SVM Classification Method Based Marginal Points of Representative Sample Sets

Wencang Zhao[1], Guangrong Ji[2], Rui Nian[2], and Chen Feng[2]

[1] College of Automation and Electronic Engineering, Qingdao University of Science & Technology, 53 ZhengZhou Road, Qingdao, 266042, China
zhaocenter-journal@yahoo.com.cn

[2] College of Information Science and Engineering, Ocean University of China, 5 YuShan Road, Qingdao, 266003, China
grji@ouc.edu.cn, nianrui_80@163.com, fccjg@sdu.edu.cn

Abstract

For the sake of getting the training data from the data itself without other previous knowledge, we present a method to calculate the representative sample sets of each cluster only based on the intrinsic character of high dimensional data, according to different objects influencing the assembling state in some dimensions of high dimensional data set. In order to classify the other samples of the data set, we adopt Support Vector Machine (SVM) as the classifier for its ability to analyze the small sample sets. During the SVM's training course, we put forward a method to select the marginal points of the representative sample sets as the different clusters' approximate support vectors to train the SVM to boost the machine's classification capability. Lastly, we analyzed the hyperspectral data to detect red tide by this method, which proved the method could classify the data effectively and the selecting method of the SVM's training samples could quickly and efficaciously boost the machine's training course.

Keyword: Representative Sample Set, Training Data, Marginal Point, Support Vector Machine, High Dimensional Data Set

I. Introduction

The dimension of the remote sensing data is more and more along with the developing of the science and technology. For example the aerial remote sensing data whose dimensions are 124 or 256 is a kind of typical high dimensional data and each dimension can produce an image. The regular data analyzing methods will encounter the dimensionality curse problem when recognizing the objects using this kind of data. At the same time, the data classification is the main analyzing method for the information extracting.

When we classify the high dimensional data using supervised learning method, the selection of the training data influence the classification result. How can we get the training data from the data itself without other previous knowledge? Various classification methods have been studied in order to

solve this problem. Because the different objects result in the data's different assembling state in some different dimensions, [1] given a method to get the objects' information through analyzing their relative dimensions blindly; [2] designed a scheme for clustering points in a high dimensional data sequence for the subsequent indexing and similar search; a new neural networks architecture (PART) and a resulting algorithm were proposed in [3] to find the projected clusters for data sets in high dimensional spaces.

Support Vector Machine (SVM) is a new popular supervised classification algorithm, because of its ability to "learn" classification rules from a set of training data [4], and moreover it is fit for processing the high dimensional data [10]. However, SVM runs the risk of learning the training data too closely, resulting in SVM do not perform well when presented with new, unknown data. So, how to select the training data is an important work during the application course of the SVM, usually it mainly depends on user's experiences. Some methods have been discussed to guarantee the quality of the training data. [5] proposed an efficient caching strategy to accelerate the decomposition methods and the Platt's Sequential Minimization Optimization (SMO) algorithm was improved by the caching strategy, which could increase classification accuracy also; the training with jitter method was presented in [4,6,7] to boost the SVM's generalization on new data.

In order to classify the large numbers of data set with high dimension, this paper presents an effective method to get the representative sample sets based the data's intrinsic character to train SVM. In order to get the training samples from the data set, we use parzen window method to calculate the probability density of every dimension's data to gain the sensitive dimensions, then intersect the same class of different sensitive dimensions to get the representative sample sets and the cluster number, and last select the marginal points in those sets as the cluster's approximate support vectors to train SVM to classify the other data.

II. Calculating Method of Representative Sample Sets

A. Probability Density Function (PDF) Estimating by Parzen Window Method

The probability density function estimating method of parzen window assumes that the scope R_n is a super-cube with d dimensions [8]. If one border of the super-cube is h_n, the cubage of it is:

$$V_n = h_n^d \tag{1}$$

We can gain the number of the samples, which drop into the super-cube, through the following window function:

$$\phi(u) = \begin{cases} 1 & \left|u_j\right| \le 1/2; \quad j = 1,..., d \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Where $\phi(u)$ is a unit whose center locates at the origin, and if x_i drops into the super-cube, $\phi((x - x_i)/h_n) = 1$; otherwise, $\phi((x - x_i)/h_n) = 0$. So the number of the samples falling into the super-cube is:

$$k_n = \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h_n}\right) \tag{3}$$

If $p_n(x)$ denotes the *nth* estimation of the PDF $p(x)$, we could gain the normal probability density function estimating method as equation (4):

