Robust Hyperspectral Classification Using Relevance Vector Machine

Journal:	*Transactions on Geoscience and Remote Sensing*
Manuscript ID:	TGRS-2009-00808.R1
Manuscript Type:	Hyperspectral Image and Signal Processing Special Issue
Date Submitted by the Author:	n/a
Complete List of Authors:	A. Mianji, Fereidoun; Harbin Institute of Technology, Information Engineering Zhang, Ye; Harbin Institute of Technology, Information Engineering Zhang, Junping; Harbin Institute of Technology, Information Engineering
Keywords:	Remote sensing, Image classification, Target identification, Image processing

RESPONSE TO EDITORS COMMENTS

Editor's Comments:
In order to increase the relevance of the paper for readers of the IEEE Trans. on Geoscience and Remote Sensing (TGRS), please add recent TGRS references (in particular from the years 2007 - 2009), if at all possible.

Answer: Thank you for your comment which also encouraged us to widen our knowledge on the very recent related works. We go through literature and add the very recent related ones to the article (17 references, 13 out of them from TGRS and GRSL).

Sincerely Yours
Mianji

Response to Dear Associated Editor Comments

Associate Editor Comments to the Author:
Relevance Vector Machine have already been investigated in the literature and may indeed be an appropriate tool for the robust classification of hyperspectral data. Please, address carefully all the comments raised by the reviewers when preparing a new version of your manuscript (rationale behind dimensionality reduction, tuning of the parameters, clearly stress your contribution...)

Answer: Thank you for the key comments. We tried our best to improve the rationality and experiments on all the highlighted points by dear reviewers. Necessary benchmarking techniques are added and supplementary experiments are carried out to enhance the article.

Sincerely Yours
Mianji

Response to Dear Reviewer 1 Comments

We are very grateful for the valuable comments. The paper is thoroughly revised with supplementary experiments. The answers to the comments are as follows:

(1) Can you explain why only classes C3, C5, C6, and C8 were selected for figures 6 through 10? The other classes did not do so well in classification when using the RVM+FLDA method.

Answer: The proposed method works very well for the other classes too. We added the related figures and necessary explanations for all the other classes to depict the obtained improvement for them. Just for class C1 which is an integrated relatively large area, the results are very comparable over all the train-to-test sample ratios, so, we did not show it, but this fact is referred to in the article.

(2) Can you explain why the RVM+FLDA did not do as well as the other methods for class C1. This was seen across all train-to-test sample ratios

Answer: C1 is a large and integrated area with a probably unimodal spectral distribution and negligible border effect, therefore, even in low train-to-test sample ratio, where there is still plenty of training samples available for this large area, FLDA+RVM can't show any superiority on C1, Although, as is shown in tables 2-5, all the techniques are comparable on C1 with just a small difference in the accuracy.

(3) Please correct all references "Sandiego' with 'San Diego'. The city is composed of two words not one.

Answer: Is edited.

Sincerely Yours
Mianji

Response to Dear Reviewer 2 Comments

We are really grateful for the very valuable comments by dear reviewer which helped us to realize that some important parts need to be improved. We did our utmost effort to enhance the theoretical and experimental parts accordingly and hopefully it is of an acceptable level now.

1. It is not clear why to use linear transformation (dimensionality reduction) for the non-linear classifier. The important information on class separation can be lost. Additionally, it is known that these methods (LDA and PCA) suffer from some problems e.g. the analysis is done based not on the classification accuracy but using some other criteria; since the lumped covariance matrix is used, LDA can loose some of the important information, LDA based feature selection can be unreliable if class mean vectors are near to one another and so on. SVM, on the other side, uses non-linear transformation which should be more appropriate for the hyperspectral image classification.

Answer: This is absolutely true and LDA is not the ideal feature reduction for many cases. The reason we have adopted it in our technique was just because of its simplicity and nonparametric nature. Furthermore, we were wondering to show the capability of our approach when it is used even with a standard simple feature reduction technique. As per dear reviewer's comment, we carried out necessary new experiments and revised the paper as follows:

1- limitations of LDA is highlighted in the beginning of section II (the first paragraph);
2- In fact any appropriate nonlinear alternatives for LDA (such as GNDA which applies kernel trick and we adopted it in our work too) that doesn't suffer its drawbacks, i.e., lose of nonlinear class information and possible singularity in ill-posed problems can be adopted into our method. We added necessary explanations to the beginning of section II. We also referred to the most recently proposed techniques for feature reduction, as alternatives for LDA, in the paper (beginning of section II, the first paragraph);
3- To validate our method when the feature reduction is performed in a nonlinear fashion, an efficient, yet relatively noncomplex in application, nonlinear feature reduction technique, i.e., GNDA is adopted to our method. The experimental results for this nonlinear feature reduction technique are presented and compared with the other techniques. Necessary introduction to this technique in brief, is presented in the beginning of section II and subsection II-B (in fact we added a subsection to section II, i.e., II-B is added)

2. The need of dimensionality reduction for RVM is not explained; The well known LDA

is explained at length, I would recommend to reduce this section and explain more RVM algorithm. In its present form the description appears as compilation of formulas and text from the original publications. For example, on page 11, l37, what are random and systematic components? Why do you need to link them? There are some statements on page 10, l20-37, which are neither well explained or supported by your experiments (e.g. what is the number of SV and RV in your case, are weights distributed around zero in your case?, and so on). Generally, a good explanation of RVM and differences between RVM and SVM is missing.

Answer: We improved the article by a thorough revision as follows accordingly:

1- RVM has the advantages of probabilistic prediction and sparsity over SVM but is relatively sensitive to small training sets in hyperdimensional spaces. The reason we adopted an appropriate feature reduction technique to RVM is to prepare the appropriate condition for exploiting its nice properties and improving its performance in particular on small and scattered land covers through a robust and accurate classifier. This explanation is added to the end of subsection II-C;
2- The LDA subsection is shortened (II-A), though we had to insert a brief on a nonlinear feature reduction, i.e., GNDA as new II-B, but we also keep it short by just referring to the related paper for detail;
3- Subsection on RVM (previously II-B, now II-C) is rewritten thoroughly. The link between the random (weights) and systematic (model parameters) components are known as the link function which is necessary to define the posterior distribution for the learning process.
4- In the revised subsection II-C, the role and importance of RVs and SVs are explained more. Also we defined no. of them in the corresponding experiments. In all of our experiments with RVM, most of weights are distributed closely to zero so that this fact is reflected in the sparsity of the RVs which are typical pixel vectors representing the classes.
5- The differences between SVM and RVM are tried to be highlighted more precisely and clearly.

3. Experimental part:
The parameter selection is unconvincing. Please use at least greed search and cross validation for C and gamma. Using the same training data set for the dimensionality reduction and algorithm training is dangerous in terms of overfitting. Can RVM be used without dimensionality reduction, please explain. It seems that your feature space is "simpler" after the dimensionality reduction, can a simple nearest neighbor classification algorithm be used in projected space instead of RVM, please compare. Please make more links, supported by your experimental results, between your experimental and theoretical parts.

Answer: In order to improve the article according to the comments we did as follows:

1- For a more precise experimental results and parameter selection, we adopted a cross validation experimental framework such that 10 separate randomly selected

training sets are used for each technique in every train-to-test sample ratio. Using this framework and a grid search, the optimum parameters for all of the techniques including SVM are estimated from a wide-enough range of values. These explanations are added to the last paragraph of subsection III-B.

2- To avoid overfitting, we used different training sets for training of the feature reduction and RVM. It is referred to in the second paragraph of subsection III-C.

3- RVM can work on hyperdimensional data without difficulty. For completion of the experiments we added the results with RVM on the original data to the tables too. It also provides a reference to compare SVM and RVM in their original form, which shows comparability on high and medium train-to-test sample ratios and superiority of SVM on low train-to-test sample ratios in particular for small land covers where less training sample are available. This explanation is added to the end of subsection II-C (last paragraph) and also the first paragraph of section III and subsection III-B. The purpose of the proposed method is combining the advantages of RVM and an appropriate feature reducing technique to obtain a robust and accurate classifier and the results show that the proposed method greatly outperforms standard RVM.

4- A nearest neighbor classifier is also applied on the simple reduced data space and the results and explanations are added to section III.

5- We improved all sections, especially II and III, to tackle this important shortage. Hopefully it is of appropriate link now.

Sincerely Yours
Mianji

Robust Hyperspectral Classification Using Relevance Vector Machine

Fereidoun A. Mianji, Student Member, IEEE, Ye Zhang, and Junping Zhang, Member, IEEE

Abstract-Curse of dimensionality is the main reason for the computational complexity and the Hughes phenomenon in supervised hyperspectral classification. The proposed approaches to tackle these problems seldom consider the real situation of insufficiency of available training samples, especially for small landcovers often containing the key information of the scene, and the problem of complexity, in a simultaneous fashion. In this paper, the capabilities of a discriminating feature reduction technique is combined with the advantages of a Bayesian learning-based probabilistic sparse kernel model, relevance vector machine (RVM), to merge a new supervised hyperspectral classification method with low computational complexity, high classification accuracy, and robustness to the Hughes phenomenon. The hyperdimensional data is firstly transformed to a class-oriented low feature space using an appropriate supervised feature reduction technique. Then, the transformed data is processed by a multiclass RVM classifier, designed based on the parallel architecture and one-against-one strategy. Necessary experiments are carried out on real hyperspectral data and the results are compared with the most efficient supervised classification techniques such as support vector machine using appropriate performance indicators. The proposed method outperforms the other approaches in all the aspects in particular for small and scattered landcover classes which are harder to be precisely classified.

Index Terms-Computational complexity, Hughes phenomenon, hyperspectral data, key information preserving, relevance vector machine, supervised classification, support vector machine.

I. INTRODUCTION

Hyperspectral imaging, also known as imaging spectroscopy, generates hundreds of images, corresponding to different wavelength channels, for the same area on the surface of the earth [1]. A hyperspectral image is a three dimensional array or cube of data with the width and length of the array corresponding to spatial dimensions and the spectrum of each point as the third dimension. Hyperspectral imaging differs from multispectral imaging in that the number of bands is much higher (20 or more) and the spectral bands are contiguous. Owing to such spectral resolution, the ability of hyperspectral imaging to detect and identify individual materials or landcover classes is considerably better than other techniques such as multispectral imaging [2].

