Customizing Kernel Functions for SVM-Based Hyperspectral Image Classification

Baofeng Guo, Steve R. Gunn, R. I. Damper, *Senior Member, IEEE*, and James D. B. Nelson

***Abstract*—Previous research applying kernel methods such as support vector machines (SVMs) to hyperspectral image classification has achieved performance competitive with the best available algorithms. However, few efforts have been made to extend SVMs to cover the specific requirements of hyperspectral image classification, for example, by building tailor-made kernels. Observation of real-life spectral imagery from the AVIRIS hyperspectral sensor shows that the useful information for classification is not equally distributed across bands, which provides potential to enhance the SVM's performance through exploring different kernel functions. Spectrally weighted kernels are, therefore, proposed, and a set of particular weights is chosen by either optimizing an estimate of generalization error or evaluating each band's utility level. To assess the effectiveness of the proposed method, experiments are carried out on the publicly available 92AV3C dataset collected from the 220-dimensional AVIRIS hyperspectral sensor. Results indicate that the method is generally effective in improving performance: spectral weighting based on learning weights by gradient descent is found to be slightly better than an alternative method based on estimating "relevance" between band information and ground truth.**

***Index Terms*—Hyperspectral image processing, mutual information (MI), remote sensing, support vector machines (SVMs).**

I. Introduction

HYPERSPECTRAL sensors simultaneously capture hundreds of narrow and contiguous spectral images from a wide range of the electromagnetic spectrum. For instance, the AVIRIS hyperspectral sensor [1] has 224 spectral bands (or images) ranging from visible light to mid-infrared areas (0.4 to 2.5μm). Such a large number of bands or images implies high-dimensionality data, presenting several significant challenges to image classification [2]–[6]. It is well known that the dimensionality of input space strongly affects performance of many classification methods (e.g., the Hughes phenomenon [7]). This requires the careful design of new algorithms that are able to handle hundreds of such spectral images at the same time minimizing the effects from the "curse of dimensionality." Kernel methods, such as support vector machines (SVMs) [8]–[10], are less sensitive to the data's dimensionality [11] and have already shown superior performance in many machine learning applications. Recently, SVMs have attracted increasing attention in remote-sensed multi/hyperspectral communities [11]–[19]. Previous literature applying SVMs to hyperspectral image classification [12], [14], [16], [17] has shown competitive performance with the best available classification algorithms. However, the full potential of SVMs—such as developing customized kernels to integrate *a priori* domain knowledge—has not been fully explored.

In this paper, spectrally weighted (SW) kernels are proposed to take better advantage of SVM techniques for hyperspectral image classification. We first illustrate a well-known phenomenon in hyperspectral imagery, i.e., the nonuniform distribution of discriminatory information across different spectral bands. Based on the AVIRIS 92AV3C dataset, some examples regarding this application-dependent distribution are given in Fig. 1. To address the characteristic that certain parts of the spectrum will provide a much richer descriptor for classification than other parts, some approaches such as a straightforward feature selection [20], [21] or a block-based approximation to the covariance matrix can be applied. Here, we propose a modification to the kernel functions that can take into account the difference of the relative utility of each spectral band by imposing a series of spectral weights. We subsequently show that the spectral weights of the SW kernels can be chosen by a gradient-descent-based automatic tuning that optimizes the SVMs' generalization error. By analyzing the relationship between the automatic tuning and the "relevance" evaluation of each band (the "relevance" can be seen as an index of importance or utility of a band to classification), we further reveal that the spectral weights can actually be more effectively derived from the mutual information between the spectral bands and the ground-truth reference map. This finding can improve the approach by reducing computational cost and saving training time.

The remainder of this paper is organized as follows. After a brief introduction to the AVIRIS 92AV3C dataset in Section II, we discuss the nonuniform information distribution across spectral bands in Section III. In Section IV, we propose spectrally weighted kernels for hyperspectral image classification, and in Section V, we investigate how to use a bound of the generalization error and mutual information to decide the spectral weights. Experiments are carried out to assess the performance of the proposed method, which are presented in Section VI. Finally, we end this paper with conclusions and a proposal for future work.

II. AVIRIS 92AV3C Dataset

The public AVIRIS 92AV3C hyperspectral dataset has been researched extensively. The dataset is illustrative of the problem

Manuscript received March 12, 2007; revised January 14, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dan Schonfeld.

B. Guo, S. R. Gunn, and R. I. Damper are with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail: bg@ecs.soton.ac.uk; srg@ecs.soton.ac.uk; rid@ecs.soton.ac.uk).

J. D. B. Nelson is with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: jdbn2@eng.cam.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.918955