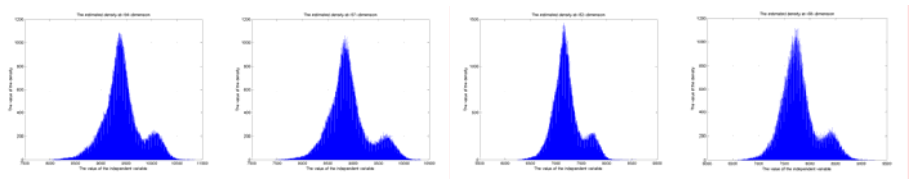
$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{v_n}\phi(\frac{x-x_i}{h_n}) \tag{4}$$

B. Calculating Representative Sample Sets of High Dimensional Data

The hyperspectral data has more than 100 dimensions. If the data could be classified into several clusters, it should contain some dimensions that are sensitive to be classified. So, we could select the sensitive dimensions to analyze the clusters' information.

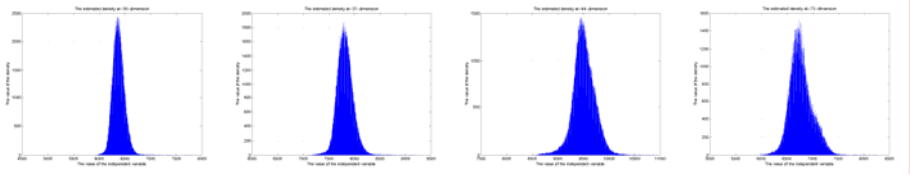
In order to save time and analyze expediently, we use parzen window method to compute every dimension's PDF, and then analyze the information of different clusters through the relative dimensions' PDF. According to the analysis of different dimensions' PDF, we can divide the data's dimensions into two classes: the sensitive dimensions set and the non-sensitive dimensions set. The first one is composed of the dimensions that have more than one rallying points which reflects the different clusters' assembling information, and the dimensions that have one rallying point constitute the another class. For example, Fig. 1 shows the PDFs of part dimensions in the data set (note: the data set used in this paper is a 400×400 data block with 124 dimensions), so, the dimensions in Fig. 1 (a) have more separable information and are elements of the sensitive dimensions set, and the dimensions in Fig. 1 (b) can not easily separable and they are elements of the non-sensitive dimensions set.

In order to select the sensitive dimensions automatically, we do the following two steps to enhance their optional ability.



(54) (57) (62) (68)

(a) The multi-rallying point dimensions



(30) (37) (44) (75)

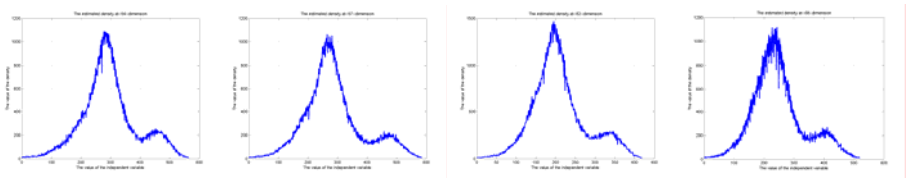
(b) The one rallying point dimensions

Fig. 1. All the data's dimensions can be separated into two kinds. (The number under each figure is the dimension No. in the data set.) Fig. (a) shows the dimensions that can be separated easily and they are called the sensitive dimensions of the high dimensional data set, and the ones in Fig. (b) have little separable information.

Step 1: Extracting 0 Processing

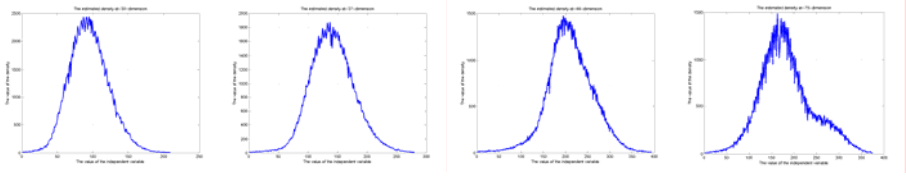
We check all the values of each dimension's PDF and delete the points whose values are 0. At the same time, we design the two dimensions vectors, $NONZero_i, i = 1,2,...,n$ (n is the number of

the sensitive dimensions), to record the non-zero points' real location, and then get the PDFs after the extracting processing, as Fig. 2 shows. From the results, we could find that the dimensions' PDFs are recognized easily than the original ones, but they have some fluctuations yet.