Analysis of hyperspectral data for defining the landcover classes through classification techniques is not a trivial task. Factors such as the curse of dimensionality and high spatial variability of landcover signatures make this task challenging. One of the main difficulties in supervised classification of hyperspectral data is that the ratio of the number of available training samples to the number of features is usually too small so the class-conditional hyperdimensional probability density functions used in standard statistical classifiers are unapproachable [3]. Hughes phenomenon is the consequence of this ill condition where the classification accuracy decreases as the number of features given as input to the classifier increases over a given threshold [4]. Variety of techniques has been applied to supervised classification of remotely sensed hyperspectral imagery to deal with the curse of dimensionality issue in classification of hyperspectral imagery in last decade. Among them five categories can be referred as the main approaches [3]:

1) A group of approaches proposes feature reduction techniques to tackle the computational cost and Hughes phenomenon as well. It includes two main methods; namely, feature selection and data transformation. Feature selection performs a reduction of spectral channels by selecting a representative subset of original features. To this end, the relative worth of features are assessed in a quantitative and rigorous way. The commonly used procedure is to determine the mathematical separability of classes using methods like Bhattacharyya distance, Jeffries–Matusita distance, and the transformed divergence measure [5], [6]. Since the identification of the best subset of features (the optimal solution) is computationally unfeasible, techniques that lead to suboptimal solutions like the basic sequential forward selection (SFS) [7], sequential forward floating selection [8], and the steepest ascent (SA) techniques [9] are normally used.

It is also possible to effect feature reduction by transforming the data to a new space of a lower dimensionality in which the separability is higher in a subset of the transformed features than in any subset of the original data. In the context of feature transformation, the most commonly used in remote sensing are the principal components analysis (PCA) or Karhunen-Loeve transform [10], canonical analysis [11], decision boundary feature extraction (DBFE) method [12], and projection pursuit algorithm [13]. The main drawback of feature reduction techniques is that the loss of information is often unavoidable and may have a negative impact on classification accuracy [14].

2) The second group of techniques proposes to use the semilabeled samples obtained through classification process in order to enhance statistics estimation and to improve classification accuracy in an iterative fashion [15]-[24]. To this end, after initial classification of test samples using the training samples, the class statistics is strengthened using iterative exploitation of the classified samples and also the training ones. To tackle the problem of maximum likelihood estimation of statistics in the presence of incomplete data, these approaches apply expectation–maximization algorithm for

integration between the training and the semilabeled samples. These methods yield higher classification accuracy since they utilize a larger portion of the image for estimation process. But there are two main drawbacks with this category of approaches: 1) in terms of computational cost it is costly and 2) in order to avoid divergence of the estimation process, it requires a precise match between the initial class model estimated and the training samples.

3) The third category of approaches, i.e., regularization of the sample covariance matrix, adopts the multivariate normal (Gaussian) probability density model. When the ratio of training samples to number of features is too low for a class of information, the traditional methods for estimating the covariance matrix may lead to inaccurate estimations of *first*- and *second*-order statistics, which in turn would make it impossible to invert the covariance matrix in maximum-likelihood classifiers. In order to reduce the variance of the estimations for low ratios of training samples to number of features, a couple of improved covariance matrix estimators have been proposed based on the idea of regularization of the sample covariance matrix [25]-[27]. The main drawback of this category of approaches is that the few available training samples may get over fitted by the estimated covariance matrices, i.e. there is the risk of poor classification due to poor approximation of statistics.

4) Support vector machines (SVMs) are a new approach for the classification of multispectral remote sensing images [28]-[47]. Compare to traditional clustering and classification techniques such as neural networks, minimum distance and maximum likelihood classifiers, they often provide higher classification accuracy [28]-[38]. One of the approved features of SVMs is their applicability on the original hyperspectral feature space, i.e. they yield accurate classification results without feature reduction process. Furthermore, SVMs outperform traditional classifiers in the presence of heterogeneous classes for which only few training samples are available [36]. This characteristic is owing to the fact that SVMs implement a classification strategy that exploits a margin-based

"geometrical" criterion rather than a purely "statistical" criterion [39]–[45]. However, SVM suffers some drawbacks such as limited sparsity and being not an ideal approach when the posterior probabilities of class membership are necessary, which will be discussed in more detail in subsection II-C. The SVM is a binary classifier in the original form, hence, in classification of real hyperspectral information which is basically designed to discriminate among a broad range of landcover classes that may be very similar from a spectral viewpoint, an appropriate design and parameter selection is necessary. Some techniques for application of SVMs in multiclass classification problems are proposed in the literature [35], [36], [46], [47].

5) Another interesting approach to overcome the curse of dimensionality problem is based on taking the spectra of each pixel in the hyperspectral image as a one-dimensional signature into account. The shape of each pixel spectra (signature) is used to model the information of related class via appropriate descriptors. The method is firstly introduced in analytical chemistry [48], [49]. The procedure aims at identifying derivative features that are more effective at separating target classes and then add them to a base subset of features for classification [50]. The point of using derivatives is to capture important spectral details in order to create the smallest set of features that will result the best classification accuracy. Its main advantage is that it simplifies the formulation of hyperspectral classification problem. But it suffers the necessary extra works to develop the appropriate shape descriptors to model the variables of the spectra of different data classes.

Although the aforementioned techniques have obtained many achievements, they rarely take the real situation of insufficiency of available training samples especially for small landcovers into account. Furthermore, their performance in simultaneously tackling the problems of complexity and preserving the key information of hyperspectral data (high classification accuracy) is limited. This

paper proposes a new robust classification approach to efficiently overcome the problems of complexity and accuracy, in particular for small landcovers, through relevant vector machine (RVM). RVM is recently used for regression and supervised classification applications [51]-[53]. In standard form, the RVM hyperspectral classification is less complex compared to SVM but is slightly less accurate for reduced training sets [54], [55]. The proposed approach adopts a combined system using an appropriate data transformation technique with a multiclass RVM design to realize a low-complex and accurate classifier which is also robust to Hughes phenomenon. It is shown that the proposed method outperforms the SVM-based approaches even on reduced training sets. More importantly, its capability to overcome the problem of morphological lack of classification accuracy, especially on small and scattered landcover classes, is validated.

The reminder of this article is organized as follows. In section II the theoretical and structural basis of the proposed method is presented. Section III describes the experiments and results in detail, and finally a discussion on the observations and the conclusion is drawn in section IV.

II. THEORETICAL AND STRUCTURAL BASIS OF THE PROPOSED METHOD

In this section the layout of the proposed new supervised hyperspectral classification method and the theoretical basis of the applied techniques are discussed. It describes how we combine a class-oriented data transformation with a parallel architecture of binary classifiers to superimpose their potentials for classification of hyperspectral data in a high performance fashion. Two feature reduction techniques are adopted in this paper. The first subsection introduces the Fisher linear discriminant analysis (FLDA or LDA) and its application for a straight forward transformation of hyperspectral data to a reduced feature space with preserved key information. FLDA is an effective subspace technique which does not

require the tuning of free parameters, hence, is simple and fast in application [56]. It is of extensive use and practical exploitation in remote sensing applications which are mainly focused on image classification and feature reduction [57]. Appreciating its good capabilities and simplicity, we adopted it in this work to show the good performance of the proposed method even with a simple linear feature reduction technique. Although the standard FLDA presents a reasonable performance in many applications, in excess of number of features to number of training samples leading to singularity of within-class matrix (ill-posed problems), it is not an appropriate choice. Furthermore, it can be unreliable if the classes follow a nonnormal-like or multimodal mixture distribution. Several strategies are proposed to overcome these obstacles [27], [58]-[62]. Subsection B presents one of these modified feature reduction approaches, i.e., generalized nonlinear discriminant analysis (GNDA). GNDA is an effective nonlinear alternative for FLDA to tackle the drawback arising of possible loss of nonlinear features in a linear transformation. Subsection C provides the mathematical background of the RVM as the core of the designed classifier and compares it with the SVM. Finally, in the last subsection the constructed multiclass classifier and the working algorithm are explained.

A. Linear Key Information Preserving Transformation of Hyperspectral Data

Discriminant analysis is the method of finding a function or model for good discrimination between different classes of the input data. These methods can be used to train e.g. a classifier or to visualize certain aspects of the data. Probably the most well known example of linear discriminant analysis method is FLDA, which is a standard supervised technique for dimension reduction in pattern recognition [56].

Fig. 1

The idea of the FLDA is to look for a direction w that separates the class means well (when projected onto the found direction) while achieving a small variance around these means [61]. The quantity measuring the difference between the means of classes is called between class variance, and the quantity measuring the variance around every class mean is called within class variance, respectively [62]. In the linear case, the goal is to find a direction that maximizes the between class variance to within class variance ratio. This is illustrated in Fig. 1. Let χ denote the space of observations ($\chi \subseteq \mathbb{R}^N$) and y the set of possible labels (here y = {+1,−1}). Also, let $C = \{(x_1, y_1),...,(x_M, y_M)\} \subseteq x \times y$ denote the training sample of size n for which $C_1 = \{(x, y) \in C \mid_{y=1}\}$ and $C_2 = \{(x, y) \in C \mid_{y=-1}\}$ split the training samples into two classes. With the means of the data projected onto the direction w represented by μ_1, μ_2 and related variances by σ_1, σ_2, we may describe FLDA according the following mathematical form.

$$\text{maximize } J(w) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1 + \sigma_2} \tag{1}$$

The goal is maximizing the ratio of between class variance to within class (intraclass) variance, i.e., J(W) which is often referred to as Rayleigh coefficient. For a p-class case we have

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{2}$$

where we define the between and within class scatter matrices S_B and S_W as

$$S_B = \sum_{j=1}^{p} n_j (m_j - m)(m_j - m)^T \tag{3}$$

$$S_W = \sum_{x_i \in C_j} (x_i - m_j)(x_i - m_j)^T \tag{4}$$

where m_j and n_j are the mean and the number of training samples in class C_j, and m is the mean of entire classes. The **w** can be determined by the generalized eigen-problem specified by

$$S_B w = \lambda S_W w \tag{5}$$

where λ is a generalized eigenvalue. Since the rank of S_B is p-1, the original L-dimensional data can be transformed into a (p-1)-dimensional space using p-1 eigenvectors associated with p-1 non-zero eigenvalues.

