Semi-supervised Hyperspectral Image Classification with Graphs

Tatyana V. Bandos*, Dengyong Zhou† and Gustavo Camps-Valls*

*Dept. Enginyeria Electrònica. Universitat de València.
C/ Dr. Moliner, 50. 46100. Burjassot, València. Spain.
E-mail: {tatyana.bandos,gustavo.camps}@uv.es

†NEC Laboratories America (Princeton)
4 Independence Way, Suite 200, Princeton NJ 08540. USA.
E-mail: dzhou@nec-labs.com

***Abstract*— This paper presents a semi-supervised graph-based method for the classification of hyperspectral images. The method is designed to exploit the spatial/contextual information in the images through composite kernels. The proposed method produces smoother classifications with respect to the intrinsic structure collectively revealed by known labeled and unlabeled points. Good accuracy in high dimensional spaces and low number of labeled samples (ill-posed situations) are produced as compared to standard inductive support vector machines.**

I. Introduction

In the remote sensing literature, many supervised and unsupervised classifiers have been developed to tackle the multi- and hyperspectral data classification problem [1]. The main difficulty with supervised methods is that the learning process heavily depends on the quality of the training dataset, which is only useful for simultaneous images, or for images with the same classes taken under the same conditions; and, even worse, the training set is frequently not available, or in a very reduced number, given the high cost of true sample labeling. On the other hand, unsupervised methods are not sensitive to the number of labeled samples since they work on the whole image, but the relationship between clusters and classes is not ensured. The use of semi-supervised classifiers can yield improved performance in these situations.

In semi-supervised learning (SSL), the algorithm is provided with some supervised information in addition to the unlabeled data. Three different classes of SSL algorithms are encountered in the literature: (1) *generative* models, which involve estimating the conditional density $p(x|y)$ (e.g. expectation-maximization (EM) algorithms with finite mixture models [2]); (2) *low density separation* algorithms, which maximize the margin for labeled and unlabeled samples simultaneously (e.g. Transductive SVM [3]); and (3) *graph-based* methods [4], in which each sample spreads its label information to its neighbors until a global stable state is achieved on the whole data set.

Graph-based methods have been lately paid attention because of their solid mathematical background, their relationship with kernel methods, sparseness properties, model visualization, and good results in many areas. In this paper, we introduce a semi-supervised graph-based method, previously presented in [5], in the context of hyperspectral image classification. In order to improve its performance, we include in the formulation the contextual information through the use of composite kernels, which have been recently revealed very useful to improve inductive support vector machines (SVMs) [6], [7]. Finally, noting that the method relies on building large kernel matrices, we reformulate the algorithm using the Nyström method to speed up the solution [8]. The method is evaluated in the real-like scenario of ill-posed classification, i.e. low number of high dimensional labeled samples.

The paper is outlined as follows. Section II reviews the main ideas underlying graph methods. Section III presents the proposed semi-supervised graph-based composite kernel classification method. Section IV discusses the classification results compared to standard SVMs in ill-posed classification. Finally, section V includes some concluding remarks and indications on further work.

II. Learning with graphs

Graph-based methods rely upon the construction of a graph representation, where the vertices are the (labeled and unlabeled) samples, and edges represent the similarity among samples in the dataset (see Fig. 1).

Typically, graph methods utilize the graph Laplacian, which is defined as follows. Let $G = (V, E)$ be a graph with a set of vertices, V, connected by a set of edges, E. The edge connecting nodes (or samples) i and j have an associated weight, $\{W_{ij}\}$. Then, the weight (or affinity) matrix W is constructed among all labeled and unlabeled samples. The (normalized) graph Laplacian is defined as

$$\mathcal{L} = I - D^{-1/2} W D^{-1/2}, \tag{1}$$

where D is a diagonal matrix defined by $D_{ii} = \sum_{ij} W_{ij}$. See [9] (Ch. 11) for more details on different families of graph-based methods.

At this point, it is worth noting that prediction consists in labeling the unlabeled nodes, and thus, these are intrinsically

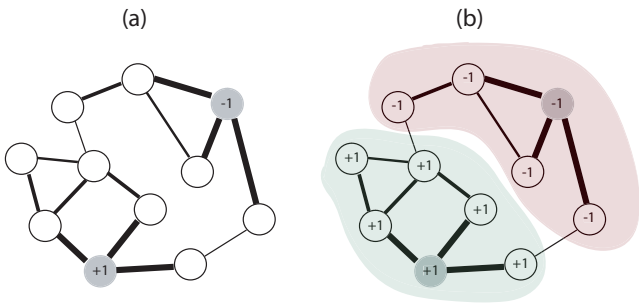

Fig. 1. Graph classification on a toy graph. (a) The two shaded circles are the initially labeled vertices (± 1), while the white nodes represent unlabeled samples. The thickness of the edges represent the similarity among samples. (b) Graph methods classify the unlabeled samples according to the weighted distance, not just to the shortest path lengths, the latter leading to incorrectly classified samples. The two clusters (shaded in green and red) are intuitively correct, even being connected by (thin weak) edges.

