Hyperspectral image classification and application based on relevance vector machine

DONG Chao, ZHAO Huijie

Precision Opto-mechatronics Technology, *Key Laboratory of Education Ministry*, *School of Instrument Science and Opto-electronics Engineering*, *Beijing University of Aeronautics and Astronautics*, *Beijing* 100191, *China*

Abstract: The relevance vector machine (RVM) is used to process the hyperspectral image in this paper to estimate the classifiers precisely in the high dimensional space with limited training samples. The detail of RVM is firstly discussed based on the sparse Bayesian theory. Then four multi-class strategies are analyzed, including One-vs-All (OAA), One-vs-One (OAO) and two direct multi-class strategies. In the experiments, the multi-class strategies are compared and RVM is further compared with several classical classifiers, including the support vector machine (SVM). The experiments show that two direct multi-class strategies occupy too much memory space with low efficiency. OAA has the highest precision, but is low in efficiency. OAO is the best in efficiency and the precision approximates to OAA. Compared with SVM, RVM is low in precision, but sparser than SVM. The sparse property is important when the test set is large, which makes RVM suitable for classifying the large-scale hyperspectral image.

Key words: remote sensing, classification, relevance vector machine, hyperspectral

CLC number: TP79 **Document code:** A

Citation format: Dong C, Zhao H J. 2010. Hyperspectral image classification and application based on relevance vector machine. *Journal of Remote Sensing*. **14**(6): 1273—1284

1 INTRODUCTION

The high spectral resolution of the hyperspectral sensor results in huge volume data and brings out new challenge of processing the remote sensing image. Affected by the Hughes phenomenon (Hughes, 1968), traditional classifiers could not be estimated precisely in the high dimensional space with limited training samples, the maximum likelihood and artificial neural network for instance. To solve the problem, sufficient training samples are required. The collection of the training samples is resource-consuming and not recommended in the applications, which calls urgently for designing the limited-training-samples classifiers. Recent development of this field can be roughly classified into four categories: (1) regularization of the covariance matrix (Tadjudin & Landgrebe, 1999); (2) feature extraction and feature selection (Kuo & Landgrebe, 2004); (3) semi-supervised methods (Dundar & Landgrebe, 2004; Jackson & Landgrebe, 2001); (4) low-complex classifiers (Melgani & Bruzzone, 2004), support vector machine (SVM) for instance. SVM is the best supervised learning method at present and can deal with the limited-trainingsamples problem in the high dimensional space. SVM maximizes the margin between classes, contributing to minimizing the training error and guaranteeing the generalization ability. In addition, the margin maximization rule ensures that the informative samples, also named support vectors, always appear nearby the boundary of classes. The non-informative samples are not involved in predicting the label of the test samples, thus the solution of SVM is sparse. However, the sparsity of SVM is not obvious in applications. The quantity of the support vectors is proportional to that of the training samples, which has a negative impact on the efficiency of classifying the large-scale hyperspectral image. Additionally, SVM contains the following defects.

(1) SVM could not output the probability of the prediction. The probability density function is desired in the applications, which can be used to measure the uncertainty of the prediction.

(2) The grid search and cross validation methods are used to estimate the excessive parameters of SVM. It is a waste of the computing resource.

(3) The kernel functions must satisfy the Mercer condition.

To avoid the defects above, Tipping (2000, 2001) derived the relevance vector machine (RVM) from the sparse Bayesian learning theory. RVM has been widely applied in the pattern recognition fields, including the electronic nose monitoring (Wang *et al*., 2009), spam classification (Yu & Xu, 2008) and visual tracking (Williams *et al*., 2005). The application of RVM in the remote sensing community emerged in the recent two years. Demir and Ertürk (2007) used RVM to classify the hy-

Received: 2009-11-09; **Accepted**: 2010-05-16

Foundation: China 863 project (No.2008AA121102) and China Geological Survey (No. 1212010816033).

First author biography: DONG Chao (1982—), Ph.D candidate in Bei Hang University. He majors in data processing techniques of the hyperspectral remote sensing image. E-mail: dongchaoxj888@126.com

perspectral image. The experiments showed that RVM is slightly worse than SVM in precision, but the sparsity property makes it efficient in classifying the large-scale hyperspectral image. Foody (2008) discussed the multi-class ability of RVM and compared it with decision analysis, decision tree, neural network and SVM. Camps-Valls (2006) retrieved the oceanic chlorophyll concentration with RVM, to monitor the water quality of the coast. The aforementioned researches are exploratory and leave lots of unsolved problems, including the training efficiency, the multi-class strategies and the under-fitting problem. Beginning at the sparse Bayesian theory, we analyze the detail of RVM and discuss four kinds of multi-class methods in this paper. In the experiments, RVM is compared with SVM in the aspects of precision and sparsity, to reveal its ability of classifying the hyperspectral image.

2 SPARSE BAYESIAN LEARNING

For the samples $\{\boldsymbol{x}_n\}_{n=1}^N$ and outputs $\{t_n\}_{n=1}^N$, the supervised learning algorithms estimate the function $t_n=f(\boldsymbol{x}_n)$ to describe the input-to-output relationship, which is further used to predict the outputs of new samples. The outputs are real values in regression and class labels in classification. The function could be defined as the linear combination of the basis functions in the input space.

$$y(\boldsymbol{x};\boldsymbol{\omega})=\sum_{n=1}^{N}\omega_n K(\boldsymbol{x},\boldsymbol{x}_n)+\omega_0 \tag{1}$$

where K(.,.) are the basis functions and $\boldsymbol{\omega}=\{\omega_n\}_{n=0}^N$ are the weights. The training procedure seeks the optimal parameters $\{\boldsymbol{\omega}_n\}_{n=0}^N$, which can both reveal the characteristic of the training samples and be helpful for predicting the outputs of the test samples precisely.

The parameters determine the complexity of the function *f*. If $\boldsymbol{\omega}$ is dense, *f* will be complex enough to approximate the training samples. However, because of the noise (regression) and overlap (classification), the approximation of the training samples could not guarantee the predicating accuracy of the test samples. If the over-complex system is adopted to describe limited training samples, it will usually cause over fitting and the prediction will be unreliable. The learning system should match the samples in complexity. It should be neither too simple (under fitting) nor too complex (over fitting).

The sparse Bayesian learning theory derives from the statistical method. It adjusts the complexity of the function by adding constraint on ω, such as RVM, sparse multinomial logistic regression (SMLR) and joint classification and feature optimization (JCFO). RVM uses the automatic relevance determination (ARD) framework (Tipping, 2001). It assumes that ω_n obeys the Gaussian distribution with mean zero and covariance α_n^{-1}. SMLR adopts the Laplacian distribution and solves the multi-class problem through the multinomial logistic regression (Krishnapuram *et al*., 2005). JCFO adds the constraints on both the parameters and features (Krishnapuram *et al*., 2004). Except for the sparse property, it can also select the optimal features for the classification.

3 RELEVANCE VECTOR CLASSIFICATION

For binary classification, the outputs are either 0 or 1. The Bernoulli distribution is adopted to construct the conditional probability density function $p(\boldsymbol{t}|\boldsymbol{\omega})$ and $y(x)$ is mapped into [0, 1] by the Sigmoid link function. Based on the definition of the Bernoulli distribution, the likelihood function is shown in the following. Equations.

$$p(\boldsymbol{t}\mid\boldsymbol{\omega})=\prod_{n=1}^{N}y_n^{\,t_n}(1-y_n)^{1-t_n} \tag{2}$$

where $\boldsymbol{t}=(t_1, t_2, \cdots, t_N)^{\mathrm{T}}$, $\boldsymbol{\omega}=(\omega_1, \omega_2, \cdots, \omega_N)^{\mathrm{T}}$, $y_n=\sigma\{y(x_n;\omega)\}$ and $\sigma(y)$ is the Sigmoid function.

$$\sigma(y)=1/(1+e^{-y}) \tag{3}$$

Calculating the derivative of Eq. (2) with respect to $\boldsymbol{\omega}$, the maximum likelihood estimation of weights can be obtained. However, this will cause the over-fitting problem. To ensure the generalization ability, the weights are supposed to satisfy the normal distribution.

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha})=\prod_{n=0}^{N}N(\omega_n\,|\,0,\alpha_n^{-1}) \tag{4}$$

Based on likelihood function and prior probability, the posterior probability density function $p(\boldsymbol{\omega}|\boldsymbol{t}, \boldsymbol{a})$ of $\boldsymbol{\omega}$ can be obtained by the Bayes' rule.

$$p(\boldsymbol{\omega}\mid\boldsymbol{t},\boldsymbol{\alpha})=\frac{p(\boldsymbol{t}\mid\boldsymbol{\omega})p(\boldsymbol{\omega}\mid\boldsymbol{\alpha})}{p(\boldsymbol{t}\mid\boldsymbol{\alpha})} \tag{5}$$

where $p(\boldsymbol{t}|\boldsymbol{a})$ is the evidence function. Maximizing the posterior probability, the optimal weights $\{\omega_n\}_{n=0}^N$ and hyperparameters $\{\alpha_n\}_{n=0}^N$ can be found. In classification, the likelihood is not Gaussian. Therefore the posterior probability density function in Eq. (5) is not Gaussian too and is analytically intractable. The posterior distribution could be approximated by the Laplacian method (Tipping, 2001) and the flow is as follows.

