposed three consensus functions based on hyper-graph. After that, many clustering ensemble algorithms have proposed [6-11]. In addition, some researchers are interested in the selection of the cluster members. Zhou et al.[12] proposed a clustering ensemble selection method based on mutual information. And Fern et al.[13] proposed three clustering ensemble selection algorithms which are JC, CAS and Convex Hull, respectively. Hong et al.[14] proposed re-sampling based selective clustering ensembles. Moreover, Yang et al.[15] studied the cluster members weighting problem. This paper is organized as follows. In section 2, we describe the commonly used ensemble methods. In section 3, cluster members' selection method is presented in detail. In section 4, we choose some commonly used datasets from UCI to test the algorithm's performance and compare with other algorithms. In the final section, we give some conclusions.

2. Ensemble method

Up to now, many clustering ensemble methods have been proposed. In the following, we briefly describe them. For a given dataset, we may use all data or partial data to obtain different clustering result. In addition, Yu et al. [16] pointed that choosing a subset

of features can get different clustering result in order to improve clustering ensemble accuracy. In the process of clustering, it can be used a single clustering algorithm with different parameters that run many times to generate the clustering results. To solve the single algorithm's some drawbacks; researchers used different clustering algorithms to generate the clustering members. The clustering members are usually some integer values which represent different class. In clustering ensemble, class labels need to be unified. For example, Minaei-Bidgoli et al. [17] studied comparison of different ensemble method. However, for the fuzzy clustering, we may directly use fuzzy membership to combine these different clustering results. In addition, researchers presented some selection method or weighted method for clustering members. For selection method, some people used mutual information as the selection criteria. And Yu et al.[18] used clustering validation evaluation criteria to select the clustering members. For the weighted method, many researchers used some criteria to weigh the clustering members. Once obtaining final clustering members, researchers convert the clustering member to 0-1 matrix, similarity matrix or weighted matrix. Here, this matrix can be regarded as a new dataset. For example, CSPA algorithm regards the matrix as a graph and obtains a new similarity matrix through its matrix multiply by its transpose matrix. Some people regarded the matrix as a new data set to clustering in order to obtain final clustering result. And some researchers regarded the matrix as a graph and used graph partitioning algorithm to process it. Recently, Lu et al.[19] proposed a method based on covariance to measure the diversity between two base clustering results and used CSPA to combine them. For the similarity matrix or weight matrix, it can also use a spectral segmentation algorithm such as SPEC.

3. Selection of clustering ensemble members and its combination

Generally speaking, clustering ensemble mainly includes selection of members and their combination. For members of clustering ensemble, their clustering results may be class labels or fuzzy memberships. When using class labels, a unified process needs to be performed before clustering combination. So we use fuzzy memberships as clustering results and this can also eliminate the process of unified class labels. In this paper, the clustering ensemble framework presented is shown in Figure 1. It mainly includes three parts: 1) Select the basis clustering algorithm and use it to generate many clustering results. 2) Use the presented measure to select further clustering members and features, and weight selected features; 3) use the generalized consensus method to combine these clustering members in order to obtain final clustering results. In the following, we introduce the last two steps in detail.

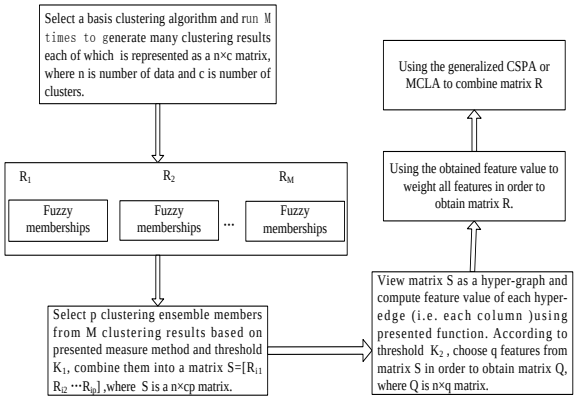

Figure 1: Framework of Selective Clustering Ensemble Method

Selection with clustering ensemble members

After using basis clustering algorithm to generate many clustering results, we construct a measure φ_k to evaluate the obtained clustering results. Suppose that there be M fuzzy clustering results. Let R_k be kth clustering result and u_{ih} represents the membership for ith data sample attributing to the hth cluster ($0 \le i \le n$, $1 \le h \le c$). Thus, the presented measure function for each clustering result is defined as follows:

$$f_k = \sum_{i=1}^{n} \sum_{j=1}^{c-1} \sum_{l=j+1}^{c} | u_{i,j} - u_{i,l} | . \qquad (1)$$

