Semi-Supervised Self Learning for Hyperspectral Image Classification

Inmaculada Dópido, *Student Member, IEEE*, Jun Li, Prashanth Reddy Marpu, Antonio Plaza, *Senior Member, IEEE*, José M. Bioucas Dias, *Member, IEEE*, and Jon Atli Benediktsson, *Fellow, IEEE*

Abstract

Remotely sensed hyperspectral imaging allows for the detailed analysis of the surface of the Earth using advanced imaging instruments which can produce high-dimensional images with hundreds of spectral bands. Supervised hyperspectral image classification is a difficult task due to the unbalance between the high dimensionality of the data and the limited availability of labeled training samples in real analysis scenarios. While the collection of labeled samples is generally difficult, expensive and time-consuming, unlabeled samples can be generated in a much easier way. This observation has fostered the idea of adopting semi-supervised learning techniques in hyperspectral image classification. The main assumption of such techniques is that the new (unlabeled) training samples can be obtained from a (limited) set of available labeled samples without significant effort/cost. In this paper, we develop a new approach for semi-supervised learning which adapts available active learning methods (in which a trained expert actively selects unlabeled samples) to a self-learning framework in which the machine learning algorithm itself selects the most useful and informative unlabeled samples for classification purposes. In this way, the labels of the selected pixels are estimated by the classifier itself, with the advantage that no extra cost is required for labeling the selected pixels using this machine-machine framework when compared with traditional machine-human active learning. The proposed approach is illustrated with two different classifiers: multinomial logistic regression (MLR) and a probabilistic pixel-wise support vector machine (SVM). Our experimental results with real hyperspectral images collected by the NASA Jet Propulsion Laboratory's Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) and the Reflective Optics Spectrographic Imaging System (ROSIS), indicate that the use of self learning represents an effective and promising strategy in the context of hyperspectral image classification.

Index Terms

Hyperspectral image classification, semi-supervised self learning, multinomial logistic regression (MLR), probabilistic support vector machine (SVM).

I. Dópido, J. Li and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres, E-10071, Spain. P. R. Marpu is with the Masdar Institute of Science and Technology, Abu Dhabi, 54224, United Arab Emirates. J. M. Bioucas-Dias is with the Telecommunications Institute, Instituto Superior Técnico, Lisbon, 1049-1, Portugal. J. A. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, Sæmundargata 2, 101 Reykjavik, Iceland.

I. INTRODUCTION

Remotely sensed hyperspectral image classification [1] takes advantage of the detailed information contained in each pixel (vector) of the hyperspectral image to generate thematic maps from detailed spectral signatures. A relevant challenge for supervised classification techniques (which assume prior knowledge in the form of class labels for different spectral signatures) is the limited availability of labeled training samples, since their collection generally involves expensive ground campaigns [2]. While the collection of labeled samples is generally difficult, expensive and time-consuming, unlabeled samples can be generated in a much easier way. This observation has fostered the idea of adopting semi-supervised learning techniques in hyperspectral image classification. The main assumption of such techniques is that new (unlabeled) training samples can be obtained from a (limited) set of available labeled samples without significant effort/cost [3].

The area of semi-supervised learning has experienced a significant evolution in terms of the adopted models, which comprise complex generative models [4]–[7], self learning models [8], [9], multi-view learning models [10], [11], transductive support vector machines (SVMs) [12], [13], and graph-based methods [14]. A survey of semi-supervised learning algorithms is available in [15]. Most of these algorithms use some type of regularization which encourages the fact that "similar" features are associated to the same class. The effect of such regularization is to push the boundaries between classes towards regions with low data density [16], where the usual strategy adopted first associates the vertices of a graph to the complete set of samples and then builds the regularizer depending on variables defined on the vertices. This trend has been successfully adopted in several recent remote sensing image classification studies. For instance, in [17] transductive SVMs (TSVMs) are used to gradually search a reliable separating hyperplane (in the kernel space) with a transductive process that incorporates both labeled and unlabeled samples in the training phase. In [18], a semi-supervised method is presented that exploits the wealth of unlabeled samples in the image, and naturally gives relative importance to the labeled ones through a graph-based methodology. In [19], kernels combining spectral-spatial information are constructed by applying spatial smoothing over the original hyperspectral data and then using composite kernels in graph-based classifiers. In [20], a semisupervised SVM is presented that exploits the wealth of unlabeled samples for regularizing the training kernel representation locally by means of cluster kernels. In [21], [22], a new semi-supervised approach is presented that exploits unlabeled training samples (selected by means of an active selection strategy based on the entropy of the samples). Here, unlabeled samples are used to improve the estimation of the class distributions, and the obtained classification is refined by using a spatial multi-level logistic prior. In [23], a novel context-sensitive semi-supervised SVM is presented that exploits the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase to improve the robustness to possible mislabeled training patterns. In [24], two semi-supervised one-class (SVM-based) approaches are presented in which the information provided by unlabeled samples present in the scene is used to improve classification accuracy and alleviate the problem of free-parameter selection. The first approach models data marginal distribution with the graph Laplacian built with both labeled and unlabeled samples. The second approach is a modification of the SVM cost function that penalizes more the errors made when classifying samples of the target class. In [25] a new method to combine labeled and unlabeled pixels to increase classification reliability and accuracy, thus addressing the sample selection bias problem, is presented and discussed. In [26], an SVM is trained with the linear combination of two kernels: a base kernel working only with labeled examples is deformed by a likelihood kernel encoding similarities between labeled and unlabeled examples, and then applied in the context of urban hyperspectral image classification. In [27], similar concepts to those addressed before are adopted using a neural network as the baseline classifier. In [28], a semi-automatic procedure to generate land cover maps from remote sensing images using active queries is presented and discussed.

In contrast to supervised classification, the aforementioned semi-supervised algorithms generally assume that a limited number of labeled samples are available *a priori*, and then enlarge the training set using unlabeled samples, thus allowing these approaches to address ill-posed problems. However, in order for this strategy to work, several requirements need to be met. First and foremost, the new (unlabeled) samples should be generated without significant cost/effort. Second, the number of unlabeled samples required in order for the semi-supervised classifier to perform properly should not be too high in order to avoid increasing computational complexity in the classification stage. In other words, as the number of unlabeled samples increases, it may be unbearable for the classifier to properly exploit all the available training samples due to computational issues. Further, if the unlabeled samples are not properly selected, these may confuse the classifier, thus introducing significant divergence or even reducing the classification accuracy obtained with the initial set of labeled samples. In order to address these issues, it is very important that the most highly informative unlabeled samples are identified in computationally efficient fashion, so that significant improvements in classification performance can be observed without the need to use a very high number of unlabeled samples.

In this work, we evaluate the feasibility of adapting available active learning techniques (in which a trained expert actively selects unlabeled samples) to a self-learning framework in which the machine learning algorithm itself selects the most useful unlabeled samples for classification purposes, with the ultimate goal of systematically achieving noticeable improvements in classification results with regards to those found by randomly selected training sets of the same size. In the literature, active learning techniques have been mainly exploited in a supervised context, *i.e.* a given supervised classifier is trained with the most representative training samples selected after a (machine-human) interaction process in which the samples are actively selected according to some criteria based on the considered classifier, and then the labels of those samples are assigned by a trained expert in fully supervised fashion [22], [29]–[33]. In this supervised context, samples with high uncertainty are generally preferred as they are usually more informative. At the same time, since the samples are labeled by a human expert, high confidence can be expected in the class label assignments. As a result, classic (supervised) active learning generally focuses on samples with high confidence at the human level and high uncertainty at the machine level.

In turn, in this work we adapt standard active learning methods into a self-learning scenario. The main idea is to obtain new (unlabeled) samples using machine-machine interaction instead of human supervision. Our first (machine) level –similar to the human level in classic (supervised) active learning– is used to infer a set of candidate unlabeled samples with high confidence. In our second (machine) level –similar to the machine level for supervised active learning– the machine learning algorithm itself automatically selects the samples with highest uncertainty from the obtained candidate set. As a result, in our proposed approach the classifier replaces the human expert. In other words, here we propose a novel two-step semi-supervised self learning approach:

- The first step infers a candidate set using a self learning strategy based on the available (labeled and unlabeled) training samples. Here, a spatial neighborhood criterion is used to derive new candidate samples as those which are spatially adjacent to the available (labeled) samples.
- The second step automatically selects (and labels) new samples from the candidate pool by assuming that those pixels which are spatially adjacent to a given class can be labeled with high confidence as belonging to the same class.

As a result, our proposed strategy relies on two main assumptions. The first assumption (global) is that training samples having the same spectral structure likely belonging to the same class. The second assumption (local) is that spatially neighboring pixels likely belong to the same class. As a result, our proposed approach naturally integrates the spatial and the spectral information in the semi-supervised classification process.

The remainder of the paper is organized as follows. Section II describes proposed approach for semi-supervised self learning. We illustrate the proposed approach with two probabilistic classifiers: multinomial logistic regression (MLR) and a probabilistic pixel-wise support vector machine (SVM), which are both shown to achieve significant improvements in classification accuracy resulting from its combination with the proposed semi-supervised self learning approach. Section III reports classification results using two real hyperspectral images collected by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) [34] and the Reflective Optics Spectrographic Imaging System (ROSIS) [35] imaging spectrometers. Finally, section IV concludes the paper with some remarks and hints at plausible future research lines.

