Title: Active and Semisupervised Learning for the Classification of Remote Sensing Images

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2014

Author(s): Claudio Persello and Lorenzo Bruzzone.

Volume: 52, Issue: 11

Page(s): 6937 - 6956

DOI: 10.1109/TGRS.2014.2305805

Active and Semi-Supervised Learning for the Classification of Remote Sensing Images

Claudio Persello, *Member, IEEE,* Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—This study aims at analyzing and comparing Active Learning (AL) and Semi-Supervised Learning (SSL) methods for the classification of Remote Sensing (RS) images. We present a literature review of the two learning paradigms and compare them theoretically and experimentally when addressing classification problems characterized by few training samples (w.r.t. the number of features) and affected by samples selection bias. Commonalities and differences are highlighted in the context of a conceptual framework used to describe the workflow of the two approaches. We point out advantages and disadvantages of the two approaches, delineating the boundary conditions on the applicability of the two paradigms with respect to both the amount and the quality of available training samples. Moreover, we investigate the integration of concepts that are in common between the two learning paradigms for improving state-of-the-art techniques and combining AL and SSL in order to jointly leverage the advantages of both approaches. In this framework, we propose a novel SSL algorithm that improves the Progressive Semi-Supervised Support Vector Machine (PS^{3}VM) by integrating concepts that are usually considered in AL methods. We performed several experiments considering both synthetic and real multispectral and hyperspectral RS data, defining different classification problems starting from different initial training sets. The experiments are carried out considering classification methods based on Support Vector Machines (SVMs).**

***Index Terms*—Active Learning, Semi-Supervised Learning, Sample Selection Bias, Support Vector Machine, Image Classification, Remote Sensing.**

I. Introduction

MACHINE learning methods have been widely applied to the analysis of Remote Sensing (RS) data in the past decade. Supervised classification methods like Support Vector Machines (SVMs) [1], [2], and kernel methods in general [3]–[5], have gained increasing attention in different fields of data mining, pattern recognition, computer vision, as well as in RS, where nowadays they are considered state-of-the-art methods. The development of the above-mentioned supervised classification techniques has been followed by the definition of novel learning paradigms that have recently gained relevance also in the RS community. Among them, an important role is played by Semi-Supervised Learning (SSL) and Active Learning (AL) methods. The automatic classification of RS images is typically performed by using supervised classification techniques, which require the availability of labeled samples for training the supervised algorithm. However, the collection of labeled samples to be used in the learning is usually time consuming and expensive. The amount and the quality of the available training samples are of central importance for obtaining accurate classification maps. Nevertheless, in many real world problems the available training samples are not sufficient in number and not adequate in quality for properly training the classifier. Thus, in order to enrich the information given in input to the supervised learning algorithm and to improve the classification accuracy, SSL techniques have been adopted for jointly leveraging the information of both labeled and unlabeled samples in the training of the classifier. SSL approaches based on SVMs have been successfully applied to the classification of multispectral and hyperspectral RS images, where the ratio between the number of training samples and the available spectral channels is small [6], [7]. However, the convergence to the correct solution is not always guaranteed with SSL techniques. An alternative approach for improving the learning of the classifier is AL, which assumes that few new samples can be labeled and added to the original training set. With this paradigm, the original training set is iteratively expanded according to an interactive process that involves a supervisor (usually a human expert), who is able to assign the correct label to any queried sample. This approach has been effectively applied for optimizing the collection of training samples in different application domains including the classification of RS images [8]–[11]. In AL: 1) the learning process iteratively queries the labels of the samples that are expected to be the most informative for an effective training of the classifier, 2) the supervisor annotates the selected samples, and 3) the classifier is re-trained using the updated training set. In this way, the unnecessary and redundant labeling of non-informative samples is avoided, greatly reducing the cost and the time of the training sample collection. The use of the two learning paradigms for the classification of RS images has been quite intensively investigated in the last decade. However, despite their commonalities, detailed analysis and comparison of the two approaches has not been done.

