Chinese Journal of Electronics
Vol.18, No.4, Oct. 2009

An Ensemble Classifier Based on Selective Independent Component Analysis of DNA Microarray Data*

WANG Xuesong[1,2], GU Yangyang[1], CHENG Yuhu[1] and ZHANG Zheng[1]

(1.*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China*)
(2.*The Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China*)

Abstract — Our study tried to deal with a gene expression problem from the view of factor analysis. In order to overcome the instability problem caused by using traditional Independent component analysis (ICA), an ensemble classifier of DNA microarray data based on a selective ICA method was proposed. At first, we analyzed the reconstruction error of each gene and selected a subset of independent components, which contributed relatively small reconstruction errors, to reconstruct new samples. After that, several Support vector machine (SVM) sub-classifiers were trained simultaneously. Finally, the best SVM sub-classifiers with high correct rates were selected to participate in the ensemble, using a majority voting method. Results on three publicly available DNA microarray datasets show the feasibility and validity of our proposed method.

Key words — Selective, Independent component analysis, DNA microarray, Ensemble learning, Support vector machine.

I. Introduction

As an advanced experimental technique, DNA microarray experiments allow the recording of expression levels of thousands of genes simultaneously[1]. It helps medical experts with access to comprehensive information about diseases. However, this highly dimensional technique, the small number of samples as well as the uncertainty caused by experiments and organisms make data mining from gene expression data a very challenging investigation[2].

Factor analysis is a practical statistical technique. It condenses complex variables obtained in the experiment to fewer variables or factors and helps to extract useful information from abundant data. Besides Principal component analysis (PCA), a long established technique, more recent factor analytical techniques include Independent component analysis (ICA), Nonnegative matrix factorization (NMF) and so on. ICA is developed from the blind source separation field. Not only does it eliminates correlations among signals but also makes full use of the high-order statistical information contained in the original data. It has been proven that ICA can deal with gene expression data effectively. Liebermeister[3] pointed out that each gene expression level is a linear combination of different Independent components (ICs). Chiappetta *et al.*[4] interpreted the ICs as latent gene control routes. Huang and Zheng[5] employed ICA to model the gene expression data and then applied optimal scoring algorithm to classify them. This strategy not only could make full use of the high-order statistical information contained in the microarray data, but also used regularized regression models to deal with the situation of large number of correlated predictor variables. Liu and Huang[6] proposed a rotation forest classifier to mine DNA microarray data, in which ICA, PCA and random projections were used as three feature transformation techniques. Liu and Huang also pointed out that, compared with other feature transformation techniques, ICA performs best. Kim *et al.*[7] applied three ICAs, *i.e.*, sICA, tICA and stICA, to gene expression time series data analysis and compared their advantages in different assignments.

The basic principle of the ICA method is a decomposition of multi-path observational signals into statistically independent components, *i.e.*, source signals[8]. However, the number of source signals is unknown in practice, which inevitably leads the ICA method to become unstable. Therefore, a novel method called Selective ICA (SICA) is proposed by us to solve this problem. In stead of exploring the number of source signals, we choose to select a subset of the ICs whose reconstruction errors are relatively small to reconstruct the samples for classifiers. In addition, because only a few samples are obtained in general, we use a Support vector machine (SVM) as the cancer classifier, which has been proven to be very useful in the classification of gene expression data[9]. As a kind of kernel-based machine learning method, the SVM is inevitably confronted with an optimal selection of a kernel function and parameters, which will affect the classification accuracy and

*Manuscript Received Oct. 2008; Accepted Dec. 2008. This work is supported by the National Nature Science Foundation of China (No.60804022), Nature Science Foundation of Jiangsu Province (No.BK2008126), Specialized Research Fund for the Doctoral Program of Higher Education of China (No.20070290537, No.200802901506).

generalization property of the SVM. As was pointed out by Zhou *et al.*[10], it is generally easier to obtain much better performances through ensemble learning than from a single learning machine, since ensemble learning results are not sensitive to the setting of the kernel function and parameters. In order to improve the classification accuracy and to avoid the optimal selection of kernel function and parameters of the SVM, a SVM ensemble classifier based on SICA is proposed to analyze the DNA microarray data.