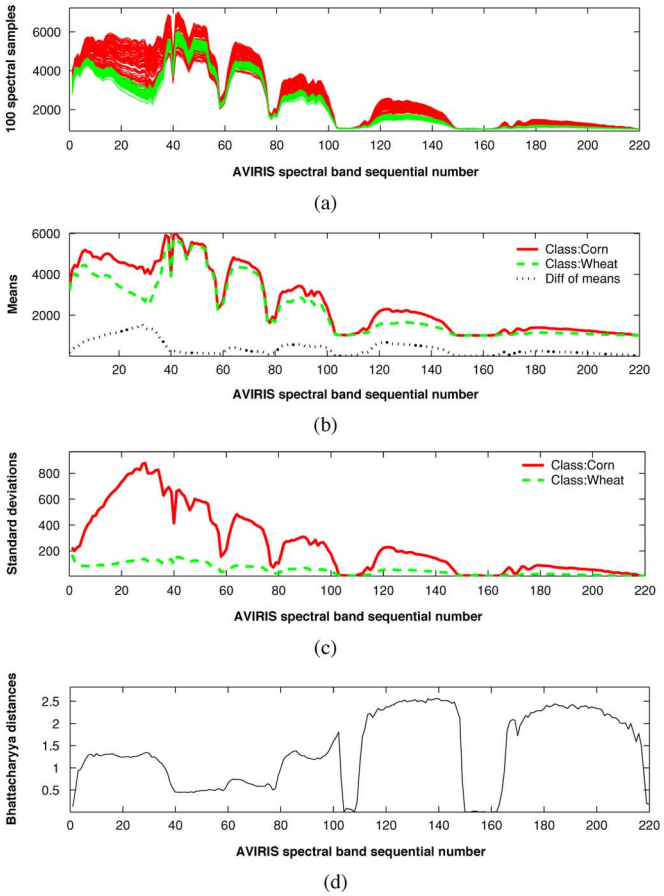


Fig. 1. Nonuniform information distribution. (a) 100 samples of spectral responses for two classes of vegetation in AVIRIS 92AV3C dataset: corn (red in the online version; dark in the print version) and wheat (green in the online version; light in the print version); the statistical features of spectral reflectance values in each spectral band: (b) the means, (c) the standard deviations, and (d) the Bhattacharyya distances between the two classes.

of hyperspectral image analysis to determine land use. It can be downloaded from ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/. Although the AVIRIS sensor collects nominally 224 bands (or images) of data, four of these contain only zeros and so are discarded, leaving 220 bands in the 92AV3C dataset. At certain frequencies, the spectral images are known to be adversely affected by atmospheric water absorption. This affects some 20 bands. Each image is of size 145×145 pixels. The data-cube was collected over a test site called Indian Pine in north-western Indiana [3], [12].

The database is accompanied by a reference map, indicating partial ground truth, whereby pixels are labeled as belonging to one of 16 classes of vegetation or other land types (see examples in Table I). Not all pixels are so labeled (e.g., highway, rail track, etc.), presumably because they correspond to uninteresting regions or were too difficult to label.

III. Nonuniform Information Distribution

Hyperspectral sensors capture signals in a wide spectrum, and it can be expected that different parts of the spectrum will have differing representative capabilities for distinguishing the objects of interest. The intrinsic spectral-distinctness of different objects might not necessarily coincide in the same wavelengths or bands. In some parts of the spectrum, materials may have a much more distinctive spectral reflectance than in other parts of the spectrum. Moreover, complex transmission conditions in the atmosphere, such as water and CO_2 absorption, also play a role in this phenomenon.

TABLE I
Number of Training and Testing Pixels in Each Class

Class	Pixels in training set	Pixels in testing set
A. Corn-notill	287	1147
B. Corn-min	167	667
C. Grass/Trees	149	598
D. Soybeans-notill	194	774
E. Soybeans-min	494	1974
F. Soybean-clean	123	491
G. Woods	259	1035

Fig. 1(a) shows 100 samples (pixels) of spectral reflectance of corn and wheat, extracted from the AVIRIS 92AV3C hyperspectral imagery. The x-axis shows the number of spectral bands (1–220), and the y-axis depicts the pixel value measured in the different bands. It is seen that substantial overlap between the two classes occurs in some bands due to the natural similarity and the variability of spectral reflectance. To separate them, we have to consider their statistical features, such as the means [see Fig. 1(b)] and standard deviations [see Fig. 1(c)] for each spectral band. If we ignore the second- or higher-order statistics [i.e., only using the difference between the two classes's means, see the dashed line in Fig. 1(b)], the two classes appear to be more separable in the bands 15–35, 80–100, and 120–140 than other heavily overlapped bands such as bands 40–80. Only considering the means of course implies a loss of information, so a better measurement of statistical separability is given by the Bhattacharyya distance, which takes account of the second-order statistic, i.e., variance. The Bhattacharyya distances between the two classes in each spectral band are presented in Fig. 1(d), where the bands 110–150, 165–215 are revised as the higher-value ones due to their lower variances. In the following discussion of customized kernels, it is not necessary to evaluate the separability of a group of bands, so the covariances among bands are not calculated.

Fig. 1 clearly shows that irrespective of using the simplest statistics [see Fig. 1(b) for the mean and Fig. 1(c) for the standard deviation] or the Bhattacharyya separability measure [see Fig. 1(d)], their values vary across bands. This indicates that in hyperspectral imagery the discriminatory information is nonuniformly distributed across the spectrum. Among the set of spectral bands, some may contain more useful information for classification than others, and have larger separability indexes accordingly. Considering that the separability measure gives an estimate of the probability of correct classification, it would be expected that classification performance can benefit from placing greater emphasis on the more informative bands.

Hence, two different strategies may be considered:

- to select effective spectral bands with spectral management algorithms, such as feature selection by a filtering approach;

- to customize directly the classifier by integrating this *a priori* knowledge (i.e., feature selection by an "embedding" approach).

The first strategy has been discussed in [20] and [21]; however, in this research, we focus on the second one. In the SVM-based classification framework, a straightforward approach is to modify the kernel functions by assigning different weights to different bands, adaptively embedding the amount of useful classification-information contained within that band. To this end, we propose the use of spectrally weighted kernels to exploit better this specific characteristic of hyperspectral imagery.

IV. Spectrally Weighted Kernels

To present the proposed spectrally weighted kernels, we first introduce several relevant SVM formulas. A full introduction to SVMs can be found in [9] and [10].

