An Improved Correlation Measure-based SOM Clustering Algorithm for Gene Selection

Jiucheng Xu, Tianhe Xu, Lin Sun, Jinyu Ren
College of Computer & Information Engineering, Henan Normal University, Xinxiang 453007, P. R. China
E-mail: tianhe1107@163.com

Abstract—**Among the large amount of genes presented in microarray gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. For this reason, reducing the dimensionality of gene expression data is imperative. Self-organizing map (SOM) is a type of mathematical cluster analysis which particularly well suited for recognizing and classifying features in complex, multidimensional data. This paper proposes an improved Self-organizing map clustering algorithm which based on neighborhood mutual information correlation measure. To evaluate the performance of the proposed approach, we apply it to six well-known gene expression datasets and compare our results with those obtained by other methods. Finally, the experimental results show that the proposed approach to gene selection is indeed efficient.**

Index Terms—**self-organizing map, neighborhood mutual information, correlation measure, clustering algorithm**

I. Introduction

In recent years, gene expression profiles (GEP) based molecular diagnosis of tumor have attracted a great number of medical researchers and computer scientists for the goal of realizing precise and early tumor diagnosis [1-9]. However, the curse of dimensionality caused by high dimensionality and small sample size of tumor dataset seriously challenges the tumor classification. So how to select important gene subsets from thousands of genes in GEP dataset to drastically reduce the dimensionality of tumor dataset is the first key step to address this problem [10].

Clustering is a main task of explorative data mining and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics, etc. [3]. When applied to gene expression data, conventional clustering algorithms may encounter a problem which is a huge number of genes (attributes) versus a small number of samples [4]. Well-known conventional clustering algorithms are: K-means algorithms [11,12], Bayesian clustering [13], and various Hierarchical clustering algorithms [14-16]. K-Means is among the most widely used partitioning algorithm. It is an unsupervised clustering algorithm that is simple and robust. However, this algorithm needs to fix the number of clusters and their initial centroids before assigning the input data. Bayesian clustering is a highly structured approach when a strong prior distribution on the data is available [13]. Hierarchical clustering has a number of shortcomings for the study of gene expression. Strict phylogenetic trees are best suited to situations of true hierarchical descent (such as in the evolution of species) and are not designed to reflect the multiple distinct ways in which expression patterns can be similar [13]. In order to deal with the high dimension problem of genes well, we propose a novel SOM algorithm in this paper.

SOM [13,17] has a number of features that make them particularly well suitable for clustering and analyzing of gene expression patterns. They are ideally suited to explore data analysis, allowing one to impose partial structure on the clusters (in contrast to the rigid structure of hierarchical clustering, the strong prior hypotheses used in Bayesian clustering, and the non-structure of K-means clustering) and facilitating easy visualization and interpretation. As for correlation measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering [13]. However, for measuring the correlation between genes, Euclidean distance is not effective enough to describe functional similarity such as positive or negative correlation in values. Empirical studies have shown that it may assign a high similarity score to a pair of dissimilarity genes [18]. Furthermore, most clustering methods are not able to effectively cope with continuous attributes which is also a distinctive characteristic of gene expression data. When applied to the continuous attributes, conventional methods commonly discretize the continuous data into a finite number of intervals for data mining. But discretization may lead to information loss. As we know, SOM utilizes Euclidean distance measuring the correlation between genes. In this paper, we use neighborhood mutual information instead of Euclidean distance to evaluate the correlation between genes. By applying improved SOM to gene expression data, clusters of genes based on their neighborhood mutual information correlation can be discovered.

The structure of the rest of this paper is as follows: Section 2 introduces the concepts of SOM clustering and neighborhood mutual information. An effective and efficient attribute clustering method ICMSOM is proposed in Section 3. To evaluate the performance of the proposed algorithm, we apply it to six gene expression datasets. The experimental results are presented in Section 4. Finally, the conclusion is drawn in Section 5.

doi:10.4304/jsw.8.12.3082-3087

II. Related Work

A. *Self-Organizing Map*

SOM [19,20,21] is a clustering tool which can convert the non-linear statistical relationships between high dimensional data into simple geometric relationships of their image points on a low (often one or two) dimensional grid. By that way, the data points which show similar properties are placed close to each other within the output of SOM algorithm [17].