In this paper we present a comparative study in order to analyze AL and SSL in relation to the classification of RS images. We analyze the two approaches for addressing classification problems with limited amount of training samples and under *sample selection bias*. We present a conceptual framework in order to describe the workflow of AL and iterative SSL methods and to point out their main differences and commonalities. On the basis of this analysis, we investigate different strategies for combining AL and SSL in order to take advantage of both paradigms in real classification problems. Moreover, we propose a novel SSL algorithm that improves the Progressive Semi-Supervised Support Vector Machine (PS^{3}VM) by integrating concepts that are usually considered in AL methods. The two approaches are studied here in the context of SVM-based classification methods considering different AL and SSL methods. We present a comparison that aims at identifying advantages and disadvantages of the two approaches, and also discuss the boundary conditions on

the applicability of these methods both in terms of available labeled samples and reliability of classification results. The experimental analysis is carried out on both toy data sets and on real RS data. Limits and potentials of both approaches are critically analyzed in the light of possible applications to RS scenarios characterized by different kinds of data and classification problems. The main novel contributions of this paper are: 1) the description of a unified conceptual framework for AL and SSL, 2) the investigation of different approaches for combining AL and SSL for the classification of real RS data, and 3) the presentation of an improved Progressive Semi-Supervised SVM (PS^{3}VM), which exploits concepts that are usually considered for AL.

The paper is organized into seven sections. In the next section, classification problems characterized by few and biased training samples are presented and formalized. Section III reviews the main categories of SSL classification and gives a brief description of the PS^{3}VM algorithm used in our study. In Section IV, AL is presented and analyzed both in the general machine-learning framework and in the context of RS applications. Moreover, we give a brief description of the SVM-based AL methods that are used in our comparison. In Section V, the two considered AL and SSL paradigms are analyzed and compared on the basis of a common conceptual framework: the main commonalities and differences are pointed out. A novel algorithm that integrates AL concepts in SSL is proposed. Different strategies for combining AL and SSL are investigated. Section VI illustrates the experimental analysis. Finally, section VII presents a discussion on the two considered strategies and draws the conclusion of the paper.

II. Problem Formulation: Classification with Small and Biased Training Sets

AL and SSL can be considered two different approaches to address ill-posed classification problems, where the available training samples are too few with respect to the number of features and do not allow one to correctly estimate the true underlying distribution of the classes. Ill-posed problems are very likely to occur in real RS classification problems, especially in the classification of the last generation of RS images, e.g., very high resolution (VHR) and hyperspectral images, where the ratio between the available training samples and the number of features is usually small. In the classification of hyperspectral images, it is clear that the high number of spectral bands leads to define the classification problem in a high dimensional feature space. In the classification of VHR images the available spectral bands provide poor spectral resolution, therefore the extraction of several textural and geometric features is usually necessary to characterize the objects present in the scene under investigation and to obtain good classification accuracies. This leads again to defining the classification problem in a high dimensional feature space.

Not only the "quantity", but also the "quality" of the available training samples is important for obtaining accurate classification results. With "quality" of training samples, we mean their capability to model the real underlying distribution of the classes. A common assumption in the design of learning algorithms is that the training data consists of examples drawn independently from the underlying distribution. In many real-world classification problems, this assumption is often violated because training points are manually selected through surveys and they don't represent a random set of samples of the general population.

This problem in known as *sample selection bias* [12], [13]. Adopting the notations introduced in [13], we can formalize the problem by considering the selection variable s which takes binary values: $s = 1$ denotes that the labeled sample $(\mathbf{x}, y)$ is included in the training set T, where $\mathbf{x}$ is the feature vector and y is the class label, while $s = 0$ denotes that $(\mathbf{x}, y)$ is not selected. The label of not selected samples is not available for the training of the classifier, however we assume here that the set of unlabeled feature vectors (called pool) is available and can be used by a SSL method. Note that this assumption is usually satisfied in the classification of RS images, where several unlabeled samples are generally available. According to [13], four cases can be considered regarding the dependance of s on the example $(\mathbf{x}, y)$:

1) If s is independent of $\mathbf{x}$ and y, the selected training set is not biased, i.e., the examples constitute a random sample set of the true underlying distribution.
2) If s is independent of y given $\mathbf{x}$, i.e., $P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x})$, the selected samples are biased, but the bias depends only on the feature vector $\mathbf{x}$. This problem is also called *covariate shift*.
3) If s is independent of $\mathbf{x}$ given y, i.e., $P(s = 1|\mathbf{x}, y) = P(s = 1|y)$, the selected samples are biased, but the bias depends only on the label y. This corresponds to a change in the prior probabilities of the classes.
4) If no independence assumption holds between $\mathbf{x}$, y and s, the selected samples are biased and no further simplification is possible.

Clearly, a training set obtained under sample selection bias (or *biased training set* for brevity) leads to skewed estimations of the true underlying distributions of the classes. Let us denote $P(\mathbf{x}, y) = P(y)P(\mathbf{x}|y)$ the joint probability of the feature vector $\mathbf{x}$ and the class label y, which represents the true underlying generative model that defines the classification problem. Let $P^{tr}(\mathbf{x}, y) = P^{tr}(y)P^{tr}(\mathbf{x}|y)$ be the joint distribution estimated from the available training samples (using a supervised approach). In problems affected by sample selection bias, we have that $P^{tr}(\mathbf{x}, y) \neq P(\mathbf{x}, y)$.