II. Materials and Methods

1. DNA microarray datasets

In our study, three DNA microarray datasets were used to compare the performances of different classification methods. They are the Leukemia, Hepatocellular and SRBCT datasets, which are publicly available at http://www.ailab.si/orange/ and http://homes.esat.kuleuven.be/~npochet/Bioinformatics/ respectively. An overview of the characteristics of all the datasets can be found in Table 1. The first two datasets deal with binary cancer classification problems. The last one consists of four types of childhood cancers. The Leukemia dataset contains 72 samples, including 47 Acute Lymphoblastic Leukemia (ALL) and 25 Acute Myeloid Leukemia (AML) cases. The Hepatocellular dataset contains 60 samples, including 40 Hepatitis C Virus (HCV) and 20 Hepatitis B Virus (HBV) cases. The SRBCT dataset contains 63 samples, including 23 Ewing's Sarcoma (EWS), 8 Burkitt's Lymphoma (BL), 12 Neuroblastoma (NB) and 20 Rhabdomyosarcoma (RMS) cases. The numbers of gene expression levels of the three datasets are 7129, 7129 and 2308 respectively.

Table 1. Summary of microarray datasets

Dataset	Number of samples	Number of gene expression levels	Sorts of cancers and the distributions
Leukemia	72	7129	ALL(47), AML(25)
Hepatocellular	60	7129	HCV(40), HBV(20)
SRBCT	63	2308	EWS(23), BL(8), NB(12), RMS(20)

2. Kruskal-Wallis test

DNA microarray experiments allow the simultaneous recording of expression levels of thousands of genes. But usually only a small set of genes are suitable in cancer identification and too large a dataset will increase computational complexity and reduce classification speed[11]. Therefore, it is necessary to select a subset of useful genes before cancer classification. In our study, we used the Kruskal-Wallis test[12] to select those genes whose expression levels fluctuate markedly.

Considering a $p_{int} \times n$ data matrix $X_{int} = (x_{ij})_{p_{int} \times n}$, where n is the number of samples, p_{int} is the number of original genes, x_{ij} is the expression level of the ith gene in the jth sample. Suppose there are k independent classes of samples in X_{int}, *i.e.*, $X_c \sim F(x-\theta_c)$, $c = 1, 2, \cdots, k$. These distributions F are continuous functions with the same form but different location parameters θ_c. Suppose $x_1^c, \cdots, x_{n_c}^c$ are samples from X_c, then n can be represented as $n = \sum_{c=1}^{k} n_c$ and the rank of x_q^c in X_{int} is R_{cq}. If the sum and average rank of X_c are denoted as, respectively, $R_c = \sum_{q=1}^{n_c} R_{cq}$ and $\bar{R}_c = R_c/n_c$, then the average rank of X_{int} is $\bar{R} = \sum_{c=1}^{k} \frac{R_c}{n} = \frac{n+1}{2}$. The Kruskal-Wallis test uses $H = \frac{12}{n(n+1)} \sum_{c=1}^{k} n_c(\bar{R}_c - \bar{R})^2$ to illustrate the expression diversity of the same gene among different classes. We calculated the H value of each gene and selected p genes with relatively large H values for later operations.

3. Independent component analysis

ICA is a signal processing method based on high-order statistical information. Its basic principle is the decomposition of the multi-path observational signals into statistically independent components, *i.e.*, source signals[5,13]. Extracting ICs can reduce data noise and make the ICs effectively interpret the high-order statistical information of the data.