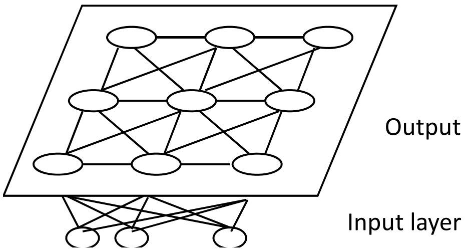

Fig.1 Structure of SOM network

SOM consists of a set of neurons which usually arranged in a two dimensional structure, the neighborhood relations exist among the neurons, which dictates the topology and structure. As shown in Fig.1, the network of SOM usually consists of two layers of neurons: input layer and output layer. Although, the neurons on input layer are fully connected to the neurons that on output layer, the neurons on each layer have no connection to each other. Neurons on the output layer are arranged in either a rectangular or hexagonal lattice. The selection of lattice shape affects the performance of the SOM due to the number of neighbors of neurons. The performance of the SOM by using a hexagonal shape is expected to be higher since more neighbors are modified in the hexagonal lattice when it is compared with a rectangular lattice [19]. Each neuron is represented by a n-dimensional weight vectors. The algorithm of SOM is initialized by assigning the values of weight vectors of each output neuron linearly or randomly. Training process of SOM starts by representing a data point randomly in the network. The distances between these data points and the weight vectors of all neurons are computed by using distance measures such as Euclidean distance. By comparing these distances, the nearest Kohonen neuron, is identified as the 'winning' neuron. The weights of the winning neuron are adjusted in order to get close to the actual data point. The weights of neighboring neurons are also updated, so that the order of the input space can be satisfied [22]. In the SOM network, the character mapping is topologically ordered and character choosing. Topologically ordered means the space position of the neuron in the network being is mapped from the character or some territory of the input sample. Character choosing means that when a data is given in the input space of non-linear distributed, SOM can choose the best character for the approach [20,23]. The above two characters determine the SOM network suitable for the disease feature.

B. *Neighborhood Mutual Information Measure*

Hu et al. [24] proposed neighborhood mutual information to cope with continuous gene data, evaluating the relevance between attributes.

There is a problem to employ mutual information in gene evaluation due to the difficulty in estimating probability density of genes. So neighborhood mutual information combines the concept of neighborhood with information theory, and generalizes Shannon's entropy to numerical information. Training samples are usually given as vectors of attribute values and the attributes are numerical, as shown in Table 1, where A_1 and A_2 are two attributes, while C is the decision label of samples.

Let $U=\{x_1,x_2,\cdots,x_n\}$ be a set of samples described with gene set *F*, and $x_i\in R^N$, Δ is a distance function on *U*, $\delta\geq 0$ is a constant, then the neighborhood of sample *x* is denoted by

$$\delta(x)=\{x_i \mid \Delta(x,x_i)\leq\delta\} \tag{1}$$

Given $S \subseteq F$ is a subset of genes, the neighborhood of sample x_i in *S* is denoted by $\delta_S(x_i)$. The neighborhood uncertainty of x_i is denoted by

$$NH_\delta^{x_i}(S)=-\log\frac{\|\delta_S(x_i)\|}{n} \tag{2}$$

and the average uncertainty of the set of samples is computed as

$$NH_\delta(S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n} \tag{3}$$

TABLE I.
Classification Task Described By Numerical

Sample ID	A_1	A_2	C
1	0.52	0.36	1
2	0.31	0.00	1
3	0.18	0.74	1
4	0.50	0.23	1
5	0.42	0.47	2
6	0.01	0.52	2
7	0.30	0.71	2
8	0.51	0.03	2
9	0.35	0.38	3
10	0.64	0.33	4

Example1. Assume $\delta=0.2$ then $\delta_S(x_1)=\{x_1, x_4, x_5, x_9, x_{10}\}$, $NH_\delta^{x_1}(S)=-\log(\frac{5}{10})=1$, where $S=\{A_1,A_2\}$.