It can be seen that the larger value of f_k is, the harder the clustering result is. For selecting better clustering members, a threshold K_1 is introduced. If $f_k > K_1$ then we select this clustering member and put it into set S otherwise discard it. According to this method, we obtain p(p<M) clustering ensemble members represented as matrix S, where S is a $n \times cp$ matrix. Next, we regard S as a hyper-graph G and select the more influential of the hyper-edge in the hyper-graph G. For this purpose, we define function L_g as follows:

$$L_g = \sum_{i=1}^{n} (-u_{i,g} \log(u_{i,g})) , \qquad (2)$$

where $u_{i,g}$ represents the gth feature for the ith data sample. Our purpose is to select some hyper-edges whose membership u are more closer to 1 or 0. Then,

we determine a threshold K_2. If $L_g < K_2$, then we select the gth clustering member and put it to matrix Q. After repeating this process, we obtain q features. Moreover, from (2) we know that the small value of the η_g is, the better feature is. For example, suppose that there are three hyper-edges A, B and C, where A=[0.1 0.1 0.8 0.8]^T, B=[0.6 0.6 0.4 0.4]^T, and C=[0.2 0.2 0.9 0.9]^T. According to above criterion, we have L_A=0.356, L_B=0.584, and L_C=0.362. If we set K_2=0.4, then hyper-edges A and C are selected. Now, suppose that Q is a q ($1 \le q \le p$) dimensional matrix. Let u_g ($1 \le t \le q$) represent each feature in matrix Q. Then Q can be expressed as $Q=[u_1\ u_2\ \ldots u_q]$,where u_g ($1 \le g \le q$) is column vector. And let L_g($1 \le g \le q$) represent each feature value using (2) for matrix Q. To further determine each feature's role, we use L_g to weight each feature for matrix Q. According to the distribution principle, the smaller value of L_g is, the larger weight this feature is of. So we define the following weighting function

$$w_t = (1 - \frac{L_t}{\sum_{t=1}^{q} L_t}) \Big/ \sum_{t=1}^{q} (1 - \frac{L_t}{\sum_{t=1}^{q} L_t}) . \qquad (3)$$

Then the weighted fuzzy membership matrix Q can be expressed as

$$R = [w_1u_1 \quad w_2u_2 \quad w_3u_3 \quad \cdots \quad w_qu_q] .$$

For the sake of convenience, we recall above method of member selection as 2SW (Two Selection and a Weighting) in this paper.

·Generalized consensus method

To deal with matrix R, we generalize the consensus function presented by Strehl and Ghosh. In this paper, we mainly use two the generalized hyper-graph partitioning algorithms which are CSPA and MCLA, respectively. For simplicity, we still write them as CSPA and MCLA. When using these functions to process the feature of the matrix R, then METIS algorithm is used to obtain the final clustering result.

·Ensemble algorithm

In this paper, we refer to ensemble algorithm presented by us as 2SWC (Two Selection, a Weighting and Combination). If combination method is CSPA or MCLA then ensemble algorithm is written as 2SWC (CSPA) or 2SWC (MCLA). Detailed algorithm is described as follows.

Step 1 Initialize threshold K_1 and K_2.

Step 2 Choose a base clustering algorithm and use it to generate M clustering members (clustering results).

Step 3 Calculate f_k using (1) for all clustering members. If $f_k > K_1$, then we select this clustering member and add it into matrix S.

Step 4 Use (2) to compute L_g for all columns. If $L_g < K_2$, then we select this feature and put it to matrix Q.

Step 5 Utilize above computational results to compute w_t in order to obtain matrix R.

Step 6 Use the generalized CSPA or MCLA and METIS algorithm to combine the selected clustering members in order to get final clustering result.

4. Experiments

To test performance of clustering ensemble algorithm 2SWC, we choose five commonly used UCI data sets whose information is listed in Table 1. The selected base clustering algorithm is GEFCMNN which refers to several parameters which are fuzzy index m, the generalized entropy index a and ratio coefficient b, respectively. To obtain better clustering result, these parameters' optimal values need to be determined. In effect, finding their optimal values is very difficult and time consuming. In the following experiments, we use ensemble method to solve it. For generating many clustering results, we give these parameters' intervals. Specifically, fuzzy index m, the generalized entropy index a and ratio coefficient b are set as [2, 12], [1, 30] and [-10000, -500], respectively. Threshold K_1 is set as 60 or so, whereas K_2 is set as about 0.3.