II. Proposed Approach

First, we briefly define the notations used in this paper. Let $\mathcal{K} \equiv \{1, \ldots, K\}$ denote a set of K class labels, $\mathcal{S} \equiv \{1, \ldots, n\}$ a set of integers indexing the n pixels of an image, $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ an image of d-dimensional feature vectors, $\mathbf{y} \equiv (y_1, \ldots, y_n)$ an image of labels, $\mathcal{D}_l \equiv \{(y_{l_1}, \mathbf{x}_{l_1}), \ldots, (y_{l_n}, \mathbf{x}_{l_n})\}$ a set of labeled samples, l_n the number of labeled training samples, $\mathcal{Y}_l \equiv \{y_{l_1}, \ldots, y_{l_n}\}$ the set of labels in $\mathcal{D}_l$, $\mathcal{X}_l \equiv \{\mathbf{x}_{l_1}, \ldots, \mathbf{x}_{l_n}\}$ the set of feature vectors in $\mathcal{D}_l$, $\mathcal{D}_u \equiv \{\mathcal{X}_u, \mathcal{Y}_u\}$ a set of unlabeled samples, $\mathcal{X}_u \equiv \{\mathbf{x}_{u_1}, \ldots, \mathbf{x}_{u_n}\}$ the set of unlabeled feature vectors in $\mathcal{D}_u$, $\mathcal{Y}_u \equiv \{y_{u_1}, \ldots, y_{u_n}\}$ the set of labels associated with $\mathcal{X}_u$, and u_n the number of unlabeled samples. With this notation in mind, the proposed semi-supervised self learning approach consists of two main ingredients: semi-supervised learning and self learning, which are described next.

A. Semi-Supervised Learning

For the semi-supervised part of our approach, we use two different probabilistic classifiers to model the class posterior density. The first one is the MLR, which is formally given by [36]:

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) = \frac{\exp(\boldsymbol{\omega}^{(k)^T}\mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K}\exp(\boldsymbol{\omega}^{(k)T}\mathbf{h}(\mathbf{x}_i))}, \tag{1}$$

where $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), ..., h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed features; $\boldsymbol{\omega}$ are the regressors and $\boldsymbol{\omega} = [\boldsymbol{\omega}^{(1)^T}, ..., \boldsymbol{\omega}^{(K)^T}]^T$. Notice that, the function $\mathbf{h}$ may be linear, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, x_{i,1}, ..., x_{i,d}]^T$, where $x_{i,j}$ is the j-th component of $\mathbf{x}_i$; or nonlinear, *i.e.*, $\mathbf{h}(\mathbf{x}_i) = [1, K_{\mathbf{x}_i,\mathbf{x}_1}, ..., K_{\mathbf{x}_i,\mathbf{x}_l}]^T$, where $K_{\mathbf{x}_i,\mathbf{x}_j} = K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\cdot, \cdot)$ is some symmetric kernel function. Kernels have been largely used because they tend to improve the data separability in the transformed space. In this paper, we use a Gaussian Radial Basis Function (RBF) $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$ kernel, which is widely used in hyperspectral image classification [37]. We selected this kernel (after extensive experimentation using other kernels, including linear and polynomial kernels) because we empirically observed that it provided the best results. From now on, d denotes the dimension of $\mathbf{h}(\mathbf{x})$. Under the present setup, learning the class densities amounts to estimating the logistic regressors. Following the work in [38], [39], we can compute $\boldsymbol{\omega}$ by obtaining the maximum a posteriori (MAP) estimate:

$$\widehat{\boldsymbol{\omega}} \ = \arg\max_{\boldsymbol{\omega}} \quad \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{2}$$

where $p(\boldsymbol{\omega}) \propto \exp(-\lambda\|\boldsymbol{\omega}\|_1)$ is a Laplacian prior to promote sparsity and λ is a regularization parameter controlling the degree of sparseness of $\widehat{\boldsymbol{\omega}}$ in [38], [39]. In our previous work [39], it was shown that parameter λ is rather insensitive to the use of different datasets, and that there are many suboptimal values for this parameter which lead to very accurate estimation of parameter $\boldsymbol{\omega}$. In our experiments, we set $\lambda = 0.001$ as we have empirically found that this parameter setting provides very good performance [40]. Finally, $\ell(\boldsymbol{\omega})$ is the log-likelihood function over the training samples $\mathcal{D}_{l+u} \equiv \mathcal{D}_l + \mathcal{D}_u$, given by:

$$\ell(\boldsymbol{\omega}) \quad \equiv \quad \sum_{i=1}^{l_n+u_n} \log p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}). \tag{3}$$

As shown by Eq. (3), labeled and unlabeled samples are integrated to learn the regressors $\boldsymbol{\omega}$. The considered semi-supervised approach belongs to the family of self learning approaches, where the training set $\mathcal{D}_{l+u}$ is incremented under the following criterion. Let $\mathcal{D}_{\mathcal{N}(i)} \equiv \{(\widehat{y}_{i_1}, \mathbf{x}_{i_1}), \ldots, (\widehat{y}_{i_n}, \mathbf{x}_{i_n})\}$ be the set of neighboring samples of $(y_i, \mathbf{x}_i)$ for $i \in \{l_1, \ldots, l_n, u_1, \ldots, u_n\}$, where i_n is the number of samples in $\mathcal{D}_{\mathcal{N}(i)}$ and $\widehat{y}_{i_j}$ is the maximum a posteriori (MAP) estimate from the MLR classifier, with $i_j \in \{i_1, \ldots, i_n\}$. If $\widehat{y}_{i_j} = y_i$, we increment the unlabeled training set by adding $(\widehat{y}_{i_j}, \mathbf{x}_{i_j})$, *i.e.*, $\mathcal{D}_u = \{\mathcal{D}_u, (\widehat{y}_{i_j}, \mathbf{x}_{i_j})\}$. This increment is reasonable due to the following considerations. First, from a global viewpoint, samples which have the same spectral structure likely belong to the same class. Second, from a local viewpoint, it is very likely that two neighboring pixels also belong to the same class. Therefore, the newly included samples are reliable for learning the classifier. In this work, we run an iterative scheme to increment the training set as this strategy can refine the estimates and enlarge the neighborhood set such that the set of potential unlabeled training samples is increased.

It is important to mention that problem (2), although convex, is very difficult to compute because the term $\ell(\boldsymbol{\omega})$ is non-quadratic and the term $\log p(\boldsymbol{\omega})$ is non-smooth. The sparse MLR (SMLR) algorithm presented in [38] solves this problem with $O((d(K-1))^3)$ complexity. However, most hyperspectral data sets are beyond the reach of this algorithm as their analysis becomes unbearable when the number of classes increases. In order to address this issue, we take advantage of the logistic regression via variable splitting and augmented Lagrangian (LORSAL) algorithm [41] which allows replacing a difficult non-smooth convex problem with a sequence of quadratic plus diagonal l_2-l_1 problems with practical complexity of $O(d^2(K-1))$. Compared with the figure $O((d(K-1))^3)$ of the SMLR algorithm, the complexity reduction of $d(K-1)^2$ is quite significant [39], [41].

Finally, we have also used an alternative probabilistic classifier for the semi-supervised learning part of our methodology. This is the probabilistic SVM in [12], [42]. Other probabilistic classifiers could be used, but we have selected the SVM as a possible alternative to MLR since this classifier is already widely used to analyze hyperspectral data [17], [18], while the MLR has only recently emerged as a feasible technique for this purpose. It should be noted that the standard SVMs do not provide probability estimates for the individual classes. In order to get these estimates, pairwise coupling of binary probabilistic estimates is applied [42], [43], which has been applied for hyperspectral classifications [44].

B. Self learning

The proposed semi-supervised self learning approach is based on two steps. In the first step, a candidate set (based on labeled and unlabeled samples) is inferred using a self learning strategy based on spatial information, so that high confidence can be expected in the class labels of the obtained candidate set. This is similar to human interaction in classic (supervised) active learning, in which the class labels are known and given by an expert. In a second step, we run standard active learning algorithms on the previously derived candidate set, so that they are adapted to a self learning scenario to automatically (and intelligently) select the most informative samples from the candidate set. Here, the goal is to find the samples with higher uncertainty.

As a result, in the proposed semi-supervised self learning scheme our aim is to select the most informative samples without the need for human supervision. The class labels of the newly selected unlabeled training samples are predicted by the considered semi-supervised algorithm as mentioned in subsection II-A. Let $\mathcal{D}_c$ be the newly generated unlabeled training set at each iteration, which meets the criteria of the considered semi-supervised algorithm. Notice that the self learning step in the proposed approach leads to high confidence in the class labels of the newly generated set $\mathcal{D}_c$. Now we can run standard active learning algorithms over $\mathcal{D}_c$ to find the most informative set $\mathcal{D}_u$, *i.e.*, samples with high uncertainty, such that $\mathcal{D}_u \subseteq \mathcal{D}_c$. Due to the fact that we use discriminative classifiers and a self learning strategy for the semi-supervised algorithm, algorithms which focus on the boundaries between the classes are preferred. In our study, we use four different techniques to evaluate the proposed approach [26]: 1) margin sampling (MS), 2) breaking ties (BT), 3) modified breaking ties (MBT) [39], and 4) normalized entropy querying by bagging (nEQB) [30], in addition to random selection (RS) in which the new samples are randomly selected from the candidate set. In the following we briefly outline each method (for a more detailed description of these approaches, we refer to [22], [45]):

- The MS technique [45] samples the candidates lying within the margin by computing their distance to the hyperplane separating the classes. In other words, the MS minimizes the distance of the sample to the optimal separating hyperplane defined for class in a one-against-all setting for multiclass problems.

- The BT algorithm [46] relies on the smallest difference of the posterior probabilities for each sample. In a multi-class setting, the algorithm can be applied (independently of the number of classes available) by calculating the difference between the two highest probabilities. As a result, the algorithm finds the samples minimizing the distance between the first two most probable classes. In previous work [39], it has been shown that the BT criterion generally focuses on the boundaries comprising many samples, possibly disregarding boundaries with fewer samples.
- The MBT scheme [39] was originally proposed to include more diversity in the sampling process as compared to the BT approach. It finds the samples maximizing the probability of the largest class for each individual class. This method takes into account all the class boundaries by conducting the sampling in cyclic fashion, making sure that the MBT does not get trapped in any class whereas BT could be trapped in a single (complex) boundary.
- The nEQB approach [30] is a form of committee-based sampling algorithm that quantifies the uncertainty of a pixel by considering a committee of learners. Each member of the committee exploits different hypotheses about the classification problem and consequently labels the pixels in the pool of candidates. The algorithm then selects the samples showing maximal disagreement between the different classification models in the committee. Specifically, the nEQB approach uses bagging [47] to build the committee and Entropy maximization as the multiclass heuristic, which provides a measure that is then normalized in order to bound it with respect to the number of classes predicted by the committee and avoid hot spots of the value of uncertainty in regions where several classes overlap. The version of nEQB used in this work is the one implemented in[1].