Given *R*, $S \subseteq F$ two subsets of genes, the neighborhood of sample x_i in gene subspace $S\bigcup R$ is denoted by $\delta_{S\bigcup R}(x_i)$, and the joint neighborhood entropy of $S\bigcup R$ is computed as

$$NH_\delta(R,S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\bigcup R}(x_i)\|}{n} \tag{4}$$

Let *R*, $S \subseteq F$ be two subsets of genes, then the neighborhood mutual information of R and S is denoted by

$$NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|\cdot\|\delta_S(x_i)\|}{n\|\delta_{S\cup R}(x_i)\|} \quad (5)$$

III. Efficient Gene Selection Algorithm

A. The Novel Correlation Measure

As for correlation measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering [15]. However, for measuring the correlation between genes, Euclidean distance is not effective enough to describe functional similarity such as positive or negative correlation in values. Empirical studies have shown that it may assign a high similarity score to a pair of dissimilarity genes [18]. Conventional SOM uses Euclidean distance measuring the correlation between genes, it is not able to cope with continuous attributes effectively which is also a distinctive characteristic of gene expression data. Au et al. [18] presented an information measure to evaluate the correlation between attributes. It is called the interdependence redundancy measure between two attributes, A_i and A_j, $i,j\in\{1,\cdots m\}$, which is denoted by

$$R(A_i:A_j)=\frac{I(A_i:A_j)}{H(A_i,A_j)} \quad (6)$$

where $I(A_i:A_j)$ is the mutual information between A_i and A_j, and $H(A_i,A_j)$ is the joint entropy of A_i and A_j.

Definition 1. Let A_i and A_j, $i,j\in\{1,\cdots m\}$ are two attributes, then the neighborhood interdependence redundancy measure between two attributes is defined as

$$NR_\delta(A_i;A_j)=\frac{NMI_\delta(A_i;A_j)}{NH_\delta(A_i,A_j)} \quad (7)$$

where $NMI_\delta(A_i;A_j)$ is the neighborhood mutual information between A_i and A_j, and $NH_\delta(A_i,A_j)$ is the joint neighborhood entropy of A_i and A_j.

Definition 2. Let $X = [x_1, x_2, \ldots, x_L]^T$ be an input data, where X is a sample. Choose a random value for the initial weight vector $w_j = [w_{j1}, w_{j2}, \ldots, w_{jL}]^T$, $j = 1, 2, \ldots, N$, N is the number of output neurons. $w_{ji}(t)$ is the ith component weight vector of the output neuron node j at time t, $x_i(t)$ is the ith input of X at time t. In this paper, we use neighborhood mutual information instead of Euclidean distance to evaluate the correlation between genes. The correlation between the input X and neuron j is defined as

$$NR_\delta(x_i(t);w_{ji}(t))=\frac{NMI_\delta(x_i(t);w_{ji}(t))}{NH_\delta(x_i(t),w_{ji}(t))} \quad (8)$$

where $w_{ji}(t)$ is the ith component weight vector of the output neuron node j at time t, $x_i(t)$ is the ith input of X at time t. $NMI_\delta(x_i(t);w_{ji}(t))$ is the neighborhood mutual information between $x_i(t)$ and $w_{ji}(t)$, and $NH_\delta(x_i(t),w_{ji}(t))$ is the joint neighborhood entropy of $x_i(t)$ and $w_{ji}(t)$.

B. The Description of The Improved SOM Clustering Algorithm (ICMSOM)

There are several steps in the application of the algorithm. To get the winner, competition and learning are needed in the process. The steps above are repeated until the feature mapping is formed. The process is as follow:

(1) Initialization: Choose random values for the initial weights w_j.

(2) Winner Finding: Find the winning neuron C at time t, using the correlation criterion:

$$C=\arg\max\frac{NMI_\delta(X;w_j)}{NH_\delta(X,w_j)},\ j=1,2,\ldots,N \quad (9)$$

where $X=[x_1,x_2,\ldots,x_l]^T$ represents an input vector at time t, N is the total number of neurons.