(54) (57) (62) (68)

(a) The multi-rallying point dimensions



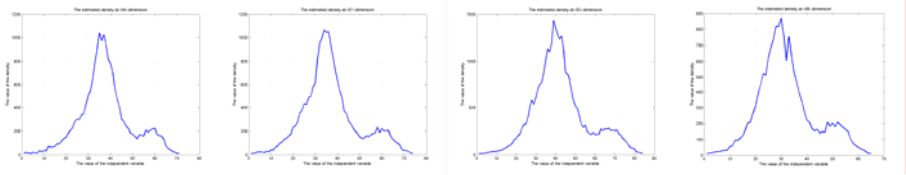
(30) (37) (44) (75)

(b) The one rallying point dimensions

Fig. 2. The PDFs results after the extracting 0 processing

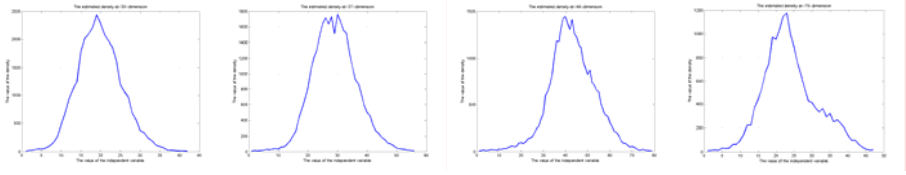
Step 2: Smoothness Processing

In order to smooth the PDF's curve, we use the down symbol method to do so. Here we down one symbol every five ones and the vectors $NONZero_i, i = 1,2,...,n$ are revised also. The finally results of the above dimensions are as Fig. 3 shows. Now we might find that it is easy to recognize the rallying points' number automatically by the configurational feature of the PDF's wave shape [9].



(54) (57) (62) (68)

(a) The multi-rallying point dimensions



(30) (37) (44) (75)

(b) The one rallying point dimensions

Fig. 3. The PDF results after the smoothness processing. (a) shows the dimensions that are sensitive and could be classified by assembling information, and (b) is the ones that can not be separated easily.

Later, we calculate the representative sample sets as follows:

Firstly, we analyze all the dimensions of the data set D using this method, and gain all sensitive dimensions in the high dimensional data set,

$$SD = D_1 + \bullet D_2 + \bullet ... + \bullet D_n \tag{5}$$

where n is the number of the sensitive dimensions and the operator "$+\bullet$" denotes the set SD is composed of the dimensions D_i , i = 1,2,..., n .

Secondly, we classify each sensitive dimension into several clusters according to single rallying points assembling state,

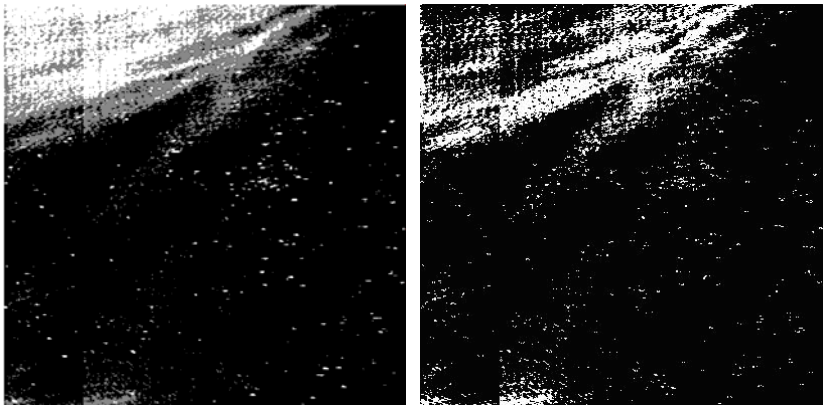
$$\begin{aligned} D_1 &= C_{11} \bigcup C_{12} \bigcup ... \bigcup C_{1m}, \\ D_2 &= C_{21} \bigcup C_{22} \bigcup ... \bigcup C_{2m}, \\ &..., \\ D_n &= C_{n1} \bigcup C_{n2} \bigcup ... \bigcup C_{nm} \end{aligned} \tag{6}$$

where m is the cluster number that the data set should be separated to and C_{ij} denotes the j th cluster of the i th sensitive dimension.