At this point, it is important to emphasize that the aforementioned sampling algorithms have been used in this work for intelligently selecting the most useful candidate samples based on the available probabilistic information. As a result, spatial information is not directly addressed by these methods, but by the strategy adopted to generate the pool of candidate samples. Since spatial information is the main criterion adopted in this stage, there is a risk that the initial pool of candidate samples may smooth out broad areas in the scene. However, we emphasize that our proposed method for generating the pool of initial candidates is not exclusively spatial as we use the probabilistic information provided by spectral-based classifiers (such as MLR or probabilistic SVM) in order to assess the similarity between the previously selected samples and the new candidates. Hence, as we have experimentally observed, no significant smoothing effects happen in broad areas and good initial candidates are generally selected. It is also worth noting that, in this work, we use two classifiers with probabilistic output that are well-suited for the aforementioned algorithms (MLR and probabilistic SVM). However, the proposed approach can be adapted to any other probabilistic classifiers.

For illustrative purposes, Fig. 1 illustrates how spatial information can be adopted as a reasonable criterion to select unlabeled samples and prevent labeling errors in a semi-supervised classification process using a probabilistic

[1] http://code.google.com/p/altoolbox

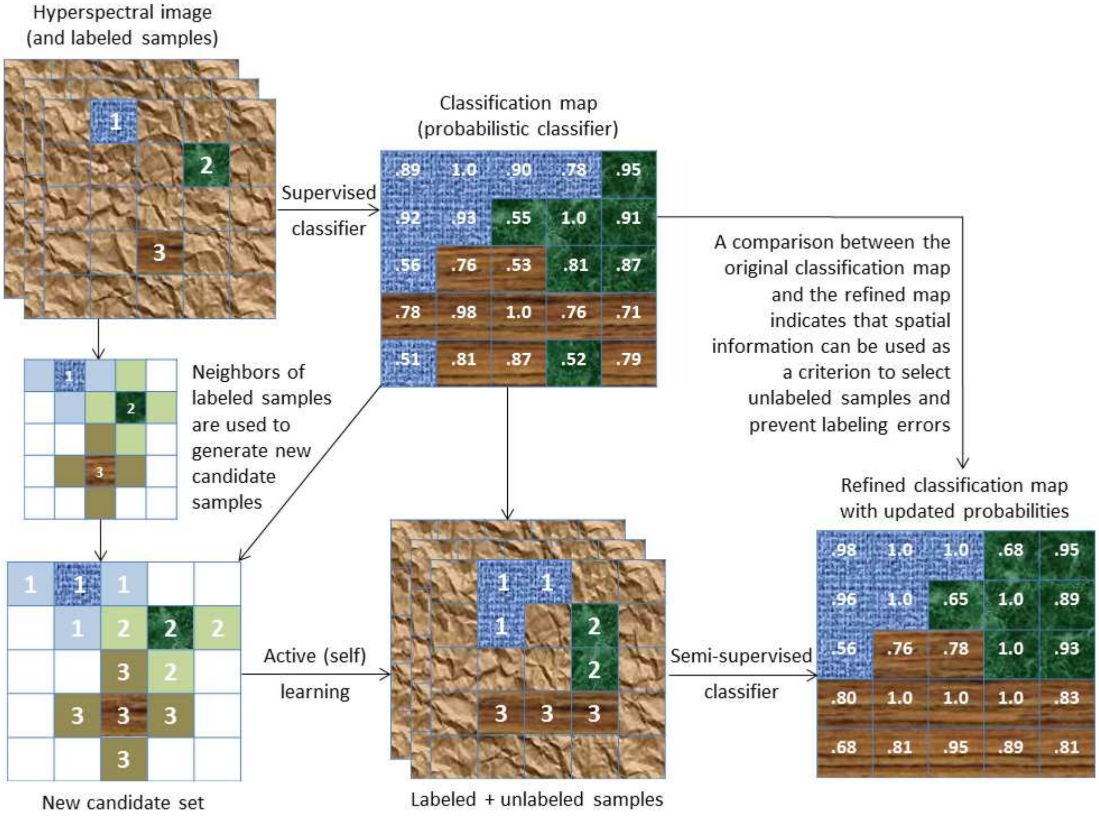


Fig. 1. A graphical example illustrating how spatial information can be used as a criterion for semi-supervised self learning in hyperspectral image classification.

classifier. As Fig. 1 shows, we use an iterative process to achieve the final classification results. First, we use a probabilistic classifier (in this work, the MLR or the probabilistic SVM) to produce a global classification map which contains the probability of each pixel to belong to each class in the considered hyperspectral image. Based on a local similarity assumption, we identify the neighbors of the labeled training samples (using first-order spatial connectivity) and then compute the candidate set $\mathcal{D}_c$ by analyzing the spectral similarity of the spatial neighbors with regards to the original labeled samples. This is done by analyzing the probabilistic output associated to each neighboring sample. In this way, the candidate set $\mathcal{D}_c$ is obtained based on spectral and spatial information and its samples are highly reliable. At the same time, it is expected that there may be redundant information in $\mathcal{D}_c$. In other words, some of the samples in the candidate set may not be useful for training the classifier as they may be too similar to the original labeled samples. This could introduce difficulties from the viewpoint of computational complexity. Therefore, after $\mathcal{D}_c$ is obtained, we run active learning algorithms on the candidate set in order to automatically select the most informative unlabeled training samples. Since the active learning algorithms are based on the available probabilistic information, they are adapted to a self-learning scenario and used to intelligently reduce possibly existing redundancies in the candidate set, thus obtaining a highly informative pool of training samples which ultimately contain only the most relevant samples for classification purposes. The newly obtained labeled and unlabeled training samples are finally used to retrain the classifier. The procedure is repeated in iterative fashion until a convergence criterion is met, for example, until a certain number of unlabeled training samples is obtained.

III. Experimental results

In this section, two real hyperspectral images are used to evaluate the proposed approach for semi-supervised self learning. In our experiments with the MLR and SVM classifiers, we apply the Gaussian RBF kernel to a normalized version of the considered hyperspectral data set[2]. We reiterate that the Gaussian RBF kernel was selected after extensive experimentation with other kernels. In all cases, the reported figures of overall accuracy (OA), average accuracy (AA), κ statistic, and class individual accuracies are obtained by averaging the results obtained after conducting 10 independent Monte Carlo runs with respect to the labeled training set $\mathcal{D}_l$ from the ground truth image, where the remaining samples are used for validation purposes. Finally, the optimal parameters C (parameter that controls the amount of penalty during the SVM optimization [12]) and σ (spread of the Gaussian RBF kernel) were chosen by 10-fold cross validation. These parameters are updated at each iteration.

In order to illustrate the good performance of the proposed approach, we use very small labeled training sets on purpose. As a result, the main difficulties that our proposed approach should circumvent can be summarized as follows. First and foremost, it is very difficult for supervised algorithms to provide good classification results as very little information is generally available about the class distribution. Poor generalization is also a risk when estimating class boundaries in scenarios dominated by limited training samples. Since our approach is semi-supervised, we take advantage of unlabeled samples in order to improve classification accuracy. However, if the number of labeled samples l is very small, increasing the number of unlabeled samples u could bias the learning process.

In order to analyze the aforementioned issues and provide a quantitative evaluation of our proposed approach with regards to the optimal case in which *true* active learning methods (i.e. those relying on the knowledge of the true labels of the selected samples) were used, we have implemented the following validation framework. Let $\mathcal{D}_{u_r}$ be a set of unlabeled samples for which true labels are available. These samples are included in the ground-truth associated to the hyperspectral image but are not used in the set of labeled samples used initially by the classifier. In order to evaluate the effectiveness of the proposed approach, we can effectively label these samples in $\mathcal{D}_{u_r}$ using their true (ground-truth) labels instead of estimating the labels by our proposed approach. Clearly, these samples will be favored over those selected by our proposed method which makes use of estimated labels. But it is interesting to quantify such an advantage (the lower it is, the better for our method). Following this rationale, the optimal case is that most samples in $\mathcal{D}_u$ have true labels available, which means that $\mathcal{D}_{u_r}$ contains most of the unlabeled samples in $\mathcal{D}_u$. In our experiments, we denote by l_r the number of unlabeled samples for which a true label is available in the ground-truth associated to the considered hyperspectral image. If $l_r = 0$, this means that the labels of all

[2] The normalization is simply given by $\mathbf{x}_i := \frac{\mathbf{x}i}{(\sqrt{\sum \|\mathbf{x}_i\|^2})}$, for $i = 1, \ldots, n$, where $\mathbf{x}_i$ is a spectral vector.

unlabeled samples are estimated by our proposed approach. If $l_r = u_r$, this means that true labels are available for all the samples in $\mathcal{D}_{u_r}$. Using this strategy, we can substantiate the deviation of our proposed approach with regards to the *optimal* case in which true labels for the selected samples are available. Typically, true labels will be only available for part of the samples as the considered hyperspectral data sets do not contain ground-truth information for all pixels. In this scenario, the *optimal* case comprises both true (whenever available) and estimated labels (the value of l_r is given in all experiments).