(3) Weights Updating: Adjust the weights of the winner and its neighbors, using the following rule:

$$w_j(t+1)=w_j(t)+\eta(t)h_{j,C}(t)[X(t)-w_j(t)] \quad (10)$$

$$h_{j,C}(t)=\exp[-\frac{NR_\delta(r_c;r_i)}{2\sigma^2(t)}] \quad (11)$$

where $X(t)$ represents an input data at time t, $h_{j,C}(t)$ is the topological neighborhood function of the winner neuron C at time t, $\eta(t)$ is a positive constant called 'learning-rate factor', $r_c\in R^2$ and $r_i\in R^2$ are the location vectors of nodes c and i, respectively. $\sigma(t)$ defines the width of the kernel. Both $\eta(t)$ and $\sigma(t)$ will decrease with time. It should be emphasized that the success of the map formation is critically dependent on the values of the main parameters (i.e., $h_{j,C}(t)$ and $\eta(t)$), the initial values of weight vectors, and the prespecified number of iterations.

(4) Repeat the steps (2) and (3) until the changes in the disease feature mapping are very small or the maximum iteration time is reached.

IV. Experimental Analysis

To validate the ICMSOM sort capability in comparison with K-means, Hierarchical and SOM, six different datasets are used to study. There are 8D5K dataset, AD400_10_10 dataset, Yeast dataset, Heart dataset, class dataset and Breast Cancer dataset (The 8D5K dataset and AD400_10_10 dataset are artificial datasets. The rest of the datasets can be got from UCI datasets.) The characters of dataset are shown in Table 2. For no test samples in database, ten-fold-cross-validation is chosen to compute the percentile of correct sort. The results are the average and standard deviation taken over ten runs.

The performance of the proposed ICMSOM method is extensively studied and compared with that of some existing algorithms, namely, K-means, Hierarchical and SOM, employing rand to measure the clustering accuracy. To analyze the performance of different algorithms, the experimentation is done on six different datasets. And each dataset is preprocessed by data cleaning and data filtering, removing the data that which expression level is negative or small, processing noise data and missing data.

TABLE II.
EXPERIMENT DATA SETS

Data-set	Sample amount	Input dimension	Output dimension	Feature
8D5K	1000	8	5	continuous
AD400_10_10	400	10	10	continuous
Yeast	93	8	5	continuous
Heart	270	13	2	continuous
class	214	9	6	continuous
Breast Cancer	683	10	2	continuous

In this experiment, the operating environment is Lenovo Windows7 PC with 3.1 GHZ CPU and 4GB RAM, and the algorithm is coded by VC++. The experimental software is MATLAB 7.1. The inputs and outputs of dataset are normalized over the interval [−1, 1]. In the training stage of SOM algorithm, the maximum iterative time is set as 500, the initial learning rate η_0 as 0.3, the initial neighborhood region $h_{j,C}(0)$ as 1. The experiment results are shown in Table 3.

According to Table 3, the accuracy of ICMSOM is higher than SOM. In four different algorithms, the accuracy of ICMSOM is the highest for six different datasets; on the other hand, the accuracy of Hierarchical is the lowest. From the Table 3, we can have the accuracies of two kind of SOM clustering algorithms are higher than other algorithms. This is because the SOM algorithm is not sensitive to initialization parameter setting, and it can gradually achieve the global optimal by adjusting the neurons in characteristics of the right value. ICMSOM algorithm is one of the improved algorithms of the traditional SOM algorithm. It not only has the advantages of traditional SOM algorithm, but also can deal with continuous data and evaluate the irrelevant and redundant genes. This method can cluster those all contains important characteristics of the related gene expression data. So the accuracy of ICMSOM is higher than SOM.

TABLE Ⅲ.
THE CORRESPONDING RAND VALUE OF DIFFERENT ALGORITHMS

Data-set	Accuracy/%			
	Algorithm			
	K-means	Hierarchical	SOM	ICMSOM
8D5K	72.3	68.5	77.6	78.1
AD400_10_10	70.4	67.2	76.2	76.5
Yeast	65.3	59.8	67.1	67.9
Heart	69.4	67.4	73.1	73.7
class	67.8	65.2	70.3	71.1
Breast Cancer	69.0	62.0	72.1	72.9

From Table 3, it can be observed that ICMSOM exhibits the best performances on the six datasets, and the performances of ICMSOM and SOM are very close, though ICMSOM performs a little better than SOM. Furthermore, the two algorithms perform markedly better than the other two algorithms, where Hierarchical has the worst performance.