Let $\mathbf{x}_i$ (and $\mathbf{x}$) be an N-dimensional hyperspectral data vector (in this research, this vector can be seen as a pixel with 220 components) with subscript i denoting the example number. The SVM classifier can be represented as

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{M} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{1}$$

The $y_i \in \{+1, -1\}$ are the classification targets (labels); $\alpha_1, \alpha_2, \ldots, \alpha_M$ are Lagrange multipliers; M is the number of examples; and b is a threshold. Furthermore, $K(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$ is an appropriate kernel function which has a corresponding inner product expansion, Φ. Commonly used functions are polynomials and Gaussian radial basis functions (RBFs), as follows:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^d \tag{2}$$

$$K(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right\} \tag{3}$$

where d is the order of polynomials and σ is a width parameter characterizing the RBFs.

For a hyperspectral data vector $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^N)$, the component x_i^j corresponds to the reflectance value of example i in the specific spectral band $j, 1 \le j \le N$. Generic kernels, e.g., (2) and (3), regard each component x_i^j with equal emphasis in their projection into feature space. However, Section III has argued that it is advantageous to moderate the spectral information according to the richness of the descriptor. For example if the component x_i^j is a reflectance value in the spectrum or bands where two classes can be clearly separated (such as the regions with the higher Bhattacharyya distance in Fig. 1), weighting this feature to have a larger effect in feature space could improve classification and, similarly, reducing it when it adds little to the description.

To modify the kernel function so as to reflect the above consideration, a weight vector $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ corresponding to each spectral band is used to scale each feature x_i^j in the hyperspectral data vector before mapping it into feature space. To simplify notation, we introduce a diagonal matrix $\mathbf{S} = \operatorname{diag}(\mathbf{s})$. Given this weighting, the SW kernels for a polynomial and an RBF can be written as

$$K_{\mathrm{sw}}(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{S}^T \mathbf{S} \mathbf{x}' + 1)^d \tag{4}$$

$$K_{\mathrm{sw}}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{\|\mathbf{S}(\mathbf{x} - \mathbf{x}')\|^2}{2\sigma^2}\right\} \tag{5}$$

where $\mathbf{S}^T\mathbf{S} = \operatorname{diag}(s_1^2, s_2^2, \ldots, s_N^2)$. In this scheme, the weights have been designed to correspond to each feature component and a simple diagonal matrix can achieve this goal.

The necessary and sufficient condition for deciding whether a function is a kernel is given by Mercer's theorem. It is easy to prove that the SW kernels still satisfy Mercer's condition, since they can be also interpreted as a scaling procedure in input space, and will not change the kernels' Mercer condition. Substituting the SW kernels into (1) gives the corresponding spectrally weighted SVMs.

Using the proposed SW kernels, *a priori* knowledge (e.g., the nonuniform information distribution) can be incorporated into the SVM learning procedure. As the weights can be considered as the part of the SVMs' model parameters, through tuning to maximize the estimate of generalization error, it should be possible to achieve better performance. SW kernels can also be seen as a form of data preprocessing, akin to feature selection. As the approach can de-emphasize less important features, it implicitly conducts feature selection. When the weights are zero, the corresponding features will be cut off equivalently [see the distance calculation in Gaussian (5) and polynomial kernels (4)]. The above procedure changes the measurement complexity of the classifiers; then, according to [7] (i.e., the Hughes phenomenon), it may affect classification accuracy for over-dimensional data, given that the number of training samples is finite (as in the case of remote-sensing applications).

V. Estimation of Spectral Weights

For SW kernels, a key problem is to choose the spectral weights s_i. On the one hand, SW kernels are motivated by the nonuniform information distribution across bands, so the spectral weights are expected to reflect the relative influence (namely the relevance to classification) of each band to the the kernel values; on the other hand, the change of the kernel value by imposing such weights should improve classification accuracy, i.e., minimizing the error $(f(\mathbf{x}) \neq y)$ on future examples. It is known that the latter goal can be achieved by optimizing a bound of generalization error $P_e = P(f(\mathbf{x}) \neq y)$. One of the well-known upper bounds of P_e is the ratio of radius to margin, $R^2/(l \cdot \gamma^2)$, where R is the radius of a sphere enclosing l mapped training examples $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_l)$, and γ is the margin between the hyperplane and the closest $\Phi(\mathbf{x}_i)$ [22], [23]. This bound may be intuitively understood as follows: the radius R indicates the compactness of data, and the margin γ implies the distance of two classes. It is similar to the Bhattacharyya separability measure, which is calculated by means (akin to the distance of two classes) and variances (akin to the compactness of data). This theorem also justifies the idea of maximizing the SVM margin $\|\mathbf{w}\|^{-1}$, or equivalently minimizing $1/(2\|\mathbf{w}\|^2)$. The factor regarding the radius may

be implicitly implemented by the choice of kernels and their parameters.

A. Scheme Based on Gradient Descent Algorithm

Based on the above discussion, if the spectral weights can be chosen to increase the SVM margin, i.e., to lower the radius-margin bound, it becomes possible to improve the classification accuracy. Thus, the first scheme for choosing the spectral weights is proposed as follows [22]–[25].