Given that p genes are selected via the Kruskal-Wallis test, then the ICA can be modelled on the following hypotheses: (1) the source signals are statistically independent; (2) the number of source signals is less than or equal to the number of observational signals; (3) the number of source signals with a Gaussian distribution is either 0 or 1 and mixing Gaussian signals are inseparable.

The ICA model of $X(t)$ can be expressed as follows[5]:

$$X(t) = A * S(t) \tag{1}$$

where $X(t) = [X_1(t), X_2(t), \cdots, X_p(t)]^T$ is a $p \times n$ data matrix whose rows correspond to observational variables and whose columns are the individuals of the corresponding variables, $A = [a_1, a_2, \cdots, a_m]$ is a $p \times m$ mixing matrix and $S(t) = [S_1(t), S_2(t), \cdots, S_m(t)]^T$ a $m \times n$ source matrix subject to the condition that the rows of $S(t)$ are as statistically independent as possible. The variables contained in the rows of $S(t)$ are called ICs. The observational variables $X(t)$ are linear combinations of these ICs. Estimating the IC can be accomplished by finding the linear combination of the observational variables, *i.e.*, estimating a matrix W to ensure that the following equation is satisfied.

$$S(t) = A^{-1} * X(t) = W * X(t) \tag{2}$$

There are a number of algorithms for performing ICA. In our study, we shall employ the FastICA algorithm[14] to address the problems of cancer classification. In general, if the number of source signals is equal to the number of observational signals, the reconstructed observational signals can comprise much comprehensive information. Therefore, we used the FastICA algorithm to obtain ICs with the same number of variables as the dimensions of the samples.

4. Selective ICA based on reconstructed sample errors

For gene expression data, each IC is of different biological significance and corresponds to a certain observational signal, which can be interpreted as the source signal of a gene expression. Therefore ICA helps to obtain meaningful gene expression information.

As the time series data of gene expression, compared with PCA, the dominant ICs obtained by ICA can reveal a larger underlying structure of the time series. Therefore, it becomes

necessary to analyze selective components independently. In order to find these dominant ICs, we hope that all ICs can be listed in an appropriate order and then a subset of components can be selected according to this order. However, since the order and variance of ICs cannot be calculated, we are unable to select the dominant ICs based on this criterion. On top of that, instability is a problem when performing ICA, especially when there exists noise in the gene expression data. As well, in practical applications, the number of ICs cannot be predicted accurately. Once this number is set larger than the number of source signals, the result may not converge. In view of these shortcomings of ICA, an optimal subset of ICs cannot assure good performance at every occasion.

It is very much necessary to remove inaccurate ICs and to select more accurate ICs to reconstruct new samples. Cheung and Xu[15] proposed a policy to eliminate that part of the ICs that create large reconstruction errors and experimental results of foreign exchange rates proved that this policy is feasible. Based on their policy, a selective ICA method is proposed in our study. The following discussion deals with the appropriate selection of ICs based on their corresponding reconstruction errors.

After the process of performing ICA, we can obtain the mixing matrix $A = [a_1, a_2, \cdots, a_m]$ and the source signal matrix $S(t) = [S_1(t), S_2(t), \cdots, S_m(t)]^T$. The ith DNA microarray gene expression level $X'_{i\bullet}$ is reconstructed by the ith independent component IC_i $(i = 1, \cdots, p)$, *i.e.*, $X'_{i\bullet} = a_i * S_i$. Actually, if the gene expression level for the ith gene of the original microarray is $X_{i\bullet}$, then the mean-square errors of the reconstructed samples are:

$$error_i = \frac{1}{n}\sum_{j=1}^{n}|X_{ij} - X'_{ij}|^2, \quad j = 1, 2, \cdots, n \tag{3}$$

We arranged these reconstructed samples according to the values of errors and selected p' ICs with relatively small errors. Assuming that IC_i is selected, then $a_i = a_i$ and $S_i = S_i$, otherwise, $a_i = 0$ and $S_i = 0$. This way, a new mixing matrix A' and a new source signal matrix S' are formed and the new sample set X_{new} can be denoted as $X_{new} = A' * S'$ based on a subset of the ICs.