B. Nonlinear Key Information Preserving Transformation of Hyperspectral Data

In recent years, the category of *kernel methods* has offered state-of-the-art performances in ill-posed classification problems associated with hyperspectral images. The strategy of kernel method is to map the data from original space into a Hilbert feature space H with higher dimensionality, where more effective hyperplanes for classification are expected to exist. Intrinsic regularization and robustness in hyperdimensional problems is the key advantage of these methods [59]-[64]. However, the effectiveness of the kernel methods depends on the selection of some critical free parameters, i.e.,

model selection, which usually needs the expert supervision to adequately conduct the model selection learning process. Therefore, complexity of these methods is a matter of the number and sensitivity of the free parameters to be tuned. Nevertheless, in many operational domains, a simple classifier which does not require the tuning of many different parameters, but still presents a competitive accuracy, is preferred.

A generalized nonlinear discriminant analysis (GNDA) is adopted in this work as an alternative for the FLDA where the linear feature reduction is not an ideal option. It is an effective yet relatively simple kernel-based method which consists of two steps. First, using a nonlinear mapping function such as Guassian kernel function, the original hyperdimensional data is mapped into a nonlinear mapping space and then the LDA is implemented in the mapped space. The detail of the approach is available in reference [60].

C. Relevance Vector Machine Classification

Supervised classification methods learn a model of relation between the target vectors $\{t_n\}_{n=1}^{N}$ and corresponding input vectors $\{x_n\}_{n=1}^{N}$ and utilize this model to predict target values for the previously unseen inputs. Input vectors (training set) and target vectors might be pixels within an image and class labels in a classification process, respectively. An important concern with this modeling is to avoid overfitting phenomena which may arise from presence of noise (in regression) or class overlap (in classification) in the training set. SVM, as a recently introduced technique to remote sensing, has many advantages over the conventional and widely used parametric classifiers such as maximum likelihood classification approach. It is relatively insensitive to the dimensionality of the

dataset [64] and usually is accurate in small training sets [65]. Further more, SVM is robust to over-fitting problem due to its between class margin maximization basis [66]. But, despite all these advantages, SVM suffers some drawbacks as follows [51]-[55]:

• The output of the SVM classification is just a class label prediction which conveys no information on the uncertainty of the class allocations, i.e. the predictions are not probabilistic. Therefore, in classification tasks where posterior probabilities of class membership are necessary, SVM can't be an ideal approach.

• The kernel function $K(.,.)$ in SVM must satisfy Mercer's condition, i.e., it must be a continuous symmetric kernel of a positive integral operator.

• Although SVMs are relatively sparse but they make liberal use of kernel functions. Consequently, the number of support vectors (SVs) steeply grows with the size of the training set. Consequently, it gets complex as the size of training sets increases.

• Estimation of the error/margin tradeoff parameter C and kernel specific parameter (e.g. gamma, which controls the width of the widely used radial basis function kernel), often entails a cross validation procedure which can be a waste of data and computation.

A recent development of the SVM, the relevance vector machine (RVM), offers an attractive to image classification applications [51]. RVM is a probabilistic sparse kernel model identical to the SVM in functional form which doesn't suffer any of the SVM's drawbacks. RVM enjoys a Bayesian approach for learning to realize the highest possible sparsity expecting the fact that a considerable number of weights get distributed around zero. It introduces a prior over the model weights governed by a set of hyperparameters, in a probabilistic framework. One hyperparameter is associated with each

weight, and the most probable values are iteratively estimated from the training data. In addition, in contrast to SVM, the non-zero weights in RVM are not associated with examples close to the decision boundary, but rather represent prototypical examples of classes, namely relevance vectors (RVs). The most important feature of RVM is that, while it has a generalization capability comparable to an equivalent SVM, it results in much fewer RVs compared with the number of SVs obtained in the SVM classification. Hence, classification can be carried out faster with the RVM compared to the SVM. The only disadvantage of RVM in comparison with SVM is that it is less robust to reduced training samples in hyperdimentional data spaces [55].

Intrinsically, RVM is a binary classifier. Let us assume a labeled dataset of n pixels $\{(\mathrm{x_i}, \mathrm{t_i})\}_{\mathrm{i=1}}^{\mathrm{n}}$, where $\mathrm{x_i} \in \mathbb{R}^{\mathrm{N}}$, being N the number of spectral channels, and a probabilistic frameworks where only two values are possible, e.g. 0 and 1 as the possible classes ($t_n \in \{0,1\}$). Dependency of the targets on the inputs from the training set in RVM can be modeled based on some function $y(\mathbf{x})$ defined over the input space in the form of

$$y(x;\mathrm{w}) = \sum_{i=1}^{M} w_i \phi_i(x) = W^T \varphi(x) \qquad (6)$$

where $\varphi(x)$ stands for the basis functions (e.g. kernels), $\varphi(\mathrm{x})=(\varphi_1(x),\varphi_2(x),...,\varphi_M(x))^T$, and $\mathrm{w} = (w_1, w_2, ..., w_M)^T$ is the model weights. To infer the function y(x) so that it generalizes from training samples to unseen data, we need to define the basis functions, i.e., type of kernels, and to estimate the weights as well. The objective is having as few nonzero weights as possible to facilitate a sparse approach with fast implementation [52].

Bayesian probabilistic framework infers a distribution over the weights rather than a point-wise estimation. According to the Bayes rule the posterior probability of w is:

$$\mathrm{p}(\mathrm{w}\,|\,\mathrm{t},\alpha)=\frac{\mathrm{p}(\mathrm{t}\,|\,\mathrm{w},\alpha)\mathrm{p}(\mathrm{w}\,|\,\alpha)}{\mathrm{p}(\mathrm{t}\,|\,\alpha)} \tag{7}$$

where $\mathrm{p}(\mathrm{t}\,|\,\mathrm{w},\alpha)$ is the likelihood, $\mathrm{p}(\mathrm{w}\,|\,\alpha)$ is the prior with weights $\alpha=[\alpha_1,...,\alpha_n]^{\mathrm{T}}$, and $\mathrm{p}(\mathrm{t}\,|\,\alpha)$ represents the evidence. Adopting a Bernoulli distribution for $p(t\,|\,x)$, the likelihood merges as:

$$\mathrm{p}(\mathrm{t}\,|\,\mathrm{w},\alpha)=\prod_{\mathrm{i}=1}^{\mathrm{n}}[\sigma(\mathrm{y}(\mathrm{x}_\mathrm{i};\mathrm{w})]^{\mathrm{t}_\mathrm{i}}[1-\sigma(\mathrm{y}(\mathrm{x}_\mathrm{i};\mathrm{w}))]^{1-\mathrm{t}_\mathrm{i}} \tag{8}$$

where $\sigma(y)=1/(1+e^{-y})$ is the logistic sigmoid link function applied to y(x) to generalize the linear model and obtain the probabilistic outputs [67]. The likelihood is complemented by a *prior* conditioned Gaussian on α over the parameters (weights) in the form of

$$p(\mathrm{w}\,|\,\alpha)=\prod_{n=1}^{N}\frac{\sqrt{\alpha_i}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha_i w_i^2}{2}\right) \tag{9}$$

The weights can not be integrated out to analytically obtain the marginal likelihood, therefore an iterative procedure like Laplacian approximation procedure [68] can be utilized as follows.

1) Find the maximum *a posteriori* weights W_{MAP}.

Owing to the fact that $p(\mathrm{w}\,|\,t,\alpha)$ is linearly proportional to $p(t\,|\,\mathrm{w})\times p(\mathrm{w}\,|\,\alpha)$, the maximum *a posteriori* (MAP) solution $\mathrm{W}_{\mathrm{MAP}}$ can be obtained through a penalized logistic log-likelihood function in an iterative fashion by maximizing:

$$\log\{p(t\,|\,\mathrm{w})\,p(\mathrm{w}\,|\,\alpha)\} = \sum_{n=1}^{N}[t_n \log y_n + (1-t_n)\log(1-y_n)] - \frac{1}{2}\mathrm{w}^T A\mathrm{w} \tag{10}$$

for W_{MAP}, with $\mathrm{y}_n = \sigma\{\mathrm{y}(\mathrm{x}_n;\mathrm{w})\}$ and $\mathrm{A} = \mathrm{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$ is being composed of the current values of α. The iteratively reweighed least-squares algorithm can be used to find W_{MAP} [51], [69].

2) Compute the Hessian at W_{MP}.

The Hessian can be derived by double differentiation of the logistic log-likelihood in the form of

$$\nabla_w \nabla_w \log p(w\,|\,t,\alpha)|_{w_{MP}} = -(\Phi^T \mathrm{B}\Phi + \mathrm{A}) \tag{11}$$

where $\mathrm{B} = diag(\beta_1, \beta_2, ..., \beta_N)$ is a diagonal matrix with $\beta_n = \sigma\{y(x_n;\mathrm{w}_{MP})\}\left[1-\sigma\{y(x_n;\mathrm{w}_{MP})\}\right]$, and Φ is the 'design' matrix with $\Phi_{nm} = K(x_n, x_{m-1})$ with $\Phi_{n1} = 1$. The posteriori is estimated over weights centered at W_{MAP} through negating and inverting of the result by a Gaussian approximation with the covariance $\sum$.

$$\sum = (\Phi^{\mathrm{T}}\mathrm{B}\Phi + \mathrm{A})^{-1} \tag{12}$$

The classification problem can be locally linearized around W_{MAP} with

$$\mathrm{w}_{\mathrm{MAP}} = \sum \Phi^{\mathrm{T}}\mathrm{B}\hat{\mathrm{t}} \tag{13}$$

$$\hat{\mathrm{t}} = \Phi \mathrm{w}_{\mathrm{MAP}} + \mathrm{B}^{-1}(\mathrm{t}-\mathrm{y}) \tag{14}$$

After obtaining $\mathbf{W}_{MAP}$, the hyperparameters α are iteratively updated using

$$\alpha_i^{new} = \frac{1-\alpha_i^{old}\sum_{ii}}{\mu_i^2} \tag{15}$$

where $\mu_i \in \mu = [\mu_1, ..., \mu_n]^T$ is the ith posterior mean weight, and $\sum_{ii}$ is the ith diagonal element of the covariance. The procedure is repeated until some suitable convergence criteria are satisfied or maximum number of iterations is reached. During the optimization process, many α_i tend to infinity, so the associated model weights w_i are effectively discarded, ultimately realizing a sparse solution. The remaining (typically very few) examples are called Relevance Vectors (RVs). It is worth noting that in RVM $p(t, w|\alpha)$ is log-concave, hence in the Bayesian treatment, when the posterior mode is unrepresentative of the overall probability mass, the Gaussian approximation is much more trustworthy compare to multilayer neural networks [52].

