IMPROVED HYPERSPECTRAL IMAGE CLASSIFICATION WITH NOISE REDUCTION PRE-PROCESS

Begüm Demir, and Sarp Ertürk

Electronics and Telecom. Eng. Dept., University of Kocaeli, 41040, Kocaeli, TR.
phone: + 902623352383, fax: + 902623352812, email: begum.demir@.kou.edu.tr, sertur@kou.edu.tr
web: http://kulis.kou.edu.tr/default_eng.asp

ABSTRACT

This paper shows that hyperspectral image classification performance using support vector machines (SVM) and relevance vector machines (RVM) can significantly be improved using a noise reduction pre-process. A wavelet domain, spatially adaptive denoising method that estimates the probabiliy that a coefficient represents a significant noise-free component is used for denoising of hyperspectral images before classification. It is shown that support vector machine and relevance vector machine classification of denoised hyperspectral images gives significantly better classification accuracy and furthermore improves sparsity.

1. GENERAL INFORMATION

Hyperspectral imaging sensors acquire data corresponding to hundreds of continuous narrow spectral bands. Therefore it becomes possible to classify regions within the scene or identify materials/objects with much higher accuracy compared to standard vision sensors. Hyperspectral image segmentation and classification approaches can basically be categorized into supervised and unsupervised methods. K-means [1] and phase correlation [2] based segmentation approaches are examples for unsupervised techniques. Vector machine based classification of hyperspectral images falls into the category of supervised techniques. Support vector machine (SVM) based approaches [3] have recently been proposed for regression and classification tasks in multispectral [4] and hyperspectral [5] images. RVM based regression and classification has been proposed in [6], and applied to hyperspectral classification in [7]. Vector machine based classification approaches stand out in that they can provide good classification performance even for objects with close spectral characteristics.

Image denoising algorithms typically provide a trade off between noise containment and preservation of actual image discontinuities and look for solutions to detect important image details. A quite recent and efficient denoising technique is wavelet thresholding (shrinkage). In this case, noise reduction is obtained from shrinking the noisy coefficient magnitudes in the wavelet domain. While noisy wavelet coefficients are reduced to insignificant values, noise free coefficients are reduced considerably less (soft-thresholding) or kept unchanged (hard-thresholding) [8].

Several techniques have been proposed to denoise hyperspectral data. In [9] a discrete Fourier transform (DFT) and wavelet based estimation scheme has been presented for the denoising of hyperspectral images. The correlation between bands has been used to denoise hyperspectral images in [10] by enforcing simultaneous sparsity on their wavelet representations.

This paper evaluates the performance of hyperspectral image classification after a denoising pre-process. For this purpose, it is proposed to utilize RVM and SVM for classification of hyperspectral images after wavelet denoising. A spatially adaptive wavelet denoising algorithm presented in [11] which is shown to provide good denoising performance is used in the denoising step for each image band of the hyperspectral image. It is shown that RVM and SVM based classification of denoised hyperspectral data can provide significantly higher classification accuracy compared with direct SVM and RVM based classification, furthermore with a reduced number of support vectors and relevance vectors.

2. VECTOR MACHINE BASED CLASSIFICATION

Supervised learning techniques make use of a training set that consists of a set of sample input vectors $\{\mathbf{x}_n\}_{n=1}^{N}$ together with the corresponding targets $\{t_n\}_{n=1}^{N}$. The targets are basically real values in regression tasks or class labels in classification problems. It is typically desired to learn a model of the dependency of the targets on the inputs from the training set, so that accurate predictions of t can be made for previously unseen values of $\mathbf{x}$. Commonly, these predications can be based on some function $g(\mathbf{x})$ defined over the input space in the form of

$$g(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) \tag{1}$$

as a linearly weighted sum of M (generally nonlinear and fixed) basis functions $\boldsymbol{\varphi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^T$. Learning is basically the process of inferring the function, or equivalently the parameters of the function $g(\mathbf{x})$. In this context, it is desired to estimate reasonable values for the parameters (or weights) $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$. Given a set of N corresponding training pairs $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$, the objective is to find values for the weights $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$ such that $g(\mathbf{x})$ generalizes well enough to new data, yet only a few elements of $\mathbf{w}$ are non-zero [6] . Having only a few non-zero weights facilitates a sparse representation with the advantage of providing fast implementation.