5. Support vector machine

The support vector machine is a machine learning method that can effectively deal with small-scale sample problems as well as with samples of large dimensions[16,17]. Therefore the SVM is suitable for analyzing gene expression data. The SVM uses a hypothetical space of linear functions in a kernel-induced feature space that respects the insights provided by a generalization theory and exploits an optimization theory[18].

Suppose that the data set is $\{(x_j, y_j)|j = 1, 2, \cdots, n\}$, where $x_j \in R^p$ is the input vector and $y_j \in \{-1, +1\}$ contains the corresponding labels. In the end, the training of the SVM is transformed into the following optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{n} y_t y_j \alpha_t \alpha_j K(x_t, x_j) - \sum_{j=1}^{n}\alpha_j \\ \text{s.t.} \quad & \sum_{j=1}^{n} y_j \alpha_j = 0, 0 \le \alpha_j \le C \end{aligned} \tag{4}$$

where α_j is a Lagrangian multiplier and C a regularization factor which determines the tradeoff between the empirical error and the complexity of the model. $K(x_t, x_j) = \varphi(x_t) \cdot \varphi(x_j)$ is a kernel function that should satisfy the Mercer theorem. There are many types of kernel functions, such as the polynomial, Gaussian and sigmoid kernels which can be used. Compared with other kernel functions, the Gaussian kernel has the advantages of a simple model, light computational burden, high computational efficiency and ease of realisation. Therefore, we have opted for the Gaussian kernel in our study.

The optimal values of α and b can be obtained by solving Eq.(4) and then the classification model of the SVM can be described as follows.

$$f(x) = \mathrm{sgn}\left[\sum \alpha_j y_j K(x_j, x) + b\right] \tag{5}$$

The SVM was initially proposed to deal with binary classification problems. Multi-classification problems can be solved by combining a number of binary SVMs using any of a number of strategies, such as 'one-Vers-one', 'one-Vers-all', 'DAG' and so on. In our study, we have used the one-Vers-one method.

III. Results and Discussion

1. Experimental operation

The architecture of a DNA microarray data ensemble classifier based on a selective ICA is shown in Fig.1.

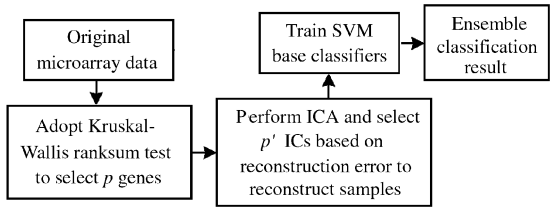

Fig. 1. Architecture of a SVM ensemble classifier based on selective ICA for DNA microarray data

The detailed operational process is described as follows:

Input The original DNA microarray matrix X_{int}, the number of genes whose expression levels fluctuate markedly among different classes p, the number of ICs that participate in reconstructing new samples p', $p' < p$, the number of SVM sub-classifiers N and the number of SVM sub-classifiers that participate in majority voting N'.

Process Adopt the Kruskal-Wallis test to select p genes whose expression levels fluctuate markedly and compose a condensed sample set X, then carry out a standardised operation on X.

for $i = 1 : N$

Perform ICA on X to get a mixing matrix A and a source signal matrix S, reconstruct each gene expression level;

compute the reconstruction errors of p ICs according to Eq.(3);

select p' ICs whose reconstruction errors are relatively small to reconstruct a new sample set X_{new};

train the SVM sub-classifier on X_{new}, use leave-one-out cross-validation to obtain the correct rate r_i.

end

The correct rates of all SVM sub-classifiers are denoted as $r = \{r_1, r_2, \cdots, r_N\}$; we selected the first N' sub-classifiers whose accuracy were relatively high and integrated them using a majority voting method on the classification results of X. We then obtained a final classification rate r_i.