(1) The evidence function is a constant. Therefore, $p(\boldsymbol{\omega}|\boldsymbol{t}, \boldsymbol{a})$ is linear proportional to $p(\boldsymbol{t}|\boldsymbol{\omega})$ $p(\boldsymbol{\omega}|\boldsymbol{a})$ and it is equivalent to maximize the logarithmic likelihood function in Eq. (6). The object function is a typical least square problem. The first item controls the fitting error of the training samples. The second item shrinks the candidates of $\boldsymbol{\omega}$ to control the complexity of the learning system and avoid over fitting. The matrix $\boldsymbol{A}$ in Eq. (6) is $\boldsymbol{A}=\mathrm{diag}\{\alpha_0,\alpha_1,\cdots,\alpha_N\}$.

$$\begin{aligned}&\log\{p(\boldsymbol{t}\mid\boldsymbol{\omega})p(\boldsymbol{\omega}\mid\boldsymbol{\alpha})\}\\&=\sum_{n=1}^{N}\{t_n\log y_n+(1-t_n)\log(1-y_n)\}-\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}A\boldsymbol{\omega}\end{aligned} \tag{6}$$

(2) Fix α and maximize $p(\boldsymbol{\omega}|\boldsymbol{t}, \boldsymbol{\alpha})$ by the iteratively reweighed least-squares (IRLS) method (Tipping, 2001). Calculate the first and second derivative of Eq. (6) with respect to ω,

the gradient vector and Hessian matrix in Eq.(7) and Eq.(8) are obtained. The matrix $\boldsymbol{B}$ equals $\mathrm{diag}\{\beta_1,\beta_2,\cdots,\beta_N\}$, where $\beta_n = y_n\ (1-y_n)$. $\boldsymbol{\Phi}$ is the design matrix, in the form of $[\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_N)]^{\mathrm{T}}$ and $\phi(\boldsymbol{x}_n)=[1,K(\boldsymbol{x}_n,\boldsymbol{x}_1),K(\boldsymbol{x}_n,\boldsymbol{x}_2),\cdots,K(\boldsymbol{x}_n, \boldsymbol{x}_N)]^{\mathrm{T}}$. The optimal weights $\boldsymbol{\omega}_{\mathrm{MP}}$ can be found through Eq. (9).

$$\boldsymbol{g}=\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{y})-\boldsymbol{A}\boldsymbol{\omega} \tag{7}$$

$$\boldsymbol{H}=-(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi}+\boldsymbol{A}) \tag{8}$$

$$\boldsymbol{\omega}_{\mathrm{MP}}^{\mathrm{new}} \leftarrow \boldsymbol{\omega}^{\mathrm{old}}-\boldsymbol{H}^{-1}\boldsymbol{g} \tag{9}$$

(3) Use the Laplacian method to approximate the posterior with the Gaussian distribution $N(\boldsymbol{\omega}|\boldsymbol{\omega}_{\mathrm{MP}},\Sigma)$, where $\Sigma=-\boldsymbol{H}^{-1}$. Thus the hyperparameter $\boldsymbol{\alpha}$ can be estimated by Eq. (10), where $\gamma_n=1-\alpha_n^{\mathrm{old}}\Sigma_{nn}$ and Σ_{nn} is the n^{th} diagonal element of the covariance matrix Σ.

$$\alpha_n^{\mathrm{new}}=\frac{\gamma_n}{(\omega_{\mathrm{MP}})_n^2} \tag{10}$$

Following the procedure above, $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$ are iteratively updated until convergence. During the optimization, many α_n have large values and posterior probability of the corresponding ω_n is zero, which guarantees the sparsity. The training samples with small-value hyperparameters α_n are the relevance vectors and used for classification.

4 MULTI-CLASS RVM

Similar to SVM, RVM is a binary classifier and can process the multi-class problem by the one-against-one (OAO) or one-against-all (OAA) methods. Additionally, RVM has the ability of direct multi-class classification. For the *K*-class problem, the likelihood in Eq. (2) can be extended into the standard multinomial form (Tipping, 2001)

$$p(\boldsymbol{x}\,|\,\boldsymbol{\omega})=\prod_{n=1}^{N}\prod_{k=1}^{K}\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}^{t_{nk}} \tag{11}$$

where the "one-of-*K*" coding method $t_n=(0,0,\cdots,1,\cdots,0)^{\mathrm{T}}$ is used for the sample $\boldsymbol{x}_n$. If $\boldsymbol{x}_n$ belongs to the k^{th} class, the k^{th} element of $\boldsymbol{t}_n$ is 1 and the rest are set 0. The classifier has *K* decision functions $\{y_k\}_{k=1}^{K}$. Each function owns private weights vector $\boldsymbol{\omega}_k$ and hyperparameters vector $\boldsymbol{\alpha}_k$. Eq. (11) is not the true likelihood, because the sum of the probabilities of any sample belonging to each class does not equal one.

$$\sum_{k=1}^{K}p(\boldsymbol{x}_n|\boldsymbol{\omega}_k)=\sum_{k=1}^{K}\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}\neq 1 \tag{12}$$

The problem can be solved by the multinomial logistic regression (Foody, 2008). The likelihood is rewritten

$$p(\boldsymbol{t}\,|\,\boldsymbol{\omega})=\prod_{n=1}^{N}\prod_{k=1}^{K}\frac{\exp\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}}{\sum_{p=1}^{K}\exp\{y_p(\boldsymbol{x}_n;\boldsymbol{\omega}_p)\}} \tag{13}$$

The multi-class RVM methods also use the hyperparameter $\boldsymbol{\alpha}$ to constraints the weights $\boldsymbol{\omega}$ and the optimizing procedure is similar to the binary. For the likelihood in Eq. (11), the posterior probability of $\boldsymbol{\omega}$ can be derived by Eq. (5). The maximization of the posterior distribution is equivalent to maximize

$$\log\{p(\boldsymbol{t}\,|\,\boldsymbol{\omega})p(\boldsymbol{\omega}\,|\,\boldsymbol{\alpha})\}=\sum_{n=1}^{N}\sum_{k=1}^{K}t_{nk}\log y_{nk}-\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\omega}+C \tag{14}$$

where $y_{nk}=\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}$, C is the constant unconcerned with $\boldsymbol{\omega}$, the vector $\boldsymbol{t}$ is $(\boldsymbol{t}_1^{\mathrm{T}},\cdots,\boldsymbol{t}_k^{\mathrm{T}},\cdots,\boldsymbol{t}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{t}_k=(t_{0k},\boldsymbol{t}_{1k},\cdots,\boldsymbol{t}_{Nk})^{\mathrm{T}}$. The structure of $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$ are similar to $\boldsymbol{t}$, the matrix $\boldsymbol{A}$ equals $\mathrm{diag}(\boldsymbol{A}_1,\cdots,\boldsymbol{A}_K)$ and $\boldsymbol{A}_k=\mathrm{diag}(\alpha_{0k},\cdots,\alpha_{Nk})$. Calculating the derivative of Eq. (14), the first and second derivative of the object functions are

$$\boldsymbol{g}=\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{y})-\boldsymbol{A}\boldsymbol{\omega} \tag{15}$$

$$\boldsymbol{H}=-(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Psi}+\boldsymbol{A}) \tag{16}$$

The structure of $\boldsymbol{y}$ is the same as $\boldsymbol{t}$. The matrix $\boldsymbol{B}$ is equals to $\mathrm{diag}(\boldsymbol{B}_1,\cdots\boldsymbol{B}_K)$ and each element $\boldsymbol{B}_k$ is $\mathrm{diag}\{y_{1k}(1-y_{1k}),\cdots,y_{Nk}(1-y_{Nk})\}$. y_{nk} is calculated by the sigmoid link function $\boldsymbol{\sigma}\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}$. The design matrix is extended into $\boldsymbol{\Psi}=\mathrm{diag}(\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_K)$, where $\boldsymbol{\Phi}_k=\boldsymbol{\Phi}$. Based on the gradient and Hessian matrix, the optimized $\boldsymbol{\omega}$ can be obtained by Eq. (9) and the hyperparameter $\boldsymbol{\alpha}$ is updated by Eq. (10). The procedure is repeated until convergence.