Fig. 2

The key idea of the proposed method is to prepare such a condition that the nice capabilities and privileges of RVM, in particular the probabilistic prediction and sparsity properties, are exploited efficiently. Regarding the relative sensitivity of RVM to the size of training set in hyperdimensional data spaces, a sound approach is an appropriate class-oriented feature reduction. To this end, we adopted FLDA and GNDA as the preclassification stages to increase the training sample to feature number ratio aiming at obtaining an efficient and robust classifier especially for small and scattered landcover patches.

D. Parallel One-Against-One Strategy Multiclass RVM Classifier

The classification of hyperspectral remote sensing data usually involves simultaneous discrimination of numerous information classes whereas RVMs are intrinsically binary classifiers. In order to deal with this requirement, four main strategies of combination of RVMs are possible. Let $C = \{C_1, C_2, ..., C_M\}$ be the set of M possible labels (information classes) associated with the L-dimensional hyperspectral image of the study area. The goal is to define the label of each L-dimensional sample x so that a predefined classification criterion is optimized. To adopt a binary classifier like RVM to this task, the general strategies are based on combining an ensemble of binary classifiers, a set of two-class problems, according to some decision rules.

In an one-against-one two-class problem, for each RVM, targets with values +1 and -1 are assigned to the samples of classes C_A and C_B, respectively. However, in one-against-all design, the comparison is between one class as C_A and all other classes together as C_B [70]. There are two main approaches to combine these subsets (binary classifiers), namely the "parallel" and the "hierarchical tree-based" approaches. Parallel approach enjoys higher discrimination accuracy because it doesn't suffer the intrinsic risk of propagation of error from the top to bottom of tree as the other approach. But in terms of computational cost, the hierarchical tree-based approach is better. On the other hand, one-against-all strategy often suffers the complex discriminant functions in discrimination task between an information class and all the others [3]. Taking all the above stated points into account, the one-against-one parallel approach as Fig. 3 is adopted for the experiments of this work.

Fig. 3

All possible pair-wise classifications are modeled through the proposed design by using $M(M-1)/2$ RVMs. Each pixel is analyzed by a discriminant function $g_{ij}(x)$ to define its belonging to one of information classes C_i, C_j where $C_i, C_j \in C, i \neq j$ and a positive score is given to the winner class of each one-against-one competition through the score function $S_i(x)$.

$$S_i(x) = \sum_{\substack{i=1 \\ j \neq i}}^{M} \operatorname{sgn}\left\{g_{ij}(x)\right\} \tag{16}$$

The final decision about the class of each pixel $C(x)$ is made on the total score obtained by each class at the end of this parallel process on the basis of the "winner-takes-all" rule, according to the following maximization

$$C(x) = \arg\max_{i=1,\ldots,M}\left\{S_i(x)\right\} \tag{17}$$

The algorithm that implements the proposed parallel one-against-one strategy is as follows:

Step 0: **Initializing**

-Set class numbers M

-Set class information labels $C = \{C_1, C_2, ..., C_M\}$

Step 1: **Analyzing**

-For i=1, …, M-1

-For j=i+1, …, M

- Analyze pixel x using discriminant function $g_{ij}(x)$, $(i \neq j)$, to define $P(x, C_i)$ and $P(x, C_j)$
- If $P(x, C_i) > P(x, C_j)$, set $S_i(x) = S_i(x) + 1$ otherwise set $S_j(x) = S_j(x) + 1$

Step 2: **Decision Making**

-Define the class label of x using $C(x) = \arg\max\{S_i(x)\}$, i=1, …, M

-Stop

III. EXPERIMENTS AND RESULTS

The aims of the experiments include: 1) to evaluate the overall classification accuracy of the proposed method in comparison with the standard RVM, applied on the original hyperdimensional data space, and the other techniques as well; 2) to investigate its performance for classification of small landcover classes; and 3) to assess the robustness of the proposed approach to varying train-to-test sample ratios for different morphologies of landcover classes. To this end, first we introduce the hyperspectral data and its ground truth map which is used for the experiments. Then the type of experiments and comparisons are explained in detail. The performance indicators which are adopted to validate our results and conclusions include the classification accuracy, the computational time for the test stage of each approach, number of SVs and RVs for the SVM-based and RVM-based approaches, and the stability or robustness of the method to train-to-test sample ratio.

A. Dataset Description

Many new hyperspectral instruments have been developed for remote sensing applications during the last decade. Among them, one of the most advanced sensors is NASA/Jet Propulsion Laboratory Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) which covers the wavelength region from 0.4 to 2.5 μm [71]. It is particularly useful to geological studies owing to 224 spectral channels at a nominal resolution of 10 nm. AVIRIS produces spectacular spectra containing an abundance of information about materials on the Earth's surface.

The test image in the experiment, namely San Diego includes a remotely sensed hyperspectral image which is collected by the AVIRIS. San Diego is a 405 $\times$400 pixel area image with a ground resolution of 3.5 meter. After removal of the water absorption regions, low SNR and bad bands 126 available bands from original data remain in the 0.4–1.8 μm wavelength range. Fig. 4(a) shows the tenth band of San Diego image.

Fig. 4

A hard classification map (ground truth) of San Diego which defines 9 classes is available and is shown in Fig. 4(b). These 9 classes are different in size and shape; some of them are integrated and large areas (classes C1, C2, C9) whereas some mainly include small and scattered areas (classes C3, C5, C6, and C8). Classes C4 and C7 include areas with both characteristics. Fig. 5 depicts this fact using single-class images.

Fig.5

B. Experimental Setup

The experiments are organized into two parts. In the first part the proposed method adopted with two different feature reduction steps, i.e., linear (FLDA+RVM) and nonlinear (GNDA+RVM), is compared to the approaches which utilize the RVM on the original hyperspectral data space and also RVM with the PCA as a widely used feature reduction technique. The RVM+PCA are applied in different numbers of reduced features, i.e., 10, 20, 30 and 40 to have a broader range of results by this method for comparison. We applied a k-Nearest Neighbor (k-NN) classifier as a standard classification algorithm to the output of the FLDA to verify the separability of the classes in the reduced data space and provide another reference for comparison as well (k-NN is not well-suited to hyperspectral data sets without dimensionality reduction due to very high complexity). Furthermore, the non-linear SVM classifier based on two types of kernel functions, namely polynomial kernel (SVM-Poly) and Gaussian radial basis function kernel (SVM-RBF), as the most efficient and accurate supervised classification approaches, are used to benchmark the results. It is worth mentioning that the linear, polynomial, and radial basis function kernels are the most popular kernels used in the RVM and SVM. Due to lower performance of linear kernel function (in most of cases), it is not used in this comparison [64].

Polynomial kernel

$$K(x_i, x_j) = (\gamma x_i . x_j)^d \quad (18)$$

Radial basis function kernel

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (19)$$

Gamma (γ) is the inner product coefficient in polynomial kernels whereas it determines the RBF width in the RBF kernels. It tunes the smoothing of the discriminant function. There is another important factor in the SVM classifier, i.e., cost of constrain violation or the regularization parameter (C) to control the tradeoff between the margin and the size of the slack variables. In order to reliably evaluate the average performance of the aforementioned approaches, we adopted a cross validation experimental framework such that ten separate randomly selected training sets are used for each approach in every train to test sample ratio. Estimated through grid search on the basis of the available training samples using this cross validation method, within a given set $\{0.1, 0.3, ..., 3\}$ for Gamma and $\{2^{-1}, 2^{1}, ..., 2^{15}\}$ for C, Gamma and C are set to 0.5 and 2^{7} in SVM-RBF and 2.3 and 2^{7} in SVM-Poly (with degree d is set to 3 in SVM-Poly), respectively, as the optimum values. Gaussian RBF kernel is applied in all the RVM-based approaches too, hence we only need to find an optimum workable Gamma for them. Gamma is set to the range of 1-3, 0.1-0.3, and 0.05-0.1 for the RVM, RVM+PCA, RVM+FLDA approaches, respectively. RVM+GNDA is composed of a nonlinear feature reduction technique and RVM, therefore, it has two adjustable parameters. Here, for simplicity only one RBF kernel is adopted in the GNDA algorithm whereas multiple Mercer kernels are adoptable in this method as well. Using same cross validation for experiment and a grid search, optimum Gamma values for the GNDA and RVM algorithms are set to the ranges of $2^{-7} - 2^{-5}$ and $2^{-12} - 2^{-10}$, respectively. The multiclass strategy which is used for all of the three approaches is one-against-one parallel architecture (Fig. 3).

C. Evaluating the Overall and Morphological Performance of the Proposed Method

The first step is transforming the hyperdimensional San Diego data into a low feature space through the FLDA. Regarding the number of classes, the transformation would be as follow:

$$(N_i, L) \Rightarrow (N_i, M-1); \quad i = 1-M \tag{20}$$

where N_i is the number of pixels in the i_{th} information class, L is the number of bands in the original hyperspectral data (here L=126), and M is the number of designated classes (here M=9). Therefore, the transformation reduces the number of features to M-1 (here 8) for all of the information classes using M-1 eigenvectors associated with M-1 non-zero eigenvalues of equation (5). Then, using the GNDA, a nonlinear transformation of the data is carried out as well.

In this experiment a ratio of 1/30 of every landcover class is used to generate sets of training samples. To prevent the risk of overfitting in combined approaches (except RVM+PCA in which PCA is an unsupervised approach), different training sets are used for the feature reduction algorithm, FLDA and GNDA, and the main classifiers, RVM and k-NN, such that the size of training sets for the feature reduction phase is about half of the one for the main classifier. Table 1 shows the number of training samples of each landcover class for training of the main classifiers as well as the number of associated test samples for the validation purpose in the experiments of this part.