transductive classifiers, i.e. the graph only returns the predicted class label for the unlabeled samples, not a decision function defined on the whole domain. These graph-based classifiers can be viewed as estimating a function F over the graph, which should be in accordance with the *smoothness* assumption, that is, a good classification function should not change too much between nearby points.

III. Graph-based composite kernel classification

A. Semisupervised graph-based method

1) Formulation: Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^N$ and a label set $\mathcal{L} = \{1, \ldots, c\}$, the first l points $\mathbf{x}_i$ $(i \leq l)$ are labeled as $y_i \in \mathcal{L}$ and the remaining points $\mathbf{x}_u$ $(l+1 \leq u \leq n)$ are unlabeled. The goal in semi-supervised learning is to predict the labels of the unlabeled points.

Let $\mathcal{F}$ denote the set of $n \times c$ matrices with nonnegative entries. A matrix $F = [F_1^\top, \ldots, F_n^\top]^\top \in \mathcal{F}$ corresponds to a classification on the dataset $\mathcal{X}$ by labeling each point $\mathbf{x}_i$ as a label $y_i = \arg\max_{j \leq c} F_{ij}$. We can understand F as a vectorial function $F : \mathcal{X} \to \mathbb{R}^c$ which assigns a vector F_i to each point $\mathbf{x}_i$. Define an $n \times c$ matrix $Y \in \mathcal{F}$ with $Y_{ij} = 1$ if $\mathbf{x}_i$ is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. Note that Y is consistent with the initial labels according to the decision rule. The algorithm can be summarized as follows:

1) Calculate the affinity matrix W defined by $W_{ij} \equiv W(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2) Construct the matrix $S = D^{-1/2} W D^{-1/2}$ in which D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W.
3) Iterate $F(t+1) = \alpha S F(t) + (1-\alpha) Y$ until convergence, where α is a parameter in $(0, 1)$.
4) Let F^* denote the limit of the sequence $\{F(t)\}$. Label each point $\mathbf{x}_i$ as a label $y_i = \arg\max_{j \leq c} F^*_{ij}$.

One can demonstrate [5] that in the limit, $F^* = \lim_{t \to \infty} F(t) = (1-\alpha)(I - \alpha S)^{-1} Y$, which is equivalent to the final estimating function:

$$F^* = (1-\alpha)(I - \alpha S)^{-1} Y, \tag{2}$$

and thus F^* can be computed directly without iterations.

2) Graph interpretation: The proposed method can be interpreted as a graph $G = (V, E)$ defined on $\mathcal{X}$, where the vertex set V is just $\mathcal{X}$ and the edges E are weighted by W. In the second step, the weight matrix W of G is normalized symmetrically, which is necessary for the convergence of the following iteration. During the third step, each sample receives the information from its neighbors (first term), and also retains its initial information (second term). The parameter α specifies the relative amount of the information from its neighbors and its initial label information. It is worth noting that *self-reinforcement* is avoided since the diagonal elements of the affinity matrix are set to zero in the first step. Moreover, the information is spread *symmetrically* since S is a symmetric matrix. Finally, the label of each unlabeled point is set to be the class of which it has received most information during the iteration process.

B. Spatio-Spectral composite kernels

Note that, in its standard use, the graph-based method proposed before only would take advantage of the spectral information. Here we propose a toolbox of composite kernels accounting for the spatial and spectral information in the affinity matrix W. For this purpose, a pixel entity $\mathbf{x}_i \in \mathbb{R}^N$ (N represents the number of spectral bands) is redefined simultaneously both in the spectral domain using its spectral content, $\mathbf{x}_i^\omega \in \mathbb{R}^{N_\omega}$, and in the spatial domain by applying some feature extraction to its surrounding area, $\mathbf{x}_i^s \in \mathbb{R}^{N_s}$, which yields N_s spatial (contextual) features. These separated entities lead to two different similarity matrices, which can be easily computed and combined. At this point, one can sum spectral and textural dedicated affinity matrices (W_ω and W_s, respectively), and introduce the cross-information between textural and spectral features ($W_{\omega s}$ and $W_{s\omega}$) in the formulation. This simple methodology yields a full family of composite methods for hyperspectral image classification [7], which can be summarized as follows:

- *The stacked features approach.* Let us define the mapping $\mathbf{\Phi}$ as a transformation of the concatenation $\mathbf{x}_i \equiv \{\mathbf{x}_i^s,\ \mathbf{x}_i^\omega\}$, then the corresponding 'stacked' affinity matrix is:

$$W_{\{s,\omega\}} \equiv W(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle, \tag{3}$$

  which does not include explicit cross relations between $\mathbf{x}_i^s$ and $\mathbf{x}_j^\omega$.
- *The direct summation kernel.* Let us assume two nonlinear transformations $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ into Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Then, the following transformation can be constructed:

$$\mathbf{\Phi}(\mathbf{x}_i) = \{\varphi_1(\mathbf{x}_i^s), \varphi_2(\mathbf{x}_i^\omega)\} \tag{4}$$

and the corresponding dot product can be easily computed as follows:

$$\begin{aligned} W(\mathbf{x}_i, \mathbf{x}_j) &= \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle \quad (5) \\ &= \langle \{\varphi_1(\mathbf{x}_i^s), \varphi_2(\mathbf{x}_i^\omega)\}, \{\varphi_1(\mathbf{x}_j^s), \varphi_2(\mathbf{x}_j^\omega)\} \rangle \\ &= W_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + W_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega) \end{aligned}$$

Note that the solution is expressed as the sum of positive definite matrices accounting for the textural and spectral counterparts, independently. Note that $dim(\mathbf{x}_i^\omega) = N_\omega$, $dim(\mathbf{x}_i^s) = N_s$, and $dim(W) = dim(W_s) = dim(W_\omega) = n \times n$.

- *The cross-information kernel.* The preceding kernel-based classifiers can be conveniently modified to account for the cross relationship between the spatial and spectral information. Assume a nonlinear mapping $\varphi(\cdot)$ to a Hilbert space $\mathcal{H}$ and three linear transformations $\mathbf{A}_k$ from $\mathcal{H}$ to $\mathcal{H}_k$, for $k = 1, 2, 3$. Let us construct the following composite vector:

$$\mathbf{\Phi}(\mathbf{x}_i) = \{\mathbf{A}_1\varphi(\mathbf{x}_i^s), \mathbf{A}_2\varphi(\mathbf{x}_i^\omega), \mathbf{A}_3(\varphi(\mathbf{x}_i^s) + \varphi(\mathbf{x}_i^\omega))\} \quad (6)$$

and compute the dot product

$$\begin{aligned} W(\mathbf{x}_i, \mathbf{x}_j) &= \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle \quad (7) \\ &= \mathbf{\Phi}(\mathbf{x}_i^s)^\top \mathbf{R}_1 \mathbf{\Phi}(\mathbf{x}_j^s) + \mathbf{\Phi}(\mathbf{x}_i^\omega)^\top \mathbf{R}_2 \mathbf{\Phi}(\mathbf{x}_j^\omega) \\ &+ \mathbf{\Phi}(\mathbf{x}_i^s)^\top \mathbf{R}_3 \mathbf{\Phi}(\mathbf{x}_j^\omega) + \mathbf{\Phi}(\mathbf{x}_i^\omega)^\top \mathbf{R}_3 \mathbf{\Phi}(\mathbf{x}_j^s) \end{aligned}$$

where $\mathbf{R}_1 = \mathbf{A}_1^\top \mathbf{A}_1 + \mathbf{A}_3^\top \mathbf{A}_3$, $\mathbf{R}_2 = \mathbf{A}_2^\top \mathbf{A}_2 + \mathbf{A}_3^\top \mathbf{A}_3$, and $\mathbf{R}_3 = \mathbf{A}_3^\top \mathbf{A}_3$ are three independent positive definite matrices.

Similarly to the direct summation kernel, it can be demonstrated that (7) can be expressed as the sum of positive definite matrices, accounting for the textural, spectral, and cross-terms between textural and spectral counterparts:

$$\begin{aligned} W(\mathbf{x}_i, \mathbf{x}_j) &= W_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + W_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega) \\ &+ W_{s\omega}(\mathbf{x}_i^s, \mathbf{x}_j^\omega) + W_{\omega s}(\mathbf{x}_i^\omega, \mathbf{x}_j^s) \quad (8) \end{aligned}$$

The only restriction for this formulation to be valid is that $\mathbf{x}_i^s$ and $\mathbf{x}_j^\omega$ need to have the same dimension ($N_\omega = N_s$).

