

Feature Selection and Classification of Hyperspectral Images With Support Vector Machines

Rick Archibald and George Fann

Abstract—Hyperspectral images consist of large number of bands which require sophisticated analysis to extract. One approach to reduce computational cost, information representation, and accelerate knowledge discovery is to eliminate bands that do not add value to the classification and analysis method which is being applied. In particular, algorithms that perform band elimination should be designed to take advantage of the structure of the classification method used. This letter introduces an embedded-feature-selection (EFS) algorithm that is tailored to operate with support vector machines (SVMs) to perform band selection and classification simultaneously. We have successfully applied this algorithm to determine a reasonable subset of bands without any user-defined stopping criteria on some sample AVIRIS images; a problem occurs in benchmarking recursive-feature-elimination methods for the SVMs.

Index Terms—Feature selection, hyperspectral images, support vector machines (SVMs).

I. INTRODUCTION

A NEW GENERATION of remote sensors is producing hyperspectral images which sample hundreds of contiguous narrow spectral bands. Hyperspectral images have been proven beneficial to many different applications of remote sensing, medical imaging, and quality assurance. The high dimensionality of these data sets has spurred the development of new techniques for analysis. Support vector machines (SVMs) [11], which are a classification paradigm developed over the last decade in machine learning theory, have been successfully applied within the remote sensing community to hyperspectral-image analysis. Recent studies comparing SVMs with other classification schemes have concluded that they provide significant advantages in accuracy, simplicity, and robustness [2].

The SVM classifies binary data by determining the separating hyperplane, or decision surface, which maximizes the margin between the two classes in the training data. Kernel functions provide SVM with the powerful additional ability of efficiently determining the nonlinear decision surfaces. The SVM structure maximizes performance attainable by training

on reduced sets, and this structure has been exploited to produce cost-reducing techniques in hyperspectral training [4]. One major drawback of SVM is that classification is binary. However, this issue can be resolved in a simple and robust procedure by training several SVMs simultaneously in an all-against-one or one-against-one scheme [11].

The computational cost of classification grows quadratically with data dimension size, making feature selection an important issue for the SVM. Feature-selection algorithms fall into three categories: filtering, wrapper, and embedded methods. Filter methods can be a fast and easy solution, but they are not usually optimal since they do not account for the mechanism of the learning algorithm utilized. Wrapper methods consist of searching for the subset of features that minimize the generalization error. This goal is the ideal, but the search is a combinatorial problem that is NP-hard [12]. Finally, in embedded methods, as the name suggests, feature selection is embedded into the learning algorithm.

This letter utilizes the underlying mechanism of SVMs to develop a band-selection embedded algorithm for hyperspectral imaging that is based on logistic boosting. The underlying methodology of boosting, which was introduced by Freund and Schapire [5], utilizes a metalearning strategy that distills the performance of many “weak” classifiers into a unified “strong” classifier. We demonstrate how to manipulate the boosting method so that an optimal distribution of bands is found for one classifier, in effect, transforming the boosting method into a feature-extraction algorithm. The proposed method is compared to the well-recognized recursive feature elimination (RFE)-SVM method [6] and found to improve convergence speed and eliminate the need to have *a priori* information about the number of important features.

The remainder of this letter is organized in the following manner. Section II presents a brief overview of SVM theory and the machinery necessary to describe the embedded-feature-selection (EFS) algorithm. The specific application of EFS with SVM classification to hyperspectral-image analysis is presented in Section III. Finally, in Section IV, we provide a summary.

II. FEATURE SELECTION WITH SVMs

Vapnik [11], at AT&T Bell Laboratories, advanced machine learning in real-world applications by developing the SVM. The decision function for SVM is expressed as

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^N y_i \alpha_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \right) \quad (1)$$

Manuscript received January 17, 2007; revised June 27, 2007. The work of R. Archibald was supported in part by the Householder Fellowship in Scientific Computing sponsored by the U.S. Department of Energy’s Applied Mathematical Sciences Program and in part by the Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

The authors are with the Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (e-mail: ArchibaldRK@ORNL.gov; fanngi@ORNL.gov).

Digital Object Identifier 10.1109/LGRS.2007.905116

U.S. Government work not protected by U.S. copyright.

where the coefficients are determined by maximizing the margin of the given training set

$$\{\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}, \quad i = 1, \dots, N \quad (2)$$

in the transformed space defined via the mapping or kernel function Φ . Specific examples of kernel functions used in this letter include

$$\begin{aligned} \text{Polynomial} : \Phi(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^n \\ \text{Gaussian(RBF)} : \Phi(\mathbf{x}_i, \mathbf{x}_j) &= e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2} \end{aligned} \quad (3)$$

for $n \in \mathbb{N}$ and $\gamma > 0$.