The flow of multi-class RVM based on the multinomial logistic regression is basically the same as that using Eq. (11). The difference only exists in the structure of $\boldsymbol{B}$ and $\boldsymbol{y}$. The matrix $\boldsymbol{B}$ is

$$\boldsymbol{B}=\begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \cdots & \boldsymbol{B}_{1K}\\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \cdots & \boldsymbol{B}_{2K}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{B}_{K1} & \boldsymbol{B}_{K2} & \cdots & \boldsymbol{B}_{KK}\end{bmatrix} \tag{17}$$

where $\boldsymbol{B}_{ij}$ equals $\mathrm{diag}\{y_{1i}(\rho_{ij}-y_{1j}),\ldots,y_{Ni}(\rho_{ij}-y_{Nj})\}$ and y_{nk} is

$$y_{nk}=\frac{\exp\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}}{\sum_{p=1}^{K}\exp\{y_p(\boldsymbol{x}_n;\boldsymbol{\omega}_p)\}} \tag{18}$$

where $k=i,j$. If $i=j$, $\rho_{ij}=1$. Otherwise, $\rho_{ij}=0$. The multi-class RVM derived from Eq. (11) and Eq. (13) are named the binary logistic regression (BLR) and multinomial logistic regression (MLR).

5 EXPERIMENTS

5.1 Data description

The hyperspectral image of Indian Pine in Indiana State, USA (AVIRIS, 1992) is used to test the methods. The image is collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) in 1992. The region contains sixteen kinds of substances, seven of which have too few samples and are not used in the test. 200 bands remain after eliminating the water absorption and low SNR bands. Nine typical classes have 8489

samples, as shown in Fig.1 and Table 1. Several classes are similar in the spectrum and hard to separate, three kinds of soybeans and two kinds of corn for instance. The image is a standard data and has been widely used to test the feature extraction and classification algorithms. The radius basis function (RBF) kernel, superior to the linear and polynomial kernels, is adopted as the basis function of RVM. The experiments are carried on a HP server, containing one Xeon 5110 processor (dual core, 1.6 G) and 2 G memory. The following two parts are carried out in the experiment.

(1) OAA, OAO, BLR and MLR are compared in the aspects of precision, sparsity and efficiency. The multi-class methods are implemented based on the Sparse Bayesian V1.1 package.

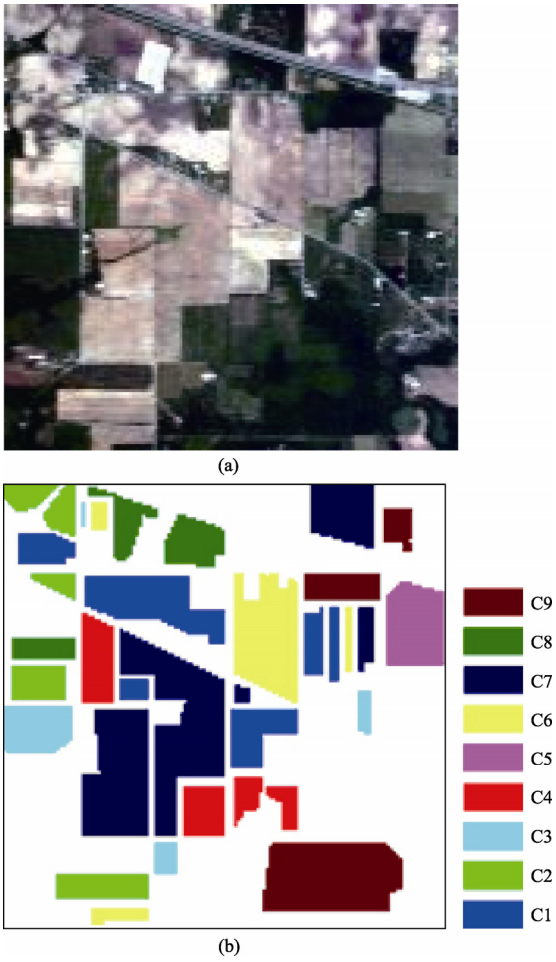

Fig. 1 Pseudo-color picture and typical classes distribution of the Indian Pine test site
(a) Pseudo-color picture of band 24-12-5; (b) Distribution of nine typical classes

Table 1 Information of the typical classes in Indian Pine

Class	Type	Number of samples
C1	Corn-notill	1265
C2	Corn-min	728
C3	Grass/Pasture	449
C4	Grass/Tress	671
C5	Hay-windrowed	456
C6	Soybeans-notill	849
C7	Soybeans-min	2268
C8	Soybeans-clean	577
C9	Woods	1226

(2) RVM is compared with radius basis functions neural network (RBFNN), K-nearest neighbor (KNN) and SVM in the aspect of precision and sparsity. RVM is implemented based on Sparse Bayesian V1.1, KNN and RBFNN are realized by the knnclassify and newrb functions in Mathlab, and libsvm-2.89 package is adopted for SVM.

5.2 Multi-class methods

The four methods are compared in precision, sparsity and efficiency. The memory consumption of BLR and MLR is $O(K^2N^2)$ and the time complexity is $O(K^3N^3)$. K is the number of classes and N is the quantity of the training samples. BLR and MLR are inefficient for large K and N, thus fewer classes and samples are first used to test the methods. Four classes are selected, including Corn-notill, Soybeans-notill, Soybeans-min and Woods. They have more samples than the rest and the crops are hard to separate. We choose 2.5%, 5% and 10% from each class for training respectively. The width β of the RBF kernel is estimated by the 5-Fold cross validation method. The result is shown in Table 2, where OA stands for the overall accuracy. OAA and BLR are the same in precision, better than the others and MLR is the worst. The methods are close in sparsity, as are all sparse. OAO is the best in efficiency, followed by OAA. BLR takes the third place and MLR is at the bottom. OAO carries out $K(K\text{-}1)/2$ times binary RVMs, which processes fewer samples each time. OAA only needs K times binary RVMs, but each classifier deals with more samples. Therefore more consumptions are involved. It is better than OAO only when dealing with the fewest training samples (2.5%). The time complexity of BLR and MLR is $O(K^3N^3)$, K^3 times that of OAA. BLR replaces matrix multiplication by dot product when computing Hessian matrix and the time consumption is saved.

Affected by BLR and MLR, the classes and samples above are not sufficient, thus the conclusion is unreliable. All nine classes of the Indian Pine test site are used to further compare

Table 2 Performance comparison of OAO, OAA, BLR and MLR

Items	2.5%				5%				10%			
	OAO	OAA	BLR	MLR	OAO	OAA	BLR	MLR	OAO	OAA	BLR	MLR
OA	0.77	0.80	0.80	0.76	0.82	0.86	0.86	0.80	0.86	0.88	0.88	0.85
RV	32	30	29	48	52	64	60	63	70	90	81	85
Time/s	5.9	5.45	21.0	53.1	7.9	17.3	78.1	231.1	21.9	62.6	488.4	1172.8
β	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

OAO and OAA. Table 3 gives the result. The percentage in the table represents the ratio of the training samples to the total in each class. The width β of the RBF kernel is estimated by the 5-Fold cross validation method and listed in the last row. OAA acquires higher precision, but OAO is better in efficiency, particularly in the large-training-set circumstance (more than 10%). The efficiency of OAA is acceptable only when the ratio is less than or equal to 10%. The Hessian matrix of OAA takes too much memory when the ratio reaches 40%, leading to memory overflow. Additionally, once the ratio exceeds a limit, the Hessian matrix of OAA may be ill-condition and the algorithm is terminated.

Table 3 Performance comparison of OAO and OAA

Items	10%		20%		30%		40%		50%	
	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA
OA	0.81	0.85	0.87	0.89	0.89	0.91	0.91		0.92	
RV	202	165	266	242	326	306	379	Memory	446	Memory
Time/s	82.3	238.4	313.5	1707.1	588.1	4742.3	1291.6	Overflow	1700.8	Overflow
β	0.5	0.25	0.1	0.25	0.1	0.25	0.1		0.25	

Summarizing the results above, the following conclusions are given.

(1) BLR and MLR are inefficient, and have no advantage in precision and sparsity. In a word, the two direct multi-classes methods are not recommendable.

(2) OAO is best in efficiency and OAA is best in precision. For fewer training samples, OAA is preferred for the higher precision. Once too many samples are involved, OAA is inefficient and may cause ill-condition Hessian, thus OAO is recommended.

5.3 Compared with SVM, KNN and RBFNN

In this experiment, RVM is compared with several classical hyperspectral classifiers in precision and sparsity, including SVM, RBFNN and KNN. First, 50% samples of each class are used to compare the performance of the methods. Based on the result in section 5.2, OAO is adopted to construct the multi-class RVM for its best efficiency. The parameters of RVM, RBFNN and SVM are optimized by the cross validation method.

The results are given in Table 4. The performance of RVM is better than KNN, equal to RBFNN and only a slightly worse than SVM. Same as the others, RVM exhibits better performance for the easy-to-separate classes, including C3, C4, C5 and C9. For the hard-to-separate classes C1, C2, C6, C7 and C8, the misclassified pixels increase. For the five kinds of hard-to separate crops, RVM is inferior to SVM, especially at C2 and C6. However, it is comparable or better than RBFNN and KNN in the circumstance. Fig.2 shows the phenomena vividly. The misclassified pixels concentrate at the crops. SVM has the fewest misclassifications, and RVM is close to RBFNN.