The support vector machine (SVM) [3] provides a successful approach to supervised learning by making predictions based on a function in the form of

$$g(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0 \tag{2}$$

where w_i shows the model weights and $K(\cdot,\cdot)$ is a kernel function effectively defining one basis function for each sample in the training set. The key feature of SVM classification is that, its target function attempts to minimize a measure of error on the training set while simultaneously maximizing the margin between the two classes that are implicitly defined in the feature space by the kernel [6]. This process results in a sparse model that depends only on a subset of kernel functions, namely those associated with training samples that lie either on the margin or on the wrong side, and the corresponding training samples are referred to as "support vectors". SVM is quite popular in supervised learning applications and has recently been applied for regression and classification of multispectral [4] as well as hyperspectral images [5] and therefore the reader is referred to these references to avoid re-phrasing the basics of SVM. The Relevance Vector Machine (RVM) has been introduced by Tipping [6] as a Bayesian treatment alternative to the SVM. The RVM 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. The most compelling feature of the RVM is that it typically utilizes significantly fewer kernel functions compared to SVM, while providing a similar performance. The application of RVM to hyperspectral image classification is recently presented in [7] and the reader is referred to the corresponding paper for detail.

3. SPATIALLY ADAPTIVE WAVELET DOMAIN NOISE REDUCTION IN HYPERSPECTRAL IMAGES

In wavelet domain thresholding (shrinkage) methods, noise reduction is obtained from shrinking the noisy coefficient magnitudes in the wavelet domain. A spatially adaptive Bayesian shrinkage approach, that uses a generalized Laplacian prior, and estimates the probability of a coefficient to contain a noise free component has been presented in [11] for image denoising. This approach has been utilized in this paper for the denoising of hyperspectral images.

In the spatially adaptive Bayesian shrinkage approach, even if two noisy coefficients have the same magnitude in the same sub-band, they can be shrunk differently according to their spatial position and local surrounding. Noise free coefficient components that exceed a threshold T are called 'signal of interest'. Two hypotheses are obtained according to a threshold value. The hypothesis H_0 is defined for the case in which the signal of interest is absent and H_1 is defined for the case in which the signal of interest is present. To model the noise free subband data, the generalized Laplacian prior is used. A local spatial activity indicator (LSAI) z_l for each spatial position l is obtained using the estimator equation

$$\hat{\beta}_l = P(H_1 | y_l, z_l) y_l = \frac{\eta_l \varepsilon_l \mu}{1 + \eta_l \varepsilon_l \mu} y_l \tag{3}$$

where the product $\eta_l \varepsilon_l \mu$ denotes the generalized likelihood ratio [11]. As shown in (4), each coefficient is shrunk according to how probable it is that it presents useful information, based on its value (via η_l), based on a measurement form the local surrounding (via ε_l), and based also on the global statistical properties of the coefficients in a given sub-band (via μ) [11].

$$\eta_l = \frac{f(y_l|H_1)}{f(y_l|H_0)},\ \varepsilon_l = \frac{f(z_l|H_1)}{f(z_l|H_0)} \text{ ve } \mu = \frac{P(H_1)}{P(H_0)} \tag{4}$$

Note that LSAI is defined in the form of locally averaged magnitudes of the coefficients in a relatively small square window $\delta(l)$ of a fixed size N within the same sub-band, as shown in (5). Hence if high magnitude coefficients are present within the neighbourhood, the LSAI value z_l will be high, and vice-versa. It is possible to formulate the LSAI in the form of

$$z_l = \frac{1}{N} \sum_{k \in \delta(l)} w_k \tag{5}$$

where w_l denotes the coefficient magnitude, i.e. $w_l = |y_l|$. For practical reasons it is assumed that all coefficients within the window are equally distributed and conditionally independent given H_0 and H_1. As Nz_l is the sum of N coefficients w_k, $f(Nz_l|H_{0,1})$ equals N convolutions of $f(w_l|H_{0,1})$ with itself, where $f(w_l|H_{0,1}) = 2f(y_l|H_{0,1})$ for $w_l > 0$, and $f(w_l|H_{0,1}) = 0$ for $w_l < 0$. $f(y|H_0)$ and $f(y|H_1)$ can be obtained from