There are many different variants of SVM, among which we use SVM^{light} [8]. Similarly, there are many approaches in extending SVM beyond binary to multiple classification [11]. The most common multiclass approach is one-against-all, where a new SVM is trained for each binary combination of classes. The decision function is determined by evaluating each of the classes SVMs with a ‘‘winner-takes-all’’ rule. This multiclass approach has proven to be robust and accurate and will be used in this letter [10].

A. Recursive Feature Elimination

A well-known property of SVM is that the generalization error, which is denoted here as G_E , is bounded by

$$G_E \leq \frac{1}{N} E \left\{ \frac{R^2}{M^2} \right\} \quad (4)$$

where R is the radius of the smallest sphere containing the transformed training data separated by a margin M , with an expectation taken over training data sets of size N [11].

The RFE focuses on minimizing the generalization error by eliminating features that maximize the margin. The predictive ability measure is inversely proportional to the margin and given by

$$W^2(\alpha) = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i, \mathbf{x}_j). \quad (5)$$

Therefore, the margin can be maximized by minimizing W , which the RFE algorithm attempts to accomplish. The extension of (5) to multiclassification problems is given as

$$W^2(\alpha)_{(-f)} = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_{i,(-f)}, \mathbf{x}_{j,(-f)}) \quad (6)$$

where $\mathbf{x}_{i,(-f)}$ is the i th hyperspectral training sample with the f th feature removed.

RFE-SVM becomes problematic when the correct number of features to remove is not known. For this case, it is standard to continue the RFE algorithm and store W (5) for each removal of a band. The ideal feature selection occurs at the iteration where (5) is a minimum. To ensure a comprehensive search, it is required that the RFE algorithm be continued until all but a

reasonable number of bands remain. In this letter, we continue the automated search until $r = 5$ bands remain.

B. Embedded Feature Selection

We begin by defining the componentwise product of two vectors ρ and $\mathbf{x} \in \mathbb{R}^d$ as

$$\rho * \mathbf{x} = (\rho_1 x_1, \dots, \rho_d x_d). \quad (7)$$

The EFS method weighs each band according to a logistically scaled measure of importance.

EFS-SVM Algorithm: Given a training set (2) and tolerance parameter $\epsilon > 0$ and attenuation parameter $\sigma > 1$, we follow the following steps.

- 1) Initialize $\rho_j = 1$, for $j = 1, \dots, N$.
- 2) Train the SVM to obtain a solution α using the modified kernel $K_\rho(\mathbf{x}, \mathbf{y}) := \Phi(\rho * \mathbf{x}, \rho * \mathbf{y})$.
- 3) Calculate

$$L(\alpha)_j = \frac{1}{1 + e^{A_M M_j + B_M}}, \quad \text{for } j = 1, \dots, N \quad (8)$$

where M_j is a measure of importance for the j th feature

$$\begin{pmatrix} A_M \\ B_M \end{pmatrix} = - \begin{pmatrix} 1 & \mu(\epsilon, M) \\ 1 & \mu(\frac{\epsilon}{\sigma}, M) \end{pmatrix}^{-1} \begin{pmatrix} \ln(\frac{1-\epsilon}{\epsilon}) \\ \ln(\frac{\sigma-\epsilon}{\epsilon}) \end{pmatrix} \quad (9)$$

for

$$\mu(\epsilon, M) = \begin{cases} \epsilon, & \text{if } \min M_j < \epsilon \\ x_\epsilon, & \text{otherwise} \end{cases} \quad (10)$$

with

$$x_\epsilon = \frac{(\epsilon - \sum_{M_j < a} M_j)(b - a)}{b} \quad (11)$$

$$a = \arg \max_{M_i} \sum_{M_j < M_i} M_j < \epsilon \quad (12)$$

and

$$b = \min_i M_i > a. \quad (13)$$

- 4) Quit if

$$|\rho - L(\alpha) * \rho| < \epsilon N$$

otherwise set

$$\rho \leftarrow L(\alpha) * \rho.$$

If $\min \rho < \epsilon/\sigma$, then remove all features that have a weight below the threshold ϵ/σ and update N accordingly. Return to step 2).