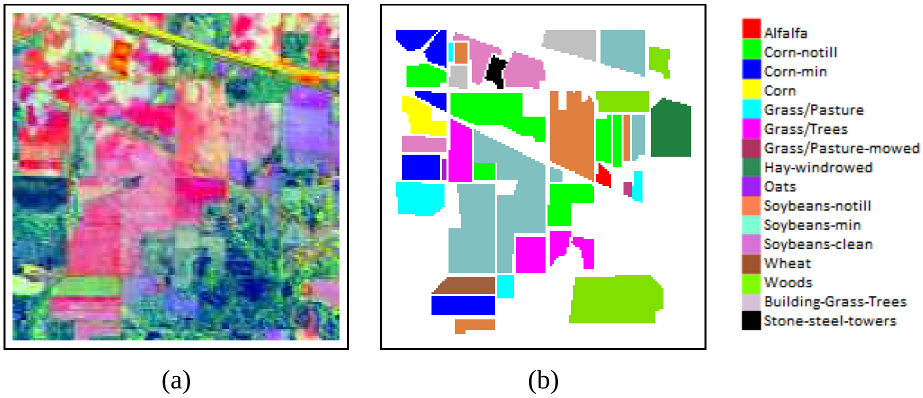
The remainder of this section is organized as follows. In subsection III-A we introduce the two datasets used for evaluation purposes in this work. In subsection III-B, we describe the experiments conducted using the first data set: AVIRIS Indian Pines. Finally, subsection III-C conducts experiments using a second data set: ROSIS Pavia University. In all cases, the results obtained by the supervised versions of the considered classifiers are also reported for comparative purposes.

A. Hyperspectral data sets

Two hyperspectral data sets collected by different instruments are used in our experiments:

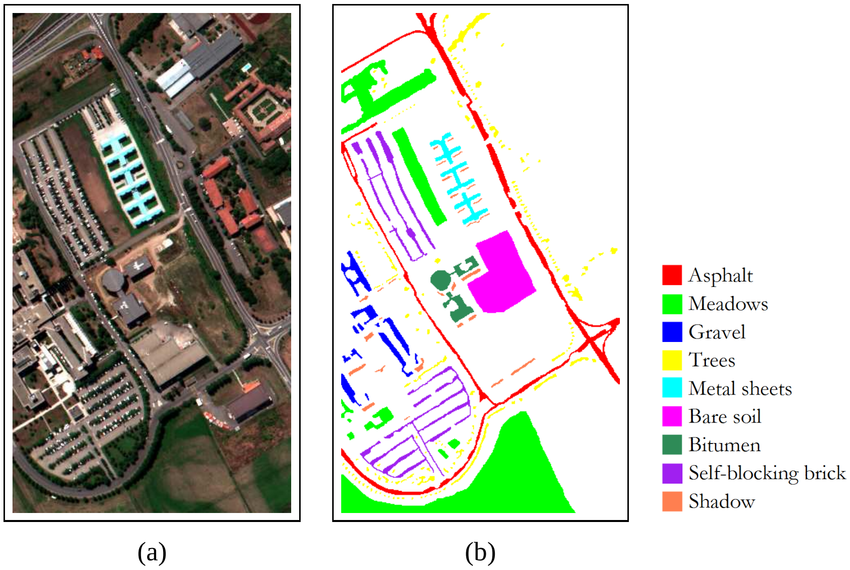
- The first hyperspectral image used in experiments was collected by the AVIRIS sensor over the Indian Pines region in Northwestern Indiana in 1992. This scene, with a size of 145 lines by 145 samples, was acquired over a mixed agricultural/forest area, early in the growing season. The scene comprises 220 spectral channels in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, moderate spatial resolution of 20 meters by pixel, and 16-bit radiometric resolution. After an initial screening, several spectral bands were removed from the data set due to noise and water absorption phenomena, leaving a total of 200 radiance channels to be used in the experiments. For illustrative purposes, Fig. 2(a) shows a false color composition of the AVIRIS Indian Pines scene, while Fig. 2(b) shows the ground-truth map available for the scene, displayed in the form of a class assignment for each labeled pixel, with 16 mutually exclusive ground-truth classes, in total, 10366 samples. These data, including ground-truth information, are available online[3], a fact which has made this scene a widely used benchmark for testing the accuracy of hyperspectral data classification algorithms. This scene constitutes a challenging classification problem due to the presence of mixed pixels in all available classes, and because of the unbalanced number of available labeled pixels per class.
- The second hyperspectral data set was collected by the ROSIS optical sensor over the urban area of the University of Pavia, Italy. The flight was operated by the Deutschen Zentrum for Luftund Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the European Union. The image size in pixels is 610×340, with very high spatial resolution of 1.3 meters per pixel. The number of data channels in the acquired image is 103 (with spectral range from 0.43 to 0.86 μm). Fig. 3(a) shows a false color composite of the image, while Fig. 3(b) shows nine ground-truth classes of interest, which comprise urban features, as well as soil and vegetation features.

[3] Available online: http://dynamo.ecn.purdue.edu/biehl/MultiSpec



(a) (b)

Fig. 2. (a) False color composition of the AVIRIS Indian Pines scene. (b) Ground truth-map containing 16 mutually exclusive land-cover classes (right).

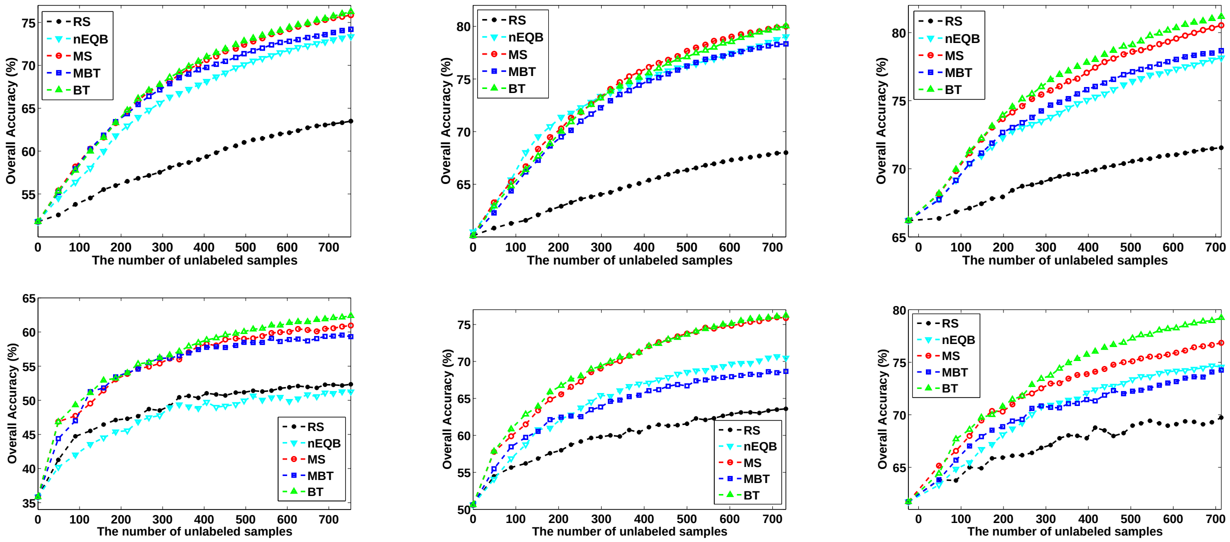


(a) (b)

Fig. 3. (a) False color composition of the ROSIS Pavia scene. (b) Ground truth-map containing 9 mutually exclusive land-cover classes.

B. Experiments with AVIRIS Indian Pines Data Set

In the first experiment we evaluated the impact of the number of unlabeled samples on the classification performance achieved by the two considered probabilistic classifiers using the AVIRIS Indian Pines data set in Fig. 2(a). Fig. 4 shows the OAs in classification accuracy as a function of the number of unlabeled samples obtained by the MLR (top) and probabilistic SVM (bottom) classifiers, respectively. The plots in Fig. 4, which were generated using estimated labels only, reveal clear advantages of using unlabeled samples for the proposed semi-supervised self learning approach when compared with the supervised algorithm alone. In all cases, the proposed strategy outperforms the corresponding supervised algorithm significantly, and the increase in performance is more relevant as the number of unlabeled samples increases. These unlabeled samples are automatically selected by the proposed approach, and represent no cost in terms of data collection or human supervision which are key aspects for self



(a) 5 labeled samples per class ($l_n = 80$) (b) 10 labeled samples per class ($l_n = 160$) (c) 15 labeled samples per class ($l_n = 240$)

Fig. 4. Overall classification accuracies (as a function of the number of unlabeled samples) obtained for the AVIRIS Indian Pines data set using the MLR (top) and probabilistic SVM (bottom) classifier, respectively. Estimated labels were used in all the experiments, i.e., $l_r = 0$.

learning. In Fig. 4 it can also be seen that using intelligent training sample selection algorithms such as MS, BT, MBT or nEQB greatly improved the obtained accuracies in comparison with simple random selection (RS). The results in Fig. 4 also reveal that BT outperformed other strategies in most cases, with MBT providing lower classification accuracies than BT. This is expected, as the candidate set $\mathcal{D}_c$ is more relevant when the samples are obtained from the class boundaries. Finally, it can also be observed that the MLR always performed better than the probabilistic SVM in terms of classification accuracies.

In order to show the classification results in more details, Table I shows the overall, average, individual classification accuracies (in percentage) and the κ statistic obtained by the supervised MLR and probabilistic SVM –trained using only 10 labeled samples per class– and by the proposed approach (based on the same classifier) using the four considered sample selection algorithms (executed using 30 iterations) in comparison with the *optimal* case for the same algorithms, in which true labels are used whenever available in the ground-truth. In all cases, we report the value of l_r to provide an indication of the number of true versus estimated labels used in the experiments. It is noticeable that, by including unlabeled samples, the classification results are significantly improved in all cases. Furthermore, it can be observed that the MLR classifier is more robust than the probabilistic SVM in our framework. For example, with $u_n = 750$ and BT sampling, only 2.24% difference in classification can be observed between the implementation using only estimated labels and the *optimal* case in which both true and estimated labels are considered. However, for the probabilistic SVM classifier the difference is 6.67%. Similar observation can be made for the other sampling algorithms considered in our experiments.