According to SVM theory, the margin γ can be derived as

$$\gamma = \frac{1}{\|\mathbf{w}\|} \tag{6}$$

where $\mathbf{w}$ is the n-dimensional vector perpendicular to the separating hyperplane, given by

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i). \tag{7}$$

From (6), maximizing γ^2 can be achieved by minimizing $\|\mathbf{w}\|^2$, or

$$\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K_{\mathrm{sw}}(\mathbf{x}_i, \mathbf{x}_j). \tag{8}$$

From (8), it is seen that given fixed α_i^0, the derivative of $\|\mathbf{w}\|^2$ is

$$\frac{\partial \|\mathbf{w}\|^2}{\partial s_p} = \sum_{i,j=1}^{l} \alpha_i^0 \alpha_j^0 y_i y_j \frac{\partial K_{\mathrm{sw}}(\mathbf{x}_i, \mathbf{x}_j)}{\partial s_p} \tag{9}$$

where s_p is a spectral weight corresponding to the pth component of a hyperspectral data vector (i.e., the pth spectral image).

Thus, the choice of spectral weights can be implemented by using a gradient descent algorithm as follows:

$$s_p(n+1) = s_p(n) - \epsilon \frac{\partial \|\mathbf{w}\|^2}{\partial s_p} \tag{10}$$

where ϵ controls the searching speed, and n is the iteration step. In this scheme, weights s_i are updated step-by-step, and a (local) minimum of $\|\mathbf{w}\|^2$ will be found after a number of iterations.

B. Scheme Based on Relevance Evaluation

The weighting scheme shown in Section V-A is based on the gradient descent algorithm, which needs a time-consuming iterative updating [see (10)], and usually it will not find the global optimal solution. On the other hand, we know that the spectral weights also reflect the relevance of each spectral band to the classification. Thus, an alternative weighting approach can be conceived as follows. First, the kernel in (5) can be rewritten as

$$K_{\mathrm{sw}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\sum_p \frac{s_p^2 \left(x_i^p - x_j^p\right)^2}{2\sigma^2}\right\} \tag{11}$$

where i and j denote the number of training examples, and $p = 1, 2, \ldots, N$ is the band number. The derivative of $K_{\mathrm{sw}}(\mathbf{x}_i, \mathbf{x}_j)$ is

$$\frac{\partial K_{\mathrm{sw}}(\mathbf{x}_i, \mathbf{x}_j)}{\partial s_p} = -\exp\left\{-\sum_p \frac{s_p^2 \left(x_i^p - x_j^p\right)^2}{2\sigma^2}\right\} \times \frac{s_p \left(x_i^p - x_j^p\right)^2}{\sigma^2}. \tag{12}$$

It can be shown that the value of (12) for different bands p is decided by the term $(x_i^p - x_j^p)^2$, given the same initialization weights s_p.

Combining (9) and (12), we get

$$\frac{\partial \|\mathbf{w}\|^2}{\partial s_p} = -\sum_{i,j=1}^{l} c_{i,j} y_i y_j s_p \left(x_i^p - x_j^p\right)^2 \tag{13}$$

where $c_{i,j} = \exp\{-\sum_p (s_p^2 (x_i^p - x_j^p)^2)/(2\sigma^2)\}(\alpha_i^0 \alpha_j^0)/(\sigma^2) \geq 0$, including all coefficients except p and y.

When the examples belong to the same class, i.e., $y_i = y_j$ and $y_i \cdot y_j = 1$, $(\partial \|\mathbf{w}\|^2)/(\partial s_p)$ will change little because $E[(x_i^p - x_j^p)^2] \to 0$. When the examples belong to different classes, i.e., $y_i \neq y_j$ and $y_i \cdot y_j = -1$, we get

$$\frac{\partial \|\mathbf{w}\|^2}{\partial s_p} \propto \sum_{i,j=1}^{l} \left(x_i^p - x_j^p\right)^2. \tag{14}$$

The expectation of variable $(x_i^p - x_j^p)^2$ is

$$E\left[\left(x_i^p - x_j^p\right)^2\right] = E\left[\left(x_i^p\right)^2\right] + E\left[\left(x_j^p\right)^2\right] - 2E\left[x_i^p x_j^p\right]. \tag{15}$$

According to the Bhattacharyya distance, given the same difference of means, the bands with the smaller variances will tend to have higher separability values (i.e., they are "good" bands, probably with lower classification errors). This assumption can be evidenced by observing the following 1-D Bhattacharyya coefficient:

$$b = \frac{1}{8}(\mu_1 - \mu_2)\left[\frac{\sigma_1 + \sigma_2}{2}\right]^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\left(\frac{\frac{\sigma_1+\sigma_2}{2}}{\sqrt{\sigma_1 \sigma_2}}\right). \tag{16}$$

So, roughly speaking, given a certain assumption (e.g., the same difference of means), the "good" bands should have lower variances, i.e., the smaller $E[(x_i^p)^2]$ and $E[(x_j^p)^2]$. Thus, we have

$$E\left[\left(x_i^p - x_j^p\right)^2\right] \propto -\lambda_p \tag{17}$$

where λ is a measure of the level of "relevance" or "goodness" of a band to classification. Combining (14) and (17), we get

$$\frac{\partial \|\mathbf{w}\|^2}{\partial s_p} \propto -\lambda_p. \tag{18}$$

Equation (18) shows that in each step of automatic tuning, the change of the spectral weight s_p is related to the level of its relevance to classification, i.e., λ_p. Therefore, to obtain a lower bound of generalization error, the spectral weights can also be found through the evaluation of relevance λ, as an alternative to the gradient descent approach discussed in Section V-A. Considering that gradient descent is usually time-consuming, a weighting scheme based on relevance evaluation is attractive. Among many possible relevance measures, we propose the use of mutual information (MI) to estimate each band's level of relevance. The advantages of employing MI are its close relation to Bayes classification errors [26] and effective implementation [27], [28].