TABLE Ⅳ.
THE CLUSTERING TIME OF DIFFERENT ALGORITHMS

Data-set	Time/s			
	Algorithm			
	K-means	Hierarchical	SOM	ICMSOM
8D5K	65.2	173.5	74.3	75.2
AD400_10_10	51.0	101.0	57.4	58.1
Yeast	41.2	98.9	56.3	57.5
Heart	48.7	120.7	59.8	60.3
class	46.2	110.3	58.1	59.3
Breast Cancer	56.7	143.1	66.0	67.2

From Table 4, it can be seen that ICMSOM and SOM have good time performance, though the clustering time of ICMSOM is a little higher than SOM. That is because it needs more procedure to compute the neighborhood mutual information of genes. However, the K-means algorithm also has a very good time performance. That is due to K-means is linear algorithm, and the time complexity of it is $O(knl)$, where k is the number of cluster, n is the number of data, l is the number of iteration. Therefore, the time complexity of SOM is $O(k^{'}mn)$, where $k^{'}$ is the number of neurons, n is the number of data, m is the number of iteration. Generally speaking, the $k^{'}$ and m of SOM are larger than k and l of K-means, so the time complexity of SOM is higher than K-means. The time complexity of ICMSOM is higher than SOM. In the four different algorithms, the

Hierarchical should be the most time-consuming and its time complexity is $O(n^2)$.

V. Conclusion

Clustering and classification are key tasks of gene identification. In virtue of the continuous attributes, our proposed attribute clustering algorithm for gene selection can directly deal with the continuous data and acquire high accuracy. Experimental results show that this algorithm outperforms than other approaches. In the recent research, the cluster configuration is studied by using qualitative analysis. In order to get through understanding about gene expression profiles, and extend our model to improve its generation, we have to investigate the cluster quality further, which refers to its shape, size and distribution, etc. That is what we want to be involved with in the further concern.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant (Nos: 60873104, 61040037), the Key Project of Science and Technology Department of Henan Province (No. 112102210194), the Science and Technology Research Key Project of Educational Department of Henan Province (No. 12A520027), the Education Fund for Youth Key Teachers of Henan Normal University.

References

[1] M. Dettling, "Bag Boosting for tumor classification with gene expression data," Bioinformatics, 2004, vol. 20, pp. 3583–3593.

[2] A.M. Bagirov, B. Ferguson, S. Ivkovic, G. Saunders, J. Yearwood, "New algorithm for multi-class cancer diagnosis using tumor gene expression signatures," Bioinformatics, 2003, vol. 19, pp. 1800–1807.

[3] Yuan Hong, Jaideep, vaidya, Haibing Lu, "Searching Engine Query Clustering using top-k Search Results," IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011, vol. 1, pp. 112-119.

[4] Martin Hopfensitz, Christoph Mussel, Christian Wawra, "Multiscale Binarization of Gene Expression Data for Reconstructing Boolean Networks," IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012, vol. 9, no. 2, pp. 487-498.

[5] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles," Bioinformatics, 2005, vol. 21, pp. 3896–3904.

[6] J.J. Dai, L. Lieu, D. Rocke, "Dimension reduction for classification with gene expression microarray data," Statistical Applications in Genetics and Molecular Biology, 2006, vol. 5 no. 6.

[7] M. Mramor, G. Leban, J. Demsar, B. Zupan, "Visualization-based cancer microarray data classification analysis," Bioinformatics, 2007, vol. 23, pp. 2147–2154.

[8] Sun Lin, Xu Jiucheng, Ren Jinyu, Xu Tianhe, Zhang Qianqian, "Granularity partition-based feature selection and its application in decision systems," Journal of Information and Computational Science, 2012, vol. 9, no. 12, pp. 3487-3500.