$$\begin{aligned} f(y|H_0) &= \int_{-\infty}^{\infty} \phi(y-\beta;\sigma) f(\beta|H_0) d\beta \\ f(y|H_1) &= \int_{-\infty}^{\infty} \phi(y-\beta;\sigma) f(\beta|H_1) d\beta \end{aligned} \tag{6}$$

where $\phi(y;\sigma)$ is the zero mean Gaussian density with standard deviation σ and conditional densities of noise-free coefficients are

$$f(\beta|H_0) = \begin{cases} \beta_0 \exp(-\lambda|\beta^v|), & \text{if } |\beta| \le T \\ 0, & \text{if } |\beta| > T \end{cases}$$
$$f(\beta|H_1) = \begin{cases} 0, & \text{if } |\beta| \le T \\ \beta_1 \exp(-\lambda|\beta^v|), & \text{if } |\beta| > T \end{cases} \tag{7}$$

with the normalization constants

$$\beta_0 = \frac{\lambda v}{2\Gamma(\frac{1}{v})\Gamma_{inc}((\lambda T)^v, \frac{1}{v})}$$
$$\beta_1 = \frac{\lambda v}{2\Gamma(\frac{1}{v})\Gamma_{inc}((\lambda T)^v, \frac{1}{v})} \tag{8}$$

where $\Gamma_{inc}(x;a) = \frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}dt$ is the incomplete gamma function and λ is the scale parameter [11]. It is possible to estimate $P(H_1)$ in the form of

$$P(H_1) = \int_{-\infty}^{\infty} f(\beta|H_1)d\beta = 1 - \int_{-T}^{T} f(\beta)d\beta \tag{9}$$

And therefore it is possible to obtain

$$P(H_1) = 1 - \Gamma_{inc}((\lambda T)^v, \frac{1}{v}) \tag{10}$$

and thus

$$\mu = \frac{P(H_1)}{P(H_0)} = \frac{1 - \Gamma_{inc}((\lambda T)^v, \frac{1}{v})}{\Gamma_{inc}((\lambda T)^v, \frac{1}{v})} \tag{11}$$

for the Laplacian prior ($v = 1$)[11].

4. EXPERIMENTAL RESULTS

RVM and SVM classification methods have been applied to a sample hyperspectral image which is taken over northwest Indiana's Indian Pine test site in June 1992 [12] because the ground truth classification result of this image is available. The data consist of 145×145 pixels with 220 bands. The number of spectral bands is initially reduced to 200 by removing bands covering water absorption as well as extremely noisy bands. The original ground truth has actually 16 classes, but some classes have a very small number of elements, and therefore, nine classes that have the highest number of elements have been selected and used to generate 4757 training samples and 4588 test samples which are shown in Table I.

In both, SVM and RVM, the one-against-one method, in which $K(K-1)/2$ binary classifiers are trained and $K(K-1)/2$ binary tests are required to make a final decision, is utilized for multi-class classification in this paper. In this case, each outcome gives one vote to the winning class. The class with the most votes is selected as the final result.

Symmlet orthogonal wavelets with eight vanishing moments are used for wavelet denoising. Four decomposition levels and a 7×7 square windows size, which have been experimentally found optimal in [11], are used in this paper. The noise standard deviation is obtained as the median absolute deviation of the coefficients in the highest frequency subband divided by 0.6745 [13] for each hyperspectral band.

Table 2 shows results for SVM based classification, and Table 3 shows results for RVM based classification of the original and denoised hyperspectral test image. It is seen from these results, that RVM and SVM classification of denoised hyperspectral images provides much higher classification accuracy and furthermore requires a significantly less number of relevance vectors and support vectors, hence the classification time is also reduced. Comparing the maximum classification accuracy, it is seen that denoised RVM and SVM provides higher (up to about 7 %) classification accuracy compared to direct RVM and SVM. It is noted in [5] that sparsity in hyperspectral data classification is an important property of a kernel method given the special characteristics of the problem, i.e., high input dimension per low number of samples; because it is possible to think of sparsity as the property that indicates the complexity (and hence computational burden) of a model.

5. CONCLUSIONS

RVM and SVM based classification after wavelet denoising of hyperspectral images is presented in this paper. SVM and RVM classification with wavelet denoised hyperspectral images is shown to provide higher classification accuracy, with a significantly smaller relevance vector rate and support vector rate and therefore much faster testing time, compared with direct SVM and RVM based classification. Hence RVM and SVM classification of denoised hyperspectral data is superior to direct SVM and RVM classification in terms of classification accuracy.