Given two random variables A and B with marginal probability distributions $p(a)$ and $p(b)$, and joint probability distribution $p(a,b), a \in A, b \in B$, the mutual information between A and B is defined as

$$I(A,B) = \sum_{a\in A}\sum_{b\in B} p(a,b)\log\frac{p(a,b)}{p(a)p(b)}. \tag{19}$$

According to Shannon's information theory, entropy measures information content in terms of uncertainty, and is defined by

$$H(A) = -\sum_{a\in A} p(a)\log p(a). \tag{20}$$

From (19) and (20), it is not difficult to find that mutual information is related to entropies by the following equations:

$$\begin{aligned} I(A,B) &= H(A) + H(B) - H(A,B) \\ &= H(A) - H(A|B) \\ &= H(B) - H(B|A) \end{aligned} \tag{21}$$

where $H(A)$ and $H(B)$ are the entropies of A and B; $H(A,B)$ is their joint entropy; and $H(A|B)$ and $H(B|A)$ are the conditional entropies of A given B and of B given A, respectively. The joint and conditional entropies can be written as

$$H(A,B) = -\sum_{a\in A}\sum_{b\in B} p(a,b)\log p(a,b) \tag{22}$$

$$H(A|B) = -\sum_{a\in A}\sum_{b\in B} p(a|b)\log p(a|b). \tag{23}$$

Treating the spectral images and the corresponding reference map as random variables, MI can be used to estimate the dependency or relevance between them. In detail, we can treat each spectral band's pixels as samples of the random variable A with possible continuous reflectance values $a \in \mathbb{R}$, and its class category as variable B with discrete vegetation labels $b \in \{\omega_1, \omega_2, \ldots, \omega_n\}$. Thus, MI between A and B can be evaluated as follows:

$$\begin{aligned} I(A,B) = &-\int_a p(a)\log p(a)\mathrm{d}a \\ &-\sum_b P(b)\log P(b) \\ &+\sum_b \int_a p(a,b)\log p(a,b)\mathrm{d}a. \end{aligned}$$

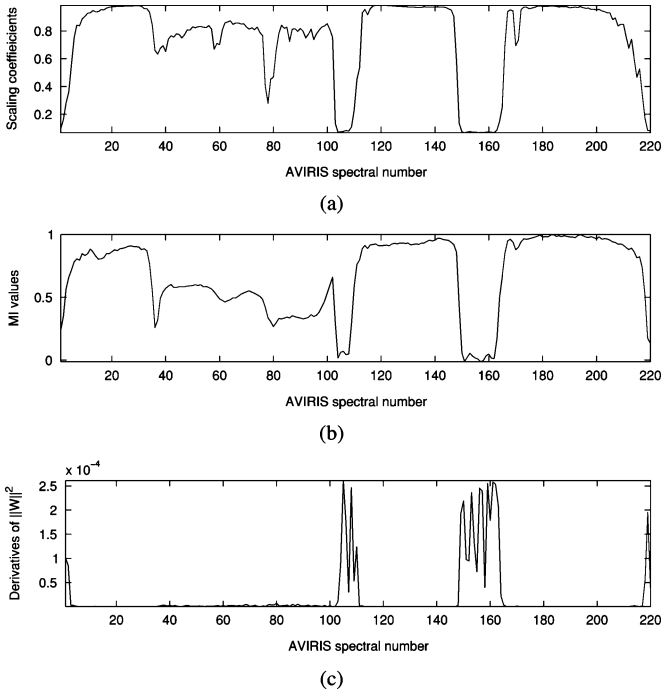


Fig. 2. Comparison of the spectral weights obtained by the two schemes: (a) the weights found by the gradient-descent tuning; (b) the weights decided by mutual information; (c) the derivatives of $\|\mathbf{w}\|^2$ in one of the iterations.

Since the reference map implicitly defines the required classification result, MI measures the relevance of each spectral band to the classification objective. Using (21), the mutual information between each of the 220 spectral images (or bands) and the corresponding reference map accompanying the 92AV3C dataset was calculated as shown in Fig. 2(b).

By comparing this MI curve to the examples of AVIRIS images, we may verify the agreement between the relevance level and the MI values. It has been found [21] that the bands most similar to the reference map (i.e., with higher relevance) are those having higher values of MI. The MI curve also reveals clearly the effect of atmospheric water absorption, giving the lowest MI values in bands 104–108 and 150–163 at precisely those frequencies where absorption occurs [21]. In this particular example, it is seen that the MI of a spectral band with respect to the reference map is consistent with visual impressions regarding the relevance or relative importance of each spectral band to classification. Moreover, it can be seen that the overall shapes of the MI curve and the Bhattacharyya distance [Fig. 1(d)] are very similar, indicating an agreement of MI with another commonly used (but more computationally expensive) separability measure.

Fig. 2 further compares the spectral weights obtained by the above two schemes; Fig. 2(a) shows the weights calculated by the gradient-descent tuning, and Fig. 2(b) is the result based on mutual information estimation. It is seen that the two sets of weights have very similar overall shape across different bands, suggesting the comparable effect of the two schemes in band-utility evaluation. Moreover, it is found that the derivatives of $\|\mathbf{w}\|^2$ are inversely proportional to the MIs, in general. Fig. 2(c) shows an example of a group of derivatives obtained in one of the iterations of (10). Recall our previous

discussion in (11)–(18): In the gradient descent algorithm, the weights are usually set up as the same coefficients over the whole bands in the initialization step, and will be gradually updated by subtracting the derivatives. Given the relation between the derivatives and the mutual information (i.e., the inverse proportion), the weights will finally converge to a result similar to the mutual information. Thus, Fig. 2 suggests that MI can effectively encode the relevance of spectral bands to classification and be employed as a spectral weight.