Output The correct rates $\{r_1^*, r_2^*, \cdots, r_{N'}^*\}$ of those SVM sub-classifiers that participated in majority voting and the correct rate of the SVM ensemble classifier $r_{ensemble}$.

For the Leukemia, Hepatocelluar and SRBCT datasets, described in Section II.1, we adopted, at first, the Kruskal-Wallis test to select 10, 10 and 20 genes respectively whose expression levels fluctuate markedly among the different classes. Then the FastICA algorithm was applied to extract ICs with the same specific numbers, *i.e.*, 10, 10 and 20. In a third step, we selected appropriate ICs according to the corresponding reconstruction errors, *i.e.*, from the first two datasets 6, 7, 8 and 9 ICs were selected and from the third set 16, 17, 18 and 19 ICs. Then 25 SVM sub-classifiers were trained on the new reconstructed samples. Finally, five SVM sub-classifiers with relatively high correct rates were chosen and the final results were integrated by the majority voting method.

2. Experimental results

Many experimental results show that the process of performing ICA and selecting a subset of ICs to reconstruct samples will make the correct rates of SVM sub-classifiers unstable. Therefore, a proper number of sub-classifiers should be trained to reflect all the possible results. We carried out four experiments on the three datasets. Tables 2-4 show the minimum and the maximum of correct rates of the 25 SVM sub-classifiers and the ensemble result. Figs.2–4 provide the boxplots of correct rates of the four experiments respectively, where the x-axis denotes the test number and the y-axis the correct rate.

Table 2. Results on the Leukemia dataset

Test number	Number of ICs used for reconstruction	Correct rate		
		Minimum	Maximum	Ensemble result
1	6	0.7778	0.9167	0.9444
2	7	0.8472	0.9306	0.9444
3	8	0.8472	0.9583	0.9583
4	9	0.8750	0.9583	0.9583

There are outliers on the first and the second box in Fig.2, the fourth box in Fig.3 and the second box in Fig.4, which shows that the SVM sub-classifier may obtain singularly excellent or poor performances in an experiment. If the correct rate of the sub-classifier is very low, it shows that too many ICs did not converge during the process of performing ICA. It will compromise the integrity of the reconstructed samples. In this case, we not only reached the goal of purifying the data, but also imported new noise to the data. Therefore it is imperative to reserve excellent classifiers and to eliminate poor ones from the ensemble.

Table 3. Results on the Hepatocellular dataset

Test number	Number of ICs used for reconstruction	Correct rate		
		Minimum	Maximum	Ensemble result
1	6	0.6333	0.7500	0.7500
2	7	0.5833	0.7500	0.7500
3	8	0.6167	0.8000	0.7667
4	9	0.6833	0.7667	0.7667

Table 4. Results on the SRBCT dataset

Test number	Number of ICs used for reconstruction	Correct rate		
		Minimum	Maximum	Ensemble result
1	16	0.6508	0.8413	0.9048
2	17	0.7302	0.9048	0.9206
3	18	0.7619	0.9683	0.9524
4	19	0.8571	0.9683	0.9524

From Figs.2–4, it can be seen that if more ICs were deleted, the five statistics in the boxplots, *i.e.*, the minimum value, the first quarter value, the median value, the third quarter value and the maximum value will fall accordingly (except the third experiment on the Hepatocellular dataset shown in Fig.3). This shows that the correct rate of the sub-classifier varies according to the number of ICs used for reconstruction. If more ICs were deleted, the correct rate of the ensemble classifier would have clearly improved relative to each sub-classifier. This result also can be obtained from Tables 2–4. We can conclude that when more ICs are deleted, the differences and mutual complementarities for each sub-classifier are enhanced. This phenomenon adequately validates the advantage of ensemble learning. It should be noted that if an excessive number of ICs were deleted, the performance of the classifier would deteriorate and the result would become unstable.