Table 1: Characteristic of Data Sets

Data set	Number of data	Number of Class	Number of features
Iris	150	3	4
Breast-w	680	2	9
Heart	270	2	13
Ionosphere	351	2	34
Haberman	306	2	3

·Comparison of different methods

We first compare the performance of 2SWC with that of CSPA and CSPA-FUZZY, where 2SWC includes 2SWC (CSPA) and 2SWC (MCLA). Moreover, CSPA or CSPA-FUZZY corresponds to 0-1labels or fuzzy memberships for clustering result. The results are shown in Figure 2 and Figure 3. We can see from above some experimental results that 2SWC, includes 2SWC (CSPA) and 2SWC (MCLA) has better clustering results compared with CSPA and CSPA-FUZZY. It is seen that when all fuzzy clustering members are directly combined, clustering

ensemble's performance is very worse. When some members are selected using 2SW method, better clustering ensemble performance is obtained. Especially, clustering performance for 2SWC (MCLA) is very clear for Ionosphere, Heart and Haberman data sets. In addition, Lu et al. used different evaluation criteria from us to obtain the clustering ensemble results for dataset Iris, Ionosphere and Breast-w. Their ensemble results are 0.8812, 0.6902 and 0.7997, respectively.

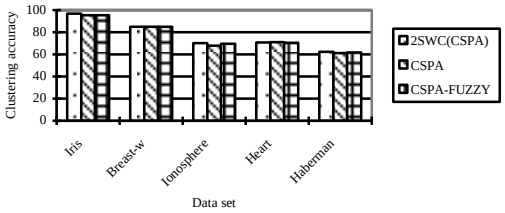

Figure 2: Experimental Results Using CSPA with Three Algorithms for Different Data Set

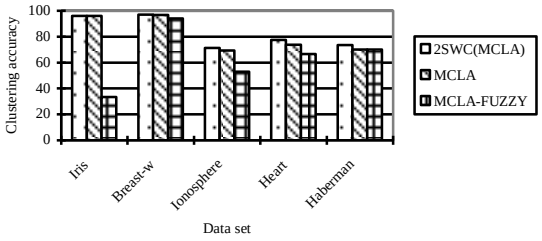

Figure 3: Experimental results using MCLA with three algorithms for different data set

·Effect of each step for 2SW selection method

In selection method of ensemble members 2SW, there are four steps: (1) Generate many fuzzy clustering results; (2) Select part members from many fuzzy clustering results; (3) Further select import features based on step 2 and (4) weight the selected features. To test the role of each step, we choose Heart and Ionosphere data sets to study them. Experimental results are seen in Figure 4 and Figure 5, where symbols "1", "2", "3" and "4" in horizontal axes represent experimental method by using first step, the previous two steps, the previous three steps and all four steps, respectively. It is seen from Figure 4 and Figure 5 that role of each step is different in the process of selection members. And when use all steps, the best clustering result can be obtained. This shows that selection method 2SW is effective to the clustering ensemble when using fuzzy membership as the clustering members.

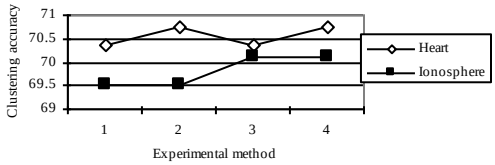

Figure 4: Experimental results of 2SW selection method using CSPA

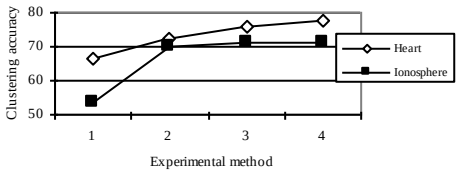

Figure 5: Experimental results of 2SW selection method using MCLA

·Impact of threshold on ensemble accuracy

In this part, we test the impacts of thresholds K_1 and K_2 on the ensemble accuracy for 2SW. In experiments, we choose Iris, Breast-w, Heart and Haberman data sets and the numbers of clustering members are set as 20 and 10. The experimental results are shown in Table 2-Talbe 5. Note that in following tables, *m*, *a* and *b* are parameters for base clustering algorithm. Their

Table 2: Experimental Results of Data Set Iris for Different Threshold

Clustering members	Threshold	Consensus function	Accuracy (%)
Number of members: 20 m∈ [2,12] b∈[-1500,-500] a∈[1,30]	K1=0 K2=1	CSPA	95.33
		MCLA	33.33
	K1=50 K2=1	CSPA	95.33
		MCLA	96
	K1=50 K2=0.358	CSPA	96
		MCLA	92.67
	K1=60 K2=1	CSPA	96.67
		MCLA	96
	K1=60 K2=0.358	CSPA	96.67
		MCLA	96
Number of members: 10 m∈ [2,12] b∈[-1500,-500] a∈[1,30]	K1=0 K2=1	CSPA	95.33
		MCLA	76
	K1=50 K2=1	CSPA	95.33
		MCLA	96
	K1=50 K2=0.36	CSPA	95.33
		MCLA	96
	K1=60 K2=1	CSPA	96
		MCLA	96
	K1=60 K2=0.32	CSPA	96.67
		MCLA	96