VI. Experiments

Following the two weighting schemes discussed in Section V, SVMs based on SW kernels are implemented to test the proposed method. The performance of the proposed method is compared with a standard SVM with no spectral weighting of the kernel as adopted in [12], [15], [17], etc., on the AVIRIS 92AV3C dataset.

In the AVIRIS 92AV3C dataset, the seven most numerous classes are chosen as the testing objects, accounting for 80.64% of all 16-class pixels. The class labels from the reference map accompanying the dataset were utilized for supervised training. Among the labeled pixels, 20% of them from each class were randomly chosen as the training set, with the remaining 80% forming the test set on which performance was assessed (see Table I). This was repeated five times to allow an estimate of the error in this sampling process. The performance measurement adopted assessed the classification error of the proposed method on a held-out set.

In the experiments, the seven classes are named as A to E, respectively (see Table I). Since SVMs are inherently binary (two-class) classifiers, it is more straightforward to evaluate performance based on each class pair. Thus, $\binom{7}{2} = 21$ classifiers were constructed based on each class-pair, named as $C_{A|B}, C_{A|C}, \ldots, C_{F|G}$, respectively. Moreover, it is more effective to learn the weights by using the two-class SVMs since the different class-pairs may have different spectral characteristics. So, in this case, the pixels from other classes will not affect the classification associated with the classifier that was not trained from those examples. The kernel function used here is the Gaussian RBF [see (5)]. The kernel parameter σ and the penalty parameter C were tested between 10^{-3} and 10^4 by a validation procedure using the training data and 0.4 and 60, respectively, were chosen as suitable values.

Fig. 3 shows the variation of classification accuracy as a function of iteration step in the weight-learning scheme discussed in Section V-A. It is seen that with each round of weight updating, the classification error changes accordingly, because of the gradual optimization of the generalization error bound. Using customized kernels, the classifier tuning or data preprocessing (e.g., scaling) required for high-dimensionality data is incorporated into the SVM learning procedure. Thus, by this embedding approach, the weights can not only de-emphasize features, but also work as the part of the SVMs' model parameters. By tuning these model parameters to maximize the estimate of the generalization error, it becomes possible to achieve a better performance. The results from four 92AV3C-based binary trials in Fig. 3 show the reduced classification errors, providing empirical evidence to support the above argument. In practical applications, the accuracy verification illustrated in Fig. 3 is not necessary, and lower classification errors can be obtained by adjusting the threshold in the gradient descent algorithm.

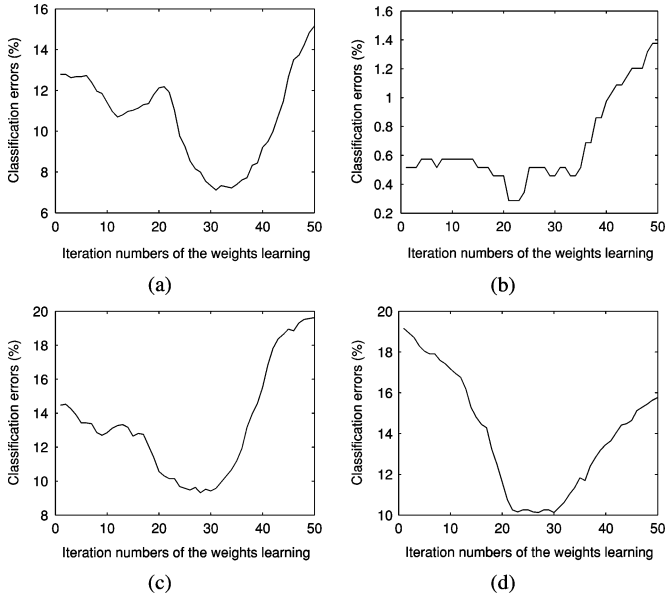


Fig. 3. Classification accuracy as a function of iteration number in the gradient descent learning; for classifiers (a) $C_{A|B}$ (Corn-notill versus Corn-min), (b)$C_{A|C}$ (Corn-notill versus Grass/Trees), (c) $C_{A|D}$ (Corn-notill versus Soybeans-notill), and (d) $C_{A|E}$ (Corn-notill versus Soybeans-min), respectively.

Results using the two weighting schemes are shown in Table II. From Table II, it can be seen that the methods based on SW kernels outperformed the unweighted method in the majority of the 21 classifiers. The improvement is especially significant when the two classes are difficult to differentiate (i.e., their classification errors are relatively higher). For example, the vegetation classes Corn-notill (A), Corn-min (B), Soybeans-notill (D), Soybeans-min (E), and Soybean-clean (F) are the most similar classes in the scene, and their classification errors are usually larger than 10% (see the numbers listed in the rows $C_{A|B}, C_{A|D}, C_{A|E}$, and $C_{A|F}$ of Table II). From the point of view of "spectral signature," it can be expected that these confusable vegetation classes will show considerable similarity in their spectral reflectances at a global level. Correspondingly, the spectral difference between them will only appear in some particular wavelengths, and the bands therein will dominate the overall discriminatory capability. Apparently, in this case the effect caused by the nonuniform information distribution is quite substantial and the SW kernels, tailor-made for this scenario, become successful.