Table 1

Table 2

Single-class and overall accuracies for the proposed method (with FLDA as well as GNDA) and the other approaches are summarized in Table 2 with maximum values in every column in bold. PCA10, …, PCA40 stand for feature reduction from the original L-dimensional feature space to the reduced feature spaces of sizes 10, …, 40, respectively, by PCA. The overall accuracy of RVM+FLDA is 2% better than RVM and RVM-PCAs and 1% better than the other approaches. RVM+GNDA even slightly outperforms RVM-FLDA. But the main advantage of the proposed technique over the other approaches is on the classification accuracy for the small or scattered landcovers. We already categorized the 9 classes into 3 groups: 1) group A includes landcover classes C1, C2, and C9 which are dominated by integrated and large areas; 2) group B includes landcover classes C3, C5, C6, and C8 which are dominated by small or scattered areas; and 3) group C includes landcover classes C4 and C7 which are a mixture of type A and B, i.e. they contain both large and scattered small areas. From Table 2 it can be concluded that the improvement of the classification accuracy through the proposed method (either with FLDA or GNDA) is particularly considerable for group B, where a gain of about 3%, 6.5%, 5.5%, and 5.5% is obtained over the other RVM approaches for landcover classes C3, C5, C6, and C8, respectively. Furthermore, the proposed approach has a gain of about 1.5%, 5%, 2%, and 2% over the best achieved results by the SVM approaches for landcover classes C3, C5, C6, and C8, respectively. Applying k-NN, instead of RVM, to the reduced feature space by FLDA presents comparable results on classes C3, C6, and C8, although it is poorer on classes C4, C5, and C7. The most critical classification accuracy belongs to C5 which is around 92% through RVM+PCAs and 93.93% through SVM-Poly, whereas the one of RVM+FLDA is the best with a value of 98.85% which is very desirable regarding the very scattered nature and few available training samples for this class. The proposed method is even slightly superior to the other approaches for group A (except class C1 for which all approaches are comparable). In Group C, the proposed approach outperforms all the others for class

C4 and is slightly better for class C7. The littleness of this superiority is probably due to the dominance of the effect of the large integrated areas in the calculation of classification accuracy, i.e. the B property of group C in class 7 is almost overshadowed by the A property.

As per the computational time, the RVM+PCA approaches have longer training times compared to the SVM-based approaches. Training time for RVM+FLDA is even longer and is about 1.5, 85 and 130 times of the RVM+PCA, SVM-Poly, and SVM-RBF, respectively. But in terms of testing time, RVM+FLDA is the fastest and is about 1.5, 2, 3.5 time faster than RVM+PCA, SVM-Poly, and SVM-RBF, respectively. The higher training time for RVM is due to the need for frequent computing and inverting of the Hessian matrix which is $O(N^3)$ for a set of N samples. The training and testing time for the GNDA version of the proposed method is about 1.5 times of the FLDA version.

D. Evaluating the Robustness of the Proposed Method to the Size of Training Set

The 9 available landcover classes are used to generate sets of training samples in different sizes for each class. The training sets are made by picking up the samples in ratios of 1/10, 1/15, 1/30, 1/45, 1/60, 1/90, and 1/120 from the landcover classes. The aim is to investigate the sensitivity of the technique to the number of training samples which is a critical variable in supervised classification. The FLDA and GNDA transformation matrices W are calculated using the training sets (different than the training sets for the RVM or k-NN) and then are applied to the training (for the RVM or k-NN) and test sets for feature reduction. For conciseness, only the results for ratios 1/10, 1/60, and 1/120 are tabulated through Tables 3 to 5, respectively. The results for the ratio of 1/30 were already shown in Table 2.

Table 3

Table 4

Table 5

The results show that Group B as the target group of this study benefits the proposed technique for all ratios of the number of training to test samples. Class C5 as the critical landcover class has a gain of about 1%, 5%, 7%, and 9% over the best results obtained by the other approaches (except k-NN for which the differences are lower) for train to test sample ratios of 1/10, 1/30, 1/60, and 1/120, respectively. The obtained results for classes of group B are depicted in Figs. 7, 9, 10, and 12 in order to provide an easy comparison of the classification accuracies for different approaches versus the train-to-test sample ratio. The superiority of the proposed technique over the other approaches for small and scattered landcover classes (group B) is clear in these figures. It shows the highest level of classification accuracy for almost whole range of the experiment. Furthermore, it outperforms the other approaches in terms of low sensitivity to the train-to-test sample ratio. This implies the robustness of the proposed method to Hughes phenomenon which concerns lose of classification accuracy for low training sample to feature ratios in supervised classification of hyperspectral images.

Fig. 6 to Fig. 13

Fig. 14

The classification accuracy versus the train-to-test sample ratio for classes C2, C4, C7, and C9 are depicted in Figs. 6, 8, 11, and 13, respectively. These curves verifies that the proposed approach is superior to the other approaches for group A (except class C1) and group C in terms of accuracy and robustness too. Due to the comparability of the obtained results for Class C1 for all the train-to-test sample ratios, corresponding curve is not shown here. This can be explained based on the morphology of the different classes, i.e. in groups B and C the contribution of neighboring classes in the collected reflectance of the target pixel is more serious compare to group A (in particular in class C1) because the pixels of the scattered landcover classes in group B are mostly (and considerably in group C) distributed in small and scattered groups surrounded by other classes. Fig. 14 gives a comparison on the overall accuracy and robustness of the proposed method with the other approaches. While the RVM+PCA approaches are a little bit lower than SVM-based approaches in terms of accuracy, both RVM+FLDA and RVM+GNDA outperforms all approaches in particular for smaller number of training samples.

Finally, the complexity of the approaches should be considered. Although, the proposed method needs more training time, as the number of training samples decreases, which is often the real case, the difference between the required training time for the RVM-based and SVM-based approaches decreases sharply. In terms of testing time, the proposed method is much faster than k-NN and is comparable to RVM and RVM-PCA. Compared to SVM, depending to the train-to-test sample ratio, it is about 1.5 to 4 times faster in these experiments. But, a more general criteria for comparison of the complexity of the RVM and SVM approaches is the number of vectors they use for the classification

task, i.e., RVs and SVs. This is a key factor in complexity of the methods, i.e., as the number of test samples increases in an experiment, higher number of SVs or RVs leads to more computational cost. Therefore, according to the obtained results SVM-based approaches are considerably more complex and vulnerable to the size of test sample set. Consequently, in classification applications with *a priori* training, the proposed method is advantageous in terms of computational cost too.

IV. DISCUSSION AND CONCLUSION

A new method for tackling the problems of curse of dimensionality and degrading performance due to limitation in the number of training samples in supervised classification of hyperspectral remote sensing images is addressed in this paper. The proposed method combines the capabilities of the recently invented binary classifier, i.e., RVM, which is a probabilistic sparse kernel model based on Bayesian approach, with the advantages of a discriminating feature reduction technique, i.e., linear or nonlinear discriminat analysis, to optimize the supervised classification model learnt from the available training samples. It performs the classification in two stages, first a discriminant transformation is applied on the original hyperdimensional data using the FLDA or GNDA to reduce the dimensionality of the feature space in a key information preserving fashion, then the transformed data are fed to a multiclass RVM classifier which is designed based on the parallel architecture and one-against-one classification strategy in order to classify the data into predetermined class labels.

The method is applied on real hyperspectral data with associated hard classification map. After transforming the test data to the reduced feature space, the classification is carried out on the transformed data. To validate the method, it is compared to the RVM as well as RVM in join with a widely used feature reduction technique, namely PCA and also to one of the most recent and efficient

supervised classification techniques, i.e., SVM with radial basis function and polynomial kernels. The performance indicators which are used to support the experimental analysis include the classification accuracy, the computational time, and the robustness to the number of training samples. It is shown that the proposed approach is superior to all the other approaches especially for landcover classes which considerably contain small and scattered areas. In terms of computational time, although it is slower than RVM+PCA and SVM-based approaches in training stage, but is faster than the others in testing stage which makes it a preferable option for applications which require low complexity such as real-time classification with *a priori* training. Furthermore, both the single-class and overall classification accuracies of the proposed method suffer the least degradation as the number of training samples reduces, i.e., it is the most robust method to the number of training samples or Hughes phenomenon among all the other approaches.

ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of China under the grants 60972143 and 60972144, and research fund for the doctorial program of higher education of China under the grant 20092302110033.

REFERENCES

[1] C.-I. Chang, *Hyperspectral Imaging: Spectral Detection and Classiffcation*, Kluwer Academic/Plenum Publishers, New York, 2003.

[2] B. S. Penn, "Using simulated annealing to obtain optimal linear end-member mixtures of hyperspectral data," *Computers & Geosciences*, vol. 28, pp. 809–817, 2002.

[3] F. Melghani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, 1778-1790, Aug. 2004.

[4] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[5] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*, Berlin, Germany: Springer-Verlag, 2006.

[6] L. Bruzzone, F. Roli, and S. B. Serpico, "Extension of the Jeffreys-Matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, no. 6, pp. 1318–1321, Nov. 1995.

[7] J. Kittler, "Feature set search algorithm," in *Patt. Recognit. Signal Process.*, C. H. Chen, Ed. Alphen aan den Rijn, Netherlands: Sijthoff and Noordhoff, 1978, pp. 41–60.

[8] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, pp. 1119–1125, 1994.

[9] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, July 2001.

[10] S.K. Jensen and F.A. Walts, "Principal component analysis and canonical analysis in remote sensing," in *Proc. American Soc. of Photogrammetry 45th Ann. Meeting*, pp. 337-348, 1997.

[11] W. Eppler, "Canonical analysis for increased classification speed and channel selection," *IEEE Trans. Geosci. Electronics*, vol. GE-14, pp. 26-33, 1976,.

[12] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 15, pp. 388–400, Apr. 1993.

[13] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.

[14] P. Zhong, P. Zhang, and R. Wang, "Dynamic learning of SMLR for feature selection and classification of hyperspectral data," *IEEE Geosc. Remote. Sens. Lett.*, vol. 5, no. 2, pp. 280-284, Apr. 2008.