C. Nyström method formulation

The proposed formulation involves three steps: firstly building the W matrix according to a composite specification, secondly, normalizing W to obtain S, and finally, solving an inversion problem given by (2). Note that direct inversion induces a computational cost of $\mathcal{O}(n^3)$, where n is the number of labeled and unlabeled samples, which in the case of remote sensing images can be very high.

The Nyström method is commonly used to produce an approximate kernel matrix $\tilde{W}$ by randomly choosing m rows/columns of the original matrix W and then making $\tilde{W}_{n,n} = W_{n,m} W_{m,m}^{-1} W_{m,n}$, $m \leq n$, where $W_{n,m}$ represents the $n \times m$ block of the W. As a result, the method simplifies the solution of the problem to the an approximated eigendecomposition of the low-rank kernel matrix $\tilde{W} = \tilde{V}\tilde{\Lambda}\tilde{V}^\top$, involving $\mathcal{O}(mn^2)$ computational cost [8].

Similarly, if we approximate the normalized matrix S by a small $p \times p$ matrix, $\tilde{S} = \tilde{V}\tilde{\Lambda}\tilde{V}^\top$, and substitute it into (2), we obtain:

$$F^* = (1-\alpha)(I - \alpha\tilde{V}\tilde{\Lambda}\tilde{V}^\top)^{-1}Y. \quad (9)$$

Now, by exploiting the Woodbury formula[1], it is straightforward to demonstrate that:

$$F^* = (1-\alpha)(Y + \tilde{V}(\tilde{\Lambda}\tilde{V}^\top\tilde{V} - \alpha^{-1}I)^{-1}\tilde{\Lambda}\tilde{V}^\top Y), \quad (10)$$

which involves inverting a matrix of size $p \times p$ (with $p \leq m \leq n$) and thus the computational cost is $\mathcal{O}(p^2 n)$, i.e. linear with the number of samples.

IV. Experimental Results

A. Data Collection

Experiments were carried out using the familiar AVIRIS image taken over NW Indiana's Indian Pine test site in June 1992 [10]. Following [11], we used a part of the 145×145 scene, called the *subset scene*, consisting of pixels [27-94]×[31-116] for a size of 68×86, which contains four labeled classes (the background pixels were not considered for classification purposes). We removed 20 noisy bands covering the region of water absorption, and finally worked with 200 spectral bands.

B. Model Development

The spectral samples $\mathbf{x}_i^\omega$ are, by definition, the spectral signature of pixels $\mathbf{x}_i$. The contextual samples, $\mathbf{x}_i^s$, were computed as the mean of a 3×3 window surrounding $\mathbf{x}_i$ for each band. In all cases, we used the RBF kernel to construct the similarity matrices W, and depending on the composite kernel used, a different σ parameter was to be tuned for each counterpart. All RBF kernel widths were tuned in the range $\sigma = \{10^{-3}, \ldots, 10^3\}$, the regularization parameter for SVM was varied in $C = \{10^0, \ldots, 10^3\}$, and the α parameter for the graph-based method was tuned in the range $\alpha = \{0.01, \ldots, 0.99\}$. A *one-against-one* multiclassification scheme was adopted in the case of SVMs.

C. Method Comparisons

In all cases, we selected the best free parameters with a reduced training set of labeled samples ($\{3, 5, 10\}$ samples per class) through 3-fold cross validation, and tested the results in the whole image. Table I shows the test results (averaged over 10 random realizations) for the composite kernels included in both the SVM and the graph-based semi-supervised classifiers.

Several conclusions can be obtained from Table I. First, the proposed graph-based method produces better classification results than the inductive SVM in all situations, and the average gain ($\sim 2\%$) remains almost constant as we increase the number of labeled samples for building the model, which confirms good robustness capabilities. It is also worth noting that the contextual classifier W_s alone produces good results, mainly due to the presence of large homogeneous classes and

[1]The Woodbury formula states the identity: $(C + AB)^{-1} = C^{-1} - C^{-1}A(I + BC^{-1}A)^{-1}BC^{-1}$, where C is an invertible $n \times n$ matrix, $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$.