Table 4 Compare RVM with SVM, KNN and RBFNN, 50% for training, 50% for test

Method	Accuracy of Each Class									OA
	C1	C2	C3	C4	C5	C6	C7	C8	C9	
RVM	0.92	0.83	0.95	0.99	0.99	0.84	0.88	0.92	0.99	0.92
SVM	0.93	0.93	0.99	1.00	1.00	0.90	0.92	0.96	0.99	0.95
KNN	0.70	0.73	0.94	0.99	1.00	0.77	0.82	0.66	0.96	0.83
RBFNN	0.89	0.78	0.95	1.00	1.00	0.84	0.91	0.93	0.98	0.92

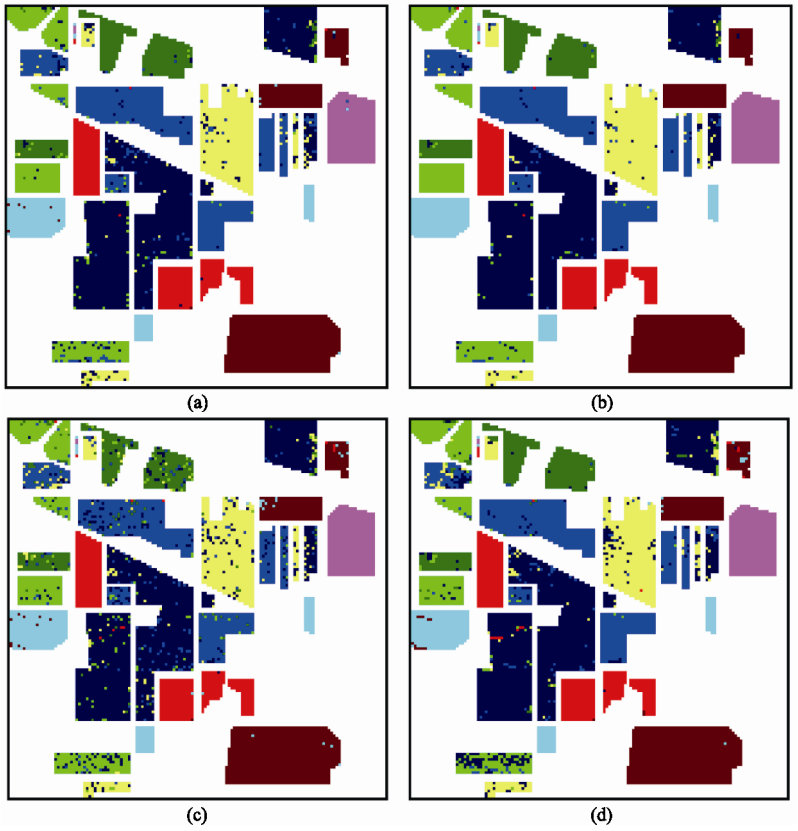

Fig. 2 Classification images of RVM, SVM, KNN and RBFNN, 50% for training in each class
(a) RVM; (b) SVM; (c) KNN; (d) RBFNN

Fig.3 compares the sparsity of RVM and SVM under different quantity of training samples. It is obviously that the relevance vectors (RV) is far less than the support vectors (SV), especially for the 50%-training-sample case. RV and SV decrease as long as less training samples are involved. SV decreases dramatically, but it is always more than RV. In the prediction of the test points, only RV and SV are involved. The less the RV or SV are, the faster the prediction takes. Therefore, RVM is more efficient in predicting the large-scale hyperspectral data set. The quantity of SV and RV can be explained from the theorem of the algorithms. SV mainly appear around the decision boundary or at the misclassified region. The overlap between different classes of the Indian Pine test site is severe. Therefore, lots of samples are at the decision boundary, resulting in more SV. RV reflects the intrinsic information of the training samples, which are far away from the decision boundary. Compared with the boundary samples, the intrinsic samples are fewer. Therefore, RVM is sparser than SVM.

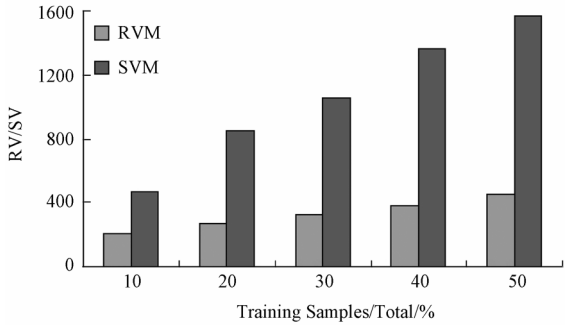

Fig. 3 Compare RVM with SVM in sparsity

Summarizing the result above, the following conclusions are extracted.

(1) For the easy-to-separate problem, the accuracy of RVM is comparable to SVM, RBFNN and KNN. For the hard-to-separate problem, RVM is slightly worse than SVM, close to RBFNN and superior to KNN.

(2) RVM is sparser than SVM. In the prediction of the large-scale hyperspectral image, RVM is more efficient. Krishnapuram (2005) pointed out that the ARD framework always prefers simple models, which guarantees the sparsity. The phenomena may cause the under-fitting problem, thus decrease the classification accuracy of RVM. The conjecture may explain why RVM is a slightly worse than SVM in the accuracy. It needs to be verified through theorem and experiments in the future.

6 CONCLUSIONS

RVM is applied to analyze the hyperspectral image in this paper, to establish a high-precision classifier in the high dimensional space with insufficient training samples. Beginning at the sparse Bayesian theory, the detail of RVM is analyzed and the multi-class methods are discussed. In the experiments, OAO, OAA and two direct multi-class methods are compared and RVM is further compared with SVM, RBFNN and KNN. The experiments show that BLR and MLR occupy too much memory and are inefficient, which make them not suitable for the real applications. OAA is best in precision, but inefficient when more training samples are involved. OAO is worse than OAA in precision, but it is more efficient and preferred for the aforementioned case. The accuracy of RVM is slightly worse than SVM. However, its solution is sparser and the prediction is faster when more test samples are involved. Generally, RVM can acquire good performance in the high dimensional space with limited training samples. Additionally, its solution is sparse, suitable for the classification of the large-scale hyperspectral image.

REFERENCES

AVIRIS. 1992. AVIRIS NW Indiana's Indian Pines 1992 Data Set. Avaiable: ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/ 92AV3C

Camps-Valls G, Gomez-Chova L, Munoz-Mari J, Vila-Frances J, Amoros-Lopez J and Calpe-Maravilia J. 2006. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sens. of Environ*., **105**: 23—33

Demir B and Ertürk S. 2007. Hyperspectral image classification using relevance vector machines. *IEEE GeoSci. Remote Sens. Lett*., **4**(4): 586—590

Dundar M M and Landgrebe D A. 2004. A cost-effective semisupervised classifier approach with kernels. *IEEE Trans. Geosci. Remote Sens*., **42**(1): 264—270

Foody G M. 2008. RVM-based multi-class classification of remotely sensed data. *Int. J. Remote Sens*., **29**(6): 1817—1823

Hughes G F. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inform. Theory*, **14**(1): 55—63

Jackson Q and Landgrebe D A. 2001. An adaptive classifier design for high dimensional data analysis with a limited training data set. *IEEE Trans. Geosci. Remote Sens*., **39**(12): 2664—2679

Krishnapuram B, Carin L, Figueiredo M A T and Hartemink A J. 2005. Sparse multinomial logistic regression: fast algorithm and generalization bounds. *IEEE Trans. Pattern Anal. Machine Intell*., **27**(6): 957—968

Krishnapuram B, Hartemink A J, Carin L and Figueiredo M A T. 2004. A bayesian approach to joint feature selection and classifier design. *IEEE Trans. Pattern Anal. Machine Intell*., **26**(9): 1105—1111

Kuo B C and Landgrebe D A. 2004. Nonparametric weighted feature extraction for classification. *IEEE Trans. Geosci. Remote Sens*., **42**(5): 1096—1105

Melgani F and Bruzzone L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens*., **42**(8): 1778—1790

Tadjudin S and Landgrebe D A. 1999. Covariance estimation with limited training samples. *IEEE Trans. Geosci. Remote Sens*., **37**(4): 2113—2118

Tipping M E. 2000. The relevance vector machine. Solla S A, Leen T K and Müller K R. Advances in Neural Information Processing Systems. vol. 2, MIT Press, Cambridge, MA

Tipping M E. 2001. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res*., **1**: 211—244

Wang X D, Ye M Y and Duanmu C J. 2009. Classification of data from electronic nose using relevance vector machines. *Sensor and Actuators B*, **140**(1): 143—148

Williams O, Blake A and Cipolla R. 2005. Sparse bayesian learning for efficient visual tracking. *IEEE Trans. Pattern Anal. Machine Intell*., **27**(8): 1292—1304

Yu B and Xu Z B. 2008. A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge Based Systems*, **21**: 355—362