In the collection of training samples for the classification of RS images, a sample selection bias is very likely to happen. Actually, we would argue that an unbiased sampling is almost impossible in practice. The selection variable s can depend on both the feature vector $\mathbf{x}$ and the class label y; nevertheless, its dependency may not be so clear. In practice, s may depend on latent variables, which are not included neither in the feature vector $\mathbf{x}$ nor in the class label y, but are correlated with both of them. For instance, it is very likely that s depends on the geographical location associated with the sample (e.g., the pixel). This type of bias is particularly common when labeled samples are collected by ground surveys. In that case, different sampling schemes are typically adopted for

field surveys: a) random sampling, b) stratified sampling, c) systematic sampling or d) cluster sampling [14]. A pure random sampling, where the sample locations are selected in a completely random fashion, allows one to obtain an unbiased statistical sampling of the true underlying distribution of the classes. However, it can be very expensive or impracticable in real applications. Moreover, it tends to undersample rarely occurring classes. In stratified sampling, a minimum number of samples are collected for each class, ensuring that every class is represented in the training set by a number of samples that reflects the prior probability of the class. However, this approach can also be impractical, because the location of the different classes is usually available only after a land-cover map has been generated. Systematic sampling is a method in which the samples are selected on the basis of specified locations (e.g., a regular grid) over the study area. In cluster sampling, the ground surveys are conducted on the basis of predefined geographical clusters, where multiple samples are collected in close proximity to one another. Clear practical advantages are offered by the last two sampling schemes, which are often preferred in real applications. However, in such cases, the geographical location of the samples clearly influences the estimated probability of the classes $P^{tr}(y)$ and the conditional probability $P^{tr}(\mathbf{x}|y)$. Thus, a selection bias that depends on the geographical location of the samples determines a bias in both $\mathbf{x}$ and y, giving rise to the general case of sample selection bias described in 4.

Another type of sample selection bias can occur when the samples are selected on the basis of *natural* classes, which are not taken into account in the classification problem. Consider as an example a classification problem where the goal is to classify a RS image according to the information classes: "vegetation" versus "rest". The two information classes contain several *natural* classes, e.g., the class "vegetation" contains the classes "grass", "forest", "agriculture fields", etc., which are not considered in the problem. The class "rest" can contain a high variety of different classes like "urban area", "water", "shadow", etc. However, the ground data collection may result in biased training samples for the class "vegetation" for instance, because only "grass" or "forest" pixels are considered, without including examples of "agriculture fields". This, will produce a biased training set that does not correctly model the real distribution of the classes. This type of selection bias directly affects the estimation of the conditional probabilities, i.e., $P^{tr}(\mathbf{x}|y) \neq P(\mathbf{x}|y)$.

In this work we investigate and compare the use of AL and SSL for addressing classification problems characterized by small and biased training set. In our experimental analysis we will compare the two learning paradigms considering different data sets and biased training sets.

III. Semi-supervised Learning Methods

Two main families of learning can be used for training a classifier: supervised learning methods (when labeled training samples are given) and unsupervised learning methods (when only unlabeled samples are available). Semi-supervised learning is between supervised and unsupervised learning, i.e., both labeled and unlabeled samples are available and are jointly leveraged by the classification algorithm [15]. The main idea of SSL is to exploit the structural information of unlabeled samples in the feature space to better model the distribution of the classes and to find a more accurate classification rule than using only labeled samples. Many semi-supervised classification techniques have been proposed in the literature so far. The main SSL methods are briefly summarized in the next subsections.

A. State of the Art of Semi-Supervised Learning

The SSL methods presented in the literature can be grouped into the following main categories: 1) self-training, 2) co-training, 3) generative probabilistic models, 4) semi-supervised SVM, and 5) Graph-based SSL. More information about semi-supervised classification can be found in [15], [16].

1) Self-training: one of the earliest ideas about using unlabeled data in the classification is self-training [15], [17]–[19]. This approach consists in an algorithm that repeatedly uses a supervised learning method. It starts by training on the labeled samples only. Then, at each iteration a part of the unlabeled samples is labeled according to the current decision function and added to the training set. Typically the most confident unlabeled samples, together with their predicted labels, are added to the training set. Then the classifier is re-trained using the additional labeled samples and the procedure is repeated. Self training is a wrapper algorithm and can be used with any supervised classifier. In [20], a self-training technique for the classification of hyperspectral images is proposed. The method operates in two steps: in the first step, confident candidate unlabeled samples are selected on the basis of spatial and spectral information; in the second step, an AL method is adopted for selecting the most informative samples among the candidate ones to be included in the training set.