Lastly, we get the representative samples of all clusters through computing the intersecting sets of the same cluster, and the samples compose the representative sample set:

$$\begin{aligned} \mathrm{RSS} = {} & C_{11} \bigcap C_{21} \bigcap ... \bigcap C_{n1} \bigcup C_{21} \bigcap C_{22} \bigcap ... \bigcap C_{n2} \bigcup ... \\ & \bigcup C_{n1} \bigcap C_{n2} \bigcap ... \bigcap C_{nm} = C_1 \bigcup C_2 \bigcup ... \bigcup C_m \end{aligned} \tag{7}$$

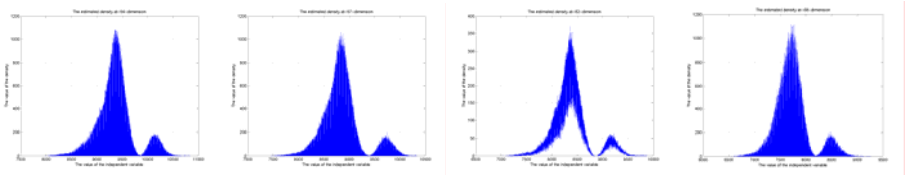
Here, we gain the representative samples of every cluster, and the other data constitutes the non-determinate data set. The results are showed as Fig. 4 (This data set can be separated into two clusters.). The white points in Fig. (a) are the representative samples of cluster 1, the black ones are the representative samples of cluster 2, and the other ones are non-determinate data, in order to observe easily, which are showed again as the white points in Fig. (b).



(a) (b)

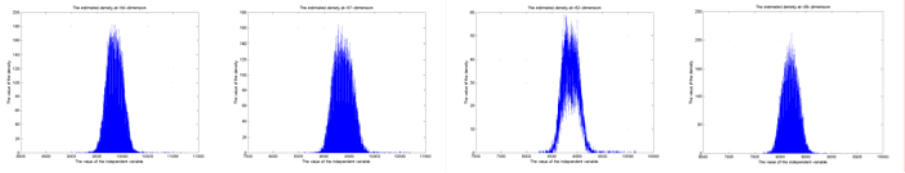
Fig. 4. The computing results of the representative sample set. In Fig. (a), the white points and black points compose representative sample set, and the other ones showed again in Fig. (b) are non-determinate data.

Most of the dimensions of the representative sample set possess the completely separable character. We show some dimensions' PDFs in the representative sample set and the non-determinate data set to indicate their separable character (Fig. 5). From Fig. 5(a), we can see the classifying mistake ratio of the representative sample set is nearly zero according to Bayesian least mistake ratio decision. But the same dimensions in the non-determinate data set show no information to classify the data easily.



(54) (57) (62) (68)

(a) The PDFs of some dimensions in the representative sample set



(54) (57) (62) (68)

(b) The PDFs of the same dimensions in the non-determinate data set

Fig. 5. (a) shows the completely separable character of the data, and (b) shows the completely non-separable feature.

III. Classifying Non-determinate Data by SVM

Now, we get the representative samples of every cluster and the clusters' number that the high dimensional data should be classified to. Although the hyperspectral data has the features of high dimension and large numbers, it is a typical small sample data set relative to its high dimension character. Because SVM is a new learning machine to deal with the small sample data, we adopt SVM to classify the data in non-determinate data set.

A. Overview of SVM

Support vector machine (SVM) is a statistical classification method proposed by Vapnik in 1998 [10]. Given *m* labeled training samples, $\{(x_i, y_i) \mid x_i \in R^n, y_i \in \{-1,1\}, i = 1,2,...,m\}$, SVM is able to generate a separation hypersurface that has maximum generalization ability. Mathematically, the decision function can be formulated as:

$$f(x) = sign\{\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) - b\} \tag{8}$$

where x_i and b are the parameters determined by SVM's learning algorithm, and $K(x_i, x)$ is the kernel function which implicitly maps the samples to a higher dimensional space. Those samples x_i with nonzero parameters α_i are called "support vectors" (SVs).

To find the coefficients α_i in the separable case, it is sufficient to find the maximum of the functional:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{9}$$

Subject to the constraints:

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1,2,...,l \tag{10}$$

This functional coincides with the functional for finding the optimal hypersurface. The learning machines that construct decision functions of the type (8) are called Support Vector Machines (SVM).

B. Classification of Non-determinate Data

In order to boost the training and recognizing efficiency, we take the following measures:

The functional (8) shows that only the support vectors judge the optimum hypersurface for classification and the other non-support vectors has little action on the classification. The representative sample sets we have calculated are the typical samples belong to each class and they do not intercross, so, the marginal samples would near the ones which would construct the support vectors of the SVM. Fig. 6 shows the schematic relations among the hypersurface, the marginal samples, and the support vectors (the solid curves, dashed curves, and the gray circles or crosses denote the hypersurface, the marginal samples, and the support vectors of two classes respectively).