[15] Q. Jackson and D. A. Landgrebe, "An adaptive classifier design for highdimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote. Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.

[16] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote. Sens.*, vol. 32, no. 5, pp. 1087–1095, Sept. 1994.

[17] M. Chi, Q. Kun, J. A. Benediktsson, and R. Feng, "Ensemble classification algorithm for hyperspectral remote sensing data," *IEEE Geosc. Remote. Sens. Lett.*, vol. 6, no. 4, pp. 762-766, Oct. 2009.

[18] S. Velasco-Forero and V. Manian, "Improving hyperspectral image classification using spatial preprocessing," *IEEE Geosc. Remote. Sens. Lett.*, vol. 6, no. 2, pp. 297-301, Apr. 2009.

[19] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosc. Remote. Sens.*, vol. 46, no. 4, pp. 1231-1232, Apr. 2008.

[20] J. A. Richards and X. Jia, "Using suitable neighbors to augment the training set in hyperspectral maximum likelihood classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 774–777, Oct. 2008.

[21] L. Bruzzone and C. Percello, "A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples," *IEEE Trans. Geosc. Remote. Sens.*, vol. 47, no. 7, pp. 2142-2154, Jul. 2009.

[22] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral classification using active label selection," in *Proc. SPIE Image Signal Process. Remote Sens. XV*, 2009, vol. 7477, pp. 74770F1-F74770F8.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 19, pp. 1–38, 1977.

[24] T. K. Moon, "The expectation-maximization algorithm," *Signal Process. Mag.*, vol. 13, pp. 47–60, 1996.

[25] Q. Du and C.-I Chang, "Linear constrained distance-based discriminant analysis for hyperspectral image classification," *Pattern Recognition*, vol.34, no. 2, pp. 361-373, 2001.

[26] C.-I Chang and B.-H. Ji, "Weighted abundance constrained linear spectral mixture analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 2, pp. 378-388, 2006.

[27] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosc. Remote. Sens.*, vol. 47, no. 3, pp. 862-873, Mar. 2009.

[28] L. Hermes, D. Frieauff, J. Puzicha, and J. M. Buhmann, "Support vector machines for land usage classification in landsat TM imagery," in *Proc. IGARSS*, Hamburg, Germany, 1999, pp. 348–350.

[29] F. Roli and G. Fumera, "Support vector machines for remote-sensing image classification," in *Proc. SPIE Int. Society Opt. Eng.*, 2001, vol. 4170, pp. 160–166.

[30] L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosc. Remote. Sens.*, vol. 47, no. 4, pp. 1108-1122, Apr. 2009.

[31] B. Demir and S. Erturk, "Clustering-based extraction of border training patterns for accurate SVM classification of hyperspectral images," *IEEE Geosc. Remote. Sens. Lett.*, vol. 6, no. 4, pp. 840-844, Oct. 2009.

[32] J. Munoz-Mari, L. Bruzzone, and G. Camps-Valls, "A support vector domain description approach to supervised classification of remote sensing images," *IEEE Trans. Geosc. Remote. Sens.*, vol. 45, no. 8, pp. 2683-2692, Aug. 2007.

[33] A. Mathur and G. M. Foody, "Crop classification by a SVM with intelligently selected training data for an operational application," *Int. J. of Remote Sens.*, vol. 29, no. 8, pp. 2227-2240, Apr. 2008.

[34] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, pp. 725–749, 2002.

[35] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. SPIE Int. Society Opt. Eng.*, 1999, vol. 3584, pp. 221–232.

[36] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," *in Summaries 8th JPL Airborne Earth Science Workshop*, 1999, JPL Pub. 99-17, pp. 217–227. Online. [Available]: ftp://popo.jpl.nasa.gov/pub/docs/workshops/99_docs/toc.html.

[37] J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *Proc. IGARSS*, Honolulu, HI, 2000, pp. 813–815.

[38] F. Melgani and L. Bruzzone, "Support vector machines for classification of hyperspectral remote-sensing images," in *Proc. IGARSS*, Toronto, ON, Canada, 2002, pp. 506–508.

[39] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[40] C. B. E. Boser, I.M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. ACM Workshop Computational Learning Theory*, 1992, pp. 144–152.

[41] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[42] Set of tutorials on SVM's and kernel methods [Online]. Available: http://www.kernel-machines.org/tutorial.html.

[43] I. El-Naqa, Y. Yongyi, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, pp. 1552–1563, Dec. 2002.

[44] J. Robinson and V. Kecman, "Combining support vector machine learning with the discrete cosine transform in image compression," *IEEE Trans. Neural Networks*, vol. 14, pp. 950–958, Jul. 2003.

[45] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 20, pp. 637–646, Jun. 1998.

[46] D. J. Sebald and J. A. Bucklew, "Support vector machines and the multiple hypothesis test problem," *IEEE Trans. Signal Process.*, vol. 49, pp. 2865–2872, Nov. 2001.

[47] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415–425, Mar. 2002.

[48] J. P. Hoffbeck and D. A. Landgrebe, "Classification of remote sensing images having high-spectral resolution," *Remote Sens. Environ.*, vol. 57, pp. 119–126, 1996.

[49] F. Tsai and W. D. Philpot, "A derivative-aided hyperspectral image analysis system for land-cover classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 40, pp. 416–425, Feb. 2002.

[50] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote. Sens.*, vol. 47, no. 8, pp. 2973-2987, Aug. 2009.

[51] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems* , vol. 12, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000.

[52] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[53] W. Liyang, Y. Yongyi, R. M. Nishikawa, M. N.Wernick, and A. Edwards, "Relevance vector machine for automatic detection of clustered microcalcifications," *IEEE Trans. Med. Imag.*, vol. 24, no. 10, pp. 1278–1285, Oct. 2005.

[54] B. Demir and S. Erturk, "Hyperspectral image classification using relevance vector machines," *IEEE Geosci. Remote Sens. Lett.,* vol. 4, no. 4, pp. 586-590, 2007.

[55] G. M. Foody, "RVM-based multi-class classification of remotely sensed data," *Int. J. of Remote Sens.*, vol. 29, no. 6, pp. 1817-1823, Mar. 2008.

[56] Q. Du, "Modified Fisher's linear discriminant analysis for hyperspectral image dimension reduction and classification," in *Proc. SPIE Chem. Bio. Sens. Ind. Environ. Monit. II,* 2006, vol. 6378, pp. 63781D1-8.

[57] M. L. Clark, D. A. Roberts, and D. B. Clark, "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales," *Remote Sens. Environ.*, vol. 96, no. 3/4, pp. 375–398, Jun. 2005.

[58] C. Percello and L. Bruzzone, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosc. Remote. Sens.*, vol. 47, no. 9, pp. 3180-3191, Sept. 2009.

[59] B-C. Kou, C-H. Li, and J-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosc. Remote. Sens.*, vol. 47, no. 4, pp. 1139-1155, Apr. 2009.

[60] L. Zhang, W. Zhuo, H. Zhang, and L. Jiao, "Generalized nonlinear discriminant analysis," in *Proc. 19th Int. Conf. Pattern Recognit. (ICPR)*, Tampa, FL, Dec. 8-11, 2008, pp. 1-4.

[61] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci Remote Sens.*, vol. 43, no. 6, pp. 1351-1362, Jun. 2005.

[62] S. Mika, "*Kernel Fisher discriminants,*" Ph.D. dissertation, Dept. Comput. Sci., Univ. Technol., Berlin, Germany, 2002.

[63] L. Bruzzone, M. Chi, and M. Marconcini, *Hyperspectral data exploitation: Theory and applications*, C.-I. Chang, Ed. Hoboken, NJ: Wiley, 2007.

[64] A. Plaza, J. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. s110-s122, 2009.

[65] A. Mathur and G. M. Foody, "Crop classification by support vector machine with intelligently selected training data for an operational application," *Int. J. of Remote Sens.*, vol. 29, no. 8, pp. 2227-2240, Apr. 2008.

[66] Z. Chen and H. Tang, "Sparse Bayesian approach to classification," IEEE Networking, In *Sensing and Control Proceedings*, pp. 914–917, 2005.

[67] G. Camps-Valls, A. Rodrigo-Gonzalez, J. Munoz-Mari, L. Gomez-Chova, and J. Calpe-Maravilla, "Hyperspectral image classification with Mahalanobis relevance vector machines," in *Proc. IGARSS*, Barcelona, Spain, 2007, pp. 3802-3805.

[68] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Comput.*, vol. 4, no. 5, pp. 720–736, 1992.

[69] I. T. Nabney, "Efficient training of RBF networks for classification," in *Proc. 9th ICANN*, Edinburgh, UK, Sep. 7-10 1999, vol. 1, pp. 210–215.

[70] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," in *Proc. Int. Conf. Patt. Recog.*, 1994, pp. 77–87.

[71] R.O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, pp. 227–248, 1998.

Fereidoun A. Mianji received the M.S. degree in medical engineering from Sharif University of Technology, Iran, in 1997. He is currently pursuing the Ph.D. degree in the Department of Information Engineering, School of Electronics and Information Technology, Harbin Institute of Technology, China. His research interests focus on digital image processing, hyperspectral image processing and application, and medical imaging.

Ye Zhang received the M.S. and Ph.D. degrees in communication and information science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1996, respectively. He is a Professor with the Department of Information Engineering, School of Electronics and Information Technique, Harbin Institute of Technology. His research interests include hyperspectral image processing and application, image and video compression, imaging scout, and vision matching. Dr. Zhang is a member of the Chinese Signal and System teaching committee and the Commission Science Technology and Industry for National Defense expert committee.

Junping Zhang (M'05) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, in 1998 and 2002, respectively. She is currently an Associate Professor with the Department of Information Engineering, School of Electronics and Information Technique, Harbin Institute of Technology. Her research interests include a wide variety of topics in the area of digital signal and image processing, such as remote sensing image processing, data fusion, change detection, pattern recognition, and wavelet theory and its application.