Table 3: Experimental Results of Data Set Breast-W for Different Threshold

Clustering members	Threshold	Consensus function	Accuracy (%)
Number of members: 20 m∈ [2,12] b∈[1,10] a∈[1,20]	K_1=0 K_2=1	CSPA	85.07
		MCLA	94
	K_1=60 K_2=1	CSPA	85.07
		MCLA	96.78
	K_1=60 K_2=0.358	CSPA	85.07
		MCLA	96.78
	K_1=120 K_2=1	CSPA	85.07
		MCLA	96.93
	K_1=120 K_2=0.358	CSPA	85.07
		MCLA	96.93
Number of members: 10 m∈ [2,12] b∈[1,10] a∈[1,20]	K_1=0 K_2=1	CSPA	85.07
		MCLA	96.63
	K_1=60 K_2=1	CSPA	85.07
		MCLA	96.63
	K_1=60 K_2=0.355	CSPA	85.07
		MCLA	96.49
	K_1=120 K_2=1	CSPA	85.07
		MCLA	96.78
	K_1=120 K_2=0.355	CSPA	85.07
		MCLA	97.07

Table 4: Experimental Results of Data Set Heart for Different Threshold

Clustering members	Threshold	Consensus function	Accuracy (%)
Number of members: 20 m∈ [2,12] b∈[-6000,-4000] a∈[1,100]	K_1=0 K_2=1	CSPA	70.37
		MCLA	66.67
	K_1=1 K_2=1	CSPA	70.37
		MCLA	74.44
	K_1=1 K_2=0.352	CSPA	75.19
		MCLA	75.93
	K_1=2 K_2=1	CSPA	70.74
		MCLA	74.81
	K_1=2 K_2=0.352	CSPA	74.81
		MCLA	76.30
Number of members: 10 m∈ [2,12] b∈[-6000,-4000] a∈[1,100]	K_1=0 K_2=1	CSPA	70.37
		MCLA	65.93
	K_1=0.01 K_2=1	CSPA	70.74
		MCLA	67.04
	K_1=0.01 K_2=0.35	CSPA	70.37
		MCLA	70.74
	K_1=1 K_2=1	CSPA	70.74
		MCLA	77.78
	K_1=1 K_2=0.35	CSPA	70.74
		MCLA	75.93

Table 5: Experimental Results of Data Set Haberman for Different Threshold

Clustering members	Threshold	Consensus function	Accuracy (%)
Number of members: 20 m∈ [2,10] b∈[-600,10] a∈[1,20]	K_1=0 K_2=1	CSPA	61.76
		MCLA	69.93
	K_1=45 K_2=1	CSPA	61.76
		MCLA	69.93
	K_1=45 K_2=0.358	CSPA	61.44
		MCLA	69.61
	K_1=50 K_2=1	CSPA	61.76
		MCLA	66.67
	K_1=50 K_2=0.25	CSPA	62.42
		MCLA	73.53
Number of members: 10 m ∈[2,10] b∈[-600,10] a∈[1,20]	K_1=0 K_2=1	CSPA	61.76
		MCLA	69.61
	K_1=5 K_2=1	CSPA	62.42
		MCLA	68.30
	K_1=5 K_2=0.35	CSPA	61.76
		MCLA	73.53
	K_1=50 K_2=1	CSPA	61.76
		MCLA	69.93
	K_1=50 K_2=0.35	CSPA	62.42
		MCLA	73.86

From the results we can see that performance of ensemble algorithm 2SWC (MCLA) strongly depends on thresholds K_1 and K_2, whereas that of 2SWC (CSPA) weakly or does not depend on them. And performance of ensemble algorithm 2SWC (MCLA) is superior to that of 2SWC (CSPA).

5. Conclusions

Aiming at many fuzzy clustering results, we present a selective clustering ensemble algorithm 2SWC which includes selection of ensemble members 2SW and generalized combination of clustering members CSPA and MCLA. To select some better members, we construct a function to measure each clustering result. Then selected clustering members are viewed as hyper-graph in order to select the more influential hyper-edges (or features) by using the defined measure and to weight the selected features. We generalize CSPA and MCLA consensus function to deal with fuzzy membership's hyper-edge. In the experiments, we choose UCI data sets to test the presented algorithm's performance. From the experimental results, we know that the proposed ensemble method can get better clustering ensemble result. And the results also tell us that using fuzzy memberships as clustering members can obtain better ensemble results. In the future, we further study the

proposed clustering ensemble algorithm and its effect of the ensemble.