For illustrative purposes, Fig. 5 analyzes the convergence of our proposed approach by plotting the obtained classification accuracies for the AVIRIS Indian Pines scene as a function of the number of unlabeled samples, using

TABLE I

Overall, average, individual classification accuracies [%], and κ statistic obtained using the MLR and probabilistic classifiers when applied to the AVIRIS Indian Pines hyperspectral data set, with 10 labeled samples per class (160 samples in total) and $u_n = 750$ unlabeled training samples. l_r denotes the number of true labels available in $\mathcal{D}_u$ (used to implement an optimal version of each sampling algorithm). The standard deviations are also reported for each test.

MLR classifier

	Supervised	MS		BT		MBT		nEQB		RS	
		$l_r = 0$	$l_r = 683$	$l_r = 0$	$l_r = 668$	$l_r = 0$	$l_r = 646$	$l_r = 0$	$l_r = 603$	$l_r = 0$	$l_r = 747$
Alfalfa (54)	83.64±5.12	84.55±6.10	86.82±5.00	85.00±6.43	84.77±5.87	87.27±2.92	89.09±3.18	82.50±3.40	81.14±4.92	79.55±4.48	80.23±5.87
Corn-Notill (1434)	48.38±6.54	71.64±6.05	75.23±6.07	72.88±4.58	74.23±4.32	72.23±3.86	72.16±5.00	77.96±4.56	73.62±3.16	60.25±7.97	61.84±9.02
Corn-Min (834)	47.65±7.33	66.36±12.63	72.73±12.55	64.60±12.79	72.28±11.97	63.86±10.46	68.50±8.56	64.82±11.64	69.14±10.11	53.39±8.47	53.18±6.63
Corn (234)	70.63±9.43	85.76±8.13	85.49±5.74	87.54±5.86	88.04±4.53	92.23±2.45	90.67±6.48	86.38±6.30	80.40±13.18	66.29±16.34	71.74±12.94
Grass-Pasture (497)	75.42±7.35	85.50±4.93	87.37±7.43	85.48±5.32	88.67±5.57	87.08±6.30	89.45±5.96	79.49±8.35	83.78±7.28	81.79±5.15	83.59±6.71
Grass-Trees (747)	86.01±4.61	96.54±1.17	96.65±1.21	95.97±2.02	97.06±1.17	96.53±1.23	97.08±1.77	91.37±5.16	93.31±2.93	94.02±2.75	94.12±2.96
Grass-Pasture-Mowed (26)	88.12±6.88	93.75±6.62	87.50±5.89	93.75±5.47	86.88±8.56	89.38±7.25	90.63±5.31	90.63±4.42	88.12±9.97	85.00±6.72	86.25±5.74
Hay-Windrowed (489)	88.89±5.41	97.45±0.82	97.43±0.89	98.27±0.55	98.16±0.64	98.77±0.39	98.60±0.61	99.19±0.33	96.43±1.75	96.74±1.33	96.35±1.38
Oats (20)	98.00±4.22	96.00±11.35	95.00±10.80	97.00±11.35	96.00±6.99	99.00±3.16	99.00±3.16	97.00±6.75	96.00±6.99	99.00±4.22	98.00±4.22
Soybeans-Notill (968)	58.68±9.18	80.87±7.17	83.39±7.99	83.36±7.39	86.03±5.47	79.84±7.40	83.25±5.37	82.00±8.82	81.86±6.29	67.47±11.43	65.50±11.99
Soybeans-Min (2468)	44.85±10.85	72.51±4.70	74.49±7.29	70.14±5.28	72.76±5.72	62.58±8.20	65.36±5.96	68.04±5.60	69.29±5.43	50.81±12.98	54.02±8.23
Soybeans-Clean (614)	52.50±9.91	80.88±10.40	85.02±7.99	82.04±9.54	86.61±6.53	85.45±8.62	85.12±9.42	83.77±10.90	87.28±6.05	61.79±12.36	65.71±11.30
Wheat (212)	98.76±1.57	99.21±0.33	99.26±0.42	99.16±0.41	99.31±0.71	99.60±0.31	99.31±0.35	98.96±0.28	97.77±0.85	99.55±0.28	99.50±0.33
Woods (1294)	75.63±9.38	92.40±3.41	93.23±3.76	94.21±5.14	94.07±2.80	94.81±3.74	93.78±3.95	86.45±10.15	82.32±7.40	88.86±6.18	89.55±6.78
Bldg-Grass-Tree-Drives (380)	50.84±7.65	66.70±7.56	65.62±6.12	67.38±11.11	68.86±7.84	66.89±7.02	67.51±7.20	78.30±12.87	72.73±7.75	55.38±8.20	54.16±9.98
Stone-Steel-Towers (95)	79.88±8.22	82.94±7.91	84.12±10.90	80.94±7.75	83.29±9.79	91.06±3.19	90.82±3.91	79.53±5.74	85.06±10.23	77.53±8.55	78.00±7.73
OA	60.12± 3.08	80.00± 1.09	82.14±5.88	80.04± 1.28	82.28±6.12	78.34± 2.11	79.68±5.28	79.02±1.53	79.64±4.88	68.01± 3.04	69.28±2.63
AA	71.74± 1.54	84.57± 1.03	85.58±3.60	84.86± 1.53	86.06±3.86	85.41± 1.12	86.27±3.84	84.15±1.24	83.64±3.05	76.09± 1.76	76.98±1.46
κ	55.43± 3.20	77.31± 1.26	79.74±6.50	77.39± 1.45	79.93±6.79	75.59± 2.29	77.08±5.85	76.31±1.66	76.85±5.40	64.01± 3.30	65.39±2.86

Probabilistic SVM classifier

	Supervised	MS		BT		MBT		nEQB		RS	
		$l_r = 0$	$l_r = 695$	$l_r = 0$	$l_r = 717$	$l_r = 0$	$l_r = 649$	$l_r = 0$	$l_r = 701$	$l_r = 0$	$l_r = 740$
Alfalfa (54)	79.77±12.70	75.23±8.67	65.23±11.19	84.32±3.78	84.77±3.72	89.77±3.08	85.91±0.96	80.00±12.21	55.45±7.74	82.05±7.68	66.14±7.98
Corn-Notill (1434)	32.32±14.21	63.90±13.67	77.46±1.89	62.97±15.49	76.54±3.16	51.33±19.49	59.70±2.85	60.72±17.53	75.67±2.12	44.56 ±18.39	55.32±3.61
Corn-Min (834)	37.17±19.56	56.70±25.76	80.24±3.09	58.12±24.62	76.58±4.23	55.98±22.21	72.34±2.15	55.42±22.33	77.97±1.64	43.28±25.34	61.77±6.22
Corn (234)	68.62±10.32	87.95±3.29	89.24±1.73	82.10±13.80	86.38±3.52	81.03±13.28	84.06±2.72	86.38±4.02	86.34±4.26	72.50±13.19	85.49±2.64
Grass-Pasture (497)	77.19±7.29	87.54±7.09	91.21±3.01	89.16±6.02	93.37±1.35	88.17±6.40	93.24±1.23	82.40±6.03	90.60±2.99	85.73±5.77	89.45±2.47
Grass-Trees (747)	65.36±14.50	93.96±2.75	91.90±2.82	95.29±2.62	94.02±2.53	90.39±4.96	88.66±2.22	87.72±7.29	92.29±2.42	88.36±5.99	82.63±4.95
Grass-Pasture-Mowed (26)	90.63±6.75	90.00±7.34	93.75±2.95	92.50±4.93	95.00±3.95	90.00±4.37	93.75±2.95	89.38±6.62	93.13±1.98	87.50 ±8.33	93.13±1.98
Hay-Windrowed (489)	78.06±8.12	95.80±1.75	97.70±0.60	97.89±0.89	98.10±0.46	98.52±1.19	98.27±0.43	93.26±3.95	97.93±1.38	93.49±4.39	97.24±0.67
Oats (20)	97.00±6.75	93.00±9.49	100.00	93.00±6.75	99.00±3.16	95.00±12.69	100.00	98.00±4.22	97.00±4.83	95.00±7.07	100.00
Soybeans-Notill (968)	49.42±18.23	80.96±7.68	88.68±3.02	82.03±8.88	91.39±2.14	72.13±24.41	87.21±2.60	71.34±27.13	85.75±2.73	65.10±18.05	84.38±3.66
Soybeans-Min (2468)	33.90±12.83	65.50±12.51	65.98±2.15	63.36±15.50	68.60±2.36	50.16±12.02	53.59±5.69	58.33±23.25	62.12±2.40	50.44±15.80	44.10±13.02
Soybeans-Clean (614)	43.31±12.88	77.90±10.32	90.79±2.09	81.42±11.08	91.42±1.24	63.00±17.91	84.39±7.02	76.71±13.10	92.04±1.71	52.91±8.92	61.94±11.52
Wheat (212)	93.61±3.96	98.37±1.07	97.82±1.40	98.66±0.81	97.52±1.34	98.22±2.40	99.01±0.52	97.28±0.91	97.48±1.00	97.38±1.51	97.62±0.45
Woods (1294)	72.39±15.02	89.24±6.07	93.90±1.92	92.94±4.58	97.34±0.40	92.10±6.25	97.81±0.55	77.73±10.45	90.73±2.72	89.36±6.60	96.94±0.74
Bldg-Grass-Tree-Drives (380)	47.84±14.90	68.11±14.08	64.95±5.97	66.81±16.28	61.97±3.04	65.46±8.72	58.51±4.37	72.54±12.16	64.86±5.76	42.35±13.44	40.00±7.62
Stone-Steel-Towers (95)	86.35±10.26	96.35±4.72	93.53±3.65	93.18±5.62	90.82±3.79	88.35±9.87	83.18±2.29	94.47±5.82	87.41±4.11	90.35±4.95	84.35±2.54
OA	50.61±5.34	75.87±3.44	81.82±7.54	76.23±5.40	82.91±0.75	68.66±5.35	75.26±1.39	70.47±5.24	79.69±0.62	63.59±5.59	68.40±2.85
AA	65.93±2.99	82.53±2.03	86.40±4.47	83.36±2.15	87.68± 0.67	79.35±2.16	83.73±0.79	80.10±2.43	84.17±0.65	73.77±2.18	77.53±0.96
κ	45.14±5.35	72.76±3.76	79.49±8.26	73.18±5.81	80.71±0.83	64.90±5.75	72.39±1.49	66.79±5.65	77.14±0.67	59.13±5.68	64.73±2.99

only 5 labeled samples per class (in total 80 labeled samples) for the MLR classifier with BT sampling approach. In the figure, we report the case in which all unlabeled samples are estimated by the proposed approach (i.e., $l_r = 0$) and also the optimal case in which true labels are used whenever possible (i.e., $l_r = u_r$). As can be seen in Fig. 5, the proposed approach achieved good performance when compared with the optimal case, with a difference of about 5% in classification accuracy when 3500 training samples were used.