2) Co-training: is based on the following assumptions: a) the features can be split into two sets, b) each sub-feature set is sufficient to train a good classifier, and c) the two sets of features are conditionally independent given the class [21]. Initially, two separate classifiers are trained with the labeled data, on the two sub-feature sets, respectively. Each classifier then classifies the unlabeled data, and provides to the other classifier the most confident unlabeled samples with their predicted labels. Each classifier is retrained with the additional training examples given by the other classifier, and the process is iterated. A variant of co-training is multi-view learning [22]. In such a setting, the original set of features is split into multiple subsets of features (called views) that are used for training different classifiers. Iterative algorithms similar to co-training can be used for SSL.

3) Generative probabilistic models: they are based on the estimation of the joint probability $P(\mathbf{x}, y|\theta)$ assuming a particular model for the data (e.g., Gaussian mixture model), where θ is the parameter vector of the model that should be estimated from the observations. The estimation of the parameter vector θ can benefit from the joint exploitation of labeled and unlabeled samples. The final classification is then performed on the basis of the Bayes rule. A popular method

for the estimation of θ is the expectation-maximization (EM) algorithm, which has been largely adopted also in RS [23]–[25]. In [25], by assuming a Gaussian mixture model, Tadjudin and Landgrebe used the iterative EM algorithm to estimate model parameters from both labeled and unlabeled samples. In terms of Fisher information, Shahshahani and Landgrebe have proved that the additional unlabeled samples are helpful for semi-supervised classification in the context of a Gaussian maximum-likelihood classifier, under a zero-bias assumption [24].

4) Semi-supervised SVM: are SSL techniques specifically developed for SVM. In [26], the author proposes the transductive SVM (TSVM) for text classification. While traditional SVM classifiers try to induce a general decision function for a learning task given a set of training points, TSVM takes into account a particular test set and try to minimize the classification errors of just those particular examples (transductive inference). This is done by using both labeled training samples and unlabeled test samples. The main idea is that the decision boundary has to pass in low density regions and this is obtained by adding to the standard SVM optimization problem an additional regularization term on unlabeled data. In [6], [27], the authors present the Progressive Semi-Supervised SVM (PS^{3}VM) classification method for addressing ill-posed problems with SVM. Such a method adopts an iterative algorithm for searching a reliable separating hyperplane in the kernel space by exploiting unlabeled samples together with their predicted labels. Given its iterative nature, the method is closely related to the self-training approach. The details of this method are presented in the next subsection. In [28], a semi-supervised SVM classification technique is proposed, where the learning phase is performed by optimizing the objective function directly in the primal formulation, without exploiting the dual representation that can be obtained with Lagrange multipliers.

5) Graph-based SSL: define a graph where the nodes are labeled and unlabeled samples, and edges reflect their similarity. These methods usually assume label smoothness over the graph to include cluster/manifold regularization. In [29], a family of semi-supervised learning algorithms based on manifold regularization is proposed. The proposed family of learners can exploit the geometry of the marginal distributions by taking advantage of the unlabeled samples. Within this general framework, two specific families of algorithms are proposed: the Laplacian Regularized Least Squares (LapRLS) and the Laplacian SVM (LapSVM). In [7], the LapSVM algorithm is applied to the classification of RS images.

B. Progressive Semi-Supervised SVM

The PS^{3}VM method is considered for our theoretical and experimental comparison with AL methods, because of its conceptual simplicity, effectiveness and the possibility to relate it easily with AL. For this reason, a more detailed description of such a method is given here. For simplicity, we refer here to binary classification problems. Generalization to the multiclass case can be obtained via the standard OAA strategy [2].

PS^{3}VM is based on an iterative algorithm, which is made of three main phases: 1) Initialization (only the original training samples are used); 2) semi-supervised learning (both the original training sample plus originally unlabeled sample with their predicted labels are considered); and 3) convergence.

1) Initialization: A standard SVM is trained using the original training samples, by solving the following constrained optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},\xi,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to: } y_i[\mathbf{w}\cdot\phi(\mathbf{x}_i)+b] \geq 1-\xi_i \quad i=1,...,n \\ &\qquad\qquad \xi_i \geq 0 \end{aligned} \tag{1}$$

where $\mathbf{w}$ is a vector orthogonal to the separating hyperplane, b is a bias term such that $b/\|\mathbf{w}\|$ represents the distance of the hyperplane from the origin, C is the regularization parameter, ϕ is function mapping the data into the feature space, ξ_i are slack variables and n is the number of training samples. According to the sign of the resulting decision function (6), *pseudo labels* are given to the unlabeled samples.