Table II also shows that there are several exceptions, e.g., the classifiers $C_{A|C}, C_{C|F}$, and $C_{C|G}$, where the weighting schemes did not successfully reduce the classification errors. Almost all of these exceptions belong to the scenario where two classes are relatively easier to separate. For example, $C_{A|C}, C_{C|F}$, and $C_{C|G}$ correspond to classification of the vegetation pairs Corn-notill versus Grass/Trees, Soybean-clean versus Grass/Trees, and Grass/Trees versus Woods, respectively. Compared to the similar crops mentioned previously, these vegetation classes are very distinct. Therefore, it can be

TABLE II
COMPARISON OF CLASSIFICATION ERROR (%) ± SAMPLING ERROR (STD); RBF KERNEL, $\sigma = 0.4, C = 60$

Classifier	Without weighting	Weights based on gradient descent	Weights based on mutual information
$C_{A\|B}$, (Corn-notill vs. Corn-min)	14.95±1.69	11.21±0.70	8.22±1.13
$C_{A\|C}$, (Corn-notill vs. Grass/Trees)	0.42±0.05	0.57±0.19	0.55±0.07
$C_{A\|D}$, (Corn-notill vs. Soybeans-notill)	15.56±0.70	12.32±1.03	13.05±0.61
$C_{A\|E}$, (Corn-notill vs. Soybeans-min)	19.45±0.62	11.46±0.79	15.54±0.51
$C_{A\|F}$, (Corn-notill vs. Soybean-clean)	17.83±0.54	11.99±1.88	10.98±0.71
$C_{A\|G}$, (Corn-notill vs. Woods)	0.10±0.04	0.10±0.05	0.09±0.05
$C_{B\|C}$, (Corn-min vs. Grass/Trees)	0.30±0.13	0.22±0.06	0.19±0.04
$C_{B\|D}$, (Corn-min vs. Soybeans-notill)	5.02±0.40	4.68±0.50	3.21±0.28
$C_{B\|E}$, (Corn-min vs. Soybeans-min)	12.06±0.52	11.43±1.88	12.35±1.05
$C_{B\|F}$, (Corn-min vs. Soybean-clean)	19.15±1.58	14.44±2.05	14.43±1.45
$C_{B\|G}$, (Corn-min vs. Woods)	0.04±0.03	0.06±0.00	0.01±0.03
$C_{C\|D}$, (Grass/Trees vs. Soybeans-notill)	0.67±0.23	0.58±0.14	0.55±0.13
$C_{C\|E}$, (Grass/Trees vs. Soybeans-min)	1.19±0.22	1.07±0.39	1.05±0.20
$C_{C\|F}$, (Grass/Trees vs. Soybean-clean)	0.48±0.20	0.50±0.05	0.53±0.12
$C_{C\|G}$, (Grass/Trees vs. Woods)	1.25±0.20	1.41±0.33	1.32±0.15
$C_{D\|E}$, (Soybeans-notill vs. Soybeans-min)	14.43±0.49	12.87±0.68	11.81±1.17
$C_{D\|F}$, (Soybeans-notill vs. Soybean-clean)	10.72±0.21	8.87±1.16	9.11±0.63
$C_{D\|G}$, (Soybeans-notill vs. Woods)	0.03±0.03	0.00±0.00	0.00±0.00
$C_{E\|F}$, (Soybeans-min vs. Corn-min)	9.32±0.41	7.89±1.24	10.95±1.19
$C_{E\|G}$, (Soybeans-min vs. Woods)	0.25±0.07	0.39±0.11	0.23±0.10
$C_{F\|G}$, (Soybean-clean vs. Woods)	0.00±0.00	0.00±0.00	0.00±0.00

expected that significant spectral differences will appear across a wide range of the spectrum, overshadowing the necessity to emphasize a particular subset of bands as the SW kernels are designed to do. As a result, the proposed scheme may lose its efficacy, but also carries the risk of over-fitting incurred by introducing extra parameters. Although no improvement has been made in these small number of exceptions, the overall effectiveness of this method in the majority of classes is still encouraging.

Comparing the two weighting schemes, the gradient-based approach appears to be slightly better than that based on mutual information. This is understandable because the former uses different sets of weights for different classifiers, which are individually optimized in each SVM's learning procedure. On the contrary, the latter scheme uses a single set of weights, i.e., the MI values, for all 21 classifiers. However, the MI-based scheme is still a useful alternative as the calculation of MI is much faster than using gradient descent.

VII. CONCLUSION

In this paper, we have proposed an extension to the SVM-based method for hyperspectral image classification using spectrally weighted (SW) kernels. This extension is motivated by the observation that the useful information for classification is not evenly distributed among each spectral band. Within the SVM framework, SW kernels can be conveniently constructed by highlighting the informative bands in the kernel mapping. We have shown that it is possible to improve the upper bound of classification error through learning the spectral weights in the customized kernels. Further research revealed that the mutual information between a spectral band and the ground truth can also be used to design spectral weights, resulting in a significant saving of computational cost. Experimental results showed that, at least for the limited binary trials based on the AVIRIS 92AV3C dataset, the classification performance can be improved to some extent by a spectral customization of the kernels using either a gradient-descent tuning or a mutual information criterion. Further work could explore the possibility of avoiding the over-fitting incurred by the multiple adjustable parameters and testing the algorithms on other labeled multiclass datasets.

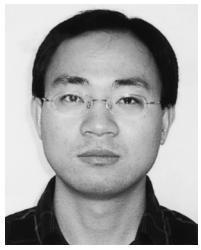
ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments which have resulted in a number of improvements in the paper.