关联向量机在高光谱影像分类中的应用

董　超, 赵慧洁

北京航空航天大学 仪器科学与光电工程学院, 教育部精密光机电一体化技术实验室, 北京 100191

摘　要: 将关联向量机应用于高光谱影像分类, 实现高维空间中训练样本不足时分类器的精确建模。从稀疏贝叶斯理论出发, 分析关联向量机原理, 探讨一对多、一对一和两种直接的多分类方法。实验环节比较了各种多分类方法, 并从精度、稀疏性两方面将关联向量机与支持向量机等经典算法比较。实验结果表明, 两种直接的多分类方法内存占用大、效率低; 一对多精度最高, 但效率较低; 一对一计算效率最高, 精度与一对多近似。关联向量机精度不如支持向量机, 但解更稀疏, 测试样本较多时实时性好, 适合处理大场景高光谱影像的分类问题。

关键词: 遥感, 分类, 关联向量机, 高光谱

中图分类号: TP79　　**文献标志码:** A

引用格式: 董　超, 赵慧洁. 2010. 关联向量机在高光谱影像分类中的应用. 遥感学报, **14**(6): 1273－1284
Dong C, Zhao H J. 2010. Hyperspectral image classification and application based on relevance vector machine. *Journal of Remote Sensing*. **14**(6): 1273－1284

1　引　言

高光谱分辨率导致传感器采集的数据量急剧膨胀, 对遥感影像处理提出巨大挑战。受 Hughes(1968)效应影响, 在训练样本数一定的情况下, 高维空间中传统分类器建模精度低, 如极大似然和人工神经网络。为保证高维空间中分类器的学习精度, 需要大量训练样本, 耗费人力和物力, 在实际应用中不可取。近年来涌现出大量研究报道, 探讨高光谱影像小样本分类问题, 主要分为以下 4 类: (1) 协方差矩阵的正规化技术(Tadjudin & Landgrebe, 1999); (2) 特征提取和特征选择(Kuo & Landgrebe, 2004); (3) 半监督学习(Dundar & Landgrebe, 2004; Jackson & Landgrebe, 2001); (4) 低复杂度分类系统(Melgani & Bruzzone, 2004), 如支持向量机(SVM)。SVM 是目前性能最优的监督学习算法, 能够在高维空间中用较少的训练样本获得较高的分类精度。SVM 以间隔最大化为优化准则, 既保证系统对训练样本的学习误差最小, 又可保证泛化能力。间隔最大化准则还保证支持向量总是出现在类别交界和错分类处, 即远离分类面的样本对分类不起决定性作用, 因此 SVM 的解是稀疏的。实际应用中 SVM 解的稀疏程度不是很高, 与训练样本数成比例增长, 影响大规模分类问题的计算效率。除此之外, SVM 包含以下缺点:

(1) 无法得到概率式预测。实际应用中, 总希望能得到预测的概率密度函数, 掌握不确定性。

(2) 未知参数较多, 需使用网格搜索和交叉验证法确定, 浪费计算资源。

(3) 核函数必须满足 Mercer 条件。

针对以上问题, Tipping(2000, 2001)从稀疏贝叶斯理论出发, 提出关联向量机(RVM)。关联向量机是近期模式识别领域的研究热点, 应用于电子鼻监测(Wang 等, 2009)、垃圾邮件检测(Yu & Xu, 2008)、机器视觉(Williams 等, 2005)等领域, 近两年遥感领域出现了 RVM 的研究。Demir 和 Ertürk(2007)将 RVM 用于高光谱影像分类, 实验结果表明 RVM 精度稍差于 SVM, 但其解更稀疏、适合大场景分类。Foody (2008)探索了 RVM 的多分类能力, 并与判别分析、决策树、神经网络和 SVM 相比较。Camps-Valls 等(2006)使用 RVM 从多光谱影像中提取叶绿素指数, 监测海岸带水体质量。上述研究成果探索性较强, 应用中仍存在较多问题, 如学习效率、多分类方法、解过稀疏等。从稀疏贝叶斯理论出发, 深入分析 RVM

收稿日期: 2009-11-09; 修订日期: 2010-05-16
基金项目: 863 计划重点项目(编号: 2008AA121102)和中国地质调查局项目(编号: 1212010816033)。
第一作者简介: 董　超(1982－　), 男, 北京航空航天大学博士研究生。研究方向为高光谱遥感影像数据处理。E-mail: dongchaoxj888@126.com。

的实现过程，探讨了 4 种不同的多分类方法。实验环节从精度、稀疏程度全面比较 RVM 和 SVM，挖掘其解决高光谱影像分类问题的优势和局限性。

2 稀疏贝叶斯学习

监督学习解决如下问题：对样本 $\{\boldsymbol{x}_n\}_{n=1}^N$ 和目标 $\{t_n\}_{n=1}^N$，估计函数 $t_n = f(\boldsymbol{x}_n)$，找出 $\boldsymbol{x}_n$ 和 t_n 间的依赖关系，以预测未知输入 $\boldsymbol{x}$ 的响应 $t = f(\boldsymbol{x})$。对回归问题，t_n 为实数；对分类问题，t_n 为类别标号。预测函数 f 可通过定义在输入空间中基函数的线性组合实现，

$$y(\boldsymbol{x};\boldsymbol{\omega}) = \sum_{n=1}^{N} \omega_n K(\boldsymbol{x},\boldsymbol{x}_n) + \omega_0 \tag{1}$$

式中，$K(\cdot,\cdot)$ 为基函数，$\boldsymbol{\omega} = \{\omega_n\}_{n=0}^N$ 为权重系数。训练过程即寻找 f 的最优参数 $\{\omega_n\}_{n=0}^N$，一方面揭示训练样本的特性，另一方面有助于准确预测未知样本的输出。

参数 $\boldsymbol{\omega}$ 决定了学习系统 f 对问题的描述能力。$\boldsymbol{\omega}$ 越稠密，f 越复杂，对训练样本的近似效果越好。然而，由于存在噪声(回归)和类间重叠(分类)，对训练样本的近似程度无法保证 f 对未知样本的预测能力。若用复杂系统描述有限样本，常会造成对训练样本的过学习，降低对未知样本的预测能力。学习系统的复杂度应与观测数据的复杂度相匹配，即不应太简单(欠学习)，也不应太复杂(过学习)。

稀疏贝叶斯学习从统计方法出发，通过对 $\boldsymbol{\omega}$ 附加约束条件，调整系统复杂度，典型算法如 RVM (Tipping, 2001)、SMLR (Krishnapuram 等, 2005)、JCFO(Krishnapuram 等, 2004)等。RVM 使用自动关联判定(ARD)框架，假设 ω_n 独立且符合零均值、方差 α_n^{-1} 的高斯分布; SMLR 采用 Laplacian 分布，并使用多项式 Logistic 回归技术实现多分类; JCFO 对权重系数和样本特征分量附加约束条件，不仅能够获得分类器的稀疏表达，同时可得到面向分类的最优子空间、实现特征选择。

3 关联向量分类

对二分类问题，目标 $t_n \in \{0,1\}$，采用 Bernoulli 分布构造条件概率密度函数 $p(t|\boldsymbol{\omega})$，通过 Sigmoid 连接函数将 $y(x)$ 映射到 $[0,1]$ 内。根据 Bernoulli 分布的定义，似然函数为:

$$p(\boldsymbol{t}\,|\,\boldsymbol{\omega}) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n} \tag{2}$$

式中，$\boldsymbol{t} = (t_1,\cdots,t_N)^{\mathrm{T}}$，$\boldsymbol{\omega} = (\omega_0,\cdots,\omega_N)^{\mathrm{T}}$，$y_n = \sigma\{y(\boldsymbol{x}_n;\boldsymbol{\omega})\}$，$\sigma(y)$ 为 Sigmoid 连接函数。

$$\sigma(y) = 1/(1+\mathrm{e}^{-y}) \tag{3}$$

对式(2)求 $\boldsymbol{\omega}$ 的导数，得权重系数的极大似然估计，这样做可能导致过学习。为保证泛化性能，假设权重系数符合式(4)正态分布。

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{n=0}^{N} N(\omega_n \,|\, 0, \alpha_n^{-1}) \tag{4}$$

根据似然函数和先验概率，由贝叶斯公式得 $\boldsymbol{\omega}$ 的后验概率密度 $p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha})$。

$$p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{t}\,|\,\boldsymbol{\omega})\,p(\boldsymbol{\omega}\,|\,\boldsymbol{\alpha})}{p(\boldsymbol{t}\,|\,\boldsymbol{\alpha})} \tag{5}$$

式中，$p(\boldsymbol{t}\,|\,\boldsymbol{\alpha})$ 为证据函数。求 $p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha})$ 的极值，可得到权重系数 $\{\omega_n\}_{n=0}^N$ 和超参数 $\{\alpha_n\}_{n=0}^N$ 的最优值。分类问题中，似然函数不符合高斯分布，导致后验概率也不符合高斯分布，故式(5)无解析解，可通过 Laplacian 方法得到其近似解(Tipping, 2001)，计算流程如下:

① 式(5)中的证据函数为常数，因此 $p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha})$ 与 $p(\boldsymbol{t}\,|\,\boldsymbol{\omega})p(\boldsymbol{\omega}\,|\,\boldsymbol{\alpha})$ 成正比，等效于求解式(6)对数似然函数的最大值。式(6)是典型的最小二乘问题，第一项保证对训练样本的拟和误差最小，第二项为惩罚项，缩小 $\boldsymbol{\omega}$ 的取值范围以控制学习系统复杂度、避免过学习。式(6)中矩阵 $\boldsymbol{A} = \mathrm{diag}\{\alpha_0,\alpha_1,\cdots,\alpha_N\}$。

$$\log\{p(\boldsymbol{t}\,|\,\boldsymbol{\omega})p(\boldsymbol{\omega}\,|\,\boldsymbol{\alpha})\} = \sum_{n=1}^{N}\{t_n \log y_n + (1-t_n)\log(1-y_n)\} - \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\omega} \tag{6}$$

② 固定 $\boldsymbol{\alpha}$，使用 IRLS 算法求解后验概率 $p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha})$ 的极值点(Tipping, 2001)。求式(6)关于 $\boldsymbol{\omega}$ 的一、二阶导数，得 $\boldsymbol{g}$ 和 $\boldsymbol{H}$，如式(7)—式(8)。其中 $\boldsymbol{B} = \mathrm{diag}\{\beta_1,\cdots,\beta_N\}$，且 $\beta_n = y_n(1-y_n)$，$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_N)]^{\mathrm{T}}$ 为设计矩阵，且 $\phi(\boldsymbol{x}_n) = [1, K(\boldsymbol{x}_n, \boldsymbol{x}_1), K(\boldsymbol{x}_n,\boldsymbol{x}_2),\ldots,K(\boldsymbol{x}_n,\boldsymbol{x}_N)]^{\mathrm{T}}$，通过式(9)可以找到权重系数最优值的近似解 $\boldsymbol{\omega}_{\mathrm{MP}}$。

$$\boldsymbol{g} = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{y}) - \boldsymbol{A}\boldsymbol{\omega} \tag{7}$$

$$\boldsymbol{H} = -(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi} + \boldsymbol{A}) \tag{8}$$

$$\boldsymbol{\omega}_{\mathrm{MP}}^{\mathrm{new}} \leftarrow \boldsymbol{\omega}_{\mathrm{MP}}^{\mathrm{old}} - \boldsymbol{H}^{-1}\boldsymbol{g} \tag{9}$$

③ 根据 Laplacian 方法将后验概率密度函数 $p(\boldsymbol{\omega}\,|\,\boldsymbol{t},\boldsymbol{\alpha})$ 近似为以 $\boldsymbol{\omega}_{\mathrm{MP}}$ 为中心的高斯分布 $N(\boldsymbol{\omega}|\boldsymbol{\omega}_{\mathrm{MP}},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma} = -\boldsymbol{H}^{-1}$，则可通过式(10)估计超参数 $\boldsymbol{\alpha}$，其中 $\gamma_n = 1 - \alpha_n^{\mathrm{old}}\boldsymbol{\Sigma}_{nn}$，$\boldsymbol{\Sigma}_{nn}$ 为协方差矩阵 Σ

对角线上的第 n 个分量。

$$\alpha_n^{\text{new}} = \frac{\gamma_n}{(\omega_{\text{MP}})_n^2} \tag{10}$$

按上述过程反复更新$\boldsymbol{\omega}$和$\boldsymbol{\alpha}$直到算法收敛。计算过程中大部分 α_n 数值很大, 对应 ω_n 的后验概率为零, 保证了解的稀疏性。数值较小 α_n 对应的样本点称为关联向量, 用于分类。

4　多分类方法

与 SVM 类似, RVM 为二分类器, 可通过一对一(OAO)或一对多(OAA)解决多类问题。此外, RVM 可直接实现多分类。对 K 类问题, 可将似然函数(2)扩展为标准的多元形式(Tipping, 2001),

$$p(\boldsymbol{t}\,|\,\boldsymbol{\omega}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}^{t_{nk}} \tag{11}$$

采用“1 对 K”方法编码样本点 $\boldsymbol{x}_n$ 的目标 $\boldsymbol{t}_n = (0,0,\cdots,1,\cdots,0)^{\mathrm{T}}$。若 $\boldsymbol{x}_n$ 属于第 k 类, 则向量 $\boldsymbol{t}_n$ 的第 k 位为 1, 其余位为 0。分类系统包含 K 个决策函数 $\{y_k\}_{k=1}^{K}$, 每个函数 y_k 均有各自的权重系数向量$\boldsymbol{\omega}_k$ 以及相应的超参数向量$\boldsymbol{\alpha}_k$。式(11)并非真正意义上的似然函数, 因为任意样本点对各类的全概率皆不为 1。

$$\sum_{k=1}^{K} p(\boldsymbol{t}_n\,|\,\boldsymbol{\omega}_k) = \sum_{k=1}^{K}\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\} \neq 1 \tag{12}$$

可采用多元 Logistic 回归解决该问题(Foody, 2008), 此时似然函数如式(13)。

$$p(\boldsymbol{t}\,|\,\boldsymbol{\omega}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\frac{\exp\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}}{\sum_{p=1}^{K}\exp\{y_p(\boldsymbol{x}_n;\boldsymbol{\omega}_p)\}} \tag{13}$$

多类 RVM 也通过超参数$\boldsymbol{\alpha}$约束权重系数$\boldsymbol{\omega}$的取值范围, 优化过程与二类问题近似。对式(11)似然函数, $\boldsymbol{\omega}$的条件后验概率通过式(5)给出, 后验分布的极值问题等效于求解下式的最大值。

$$\log\{p(\boldsymbol{t}\,|\,\boldsymbol{\omega})p(\boldsymbol{\omega}\,|\,\boldsymbol{\alpha})\} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\log y_{nk} - \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\omega} + C \tag{14}$$

式中, $y_{nk} = \sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}$, C 为与向量 $\boldsymbol{\omega}$ 无关的常数, $\boldsymbol{t} = (\boldsymbol{t}_1^{\mathrm{T}},\cdots,\boldsymbol{t}_k^{\mathrm{T}},\cdots,\boldsymbol{t}_K^{\mathrm{T}})^{\mathrm{T}}$ 且向量$\boldsymbol{t}_k = (t_{0k},t_{1k},\ldots,t_{Nk})^{\mathrm{T}}$。向量 $\boldsymbol{\omega}$ 和 $\boldsymbol{\alpha}$ 与 $\boldsymbol{t}$ 构造方法相同, 矩阵 $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{A}_1,\cdots,\boldsymbol{A}_K)$ 且 $\boldsymbol{A}_k = \mathrm{diag}(\alpha_{0k},\cdots,\alpha_{Nk})$。对式(14)求导, 目标函数的一、二阶导数分别为:

$$\boldsymbol{g} = \boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{y}) - \boldsymbol{A}\boldsymbol{\omega} \tag{15}$$

$$\boldsymbol{H} = -(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Psi} + \boldsymbol{A}) \tag{16}$$

$\boldsymbol{y}$ 与 $\boldsymbol{t}$ 结构相同, $\boldsymbol{B} = \mathrm{diag}(\boldsymbol{B}_1,\cdots,\boldsymbol{B}_K)$, $\boldsymbol{B}_k = \mathrm{diag}\{y_{1k}(1-y_{1k}),\cdots,y_{Nk}(1-y_{Nk})\}$, y_{nk} 由连接函数 $\sigma\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}$ 给出。设计矩阵扩展为 $\boldsymbol{\Psi} = \mathrm{diag}(\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_K)$, 其中 $\boldsymbol{\Phi}_k = \boldsymbol{\Phi}$。根据梯度和 Hessian 矩阵, 使用式(9)求解$\boldsymbol{\omega}_{\text{MP}}$, 根据式(10)更新超参数$\boldsymbol{\alpha}$, 重复上述过程直至收敛。

使用多元 Logistic 回归构建似然函数, 算法流程基本相同, 仅在 Hessian 矩阵中的 $\boldsymbol{B}$ 和 $\boldsymbol{y}$ 处存在差异, 矩阵 $\boldsymbol{B}$ 的结构如式(17):

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \cdots & \boldsymbol{B}_{1K} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \cdots & \boldsymbol{B}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{B}_{K1} & \boldsymbol{B}_{K2} & \cdots & \boldsymbol{B}_{KK} \end{bmatrix} \tag{17}$$

$\boldsymbol{B}_{ij} = \mathrm{diag}\{y_{1i}(\rho_{ij}-y_{1j}),\cdots,y_{Ni}(\rho_{ij}-y_{Nj})\}$, y_{nk} 由多元 Logistic 回归函数求得,

$$y_{nk} = \frac{\exp\{y_k(\boldsymbol{x}_n;\boldsymbol{\omega}_k)\}}{\sum_{p=1}^{K}\exp\{y_p(\boldsymbol{x}_n;\boldsymbol{\omega}_p)\}} \tag{18}$$