3. Comparison of experiments

In order to show the efficiency and feasibility of the proposed method, called SICA+SVM in our study, the highest correct rates, using other two methods, are listed in Table 5 for comparison. In method 1, microarray data are classified by SVM directly. In method 2, all the ICA features are used to train SVM for classification.

Table 5. Highest correct rates of the three methods on the three datasets

No.	Method	Leukemia	Hepatocellular	SRBCT
1	SVM	0.9583	0.7500	0.9206
2	ICA+SVM	0.9583	0.7500	0.9206
3	SICA+SVM	0.9583	0.7667	0.9524

It is clear that if all the ICs were used for sample reconstruction, the correct rate of the classifier would not always be

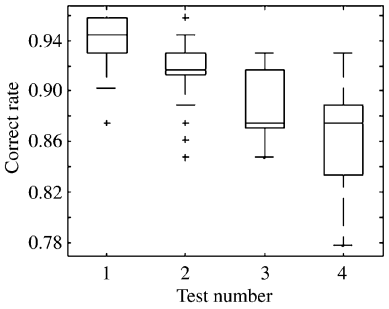

Fig. 2. Boxplot of correct rates on the Leukemia dataset

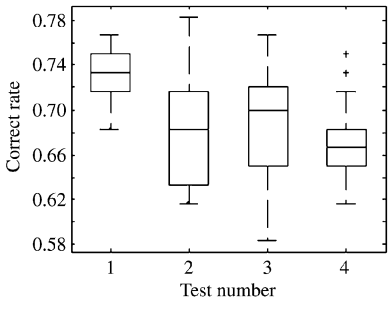

Fig. 3. Boxplot of correct rates on the Hepatocellular dataset

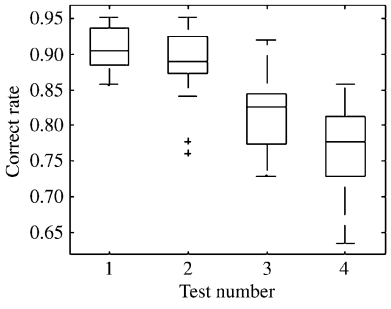

Fig. 4. Boxplot of correct rates on the SRBCT dataset

better than using the SVM directly. If we were to choose a subset of the ICs which contribute to good sample reconstruction, the results might improve.

In order to validate the rationality of the method proposed by us, we tried to select some ICs randomly for reconstruction. It turns out that this method could not obtain a stable result, even after the ensemble.

IV. Conclusion

When dealing with traditional ICA, the limitation of the algorithm itself cannot be avoided. If ICA is applied to microarray data analysis directly, the result may be unstable. In order to solve this problem, we have used the reconstruction criterion based on mean-square errors to select a subset of independent components, which has been proven feasible in the field of time series analysis. Therefore, based on our investigation, the DNA microarray data ensemble classifier, based on a selective independent component analysis, has been proposed. The samples are reconstructed by a subset of independent components and several SVM sub-classifiers are trained simultaneously. Finally, some excellent sub-classifiers with high correct rates were selected to participate in the ensemble, using a majority voting method. Our results, based on three publicly available microarray datasets, show the feasibility and validity of the method proposed by us.

References

[1] R. Bellazzi, B. Zupan, "Towards knowledge-based gene expression data mining", *Journal of Biomedical Informatics*, Vol.40, No.6, pp.787–802, 2007.

[2] X.G. Lu, Y.P. Lin, H.J. Wang *et al.*, "A relative space based cancer classification with gene expression profile", *Acta Electronica Sinica*, Vol.36, No.4, pp.614–619, 2008. (in Chinese)

[3] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis", *Bioinformatics*, Vol.18, No.1, pp.51–60, 2002.

[4] P. Chiappetta, M.C. Roubaud, B. Torresani, "Blind source separation and the analysis of microarray data", *Journal of Computational Biology*, Vol.11, No.6, pp.1090–1109, 2004.