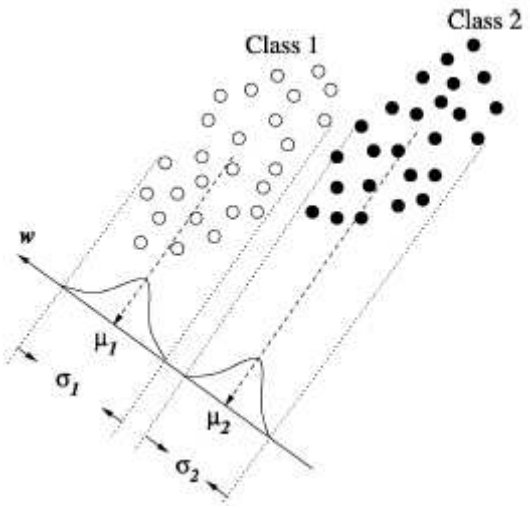


Fig. 1. Illustration of Fisher discriminant for two classes. We search for a direction w, such that the difference between the class means projected onto this directions (μ_1 and μ_2) is large and such that the variance around these means (σ_1 and σ_2) is small.

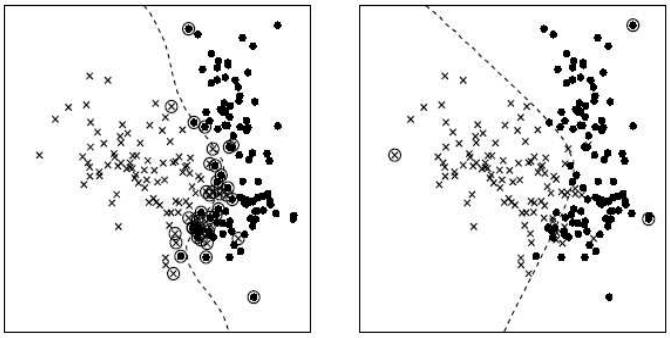


Fig. 2. A comparison of the results of training functionally of identical SVM (left) and RVM (right) classifiers on a typical synthetic dataset [51]. The decision boundary is shown dashed and the vectors are circled to highlight the dramatic reduction of vectors in RVM.

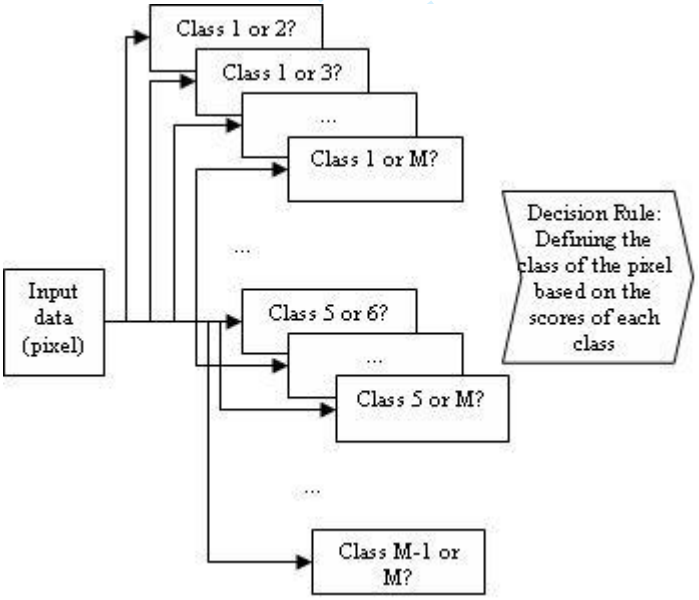
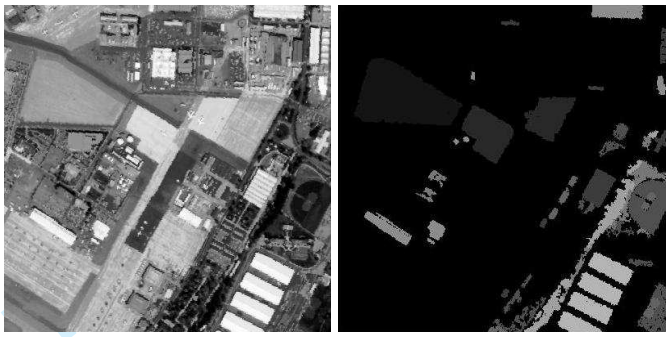
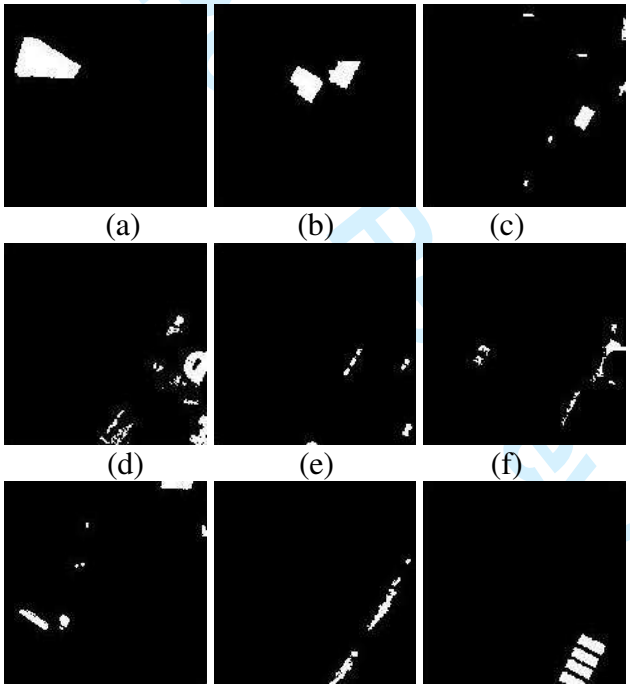


Fig. 3. Block diagram of the designed parallel one-against-one architecture for the multiclass RVM classifier. M(M-1)/2 binary RVMs are needed in this system.



(a) (b)

Fig. 4. San Diego AVIRIS-II hyperspectral image. (a) Tenth band. (b) Ground truth map.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 5. The singular hard classification map of Sandeigo image. Images (a) to (i) correspond to classes C1 to C9, respectively.

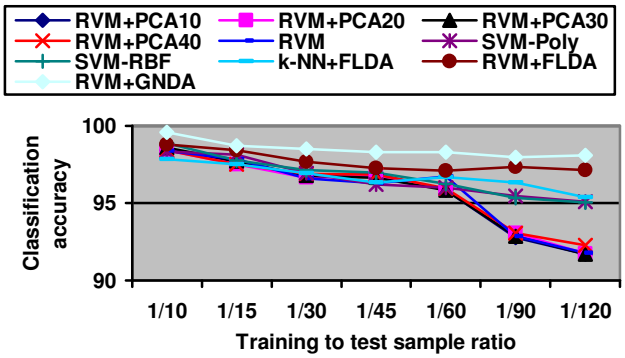


Fig. 6. Classification accuracy versus train-to-test sample ratio for landcover class C2.

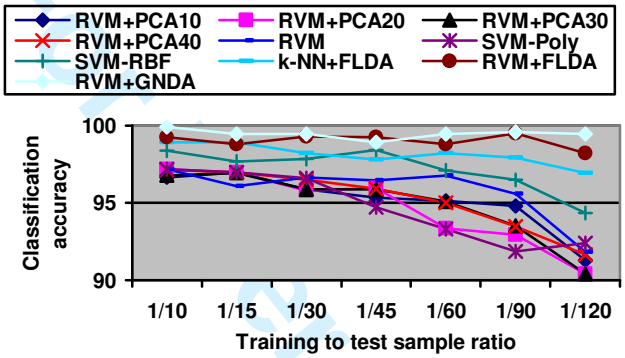


Fig. 7. Classification accuracy versus train-to-test sample ratio for landcover class C3.

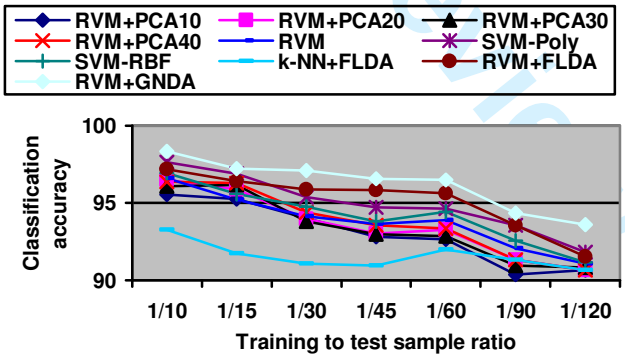


Fig. 8. Classification accuracy versus train-to-test sample ratio for landcover class C4.

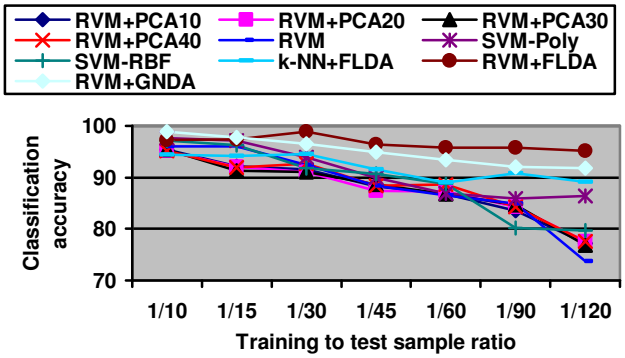


Fig. 9. Classification accuracy versus train-to-test sample ratio for landcover class C5.

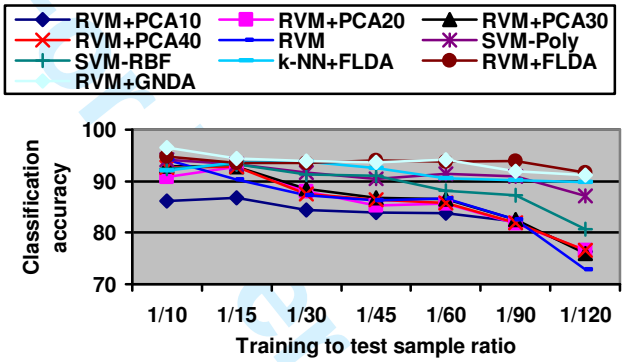


Fig. 10. Classification accuracy versus train-to-test sample ratio for landcover class C6.