2) Semi-supervised Learning: After the initialization, for any iteration until convergence, a set of samples from a pool $\mathcal{U}$ of unlabeled samples are iteratively selected and added to the training set together with their *pseudo labels* and removed from the pool. Let us define the following sets of samples that lie in the upper and lower side of the margin:

$$H_{up} = \{\mathbf{x}|\mathbf{x}\in\mathcal{U}, 0 \leq f(\mathbf{x}) \leq 1)\} \tag{2}$$

$$H_{down} = \{\mathbf{x}|\mathbf{x}\in\mathcal{U}, -1 \leq f(\mathbf{x}) \leq 0)\} \tag{3}$$

At each iteration, ρ samples are selected from each side of the margin. In particular, the ρ samples with $f(\mathbf{x})$ closer to 1 are selected from H_{up} and the ρ samples with $f(\mathbf{x})$ closer to -1 are taken from H_{down}. This results in the selection of a total of 2ρ samples, which are named *semi-labeled*. The SVM is then re-trained using both the n original training samples and the m semi-labeled ones (accumulated until the current iteration), according to the following problem:

$$\begin{aligned} &\min_{\mathbf{w},\xi,\xi^*,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{m}C_j^*\xi_j^* \\ &\text{subject to: } y_i[\mathbf{w}\cdot\phi(\mathbf{x}_i)+b] \geq 1-\xi_i \qquad i=1,...,n \\ &\qquad\qquad y_j^*[\mathbf{w}\cdot\phi(\mathbf{x}_j^*)+b] \geq 1-\xi_j^* \qquad j=1,...,m \\ &\qquad\qquad \xi_i,\xi_j^* \geq 0. \end{aligned} \tag{4}$$

The regularization parameter C_j^* for the semi-labeled patterns increases in a quadratic way, depending on the number of iterations that the associated semi-labeled sample x_j^* have been assigned to the same label (see [27] for details). If the label of a semi-labeled pattern x_j^* at iteration i is different from the one at iteration $i-1$, such a label is erased, and x_j^* is moved back to the pool $\mathcal{U}$.

3) Convergence: the iterative procedure is stopped when both the number of mislabeled training samples and the number of pseudo-labeled patterns which lie into the margin band are lower or equal than $\beta\cdot m$, where β is a user-defined parameter. When convergence is reached, the SVM is trained

for the last time, according to the following minimization problem:

$$\begin{aligned} &\min_{\mathbf{w},\xi,\xi^*,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + C^*_{max}\sum_{j=1}^{m}\xi^*_j \\ &\text{subject to: } y_i[\mathbf{w}\cdot\phi(\mathbf{x}_i)+b] \geq 1-\xi_i \qquad i=1,...,n \\ &\qquad\qquad y^*_j[\mathbf{w}\cdot\phi(\mathbf{x}^*_j)+b] \geq 1-\xi^*_j \qquad j=1,...,m \\ &\qquad\qquad \xi_i,\xi^*_j \geq 0 \end{aligned} \tag{5}$$

where the entire set of semi-labeled samples is associated with the same regularization parameter C^*_{max}.

IV. Active Learning Methods

AL is an approach to iteratively select the most informative samples for defining a training set by exploiting the classification rule [8]–[11], [30]. Using the formalism introduced in the section II, AL consists in actively controlling the selection variable s in order to select informative samples making use of the learner's feedback. In this way, the selection focuses on the most uncertain and diverse samples, therefore avoiding the labeling of redundant and non-informative samples. It is worth noting that AL does not lead to unbiased training sets, as we would obtain from a completely random selection strategy. AL aims instead at minimizing the number of training samples to be labeled in order to obtain a satisfactory classification accuracy. To this end, different AL methods aim at annotating the samples that can lead to the highest gain in classification accuracy. The focus is therefore usually on obtaining a better estimate of the posterior probability $P(y|\mathbf{x})$ rather than a good estimate of the generative model $P(\mathbf{x},y)$.