References

[1] K. Li, H.Y. Ma, and Y. Wang, "Unified model of fuzzy clustering algorithm based on entropy and its application to image segmentation", Journal of Computational Information Systems, 7(15), 476-483, 2011.

[2] K. Li, P. Li, "A fuzzy clustering algorithm with the generalized entropy based on neural network", Journal of Computational Information Systems, 9(1), 353-360, 2013.

[3] L. I. Kuncheva and S. T. Hadjitodorov, "Using Diversity in Cluster Ensembles", Proceedings. of IEEE International Conference on Systems, Man and Cybernetics, pp. 1214-1219, 2004.

[4] M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar and C. Domeniconi, "A clustering framework based on subjective and objective validity criteria", ACM Transactions on Knowledge Discovery from Data, 1(4), 18:2-18:25,2008.

[5] A. Strehl, J. Ghosh, "Cluster Ensemble-a knowledge reuse framework for combining Multiple partitions", Journal on Machine Learning Research(JMLR),3,583-617, 2002.

[6] M. Selim, E. Ertunc, "Combining multiple clusterings using similarity graph", Pattern Recognition, 44, 694-703, 2011.

[7] Vega-Pons, Sandro, Ruiz-Shulcloper, José Source, "A survey of clustering ensemble algorithms", International Journal of Pattern Recognition and Artificial Intelligence, 25, 337-372, 2011.

[8] Z. W. Yu, H. S. Wong, G. X. Yu. et al, "Hybrid cluster ensemble framework based on the random combination of data transformation operators", Pattern Recognition, 45, 1826-1837, 2012.

[9] Z. S. Hong, H. S. Wong, Y. Shen, "Generalized Adjusted Rand Indices for cluster ensembles", Pattern Recognition, 45, 2214-2226, 2012.

[10] J. Yang, X. Q. Zeng, S. M. Zhong and S.L Wu, "Effective Neural Network Ensemble Approach for Improving Generalization Performance", IEEE Transactions on Neural Networks and Learning Systems, 24(6), 878-888, 2013.

[11] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering", IEEE Transactions on Knowledge and Data Engineering, 24(3), 413-425, 2012.

[12] Z. H. Zhou, W. Tang, "Clusterer ensemble", Knowledge-Based Systems, 19, 77-83, 2006.

[13] X. Fern , W. Lin, "Cluster Ensemble Selection", Journal Statistical Analysis and Data Mining,1(3),128-141,2008.

[14] Y. Hong, S. K Wong, H. L. Wang, Q. S. Ren, "Resampling-based selective clustering ensembles", Pattern Recognition Letters, 30, 298-305, 2009.

[15] Y. Yang and K. Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE Transactions on Knowledge and Data Engineering, 23(2), 307-320, 2011.

[16] Z. Yu, H. S. Wong, J. You, Q. M. Yang, H. Y. Liao, "Knowledge Based Cluster Ensemble for Cancer Discovery From Biomolecular Data", IEEE Transactions on Nanobio Science, 10(2),76-85, 2011.

[17] B. Minaei-Bidgoli, A. Topchy and W. Punch, "A comparison of resampling methods for clustering ensembles", Proceeding of the. International Conference on Artificial Intelligence, Vol. 2, pp. 939-945, 2004.

[18] Z. W. Yu, L. Li, D. X. Wang, J. You, G. Q. Han, H. T. Chen, "Structure Ensemble Based on Fyzzy C-means", Proceedings of the 2012 International Conference on Machine Learning and Cybernetics,pp.1383-1389, 2012.

[19] X. Y. Lu, Y. Yang, H. J. Wang, "Selective clustering ensemble based on covariance", Lecture Notes in Computer Science, Vol. 7872, pp. 179-189,2013.

Kai Li received the B.S. and M.S. degrees in mathematics department electrical engineering department from Hebei University, Baoding, China, in 1982 and 1992, respectively. He received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2001.He is currently a Professor in school of mathematics and computer science, Hebei University. His current research interests include machine learning, data mining, computational intelligence, and pattern recognition.

Peng Li received the B.S. degree from Beijing University of chemical technology, Beijing, China, in 2010. He is now pursuing the M.S. degrees in school of mathematics and computer science, Hebei University. His current research interests include neural network, fuzzy clustering and clustering ensemble.