式中 $k = i, j$。若 i=j, $\rho_{ij} = 1$, 否则 $\rho_{ij} = 0$。式(11)和式(13)对应的多分类方法分别记为二元 Logistic 回归(BLR)和多元 Logistic 回归(MLR)。

5　实验结果

5.1　实验数据

实验环节使用 1992 年 AVIRIS 传感器采集的美国印第安纳州 Indian Pine 实验区高光谱影像, 包含 16 种地物, 其中 7 种地物样本点过少、未用于测试(AVIRIS, 1992)。去除水汽吸收和低信噪比波段后, 剩余 200 个波段, 9 类典型地物共包含 8489 个样本点, 如图 1 和表 1 所示。该地区多种地物光谱曲线近似、分类难度大, 如 3 种大豆和 2 种玉米, 是目前比较标准的高光谱数据, 用于多种特征提取和分类算法性能的测试。RBF 核函数性能优于线性核和多项式核, 作为测试过程中 RVM 的基函数。测试平台为 HP 服务器, 处理器为志强 5110(双核, 主频 1.6G), 内存 2G, 实验环节由如下:

(1) 从精度、解的稀疏程度和计算效率 3 个方面比较了 OAA、OAO、BLR 和 MLR4 种多分类方法。各多分类方法在 Sparse Bayesian V1.1 基础上实现。

(2) 从精度、解的稀疏程度两方面比较了 RVM、RBFNN、KNN 和 SVM 4 种方法, 其中 RVM 基于 Sparse Bayesian V1.1 软件包, KNN 和 RBFNN 分别由 Matlab 自带的 knnclassify 和 newrb 函数实现,

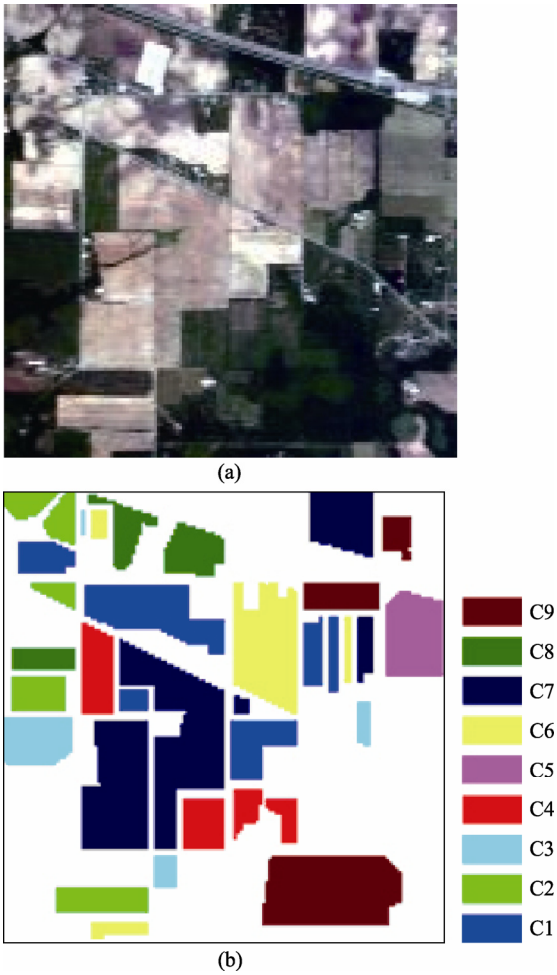

图 1 Indian Pine 实验区假彩图和典型地物分布图
(a) 24-12-5 三波段假彩合成图; (b) 典型地物分布图

表 1 **Indian Pine** 实验区典型地物信息统计表

类别	名称	总样本数
C1	玉米地 1	1265
C2	玉米地 2	728
C3	牧草	449
C4	树干	671
C5	干草	456
C6	大豆地 1	849
C7	大豆地 2	2268
C8	大豆地 3	577
C9	树林	1226

SVM 使用 libsvm-2.89 软件包。

5.2 多分类方法比较

从精度、稀疏程度和求解效率 3 方面比较了 4 种多分类方法。BLR 和 MLR 的内存占用率为 $O(K^2N^2)$、时间复杂度为 $O(K^3N^3)$, K 为类别数、N 为训练样本总数。若 K、N 较大, BLR 和 MLR 效率低, 故先比较训练样本数较少时各方法的性能。从 Indian Pine 数据集中提取大豆地 1、玉米地 1、玉米地 2 和树林 4 类, 这 4 类地物样本数多、三类农作物分类难度大, 分别从各类选取 2.5%、5%和 10%的样本点作为训练样本。采用 5Fold 交叉验证确定 RBF 核函数宽度系数 β, 实验结果如表 2, 其中 OA 为总体分类精度。OAA 和 BLR 分类精度相同, 且高于其他两种分类方法, MLR 的精度总是最差。4 种算法的稀疏程度近似, 均较稀疏。效率方面, OAO 最优、OAA 次之、BLR 第三、MLR 最差。OAO 执行 $K(K-1)/2$ 次二分类, 每次处理的训练样本数最少。OAA 执行 K 次二分类, 但每次处理的样本数较多, 总时间加长, 仅在训练样本最少(2.5%)时和 OAO 接近。BLR 和 MLR 的时间复杂度为 $O(K^3N^3)$, 理论上为 OAA 的 K^3 倍。BLR 在求解 Hessian 矩阵时采用点乘替代矩阵乘法运算, 降低了计算开销。

受 BLR 和 MLR 限制, 上述实验类别数和样本数均较少, 结果局限性大。为进一步比较 OAO 和 OAA 分类方法, 使用 Indian Pines 数据集中的全部 9 类进行比较, 实验结果如表 3, 表中百分比数字表示各类训练样本数占该类总样本数的百分比。RBF 核宽度系数 β 由 5Fold 交叉验证方法确定, 在表 3 最后一行给出。OAA 方法精度优于 OAO; 效率方面, OAO 效果较好, 尤其在训练样本较多的时候(>10%); 仅在样本数非常少(10%)时, OAA 的效率尚可接受。当训练样本占总样本数比重达到 40%, OAA 方法中 Hessian 矩阵内存占用率大, 发生内存溢出。此外, 若训练样本过多, OAA 方法下 Hessian 矩阵有时为病态矩阵, 导致算法中断。综合以上实验结果, 可得出如下结论。

表 2 **OAO、OAA、BLR** 和 **MLR** 四种多分类方法性能统计表

测试项	2.5%				5%				10%			
	OAO	OAA	BLR	MLR	OAO	OAA	BLR	MLR	OAO	OAA	BLR	MLR
OA	0.77	0.80	0.80	0.76	0.82	0.86	0.86	0.80	0.86	0.88	0.88	0.85
RV	32	30	29	48	52	64	60	63	70	90	81	85
Time/s	5.9	5.45	21.0	53.1	7.9	17.3	78.1	231.1	21.9	62.6	488.4	1172.8
β	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

表 3　**OAO** 和 **OAA** 两种多分类方法性能比较统计表

测试项	10%		20%		30%		40%		50%	
	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA
OA	0.81	0.85	0.87	0.89	0.89	0.91	0.91		0.92	
RV	202	165	266	242	326	306	379	内存	446	内存
Time/s	82.3	238.4	313.5	1707.1	588.1	4742.3	1291.6	溢出	1700.8	溢出
β	0.5	0.25	0.1	0.25	0.1	0.25	0.1		0.25	

(1) BLR 和 MLR 效率太低，在精度和解的稀疏程度上无优势。总体来看，两种直接的多分类方法不可取。

(2) OAO 效率最优，OAA 精度最佳。训练样本较少时，优先选择分类精度更好的 OAA；若训练样本较多，OAA 效率低且会导致 Hessian 病态，则应选取效率更佳的 OAO。

5.3　与 SVM 等比较

该实验以精度、稀疏程度等为标准，将 RVM 与 SVM 等 3 种典型方法比较。首先，从每类抽取 50%作为训练样本，比较各算法的性能。根据 5.2 节结果，使用效率最高的 OAO 构造多类 RVM。RVM、RBFNN 和 SVM 的参数均使用交叉验证方法确定。

表 4 为每类 50%作为训练样本 4 种分类器的实验结果。可以看出，RVM 的 OA 仅次于 SVM，和 RBFNN 相同，KNN 性能最差。和其他 3 种方法类似，RVM 在类别可分性较高的 C3、C4、C5 和 C9 类处精度高，而对可分性较低的 C1、C2、C6、C7 和 C8 五种农作物误判较多。对 5 种较难分的农作物，RVM 的各类分类精度均不如 SVM，在 C2 和 C6 两类尤其明显，但与其 RBFNN、KNN 相比 RVM 解决难分问题的效果更好。图 2 直观地说明了上述问题，误判均集中在 5 种农作物，但 SVM 的误判点最少、RVM 和 RBFNN 近似。