In this paper, we select the marginal samples of the representative sample sets by the Canny algorithm as the clusters' approximate support vectors to train the SVM , which not only reduce the amount of the training samples, but also boost the learning course. The realization courses are as follows.

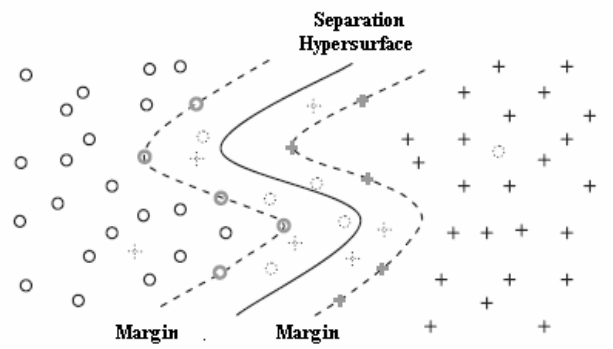
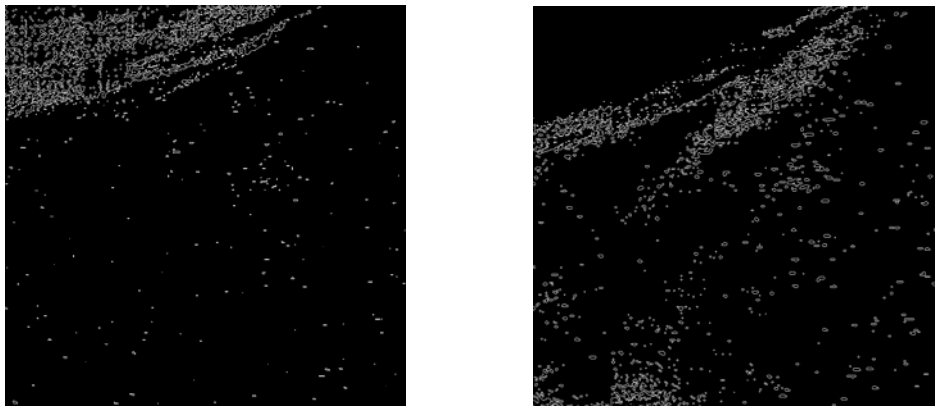


Fig. 6. Schematic explanation of the separation optimum hypersurface (solid curves), marginal samples (dashed curves) and support vectors of SVM (gray circles/crosses). The positive and the negative training samples are indicated by circles and crosses, respectively.

Firstly, we calculated the margin of the representative sample sets by Canny algorithm, which are as Fig. 7.



(a) (b)

Fig. 7. The marginal points calculated by Canny algorithm. Fig. (a) shows the red tide cluster marginal points and the Fig. (b) denotes the ocean water ones.

Secondly, we respectively selected 150 samples from different clusters' marginal points to train the SVM and stopped the training course according to the power MSE's convergent line. We stopped the training after 30,000 in this paper as Fig. 8 showed. Now the samples with the nonzero coefficients α_i construct the optimal hypersurface which can separate the data into two classes.

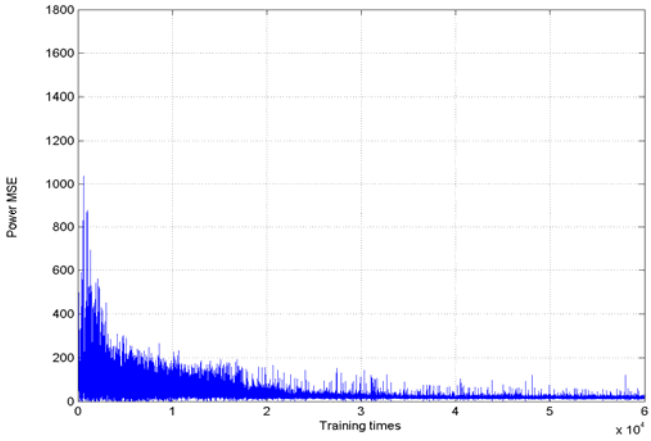


Fig. 8. The MSE of the power is constringed after 30,000 training times.

Lastly, we recognized the non-determinate data by the trained SVM and got the classification results as Fig. 9 showed, in which the white points are red tide points and the black ones are ocean water points.