6. ACKNOWLEDGMENT

This work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under the "Hyperspectral Classification, Segmentation and Recognition" project with contract no. EEEAG/107E011, and also Turkish State Planning organization project DPT 2008K-120800.

REFERENCES

[1] A. Meyer, D. Paglieroni, and C. Astaneh, "K-Means re-clustering: algorithmic options with quantifiable performance comparisons", *SPIE Photonics West, Optical Engineering at LLNL*, vol. 5001, pp. 84-92, 2003.

[2] A. Ertürk, and S. Ertürk, "Unsupervised segmentation of hyperspectral images using modified phase correlation", *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 527-531, Nov 2006.

[3] B. E. Boser, I. M. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992, pp. 144-152.

[4] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification", *Int. J. Remote Sensing*, vol. 23, no. 4, pp.725-749, 2002.

[5] G. Camps-Valls, and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1352-1362, June 2005.

[6] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine*,"* *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[7] B. Demir, and S. Ertürk,"Hyperspectral image classification using relevance vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol.4, no. 4, pp. 586-590, 2007.

[8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.

[9] I. Atkinson, F. Kamalabadi, and D. L. Jones, "Wavelet-based hyperspectral image estimation," *IEEE Int. Geosci. Remote Sensing Symp. (IGARSS 2003)*, Toulouse, France, July 2003, pp. 743-745.

[10] A. C. Zelinski and V. K. Goyal, "Denoising hyperspectral imagery and recovering junk bands using wavelets and sparse approximation," in *IEEE Int. Geosci. Remote Sensing Symp. (IGARSS 2006)*, Denver, Aug. 2006, pp. 387-390.

[11] A. Pizurica and W. Philips, "Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 654-665, March 2006.

[12] AVIRIS NW Indiana's Indian Pines 1992 data set [Online]. Available:ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C.lan (original files) and ftp://ftp.ecn.purdue.edu/biehl/PC_MultiSpec/ThyFiles.zip (ground truth).

[13] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[14] LIBSVM: A library for support vector machines, Chang, C.-C. and Lin, C.-J: (2001). [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

TABLE I
NUMBER OF TRAINING AND TEST SAMPLES

Class	Training	Test
C1-Corn-no till	742	692
C2-Corn-min till	442	392
C3-Grass/Pasture	260	237
C4-Grass/Trees	389	358
C5-Hay-windrowed	236	253
C6-Soybean-no till	487	481
C7-Soybean-min till	1245	1223
C8-Soybean-clean till	305	309
C9-Woods	651	643
Total	4757	4588

TABLE II
CLASSIFICATION ACCURACY (AC) and NUMBER OF SUPPORT VECTORS (SV) FOR SVM of ORIGINAL DATA (SVM) and DENOISED DATA (D-SVM)

Method	Kernel type	Kernel Parameter		C	AC	SV
		γ	D			
SVM	RBF	0.1	-	1000	90.88	3195
D-SVM	**RBF**	**0.1**	**-**	**1000**	**97.73**	**2061**
SVM	RBF	1	-	65	91.95	3259
D-SVM	**RBF**	**1**	**-**	**65**	**98.45**	**3064**
SVM	Poly.	1	7	60	90.03	2066
D-SVM	**Poly.**	**1**	**7**	**60**	**95.27**	**1634**
SVM	Poly	2	3	40	88.86	1993
D-SVM	**Poly**	**2**	**3**	**40**	**95.85**	**1796**

TABLE III
CLASSIFICATION ACCURACY (AC) and NUMBER OF RELEVANCE VECTORS (RV) FOR RVM of ORIGINAL DATA (RVM) and DENOISED DATA (D-RVM)

Method	Kernel type	Kernel parameter		AC	RV
		γ	d		
RVM	RBF	0.1	-	89.40	414
D-RVM	**RBF**	**0.1**	**-**	**95.42**	**377**
RVM	RBF	1	-	90.14	514
D-RVM	**RBF**	**1**	**-**	**95.40**	**439**
RVM	Poly.	1	7	89.39	412
D-RVM	**Poly.**	**1**	**7**	**94.50**	**320**
RVM	Poly	2	3	87.23	402
D-RVM	**Poly**	**2**	**3**	**94.45**	**317**