Finally, Fig. 6 shows some of the classification maps obtained by the MLR and probabilistic SVM classifiers for the AVIRIS Indian Pines scene. These classification maps correspond to one of the 10 Monte-Carlo runs that were averaged in order to generate the classification scores reported in Table I. The advantages obtained by adopting a

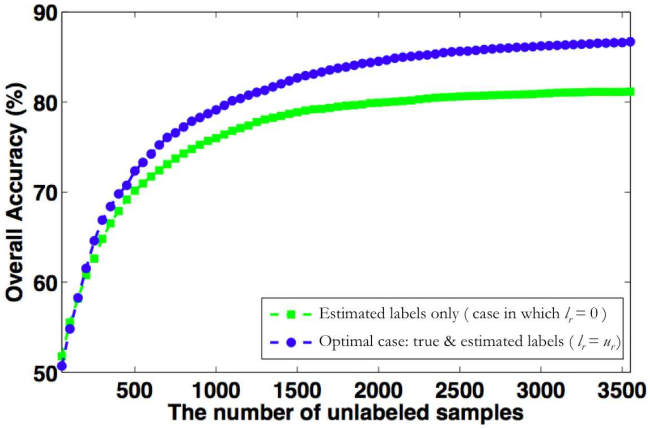
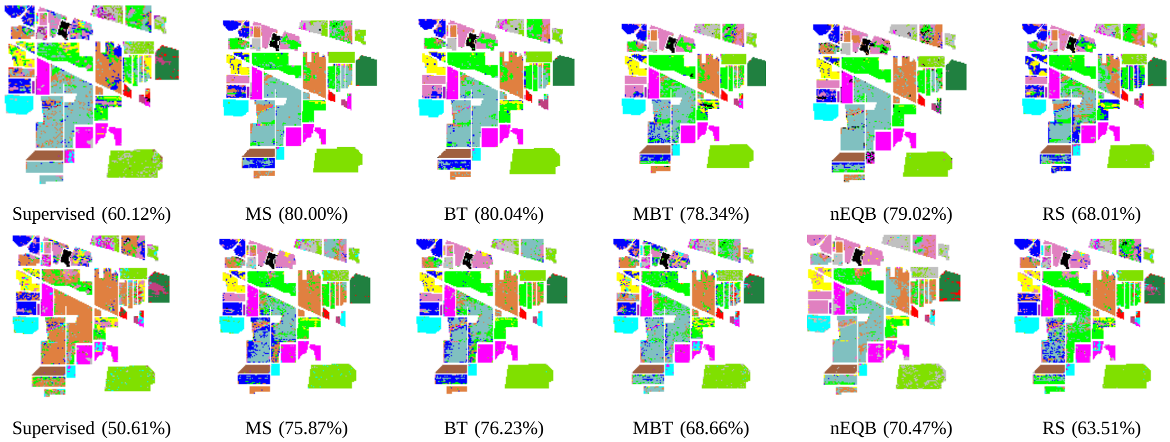


Fig. 5. Overall classification accuracies (as a function of the number of unlabeled samples) obtained for the AVIRIS Indian Pines data set using the MLR classifier with BT sampling by using 5 labeled samples per class (in total 80 samples). Two cases are displayed: the one in which all unlabeled samples are estimated by the proposed approach (i.e., $l_r = 0$) and the optimal case, in which true labels are used whenever possible (i.e., $l_r = u_r$).


Supervised (60.12%) MS (80.00%) BT (80.04%) MBT (78.34%) nEQB (79.02%) RS (68.01%)

Supervised (50.61%) MS (75.87%) BT (76.23%) MBT (68.66%) nEQB (70.47%) RS (63.51%)

Fig. 6. Classification maps and overall classification accuracies (in the parentheses) obtained after applying the MLR (top) and probabilistic SVM (bottom) classifiers to the AVIRIS Indian Pines data set by using 10 labeled training samples and 750 unlabeled samples, *i.e.*, $l_n = 160, u_n = 750$ and $l_r = 0$.

semi-supervised learning approach with regards to the corresponding supervised case can be clearly appreciated in the classification maps displayed in Fig. 6, which also report the classification OAs obtained for each method in the parentheses.

C. Experiments with ROSIS Pavia University Data Set

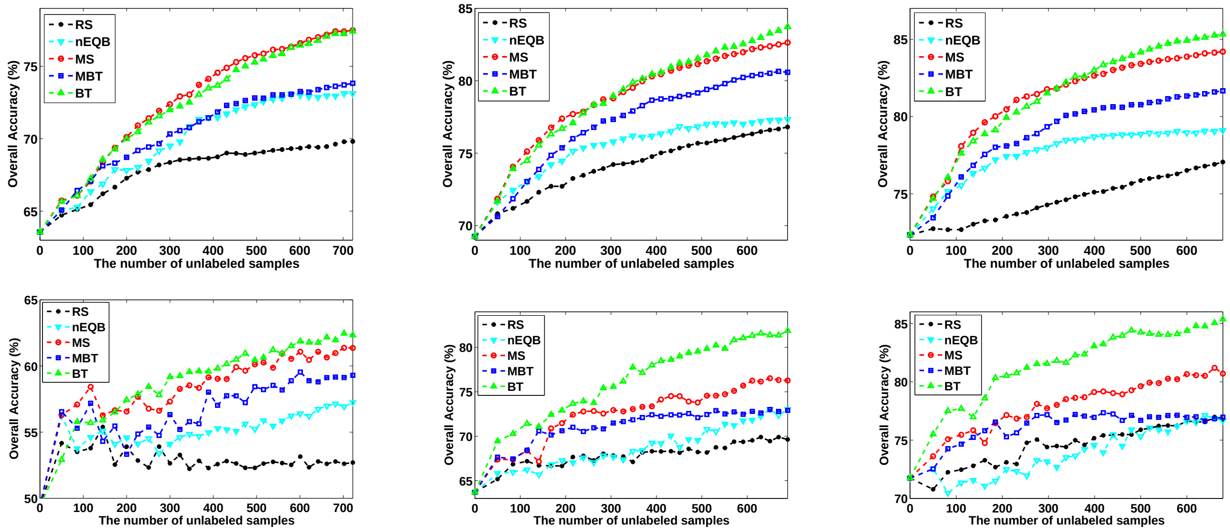
In this subsection we perform a set o experiments to evaluate the proposed approach using the ROSIS University of Pavia dataset. This problem represents a very challenging classification scenario dominated by complex urban classes and nested regions. First, Fig. 7 shows how the OA results increase as the number of unlabeled samples increases, indicating again clear advantages of using unlabeled samples for the proposed semi-supervised self learning approach in comparison with the supervised case. In this experiment, the four considered sample selection approaches (MS, BT, MBT and nEQB) perform similarly and slightly better than simple random selection. For instance, when $l_n = 45$ labeled samples were used, the performance increase observed after including $u_n = 700$ unlabeled samples with regards to the supervised case was 13.93% (for the MS), 13.86% (for the BT), 10.27% (for the MBT) and 9.56% (for the nEQB). These results confirm our introspection that the proposed semi-supervised self learning approach can greatly assist in improving the results obtained by different supervised classifiers based on limited training samples.

Furthermore, Table II shows the overall, average, individual classification accuracies (in percentage) and the κ statistic using only 10 labeled samples per class, in total, $l_n = 90$ samples and $u_n = 700$ unlabeled samples for the semi-supervised cases in comparison with the *optimal* case, in which true labels are used whenever available in the ground-truth. In all cases, we provide the value of l_r to provide an indication of the number of true versus estimated labels used in the experiments. It can be observed from Table II that the proposed approach is quite robust as it achieved classification results which are very similar to those found by the optimal case. For example, by using the BT sampling algorithm the proposed aproach obtained an OA of 83.73% which is almost the same as the one obtained the optimal case, which achieved an OA of 84.07% by using true labels whenever possible. This observation is confirmed by Fig. 8, which plots the classification accuracy obtained (as a function of the number of unlabeled samples) for a case in which 100 labeled training samples per class were used (a total 900 samples) for the MLR classifier with BT sampling approach. In the figure, we report the case in which all unlabeled samples are estimated by the proposed approach (i.e., $l_r = 0$) and also the optimal case in which true labels are used whenever possible (i.e., $l_r = u_r$). Although in this experiment the number of initial labeled samples is significant, it is remarkable that the results obtained by the proposed approach using only estimated labels are almost the same than those obtained with the optimal version using true labels, which means that the unlabeled training samples estimated by the proposed approach are highly reliable in this experiment.

For illustrative purposes, Fig. 9 shows some of the classification maps obtained by the MLR (top) and probabilistic SVM (bottom) classifiers for the ROSIS Pavia University dataset, which corresponds to one of the 10 Monte-Carlo runs that were averaged in order to generate the classification scores reported in Table II.

IV. Conclusions and Future Work

In this paper, we have developed a new approach for semi-supervised classification of hyperspectral images in which unlabeled samples are intelligently selected using a self learning approach. Specifically, we automatically select the most informative unlabeled training samples with the ultimate goal of improving classification results obtained using randomly selected training samples. In our semi-supervised context, the labels of the selected training samples are estimated by the classifier itself, with the advantage that no extra cost is required for labeling the selected samples when compared to classic (supervised) active learning. Our experimental results, conducted using two different classifiers: sparse multinomial logistic regression (MLR) and probabilistic support vector machine (SVM),



(a) $l_n = 45$ (b) $l_n = 90$ (c) $l_n = 135$

Fig. 7. Overall classification accuracies (as a function of the number of unlabeled samples) obtained for the ROSIS Pavia University data set using the MLR (top) and probabilistic SVM (bottom) classifier, respectively. Estimated labels were used in all the experiments, i.e., $l_r = 0$.