[5] D.S. Huang, C.H. Zheng, "Independent component analysis based penalized discriminant method for tumor classification using gene expression data", *Bioinformatics*, Vol.22, No.15, pp.1855–1862, 2006.

[6] K.H. Liu, D.S. Huang, "Cancer classification using rotation forest", *Computers in Biology and Medicine*, Vol.38, No.5, pp.601–610, 2008.

[7] S.J. Kim, J.K. Kim, S.J. Choi, "Independent arrays or independent time courses for gene expression time series data analysis", *Neurocomputing*, Vol.71, No.10-12, pp.2377–2387, 2008.

[8] A.S. Carpentier, A. Riva, P. Tisseur *et al.*, "The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA", *Computational Biology and Chemistry*, Vol.28, No.1, pp.3–10, 2004.

[9] M.S. Mohamad, S. Deris, R.M. Illias, "A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray", *International Journal of Computational Intelligence and Applications*, Vol.5, No.1, pp.91–107, 2005.

[10] Z.H. Zhou, J.X. Wu, W. Tang, "Ensemble neural networks: many could be better than all", *Artificial Intelligence*, Vol.137, No.1-2, pp.239–263, 2002.

[11] Q. Shen, W.M. Shi, W. Kong, "New gene selection method for multiclass tumor classification by class centroid", *Journal of Biomedical Informatics*, 2008; doi:10.1016/j.jbi.2008.05.011.

[12] G.D. Ruxton, G. Beauchamp, "Some suggestions about appropriate use of the Kruskal-Wallis test", *Animal Behaviour*, 2008; doi:10.1016/j.anbehav.2008.04.011.

[13] Y. Chen, R.S. Wang, "A method for texture classification by integrating gabor filters and ICA", *Acta Electronica Sinica*, Vol.35, No.2, pp.299–303, 2006. (in Chinese)

[14] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Transactions on Neural Networks*, Vol.10, No.3, pp.626–634, 1999.

[15] Y.M. Cheung, L. Xu, "An empirical method to select dominant independent components in ICA for time series analysis", *Proceedings of the Joint Conf. on Neural Networks*, Washington, DC, USA, pp.3883–3887, 1999.

[16] F.H. Song, P. Li, "Nonlinear internal model control based on support vector machine αth-order inverse system method", *Acta Automatica Sinica*, Vol.33, No.7, pp.778–781, 2007. (in Chinese)

[17] Xuesong Wang, Xilan Tian, Yuhu Cheng, "Value approximation with least squares support vector machine in reinforcement learning system", *Journal of Computational and Theoretical Nanoscience*, Vol.4, No.7/8, pp.1290–1294, 2007.

[18] V. Vapnik, *The Nature of Statistical Learning Theory*, New York, Springer Verlag, 1995.

WANG Xuesong received the Ph.D. degree from China University of Mining and Technology in 2002. From July 2002 to June 2004, she worked as a postdoctoral researcher at Beijing Institute of Technology. Since 2004, she has been an associate professor in the School of Information and Electrical Engineering, China University of Mining and Technology. Her main research interests include machine learning, bioinformatics, artificial intelligence and intelligent system. (Email: wangxuesongcumt@163.com)

GU Yangyang received the B.S. degree from China University of Mining and Technology in 2007. She is currently a M.S. candidate in the School of Information and Electrical Engineering, China University of Mining and Technology. Her main research interests include kernel machine learning and bioinformatics.

CHENG Yuhu received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2004. He is currently an associate professor in the School of Information and Electrical Engineering, China University of Mining and Technology. His main research interests include machine learning and intelligent system.

ZHANG Zheng received the B.S. degree from Shandong University of Science and Technology in 2005. She is currently a M.S. candidate in the School of Information and Electrical Engineering, China University of Mining and Technology. Her main research interests include support vector machine and reinforcement learning.