[9] Sun Lin, Xu Jiucheng, Li Shuangqun, Gao Yunpeng, "Rough entropy extensions for feature selection under incomplete decision information systems," Advances in Information Sciences and Service Sciences, 2011, vol. 3, no. 11, pp. 264–274.

[10] S.L. Wang, X.L. Li, "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction," Computers in Biology and Medicine, 2010, vol. 40, pp. 179–189.

[11] F.D. Smet, J.Mathys, K. Marchal, G. Thijs, B. De Moor, Y. Moreau, "Adaptive quality-based clustering of gene expression profiles," Bioinformatics, 2002, vol. 5, no. 18, pp. 735-746.

[12] Xiao Qing, Qian Xiaodong, Liao Hui, "Clustering algorithm analysis of web users with dissimilarity and SOM neural networks," Journal of Software, 2012, vol. 7, no. 11, pp. 2533-2537.

[13] P. Tamayo*, D. Slonim*, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," Proc. Natl. Acad. Sci. USA, 1999, vol. 96, pp. 2907–2912.

[14] E. Domay, "Cluster Analysis of Gene Expression Data," Statistical Physics, 2003, vol. 110, pp. 1117-1139.

[15] D.K.Y. Chiu, A.K.C. Wong, "Multiple pattern association for interpreting structural and functional characteristic of biomolecules," Information Sciences, 2004, vol. 167, pp. 23-39.

[16] W. Wu, W. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," Information Sciences, 2004, vol. 159, pp. 233–254.

[17] T. Kohonen, "Self-Organizing Maps," third ed. Berlin: Springer-Verlag, 2001.

[18] W.H. Au, Keith C.C. Chan, Andrew K.C. Wong, Yang Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, vol. 2, no. 2, pp. 83-101.

[19] C. Budayan, I. Dikmen, M.T. Birgonul, "Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping," Expert Systems with Applications, 2009, vol. 36, pp. 11772–11781.

[20] K. Zhang, Y. Chai, "Self-organizing feature map for cluster analysis in multi-disease diagnosis," Expert Systems with Applications, 2010, vol. 37, pp. 6359–6367.

[21] Li Xinwu, "A new text clustering algorithm based on improved K-means," Journal of Software, 2012, vol. 7, no. 1, pp. 95-101.

[22] B. Curry, F. Davies, P. Evans, L. Moutinho, "The Kohonen self-organizing map: An application to the study of strategic groups in the UK hotel industry," Expert Systems, 2001, vol. 18, no. 1, pp. 19-31.

[23] S.C. Chi, C.C. Yang, "Integration of ant colony SOM and K-means for clustering analyses," In 10th International conference KES2006: Knowledge-based intelligent information and engineering systems, 2006, vol. 1, pp. 1-8.

[24] Qinghua Hu, Wei Pan, Shuang An, Peijun Ma, Jinmao Wei, "An efficient gene selection technique for cancer recognition based on neighborhood mutual information," Int. J. Mach. Learn. & Cyber, 2010, vol. 1, pp. 63-74.

Jiucheng Xu is a Professor at College of Computer & Information Engineering, Henan Normal University. He received his B.S. degree in Xi'an Jiao tong University, Shanxi, China, in 1995, and 2004, respectively. Mathematics from Henan Normal University, Henan, China, in 1986, the M.S. degree and the Ph.D. degree in Computer Science and Technology from His main research interests include granular computing, rough set, data mining, intelligent information processing, bioinformatics.

Tianhe Xu received her B.S. degree in Education Science Department from Luoyang Normal University, Henan, China, in 2011. Now, She is a master graduate in Computer Science and Technology of Henan Normal University. Her main research interests include granular computing, data mining, bioinformatics.

Lin Sun received his B.S. and M.S. degree in Computer Science and Technology from Henan Normal University, Henan, China, in 2003, and 2007, respectively. His main research interests include granular computing, rough set, intelligent information processing.

Jinyu Ren received her B.S. degree in Mathematics and Information Science Department from Shijiazhuang University, Hebei, China, in 2011. Now, She is a master graduate in Computer Science and Technology of Henan Normal University. Her main research interests include granular computing, intelligent information processing.