表 4　**Indian Pine** 实验区 **RVM** 性能测试(各类 **50%**作为训练样本，**50%**作为测试样本)

方法	各类分类精度									OA
	C1	C2	C3	C4	C5	C6	C7	C8	C9	
RVM	0.92	0.83	0.95	0.99	0.99	0.84	0.88	0.92	0.99	0.92
SVM	0.93	0.93	0.99	1.00	1.00	0.90	0.92	0.96	0.99	0.95
KNN	0.70	0.73	0.94	0.99	1.00	0.77	0.82	0.66	0.96	0.83
RBFNN	0.89	0.78	0.95	1.00	1.00	0.84	0.91	0.93	0.98	0.92

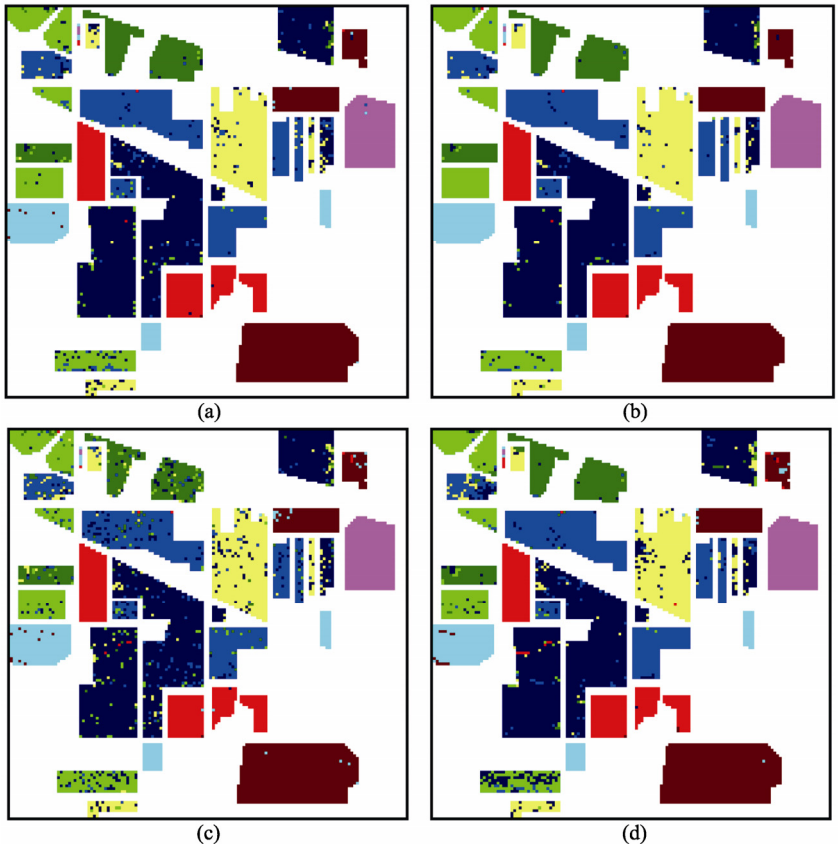

图 2　RVM 等四种方法分类效果示意图，各类 50%作为训练样本

(a) RVM; (b) SVM; (c) KNN; (d) RBFNN

图 3 比较了不同数量训练样本下, RVM 和 SVM 解的稀疏程度。可以看出, 关联向量(RV)远少于支持向量(SV), 在训练样本最多(50%)时最为明显。随训练样本减少, RV 和 SV 数量呈下降趋势, SV 减少的速度较快但总比 RV 多。SVM 和 RVM 判别过程中仅 SV 和 RV 参与运算, SV 或 RV 越少, 判别时间短, 因此稀疏特性保证 RVM 在大场景分类时效率较高。支持向量、关联向量个数间的关系由可从两算法原理进行分析。支持向量集中出现在决策面和错分类处, Indian Pine 实验区数据类间重叠区域大, 决策面附近的样本点较多, 支持向量必然多。关联向量反映训练样本集的本征信息、远离决策面, 与决策面附近样本相比, 本征样本数量相对较少, 故关联向量总量较低。

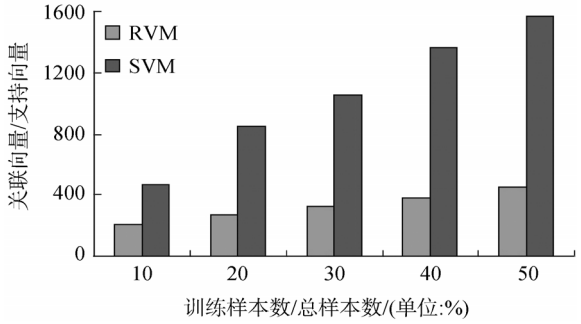

图 3 RVM 和 SVM 解稀疏程度比较

综合上述实验结果, 得出如下结论:

(1) 分类精度方面, 易分问题 RVM 性能和几种典型算法近似; 对难分问题稍差于 SVM, 与 RBFNN 近似, 优于 KNN。

(2) RVM 解的稀疏程度更高, 处理大场景高光谱影像时, 效率高、实时性好。在 Krishnapuram 等(2005)研究中发现, ARD 模型总是更倾向于简单的模型, 即 RVM 解的稀疏程度总是很高, 这样可能产生欠学习现象, 由此导致 RVM 性能上的损失。这有可能是 RVM 性能不如 SVM 的原因, 有待从理论和实验上进一步验证。

6 结 论

将 RVM 应用于高光谱影像处理, 实现高维空间中训练样本不足时分类器的精确建模。从稀疏贝叶斯理论出发, 深入分析算法原理和各种多分类方法。实验环节比较了 OAO、OAA 和两种直接多分类方法的优缺点, 并全面比较 SVM、RVM 等算法。实验结果表明: BLR 和 MLR 内存占用大、效率低, 实用性差; OAA 精度最高, 但当样本过多时效率较低; OAO 精度不如 OAA, 但计算效率较高, 当训练样本较多时优先考虑; RVM 精度不如 SVM, 但其解更稀疏, 当测试样本较多时, 实时性好。总体来看, 关联向量机能够在高维空间中用较少的训练样本获得较高的分类精度, RVM 解的稀疏性高, 适合处理大场景高光谱影像分类问题。

REFERENCES

AVIRIS. 1992. AVIRIS NW Indiana's Indian Pines 1992 Data Set. Avaiable: ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/ 92AV3C

Camps-Valls G, Gomez-Chova L, Munoz-Mari J, Vila-Frances J, Amoros-Lopez J and Calpe-Maravilia J. 2006. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sens. of Environ*., **105**: 23—33

Demir B and Ertürk S. 2007. Hyperspectral image classification using relevance vector machines. *IEEE GeoSci. Remote Sens. Lett*., **4**(4): 586—590

Dundar M M and Landgrebe D A. 2004. A cost-effective semisupervised classifier approach with kernels. *IEEE Trans. Geosci. Remote Sens*., **4**2(1): 264—270

Foody G M. 2008. RVM-based multi-class classification of remotely sensed data. *Int. J. Remote Sens*., **29**(6): 1817—1823

Hughes G F. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inform. Theory*, **14**(1): 55—63

Jackson Q and Landgrebe D A. 2001. An adaptive classifier design for high dimensional data analysis with a limited training data set. *IEEE Trans. Geosci. Remote Sens*., **39**(12): 2664—2679

Krishnapuram B, Carin L, Figueiredo M A T and Hartemink A J. 2005. Sparse multinomial logistic regression: fast algorithm and generalization bounds. *IEEE Trans. Pattern Anal. Machine Intell*., **27**(6): 957—968

Krishnapuram B, Hartemink A J, Carin L and Figueiredo M A T. 2004. A bayesian approach to joint feature selection and classifier design. *IEEE Trans. Pattern Anal. Machine Intell*., **26**(9): 1105—1111

Kuo B C and Landgrebe D A. 2004. Nonparametric weighted feature extraction for classification. *IEEE Trans. Geosci. Remote Sens*., **42**(5): 1096—1105

Melgani F and Bruzzone L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens*., **42**(8): 1778—1790

Tadjudin S and Landgrebe D A. 1999. Covariance estimation with limited training samples. *IEEE Trans. Geosci. Remote Sens*., **37**(4): 2113—2118

Tipping M E. 2000. The relevance vector machine. Solla S A, Leen T K and Müller K R. Advances in Neural Information Processing Systems. vol. 2, MIT Press, Cambridge, MA

Tipping M E. 2001. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res*., **1**: 211—244

Wang X D, Ye M Y and Duanmu C J. 2009. Classification of data from electronic nose using relevance vector machines. *Sensor and Actuators B*, **140**(1): 143—148

Williams O, Blake A and Cipolla R. 2005. Sparse bayesian learning for efficient visual tracking. *IEEE Trans. Pattern Anal. Machine Intell*., **27**(8): 1292—1304

Yu B and Xu Z B. 2008. A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge Based Systems*, **21**: 355—362