This algorithm embeds a weighting into the kernel and iteratively updates the weights by the logistic function (8), naturally bounded in the interval $[0, 1]$. Unlike RFE where the weighting can be considered as all or nothing, features can be damped

or removed based upon the measure used to gauge their importance. This letter examines two measures. First, we use the obvious choice

$$M_f = W^2(\alpha)_{(-f)} \quad (14)$$

of (6). Second, we introduce a decision function feature sensitivity measure

$$M_f = \max_{j:\alpha_j \neq 0} \sum_{i=1}^N y_i \alpha_i \Phi(\mathbf{x}_i, \mathbf{x}_{j,(+f)}) - \min_{j:\alpha_j \neq 0} \sum_{i=1}^N y_i \alpha_i \Phi(\mathbf{x}_i, \mathbf{x}_{j,(+f)}) \quad (15)$$

where $\mathbf{x}_{j,(+f)}$ is the j th hyperspectral training sample with every feature except the f th replaced by the feature average of the whole training set. This measure uses only the support vectors (rich source of information with minimal computation) to measure the sensitivity of the decision function to each feature.

C. Complexity Measure

The major cost of calculation in SVM arises in training, a process that is dominated by the complexity of solving the dual formulation which is known to be $\mathcal{O}(\max(N, d) \min(N, d)^2)$ [3]. Based on the dual formulation complexity, we define the measure

$$R_{co} = \frac{N d_s^2 + \sum_{j=1}^{N_{it}} \max(N_s, d_j) \min(N_s, d_j)^2}{N d^2} \quad (16)$$

the computational cost ratio of first performing features selection, and, then, the SVM on a reduced dimension size, as compared to directly performing SVM. Here, d_s is the number of dimension after feature selection, N_s is the size of training data subset used to determine the feature selection, and the whole data set has size N and dimension d . Here, we assume that $N > d > d_s$, which is a reasonable assumption for dimension size of hyperspectral data.

III. APPLICATIONS TO HYPERSPECTRAL DATA AND CLASSIFICATION RESULTS

We apply the RFE and the EFS algorithm to the publicly available Indian Pines data [9], consisting of 145×145 pixels by 220 bands of reflectance AVIRIS data. Ground truth is known for most of the scene. From the 16 identified classes in the scene, we use the nine largest samples with training and test set sizes reported in Table I. The type of classes and structure of training samples is specifically chosen to be similar to the report [2] so that results can be compared.

We determine the optimal penalty and free parameters for the SVM classification of this data set only once for each kernel using a fast simulated annealing process [1]. For each application of the EFS algorithm, we use tolerance parameter $\epsilon = 0.01$ and attenuation parameter $\sigma = 10$. This has the effect of attenuating

TABLE I
LAND COVER CLASSES WITH TRAINING AND TEST SET SIZES FOR THE CLASSIFICATION EXPERIMENT

Class	Land Cover Type	Train	Test
C_1	Corn-no Till	717	717
C_2	Corn-mil Till	417	417
C_3	Grass/Pasture	248	249
C_4	Grass/Trees	373	374
C_5	Hay-windrowed	244	245
C_6	Soybean-no Till	484	484
C_7	Soybean-min Till	1234	1234
C_8	Soybean-clean Till	307	307
C_9	Woods	647	647
	Total	4670	4674

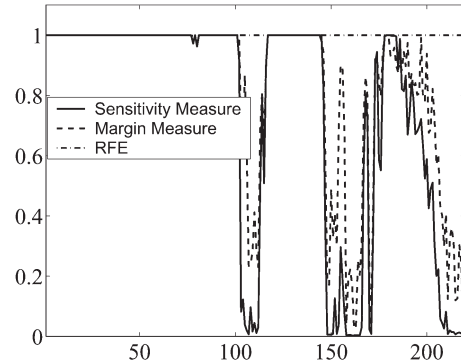


Fig. 1. Band selection using the RFE and the EFS algorithm with measures (14) and (15) on the Indian Pines hyperspectral data set for the Gaussian kernel function (3).

bands that have less than a 1% significance weighting and eliminates bands that iteratively fall below 0.1% significance. These are the only parameters of the EFS algorithm and have been found to be robust for values that correspond to modest attenuation (less than 5% significance weighting). Finally, a standard preprocessing step of zero-mean normalization is done for this data.

It is accepted that, for this data set, because of atmospheric water absorption, a total of 20 channels can be identified as noisy (104–108, 150–163, 220) and safely removed as a preprocessing step [2]. In this experiment, they are purposefully not removed and act as a natural test for feature-extraction algorithms. Fig. 1 shows the average weighting for band selection using the RFE and the EFS algorithm with measures (14) and (15) for ten different training data random samplings of size $d/5$. It is emphasized here that band selection is quickly generated by a small subset of the training data. Once band selection is determined, training proceeds on the full set. As shown in Fig. 1, only the EFS algorithm removes the bands associated with atmospheric water absorption, with the sensitivity measure (15) outperforming the margin measure (14) in consistency and accuracy.