In order to precisely describe the workflow of a general AL process, let us model it as a quintuple (G, Q, S, T, $\mathcal{U}$) [31]. G is a supervised classifier, which is trained with the training set T. Q is the query function used to select the most informative unlabeled samples from a pool $\mathcal{U}$ of unlabeled samples on the basis of the current classification results. S is a supervisor who can assign the (true) class label to any unlabeled sample of $\mathcal{U}$ (e.g., a human expert). The AL process is an iterative process, where the supervisor S interacts with the classification system by labeling the most informative samples selected by the query function Q at each iteration. At the first stage, an initial training set T made up of few labeled samples is required for the training of the classifier G. After initialization, the query function Q selects a set of samples from the pool $\mathcal{U}$ and the supervisor S assigns them the true class labels. Then, these new-labeled samples are included into T and the classifier G is retrained using the updated training set. The closed loop of querying and retraining continues until a stopping criterion is satisfied. Algorithm 1 gives a description of a general AL process.

Algorithm 1 Active Learning Procedure

1: Train the classifier G with the initial training set
2: Classify the unlabeled samples of the pool $\mathcal{U}$
3: **repeat**
4: Query a set of samples (with the query function Q) from the pool $\mathcal{U}$
5: The user S manually label the selected samples
6: the new labeled samples are added to the training set T
7: Re-train the classifier G
8: **until** a stopping criterion is satisfied

A. State of the Art of Active Learning

The query function Q constitutes the core of each AL technique. Several query functions have been proposed so far in the machine learning literature. Most of these works have focused on the selection of one sample to be labeled in each iteration. To this end, different criteria have been adopted for selecting the (expected) most informative sample. One of the first strategies introduced in the literature is based on *uncertainty* sampling [32], which aims at selecting the closest sample to the decision boundary. In the probabilistic approach presented in [32], the posterior probability of the classes is estimated for both obtaining the classification rule and to estimate the uncertainty of unlabeled samples. In the two-class case, the query of the most uncertain samples is obtained by choosing the samples associated to a posterior probability that is closest to 0.5, since this value corresponds to the classifier being most uncertain of the correct class label. The same principle has also been used in the context of SVM classification [33]–[35]. The SVM classifier is particularly suited to AL due to its intrinsic high generalization capabilities and because its classification rule can be characterized by a small set of support vectors that can be easily updated over successive learning iterations. The query strategy proposed in [35] is based on the splitting of the version space: the points which split the current version space into two halves having equal volumes are selected at each step, as they are likely to be the actual support vectors. Three heuristics for approximating the above criterion are described; the simplest among them selects the point closest to the hyperplane as in [34].

Another strategy is *query by committe* (QBC) [36], [37]. A committee of classifiers using different hypothesis about parameters is trained to label a set of unknown examples. The algorithm selects the samples where the disagreement between the classifiers is maximal. In [38], two query methods that combine the idea of query by committee and that of boosting and bagging are proposed. In [31], an approach is proposed that estimates the uncertainty level of each sample according to the output score of a classifier and selects only those samples whose outputs are within the uncertainty range. In [39], the authors present possible generalizations of the active SVM approach to multiclass problems. A survey of several existing methods is available in [40].

Other studies have focused on the selection of *batches of samples* at each iteration, which allow one to speed up the learning process. In this latter setting, the overlap of information among the selected samples has to be considered in order to evaluate their expected information content. Brinker introduced an SVM-based batch approach, which selects a batch of samples that minimizes the margin while maximizing their *diversity* [41]. The diversity is assessed considering the kernel cosine-angular distance between points. Another approach to consider the diversity in the query function is

the use of clustering [42], [43]. In [42], an AL heuristic is presented, which explores the clustering structure of samples and identifies uncertain samples avoiding redundancy. In [44], the authors select the batch of instances that maximizes the Fisher information of a classification model, which leads to a trade-off between uncertainty and diversity. In [45], batch active learning is formulated as an optimization problem that maximizes the discriminative classification performance while taking into consideration the unlabeled examples. Azimi *et. al* [46] used Monte-Carlo simulation to estimate the distribution of unlabeled examples selected by a sequential policy and query the samples that best matched such a distribution.

Active learning has been applied to a variety of real-world problem domains, including text classification, information extraction, video classification and retrieval, speech recognition [40]. In recent years, AL has attracted the interest of the remote sensing community, and it has mainly been applied to the classification of multispectral and hyperspectral images. In RS problems, the supervisor S is a human expert that can derive the land-cover type of the area on the ground associated to the selected patterns. No particular restrictions are usually considered for the initial training set T and its size, since we expect that the AL process can be started up with few samples for each class without affecting the convergence capability. This is a very important observation, because this property does not apply to the use of SSL classification techniques and thus one should be more careful on the original training set before deciding to adopt a SSL technique. Analyzing this issue is actually one of the main goals of this work and will be further discussed in this paper. The pool of unlabeled samples $\mathcal{U}$ can be associated to the whole considered image or to a portion of it (for reducing the computational time associated to the query function and/or for considering only the areas of the scene accessible for labeling). The use of AL for RS data classification has been investigated in [8]–[11], [30].