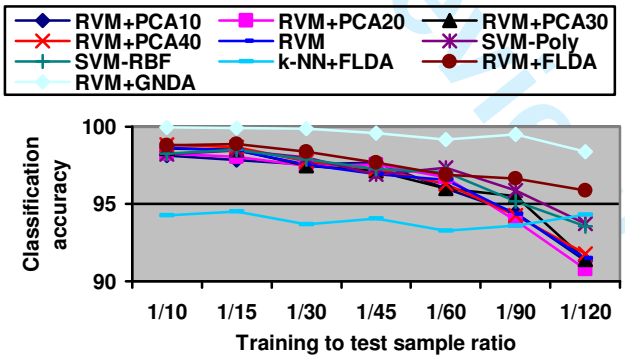


Fig. 11. Classification accuracy versus train-to-test sample ratio for landcover class C7.

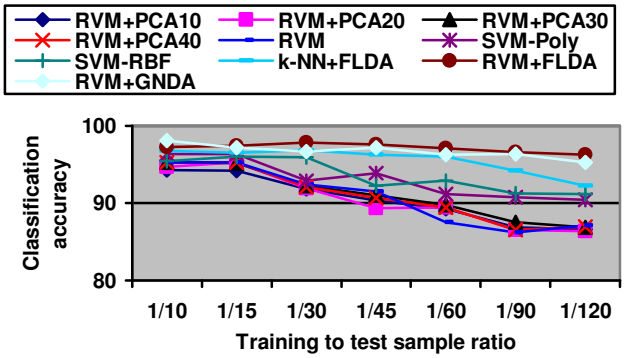


Fig. 12. Classification accuracy versus train-to-test sample ratio for landcover class C8.

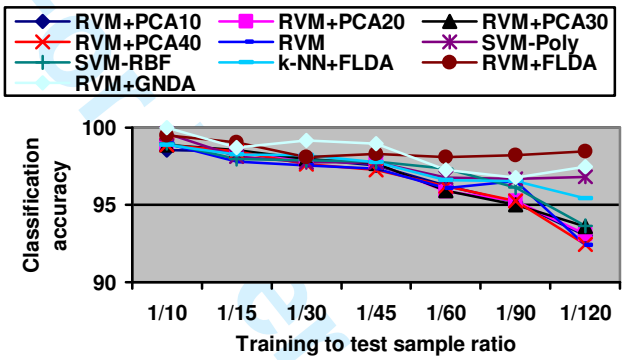


Fig. 13. Classification accuracy versus train-to-test sample ratio for landcover class C9.

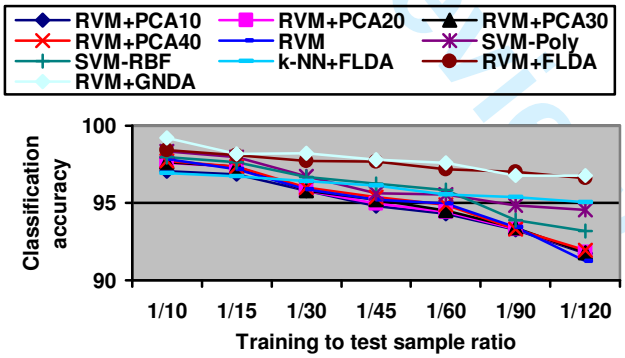


Fig. 14. Overall classification accuracy versus train-to-test sample ratio for 9 selected landcover classes of San Diego hyperspectral image.

Table 1- Number of training and test samples used in the experiments of subsection III.C.

Class Label	Training	Test
C1	235	7050
C2	157	4725
C3	62	1860
C4	162	4860
C5	29	885
C6	58	1755
C7	82	2460
C8	71	2145
C9	131	3930
Total	987	29670

Table 2. single class and overall accuracies, and the computational time achieved through the different approaches. Train-to-test sample ratio: $1/30$.

Method	Classification Accuracy (%)										#SV/ RV	Times (s)	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	OA		Train	Test
RVM+PCA10	99.47	96.72	95.85	94.02	91.52	84.35	97.56	91.85	97.98	95.87	145	105.8	1.6
RVM+PCA20	99.45	96.65	95.80	93.96	91.07	87.92	97.47	91.90	97.91	95.80	135	96.3	1.9
RVM+PCA30	99.47	96.78	95.85	93.80	91.07	88.41	97.47	92.27	97.96	95.80	149	107	2.6
RVM+PCA40	99.47	96.91	96.50	94.37	92.67	87.52	97.76	92.08	97.63	95.98	110	94	2.6
RVM	99.47	96.61	96.66	94.12	92.44	87.18	97.56	92.41	97.57	95.87	114	118.7	8.8
SVM-Poly	99.40	96.93	96.60	95.38	93.93	91.65	98.00	92.93	97.78	96.75	12852	1.9	2.3
SVM-RBF	99.43	97.16	97.84	94.74	91.87	91.36	97.88	95.93	97.88	96.68	21420	1.2	4.4
k-NN+FLDA	**99.49**	96.93	98.22	91.09	94.60	93.85	93.68	96.90	98.21	96.41			529.3
RVM+FLDA	99.35	97.67	99.30	95.85	**98.85**	93.58	98.37	**97.89**	98.10	97.72	125	165.5	1.2
RVM+GNDA	99.15	**98.52**	**99.46**	**97.11**	96.59	**93.97**	**99.89**	96.73	**99.19**	**98.21**	180	227	1.9

Table 3. Single class and overall accuracies, and the computational time achieved through the different approaches. Train-to-test sample ratio: $1/10$.

Method	Classification Accuracy (%)										#SV/ RV	Times (s)	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	OA		Train	Test
RVM+PCA10	99.87	98.43	96.66	95.55	95.42	86.16	98.14	94.33	98.55	97.07	417	1256	3.9
RVM+PCA20	99.89	98.45	97.20	96.47	95.19	90.79	98.28	94.75	98.90	97.59	246	1329	3.1
RVM+PCA30	99.87	98.58	96.77	96.08	95.42	92.91	98.60	95.27	98.90	97.66	421	1236	4.8
RVM+PCA40	99.89	98.45	97.14	96.35	95.19	92.68	98.84	95.27	98.90	97.71	252	1107	4
RVM	99.86	98.37	97.20	96.62	95.99	94.22	98.64	95.32	98.95	97.84	274	1147	10.2
SVM-Poly	99.96	98.33	97.20	97.63	97.59	94.14	98.21	96.35	99.62	98.33	22680	11.2	8.5
SVM-RBF	99.86	98.88	98.38	96.93	97.14	92.45	98.25	95.46	99.00	97.99	41454	2.2	8.3
k-NN+FLDA	**100**	97.86	98.92	93.28	94.39	92.08	94.25	96.81	98.90	96.95			1025
RVM+FLDA	99.73	98.81	99.25	97.18	97.25	94.74	98.82	97.24	99.49	98.41	129	1470	1
RVM+GNDA	99.72	**99.58**	**99.91**	**98.35**	**98.86**	**96.57**	**99.94**	**98.13**	**100**	**99.22**	197	3415	1.7

Table 4. Single class and overall accuracies, and the computational time achieved through the different approaches. Train-to-test sample ratio: $1/60$.

Method	Classification Accuracy (%)										#SV/ RV	Times (s)	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	OA		Train	Test
RVM+PCA10	99.50	95.82	95.14	92.66	86.71	83.80	96.01	89.23	96.25	94.31	109	45.7	1.3
RVM+PCA20	99.50	95.87	93.36	93.28	87.29	85.64	96.74	89.41	95.88	94.43	108	37.3	1.8
RVM+PCA30	99.50	95.84	95.09	92.87	86.71	86.44	96.01	89.74	95.90	94.51	107	36.4	2.1
RVM+PCA40	**99.83**	95.93	95.02	93.36	88.55	85.76	96.29	89.41	96.18	94.89	108	33	2.6
RVM	99.52	96.02	96.77	93.88	86.48	86.61	96.58	87.49	96.07	94.96	109	36.3	8.1
SVM-Poly	99.36	96.01	93.29	94.62	86.83	91.48	97.35	91.15	96.76	95.52	9450	1.3	2.1
SVM-RBF	99.47	96.25	97.09	94.41	88.66	88.12	97.08	92.93	97.37	95.81	14868	1.1	3
k-NN+FLDA	98.75	96.69	98.22	92.00	89.91	90.56	93.27	96.01	96.63	95.54			332
RVM+FLDA	99.29	97.12	98.81	95.60	**95.80**	93.82	96.90	**97.12**	**98.11**	97.21	113	71	1.1
RVM+GNDA	99.57	**98.31**	**99.46**	**96.49**	93.39	**94.16**	**99.19**	96.30	97.28	**97.60**	204	87.2	2.2

Table 5. Single class and overall accuracies, and the computational time achieved through the different approaches. Train-to-test sample ratio: $1/120$.

Method	Classification Accuracy (%)										#SV/ RV	Times (s)	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	OA		Train	Test
RVM+PCA10	99.40	91.70	91.30	90.68	77.21	76.37	91.32	86.60	93.02	91.75	107	27.1	1.4
RVM+PCA20	99.40	91.74	90.24	90.66	77.55	76.54	90.79	86.40	93.08	91.82	125	26.4	1.9
RVM+PCA30	99.40	91.70	90.40	90.81	76.86	75.97	91.40	86.88	93.60	91.76	106	24.9	2.3
RVM+PCA40	99.40	92.29	91.65	90.66	77.55	76.54	90.79	86.98	93.42	91.95	125	27.7	3.1
RVM	99.42	91.79	91.8	91.09	73.77	72.88	91.16	87.12	92.40	91.24	109	25.9	7.8
SVM-Poly	99.46	95.40	92.40	91.43	86.42	87.07	93.72	90.43	96.83	94.55	7056	1.1	1.5
SVM-RBF	99.62	95.04	94.34	91.16	79.61	76.72	93.56	91.16	93.62	93.19	10962	1	2.5
k-NN+FLDA	99.84	95.38	96.93	90.66	89.15	89.81	94.30	92.22	95.40	95.03			300.4
RVM+FLDA	99.42	97.16	98.22	91.51	**95.21**	**91.65**	95.88	**97.56**	**98.48**	96.61	105	41.4	1.2
RVM+GNDA	**99.72**	**98.09**	**99.46**	**93.61**	91.77	91.21	**98.37**	95.33	97.50	**96.78**	164	47.8	1.8