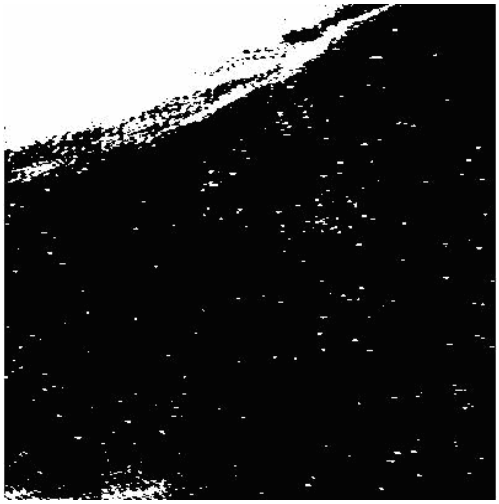


Fig. 9. The classification results of the 400×400 data block with 124 dimensions, in which the white points are red tide points and the black ones are ocean water points.

IV. Conclusions

The data classification is the main analyzing method for data information extracting; moreover, the training data and the cluster number are very important for many methods, such as SVM. But, in most cases we cannot know any previous knowledge about them, and we should extract the information from the data itself. In this paper, based on the data's intrinsic character, we put forward a method to get the representative samples of every cluster and the cluster number, and then classify

the data of the non-determinate data by SVM. In order to boost the learning and classifying efficiency, we select the marginal samples of the representative sample sets as the approximate support vectors to train the SVM. At last, we used a 400×400 aerial remote sensing data block with 124 dimensions to recognize the red tide by this method, which proved effective to classify the data set.

References

[1]. W. C. Zhao, G. R. Ji, C. Feng, R. Nian, Blindly Selecting Method of Training Samples Based Hyper-spectral Image's Intrinsic Character for Object Recognition. Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology. Suzhou, China, 2005, pp. 113-116

[2]. S. L. Lee, C. W. Chung, On the effective clustering of multidimensional data sequences. Information Processing Letters, vol. 80, 2001, pp. 87-95

[3]. Y. Q. Cao, J. H. Wu, Projective ART for clustering data sets in high dimensional spaces. Neural Networks, vol. 15, 2002, pp. 105-120

[4]. C. Nello, S. T. John, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, England, 2000

[5]. J. Sun, N. N. Zheng, Z. H. Zhang, An Improved Sequential Minimization Optimization Algorithm for Support Vector Machine Training. Journal of Software, vol. 13, 2002, pp. 2007-2013

[6]. L. Holmstrom, P. Koistinen, Using additive noise in back-propagation training. IEEE Trans. Neural Networks, vol. 3(1), 1992, pp. 24-38

[7]. G. An, The effects of adding norse during backpropagation training on a generalization performance. Neural Comprter, vol. 8, 1996, pp. 643-674

[8]. R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, second edition. Indianapolis: Wiley-Interscience, 2001, pp. 134-135

[9]. W. C. Zhao, G. R. Ji, Ocean Red Tide Recognition Method Based Absorbing and Reflecting Crest of Hyper-spectral Images. Acta Oceanologica Sinica,(In press)

[10]. V. N. Vapnik, Statistical Learning Theory. Wiley, New York, 1998

Acknowledgements

This research was supported by the National 863 Natural Science Foundation of P. R. China（2001AA636030）and the Doctoral Fund of Qingdao University of Science & Technology.

Wencang Zhao received the M.Sc and PhD degrees from ShanDong University in 2002 and Ocean University of China in 2005 respectively. He is with the College of Automation and Electronic Engineering of Qingdao University of Science & Technology as an associate professor now.

Dr Zhao is interested in the analysis of statistical signal, the image processing and pattern recognition, the artificial intelligence, and the machine learning.

Guangrong Ji is a professor of the Department of Electronic Engineering, Ocean University of China and a doctoral tutor. His major researches are the image analysis, pattern recognition, artificial neural network, the analysis of the statistical signal and so on.

Rui Nian is a doctoral student in Ocean University of China now. Her major is Ocean Information detection and processing. She earned her Bachelor and Master Degree in Information processing from Ocean University of China too. Her primary research interests include pattern recognition, image processing, neural networks and high-dimensional space theory.

Feng Chen received the M.Sc degree in signal and information processing from Shandong University .Now she is pursuing her PhD degree at Ocean University of China. Her research interests include nonlinear time series analysis and pattern recognition.