In remote sensing applications, several studies adopt an uncertainty criterion in combination with SVM classifiers for the definition of AL methods. This approach revealed very effective in real RS problems and computationally very efficient. The AL methods introduced in the machine learning community for binary SVM classification [33]–[35] have been extended to deal with multi-class problems. The technique proposed in [8] selects the most uncertain sample for each binary SVM (i.e., the one closest to the separating hyperplane) in a *One-Against-All* (OAA) multi-class architecture. In [10], different batch-mode AL techniques for the classification of RS images with SVM are investigated. The investigated batch-mode methods make use of different query functions, which are based on both the uncertainty and diversity criteria. The Multiclass-Level Uncertainty (MCLU) method is introduced, which is an effective extension of the uncertainty criterion to address multi-class problems using a OAA multi-class architecture of binary SVMs. The method proposed in [10] is based on clustering in the kernel space to select diverse samples. The most uncertain sample from each cluster is selected to be included in the batch of samples to be queried. Such a technique is named Multiclass-Level Uncertainty with Enhanced Clustering Based Diversity (MCLU-ECBD). A brief description of this method is given in the next subsection. In [30], two AL techniques for multi-class RS classification problems are proposed. The first technique is margin sampling by closest support vector, which selects the most uncertain unlabeled samples that do not share the closest support vector. The second technique follows the idea of QBC with bagging presented in [38]. Such method is extended to deal with multi-class problems by using the entropy as a measure of disagreement. The samples are selected according to the maximum disagreement between a committee of classifiers, which is obtained by bagging: different training sets are drawn with replacement from the original training data and used for training different supervised classifiers. In [9], an AL technique is presented, which selects the unlabeled sample that maximizes the information gain between the a posteriori probability distribution estimated from the current training set and the training set obtained by including that sample into it. The information gain is measured by the Kullback-Leibler (KL) divergence. This KL-Maximization technique can be implemented with any classifier that can estimate the posterior class probabilities. In [11], a cluster-assumption based AL method is proposed. Basically, it exploits the fact that if patterns are in the same cluster, they are likely to be of the same class [15]. The query function of this method aims therefore at selecting the most uncertain samples that lie in low-density regions of the feature space. It is worth noting that the cluster assumption, which is usually considered in semi-supervised classification, is seldom used in the definition of query functions in AL. Di *et al.* [47] investigate AL methods based on multiview disagreement, which exploits the idea of QBC. In this case, the committee of classifiers is derived by using multiple views, i.e., different disjoint subsets of features. The paper investigates different approaches to view generation from hyperspectral images, including clustering, random selection and uniform slicing methods. It is worth noting that most of the multiview methods are applied to SSL methods. However, the same concept can be used to develop AL methods. In [48], the multiview-based AL method is combined with a regularizer based on the manifold space that penalizes rapid changes in the classification function close to sample points (both in the spectral and spatial domain).

B. SVM-based Active Learning

We report here a more detailed presentation of SVM-based AL methods, and in particular of the MCLU and MCLU-ECBD methods [10], since they are considered in the remainder of the paper for the comparison with SSL techniques.

SVM is a binary classifier, which aims at dividing the feature space into two subspaces (one for each class) using a separating hyperplane. Given a training set T, the SVM is trained by solving a quadratic programming problem [1]. The decision rule used to classify unknown samples is based on the sign of the obtained discrimination function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$, associated to the hyperplane. An important property of SVMs is related to the possibility to project the original data into a higher dimensional feature space via a positive semidefinite kernel function [3]. The training phase of the classifier can be

formulated in a dual form as a minimization problem, which lead to the calculation of the values of Lagrange multipliers α_i associated with the original training patterns. After the training, the discrimination function is given by

$$f(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b, \tag{6}$$

where SV is the set of support vectors, i.e., the subset of the training samples associated to $\alpha_i > 0$. As mentioned before, one of the first and most effective criteria in AL is based on the evaluation of the *uncertainty* of the samples. The most uncertain samples have the lowest probability to be correctly classified by the current classification model and are therefore the most useful to be included in the training set. The implementation of such a criterion with a binary SVM, results in the selection of the sample $\mathbf{x}^* \in \mathcal{U}$ that lie closest to the separating hyperplane, i.e., $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{U}} |f(\mathbf{x})|$. In order to extend the approach to deal with multiclass problems, we consider the OAA architecture [2], which involves a parallel architecture of SVMs, one for each information class. Each SVM solves a two-class problem defined by one information class against all the others. The MCLU technique selects the expected most informative sample according to a multiclass confidence value, which is defined on the basis of the discrimination function of the binary SVMs. An effective confidence measure is defined as