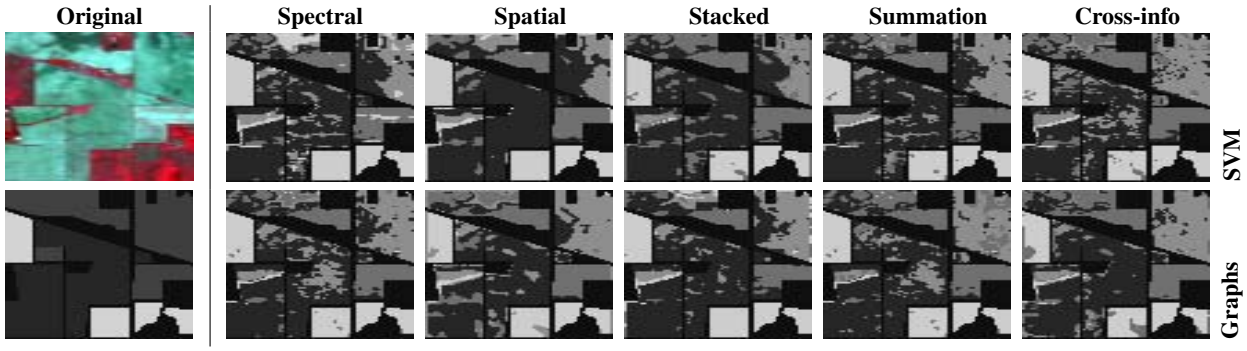

Fig. 2. **Left panel:** Original hyperspectral image: (top) three-channel false color composition ([RGB: {50,27,17}]) and (bottom) the true classification map. **Right panel:** Best thematic maps produced with the SVM-based (top) and the graph-based composite methods (bottom) with 5 training pixels by class.

TABLE I

OVERALL ACCURACY (OA[%]) AS A FUNCTION OF THE NUMBER OF LABELED SAMPLES PER CLASS†. AVERAGE RESULTS OVER 10 REALIZATIONS ARE SHOWN AS [SVM / GRAPH].

	No. training samples per class		
Composite kernel	*3*	*5*	*10*
Spectral	58.43/60.28	58.70/60.54	67.66/*69.17*
Spatial	51.77/52.42	55.96/57.69	65.49/66.60
Stacked	52.01/53.48	55.68/57.18	67.02/68.16
Summation	61.26/62.39	64.89/*66.86*	*69.43*/**71.32**
Cross-information	*64.57*/**66.09**	65.02/**67.13**	66.36/67.87

† Best results (bold) and second best (italics) are highlighted for each problem.

the high spatial resolution of the sensor. Note that the extracted textural features $\mathbf{x}_i^s$ contain spectral information to some extent as we computed them *per* spectral channel, thus they can be regarded as contextual or local spectral features. However, the accuracy is lower than the rest of methods, which demonstrates the relevance of the spectral information for hyperspectral image classification. With regard to the standard stacked approach, it is worth to note that poor results are obtained, probably due to the *curse of dimensionality* induced when working with such limited amount of labeled samples and high dimension (twice the rest of the methods). Furthermore, it is worth mentioning that all composite classifiers improved the results obtained by the usual spectral kernel, especially significant ($\sim$6%) when low number of labeled samples is used. These results confirm the validity of the presented framework.

Figure 2 shows the classified images with SVM and the graph-based method using different composite kernels for integrating the spatial and spectral information. Methods were trained with only 5 randomly selected training samples per class. The numerical results shown in Table I are confirmed by inspecting these classification maps, where better integration of the spatial information is achieved by the graph-based semi-supervised method, and smoother classification maps are obtained, more noticeable for the minority classes and class borders.

V. CONCLUSIONS

This paper proposed a graph-based method for hyperspectral image classification. The method takes advantage of both the high number of unlabeled samples present in the image, and the integration of contextual information. The obtained results suggest good robustness and accuracy to limited sized labeled datasets, as compared to the state-of-the-art inductive SVM. Next steps will consider the inclusion of more sophisticated spatial features and composite kernels.

ACKNOWLEDGMENTS

This research has been partially supported by the CICYT under Project DATASAT and by the "Grups Emergents" programme of Generalitat Valenciana under project HYPERCLASS/GV05/011.

REFERENCES

[1] J. A. Richards and Xiuping Jia, *Remote Sensing Digital Image Analysis. An Introduction*, Springer-Verlag, Berlin, Heidenberg, 3rd edition, 1999.

[2] N. M. Dempster, A. P. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, Jan 1977.

[3] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[4] M. I. Jordan, *Learning in Graphical Models*, MIT Press, Cambridge, Massachusetts and London, England, 1st edition, 1999.

[5] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems, NIPS'04*, Cambridge, MA, 2004. MIT press., Dec. 2004.

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

[7] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93–97, Jan 2006.

[8] C. K. I. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems, NIPS13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., Dec. 2001, pp. 682–688.

[9] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, Massachusetts and London, England, 1st edition, 2006.

[10] D. Landgrebe, "AVIRIS NW Indiana's Indian Pines 1992 data set," 1992, http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html.

[11] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *Proceedings of The 1999 Airborne Geoscience Workshop*, Feb. 1999.