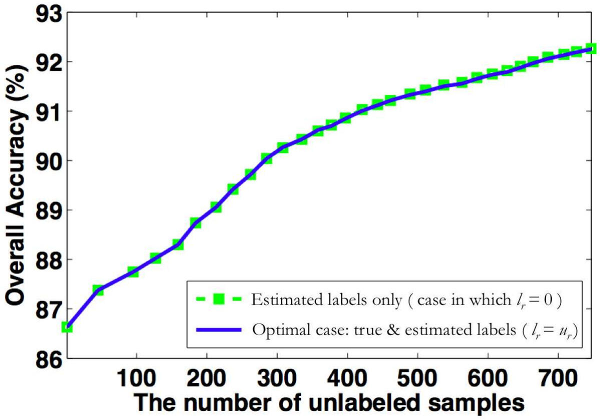


Fig. 8. Overall classification accuracies (as a function of the number of unlabeled samples) obtained for the ROSIS Pavia University data set using the MLR classifier with BT sampling by using 100 labeled samples per class (in total 900 samples). Two cases are displayed: the one in which all unlabeled samples are estimated by the proposed approach (i.e., $l_r = 0$) and the optimal case, in which true labels are used whenever possible (i.e., $l_r = u_r$).

indicate that the proposed approach can greatly increase the classification accuracies obtained in the supervised case through the incorporation of unlabeled samples which can be obtained with very little cost and effort. The obtained results have been compared to the *optimal* case in which true labels are used, and the differences observed when using estimated samples by our proposed approach were always quite small. This is a good quantitative indicator of the good performance achieved by our proposed approach, which has been illustrated using two hyperspectral scenes collected by different instruments. In future work, we are planning on combining the proposed approach with other

TABLE II

OVERALL, AVERAGE, INDIVIDUAL CLASSIFICATION ACCURACIES [%], AND κ STATISTIC OBTAINED USING THE MLR AND PROBABILISTIC SVM CLASSIFIERS WHEN APPLIED TO THE ROSIS UNIVERSITY OF PAVIA HYPERSPECTRAL DATA SET BY USING 10 LABELED SAMPLES PER CLASS (IN TOTAL 90 SAMPLES) AND $u_n = 700$ UNLABELED TRAINING SAMPLES. l_r DENOTES THE NUMBER OF TRUE LABELS AVAILABLE IN $\mathcal{D}_u$ (USED TO IMPLEMENT AN OPTIMAL VERSION OF EACH SAMPLING ALGORITHM). THE STANDARD DEVIATIONS ARE ALSO REPORTED FOR EACH TEST.

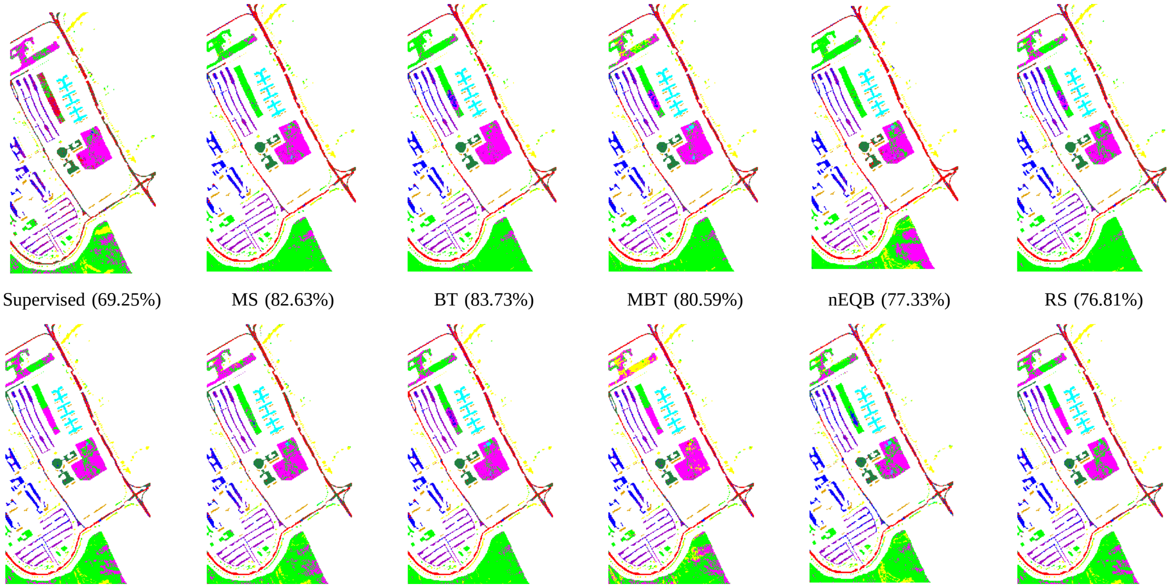
MLR classifier

	Supervised	MS		BT		MBT		nEQB		RS	
		$l_r = 0$	$l_r = 443$	$l_r = 0$	$l_r = 356$	$l_r = 0$	$l_r = 412$	$l_r = 0$	$l_r = 365$	$l_r = 0$	$l_r = 558$
Asphalt (6631)	64.05±7.34	74.57±7.48	75.41±6.01	72.62±4.97	74.72±6.14	71.43±4.75	71.54±4.57	72.91±7.37	72.40±7.63	66.40±7.55	68.85±6.03
Meadows (18649)	63.15±7.27	80.71±5.71	83.92±2.84	83.33±4.49	84.62±2.24	77.35±3.56	80.57±4.67	74.08±6.95	81.18±4.75	76.23±6.93	81.01±4.62
Gravel (2099)	66.28±9.21	80.05±9.35	80.33±8.86	82.07±9.31	81.09±9.00	79.24±9.19	81.55±7.65	81.86±7.50	82.28±7.59	73.44±7.55	77.21±10.77
Trees (3064)	84.74±11.11	84.88±9.97	85.47±8.66	88.07±8.87	83.32±9.45	94.41±3.58	88.45±6.82	91.46±4.05	85.64±8.83	82.97±9.16	85.04±5.01
Metal Sheets (1345)	98.64±0.60	99.49±0.44	98.68±1.24	99.29±0.36	99.32±0.47	99.77±0.21	99.70±0.29	98.79±0.82	98.85±0.82	99.03±0.48	98.87±0.54
Bare Soil (5029)	69.54±8.79	89.61±3.22	89.74±4.15	89.59±3.86	88.93±4.62	82.45±6.58	86.31±4.13	71.29±6.26	82.99±5.29	76.84±11.58	82.57±8.73
Bitumen (1330)	87.70±3.31	95.29±1.66	93.93±2.18	96.17±0.99	95.39±2.02	96.53±1.18	96.52±1.17	85.39±7.60	90.26±5.56	92.07±3.52	93.27±3.79
Self-Blocking Bricks (3682)	73.22±7.57	82.19±7.02	81.38±5.06	80.99±7.09	80.48±4.46	82.87±6.76	77.83±7.48	79.29±9.17	80.16±8.40	76.08±7.85	75.74±10.34
Shadow (947)	98.44±1.91	98.90±2.56	97.88±3.33	99.12±1.79	98.60±1.86	98.98±1.88	99.30±0.49	99.88±0.15	99.55±0.72	98.85±2.28	99.52±0.32
OA	69.25±3.75	82.63±2.55	84.08±0.98	83.73±1.86	84.07±1.52	80.59±1.38	81.72±1.96	77.33±3.80	81.5±51.54	76.81±3.38	80.30±2.54
AA	78.42±1.75	87.30±1.28	87.41±0.76	87.92±1.13	87.39±1.25	87.00±0.77	86.86±0.73	83.88±2.30	85.92±0.96	82.43±1.60	84.68±1.39
κ	61.69±4.01	77.78±3.08	79.50±1.14	79.12±2.23	79.45±1.90	75.44±1.61	76.70±2.27	71.27±4.53	76.36±1.74	70.45±3.86	74.75±3.03