REFERENCES

[1] *Airborne Visible/Infrared Imaging Spectrometer*, AVIRIS. [Online]. Available: http://aviris.jpl.nasa.gov/
[2] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.
[3] D. Landgrebe, "On information extraction principles for hyperspectral data: A white paper," Tech. Rep., School Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, 1997 [Online]. Available: http://dynamo.ecn.purdue.edu/landgreb/whitepaper.pdf
[4] X. Yu, L. Hoff, I. Reed, A. Chen, and L. Stotts, "Automatic target detection and recognition in multiband imagery: A unified ml detection and estimation approach," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 143–156, Jan. 1997.
[5] S. M. Schweizer and J. M. F. Moura, "Efficient detection in hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 584–584, Apr. 2001.
[6] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 12–16, Jan. 2002.
[7] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop on Computational Learning Theory*, Pittsburgh, PA, 1992, pp. 144–152.
[9] C. Cortes and V. N. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 1–25, 1995.
[10] C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowl. Disc. Data Mining*, vol. 2, no. 2, pp. 121–167, 1998.
[11] 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.

[12] J. Gualtieri and R. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. 27th AIPR Workshop Advances in Computer Assisted Recognition*, Washington, DC, 1998, pp. 121–132.
[13] M. Brown, H. G. Lewis, and S. Gunn, "Linear spectral mixture models and support vector machines for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2346–2360, May 1999.
[14] F. Roli and G. Fumera, S. B. Serpico, Ed., "Support vector machines for remote-sensing image classification," in *Proc. SPIE Image and Signal Processing for Remote Sensing VI*, 2001, vol. 4170, pp. 160–166.
[15] C. Huang, L. S. Davis, and J. R. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, Feb. 2002.
[16] M. Lennon, G. Mercier, and L. Hubert-Moy, "Classification of hyperspectral images with nonlinear filtering and support vector machines," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp.*, Toronto, ON, Canada, Jun. 24–28, 2002, vol. 3, pp. 1670–1672.
[17] C. A. Shah, P. Watanachaturaporn, M. K. Arora, and P. K. Varshney, "Some recent results on hyperspectral image classification," in *Proc. IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Greenbelt, MD, 2003, pp. 346–353.
[18] G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J. D. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.
[19] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[20] S. 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, Jul. 2001.
[21] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 3, no. 4, pp. 522–526, Apr. 2006.
[22] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 131–159, Jan. 2002.
[23] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Proceedings of Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 2000 [Online]. Available: http://books.nips.cc/papers/files/nips13/
[24] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in SVMs," in *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 553–560.
[25] N. Cristianini, C. Campbell, and J. Shawe-Taylor, "Dynamically adapting kernels in support vector machines," in *Proceedings of Advances in Neural Information Processing Systems 11*. Cambridge, MA: MIT Press, 1999, pp. 204–210.
[26] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1415–1438, 2003.
[27] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[28] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Incorporating connected region labelling into automated registration using mutual information," in *Proc. Workshop on Mathematical Methods in Biomedical Image Analysis*, 1996, pp. 23–31.



Baofeng Guo was born in Xi'an, China, in 1973, and received the B.Eng. degree in electronic engineering and the M.Eng. degree in signal processing from Xidian University, Xi'an, in 1995 and 1998, respectively, and the Ph.D. degree in signal processing from the Chinese Academy of Sciences, Beijing, in 2001, respectively.

From 2002 to 2004, he was a Research Assistant in the Department of Computer Science, University of Bristol, U.K. Since 2004, he has been Research Fellow with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. His current research interests are pattern recognition, machine learning, and image processing.

Steve R. Gunn was born in Bristol, U.K., in 1970. He received the B.Eng. degree (first class honors) in electronic engineering and the Ph.D. degree in computer vision from the University of Southampton, Southampton, U.K., in 1992 and 1996, respectively.

He is currently a Professor in the Information: Signals, Images and Systems Research Group, School of Electronics and Computer Science, University of Southampton. He has published over 80 research papers in the areas of computer vision and machine learning, and he has also published the book *Feature Extraction: Foundations and Applications* (Springer, 2006). His current research interests are in the area of sparse representations, feature selection, and subspace methods for identification of salient parts of the data space for prediction.

Prof. Gunn serves on various program committees and is the operational coordinator for the EU PASCAL Network of Excellence.

R. I. Damper (M'87–SM'89) was born in Tunbridge Wells, U.K., in 1948. He received the M.Sc. degree in biophysics and the Ph.D. degree in electrical engineering from the University of London, London, U.K., in 1973 and 1979, respectively, and the Diploma in electrical engineering from Imperial College, London.

He was appointed Lecturer in electrical engineering at the University of Abertay Dundee, U.K., in 1976, Lecturer in electronics at the University of Southampton, Southampton, U.K., in 1980, Senior Lecturer in electronics and computer science in 1989, Reader in 1998, and Professor in 2003. He has wide research interests including speech science and technology, neural computing, cognitive modeling, pattern recognition, and intelligent systems engineering. He has published approximately 300 research articles and authored the undergraduate text *Introduction to Discrete-Time Signals and Systems*.

Dr. Damper is a Chartered Engineer and a Fellow of the U.K. Institution of Engineering and Technology, a Chartered Physicist, a Fellow of the U.K. Institute of Physics, and an Honorary (Foreign) Member of the Yugoslav Engineering Academy.

James D. B. Nelson was born in York, U.K., in 1975. He received the B.Sc. degree in mathematics and the Ph.D. degree for his work on the application of Riesz bases to signal theory from Anglia Polytechnic University, U.K.

He has held research posts at Cranfield University, U.K. (three years), and the University of Southampton, Southampton, U.K. (two years), and is currently a Research Associate at the University of Cambridge, Cambridge, U.K. His research interests include signal and image processing, wavelets, and support vector machines.