Table II has the results of the EFS and RFE classification for the Indian Pines data set. The weights for band selection are determined for each algorithm by taking the average result of ten different training data random samplings of size $d/5$. These estimates for weights and band number are then used on the whole data set to test and train. Convergence of the EFS algorithm is quick, requiring only a few iterations for

TABLE II

RESULTS OF THE EFS AND RFE CLASSIFICATION FOR THE INDIAN PINES DATA SET. SVM WITH NO BAND SELECTION IS INCLUDED FOR COMPARISON. ALL COMPUTATIONS WERE PERFORMED ON A 1.4-GHz AMD DUAL-CORE NEMESIS WORKSTATION WITH A 2-GB RAM

Kernel	EFS								RFE			SVM	
	Measure (14)				Measure (15)				Error	Bands	R_{co}/CPU	Error	CPU
	Error	Bands	Iter.	R_{co}/CPU	Error	Bands	Iter.	R_{co}/CPU					
Polynomial	92.79%	208±4	2±1	0.90/248s	92.98%	197±3	2±1	0.80/221s	92.94%	220±1	1.20/328s	92.94%	280s
Gaussian	93.95%	202±13	2±1	0.85/231s	93.99%	185±16	2±1	0.71/197s	93.95%	220±1	1.20/336s	93.95%	276s

TABLE III

RESULTS OF THE EFS AND RFE CLASSIFICATION FOR THE INDIAN PINES DATA SET FOR THE INDIVIDUAL CLASSES

Kernel	Method	Individual Classes								
		C1	C2	C3	C4	C5	C6	C7	C8	C9
Polynomial	EFE eq.(14)	91.89%	93.05%	92.91%	95.97%	100%	86.58%	91.28%	90.59%	97.57%
	EFE eq.(15)	90.96%	94.00%	93.86%	95.77%	100%	86.14%	91.74%	90.61%	98.53%
	RFE	91.71%	93.81%	94.72%	95.76%	100%	86.93%	91.09%	91.07%	97.64%
Gaussian	EFE eq.(14)	92.36%	95.89%	95.59%	96.26%	100%	87.56%	92.21%	92.46%	99.03%
	EFE eq.(15)	92.54%	94.95%	95.68%	96.45%	100%	87.91%	92.39%	92.28%	99.10%
	RFE	92.44%	94.85%	95.58%	96.54%	100%	86.00%	92.29%	92.18%	98.93%

each of the tested kernels. In contrast, the automated RFE method performs little or no feature extraction, an indication that the minimum of (5) across removed features is not a good indicator for this data set. A user-defined stopping point for RFE is necessary. It is noted that this is not a breakdown in the RFE method but rather in the method to determine the optimal number of bands. The band selection and the classification rate of RFE are similar to the EFS algorithm with measure (5) if we preset the RFE method to stop at the same number of bands.

Tables II and III are consistent with the results published in [2] and show that the EFS provides, at best, a slight gain in classification accuracy. This result is expected since SVM minimizes the effect of the Hughes phenomenon [7], and therefore, as long as essential informative bands are not removed, the classification rate should be flat.

The major advantages of EFS are the significant reduction in computational time and the data-driven knowledge discovery of important bands in classification. The computational speedup is derived from the fact that the EFS method is able to determine significant bands from the greatly reduced training sets within a few iterations. The cost of finding an optimal subset of bands does not overshadow the benefit of reduced overall training time. This property would be shared with RFE if the user defined the same number of bands as determined by EFS and furthermore used an accelerated band removal scheme.

IV. CONCLUDING REMARKS

This letter develops an EFS algorithm that is specifically designed to operate with the SVM. A feature sensitivity measure is introduced that is referenced to the decision function and is demonstrated, as compared to a well-known SVM margin measure, to find a reduced subset of predictive hyperspectral bands. This letter offers a data-driven knowledge discovery and classification with an increased computational efficiency in processing and analyzing the hyperspectral images.

ACKNOWLEDGMENT

The submitted manuscript has been authored by contractors [UT-Battelle, Manager of Oak Ridge National Laboratory (ORNL)] of the U.S. Government under Contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

REFERENCES

- [1] M. Boardman and T. Trappenberg, "International Joint Conference on Neural Networks," Vancouver, BC, Canada, Jul. 2006.
- [2] 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.
- [3] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.
- [4] G. M. Foody and A. Mathur, "The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM," *Remote Sens. Environ.*, vol. 103, no. 2, pp. 179–189, Jul. 2006.
- [5] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
- [6] I. M. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.
- [7] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
- [8] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 11.
- [9] D. Landgrebe, *AVIRIS NW Indianas Indian Pines 1992 data set*, 1992. [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)
- [10] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, 2004.
- [11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [12] J. Weston, A. Elisseeff, and B. Schölkopf, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.