Probabilistic SVM classifier

	Supervised	MS		BT		MBT		nEQB		RS	
		$l_r = 0$	$l_r = 454$	$l_r = 0$	$l_r = 382$	$l_r = 0$	$l_r = 324$	$l_r = 0$	$l_r = 337$	$l_r = 0$	$l_r = 557$
Asphalt (6631)	60.43± 8.23	75.71± 12.63	76.25±9.46	74.38± 7.89	72.82±8.00	72.27± 3.13	70.68±3.72	70.16±8.34	70.01±9.05	61.14± 7.06	61.52±5.37
Meadows (18649)	54.36± 9.43	68.35± 7.10	69.95±6.72	79.57± 8.28	78.96±9.21	63.53± 11.57	64.61±12.56	66.16±13.17	66.62±7.29	62.35± 12.01	65.95±12.93
Gravel (2099)	62.23± 10.33	75.72± 14.03	75.30±11.18	80.01± 9.72	80.25±8.33	72.58± 12.90	75.14±9.40	80.82±9.35	80.05±9.60	70.97± 12.39	70.61±10.49
Trees (3064)	90.75± 7.19	88.77± 9.31	88.94±5.97	85.14± 8.41	87.80±7.00	92.31± 7.43	92.01±5.25	89.52±7.01	89.43±7.24	89.90± 5.20	85.15±7.60
Metal Sheets (1345)	96.68± 5.68	99.91± 0.08	99.90±0.11	99.84± 0.10	99.83±0.12	99.55± 0.33	99.64±0.34	99.86±0.10	99.86±0.10	99.71± 0.12	99.69±0.19
Bare Soil (5029)	62.74± 19.59	87.47± 4.81	88.08±5.22	88.60± 3.92	90.26±2.66	77.89± 12.67	78.95±9.39	73.03±10.42	76.96±7.93	75.01± 14.21	73.05±23.65
Bitumen (1330)	89.90± 5.14	92.47± 4.12	93.21±3.22	94.38± 3.55	95.67±1.97	94.84± 1.39	95.56±1.67	90.33±3.92	90.79±3.22	92.93± 4.66	92.97±3.67
Self-Blocking Bricks (3682)	66.50± 8.44	71.64± 18.83	75.52±9.44	80.89± 8.04	80.39±8.07	74.95± 24.57	81.00±7.17	72.01±5.71	71.82±7.16	70.04± 12.33	72.23±13.42
Shadow (947)	99.26± 1.62	99.77± 0.19	99.73±0.52	97.98± 2.74	98.54±1.43	97.56± 2.68	99.11±2.16	99.90±0.01	99.88±0.14	99.31± 1.47	99.77±0.26
OA	63.68± 4.97	76.27± 4.68	77.47±3.26	81.85± 4.44	81.95±4.68	72.90± 5.42	73.93±4.95	73.61±3.89	77.02±5.87	69.63± 5.25	70.88±5.20
AA	75.76± 3.74	84.42± 2.22	85.21±1.47	86.75± 1.55	87.17±1.45	82.83± 2.43	84.08±1.46	82.42±1.98	82.82±2.03	80.15± 2.92	80.11±3.46
κ	55.48± 5.55	70.40± 5.26	71.79±3.71	76.89± 5.19	76.94±5.42	66.60± 5.85	67.86±5.40	66.44±6.26	67.08±4.40	62.46± 5.57	63.70±5.70

probabilistic classifiers. We are also considering the use of expectation-maximization as a form of self learning [15]. Although in this manuscript we focused our experiments on hyperspectral data, the proposed approach can also be applied to other types of remote sensing data, such as multispectral data sets. In fact, since the dimensionality of the considered hyperspectral data sets is quite high, the proposed approach could greatly benefit from the use of feature extraction/selection methods prior to classification in order to make the proposed less sensitive to the Hughes effect [48] and to the possibly very limited initial availability of training samples. This research topic also deserves future attention. Another interesting future research line is to adapt our proposed sample selection strategy (which is based on the selection of individual pixels) to the selection and labeling of spatial subregions or boxes within the image, which could be beneficial in certain applications. Finally, another important research topic deserving future attention is the inclusion of a cost associated to the labels generated by the proposed algorithm. This may allow a better evaluation of the training samples actively selected by our proposed approach.



Supervised (69.25%) MS (82.63%) BT (83.73%) MBT (80.59%) nEQB (77.33%) RS (76.81%)

Supervised (63.68%) MS (76.27%) BT (81.85%) MBT (72.90%) nEQB (77.02%) RS (69.63%)

Fig. 9. Classification maps and overall classification accuracies (in the parentheses) obtained after applying the MLR (top) and probabilistic SVM (bottom) classifier to the ROSIS Pavia University data set (in all cases, $l_n = 90$ and $l_r = 0$).

Acknowledgment

This work has been supported by the European Community's Marie Curie Research Training Networks Programme under contract MRTN-CT-2006-035927, Hyperspectral Imaging Network (HYPER-I-NET). Funding from the Portuguese Science and Technology Foundation, project PEst-OE/EEI/LA0008/2011, and from the Spanish Ministry of Science and Innovation (CEOS-SPAIN project, reference AYA2011-29334-C02-02) is also gratefully acknowledged. This research was also supported in part by the Icelandic Research Fund and the University of Iceland Research Fund. The authors would like to thank Prof. D. Landgrebe for making the AVIRIS Indian Pines hyperspectral data set available to the community and Prof. P. Gamba for providing the ROSIS data over Pavia, Italy, along with the training and test set. Last but not least, the authors would like to take this opportunity to gratefully thank the Editors and the four Anonymous Reviewers for their outstanding comments and suggestions, which greatly helped us to improve the technical quality and presentation of this work.

References

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: John Wiley, 2003.

[2] F. Bovolo, L. Bruzzone, and L. Carline, "A novel technique for subpixel image classification based on support vection machine," *IEEE Transactions on Image Processing*, vol. 19, pp. 2983–2999, 2010.

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

[4] S. Baluja, "Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data," in *Neural Information Procesing systems (NIPS '98)*, 1998.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training." Morgan Kaufmann Publishers, 1998, pp. 92–100.

[6] T. Mitchell, “The role of unlabeled data in supervised learning,” in *Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999.

[7] A. Fujino, N. Ueda, and K. Saito, “A hybrid generative/discriminative approach to semi-supervised classifer degin,” in *AAAI’05 Proceedings of the 20th national coference on Artificial intelligence*, vol. 2, 2005.

[8] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ser. ACL’95, 1995, pp. 189–196.

[9] C. Rosenberg, M. Hebert, and H. Schneiderman, “Semi-supervised self-training of object detection models,” in *Seventh IEEE Workshop on Applications of Computer Vision*, January 2005.

[10] I. Cowan, T. G, and V. R. D. Sa, “Learning classification with unlabeled data,” 1994.

[11] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel, “Efficient co-regularised least squares regression,” in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML’06, 2006, pp. 137–144.

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

[13] T. Joachims, “Transductive inference for text classification using support vector machines,” in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML’99, 1999, pp. 200–209.

[14] A. Blum and S. Chawla, “Learning from labeled and unlabeled data using graph mincuts,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML’01, 2001, pp. 19–26.

[15] X. Zhu, “Semi-supervised learning literature survey,” Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[16] O. Chapelle, M. Chi, and A. Zien, “A continuation method for semi-supervised SVMs,” in *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press, 2006, pp. 185–192.

[17] L. Bruzzone, M. Chi, and M. Marconcini, “A novel transductive SVM for the semisupervised classification of remote-sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 11, pp. 3363–3373, 2006.

[18] G. Camps-Valls, T. Bandos, and D. Zhou, “Semi-supervised graph-based hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3044–3054, Oct 2007.

[19] S. Velasco-Forero and V. Manian, “Improving hyperspectral image classification using spatial preprocessing,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, pp. 297–301, 2009.

[20] D. Tuia and G. Camps-Valls, “Semisupervised remote sensing image classification with cluster kernels,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 224 –228, april 2009.

[21] J. Li, J. Bioucas-Dias, and A. Plaza, “Semi-supervised hyperspectral classification,” in *First IEEE GRSS Workshop on Hyperspectral Image and Signal Processing*, 2009.

[22] ——, “Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, pp. 4085–4098, 2010.

[23] L. Bruzzone and C. Persello, “A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2142 –2154, 2009.

[24] J. Muñoz Marí, F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camp-Valls, “Semisupervised one-class support vector machines for classification of remote sensing data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 8, pp. 3188 –3197, 2010.

[25] L. Gomez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, “Mean map kernel methods for semisupervised cloud classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 207 –220, jan. 2010.

[26] D. Tuia and G. Camps-Valls, “Urban image classification with semisupervised multiscale cluster kernels,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, pp. 65 –74, march 2011.

[27] F. Ratle, G. Camps-Valls, and J. Weston, “Semisupervised neural networks for efficient hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2271 –2282, may 2010.

[28] M. nandoz Marí, J., D. Tuia, and G. Camps-Valls, “Semisupervised classification of remote sensing images with active queries,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1 –12, 2012.

[29] S. Rajan, J. Ghosh, and M. M. Crawford, “An active learning approach to hyperspectral data classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 1231–1242, 2008.

[30] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, “Active learning methods for remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[31] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, “A survey of active learning algorithms for supervised remote sensing image classification,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606 –617, june 2011.

[32] W. Di and M. M. Crawford, “Active learning via multi-view and local proximity co-regularization for hyperspectral image classification,” *IEEE Journal of Selected Topics Signal Processing*, vol. 5, no. 3, pp. 618–628, 2011.

[33] S. Patra and L. Bruzzone, “A batch-mode active learning technique based on multiple uncertainty for SVM classifier,” *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 497 –501, may 2012.

[34] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis *et al.*, “Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS),” *Remote Sensing of Environment*, vol. 65, no. 3, pp. 227–248, 1998.

[35] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, “Classification of hyperspectral data from urban areas based on extended morphological profiles,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, pp. 480–491, 2005.

[36] D. Böhning, “Multinomial logistic regression algorithm,” *Annals of the Institute of Statistical Mathematics*, vol. 44, pp. 197–200, 1992.

[37] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

[38] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, “Sparse multinomial logistic regression: Fast algorithms and generalization bounds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[39] J. Li, J. Bioucas-Dias, and A. Plaza, “Hyperspectral image segmentation using a new Bayesian approach with active learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 19, pp. 3947–3960, 2011.

[40] J. Li, J. M. Bioucas-Dias, and A. Plaza, “Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, 2012.

[41] J. Bioucas-Dias and M. Figueiredo, “Logistic regression via variable splitting and augmented Lagrangian tools,” Instituto Superior Técnico, TULisbon, Tech. Rep., 2009.

[42] J. Platt, “Probabilities for support vector machines,” in *Advances in large margin classifiers*. Cambridge: MIT Press, 2000, pp. 61–74.

[43] T.-F. Wu, C.-J. Lin, and R. C. Weng, “Probability estimates for multiclass classification by pairwise coupling,” *Journal of Machine Learning Research*, no. 5, pp. 975–1005, 2004.

[44] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, “SVM- and MRF-based method for accurate classification of hyperspectral images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 736–740, 2010.

[45] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, “A survey of active learning algorithms for supervised remote sensing image classification,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[46] T. Luo, K. Kramer, D. B. Goldgof, S. Samson, A. Remsen, T. Hopkins, and D. Cohn, “Active learning to recognize multiple types of plankton,” *Journal of Machine Learning Research*, pp. 589–613, 2005.

[47] L. Breiman, *Bagging predictors*. Technical Report 421, Univ. of California at Berkeley, 1994.

[48] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*. Springer, 2006.