$$c(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x}), \tag{7}$$

where $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are the first and second highest output score of the binary SVMs in the OAA architecture. Similarly to the binary case, querying the sample that minimizes $c(\mathbf{x})$ results in the selection of the sample that is closest to the boundary between the two most probable classes. In the selection of more than one sample per iteration (batch-mode AL), a diversity criterion should be considered in order to avoid redundancy among the samples in the batch. MCLU-ECBD is an effective technique that combines MCLU with a diversity criterion based on *kernel k-means* clustering [10]. The method operates in two steps: 1) in the first step the m most uncertain samples are selected according to MCLU, 2) in the second step, $h < m$ diverse samples are selected by applying kernel k-means clustering to the uncertain samples for defining h different clusters, and finally taking the most uncertain sample from each cluster.

V. Active versus semi-supervised learning

In this Section, we present an analysis of AL and SSL, where we highlight commonalities and differences between the two learning paradigms in the context of a conceptual framework adopted for describing the workflow of the two approaches. As a result of our analysis, we also propose a novel SSL algorithm that improves the Progressive Semi-Supervised Support Vector Machine (PS^{3}VM) by integrating concepts that are usually considered in AL methods. Finally, we investigate different strategies for combining AL and SSL.

A. Analysis of Commonalities and Differences between AL and SSL

We already noted that both active and semi-supervised classification methods aim at solving ill-posed problems. However, the two approaches share commonalities not only in their ultimate goal, but often they implement similar concepts in their algorithms as well. The more relevant similarity is related to the iterative procedure, which is implicit in AL methods, and is adopted by most SSL methods as well (e.g., self-training, co-training, multiview-based methods, EM algorithm for semi-supervised generative models, and semi-supervised SVMs like PS^{3}VM). Several methods share common principles to assess the uncertainty of the samples, e.g., the distance to the classification hyperplane in the context of SVM classification, or the classifier agreement in co-training, multiview or committee of classifiers. In general, we can describe both AL and iterative SSL classification methods using a unified conceptual framework based on the quintuple (G, Q, S, T, $\mathcal{U}$) as done in Section IV. The framework is graphically represented as a block diagram in Figure 1. This will help us to analyze the two approaches synthesizing their main properties. To simplify our analysis, we will focus here on iterative SSL methods like self-training and PS^{3}VM.

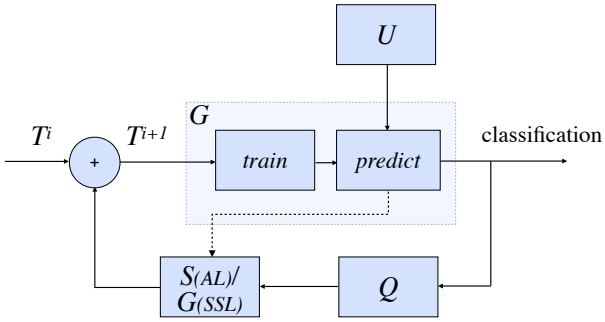

Fig. 1. Block diagram describing the conceptual framework for AL and SSL.

In such a framework, the main difference between AL and SSL approaches is in the supervisor S. In AL, the labeling of the queried samples is carried out by a human expert S, which is supposed to be able to associate the correct label to any selected pattern. The selection of the least confident samples is typically adopted, completely relying on the capability of the human expert to provide their correct labels. This is based on the observation that from the information theory, the most uncertain samples are those more informative. In SSL the labels of selected patterns are predicted on the basis of the current classification rule, i.e., the supervisor S coincides with the classifier G. However, the classifier G cannot be considered completely reliable in assigning the correct label to any sample. In general, the label reliability depends on the particular selected sample. For this reason, the query function Q used in AL and SSL are usually based on very different or even opposite criteria. In the case of self-training, the criterion used by Q usually selects samples with the most confident label given by the classifier. This means that a conservative selection is operated in order to avoid introducing

possibly wrong *semi-labeled* samples in the training set and thus decreasing the accuracy of the classification system. Very similar observations can be done about co-training or multi-view methods, which use a committee of classifiers trained on the different views to evaluate the confidence of the samples by considering the committee agreement. In SSL methods, the samples associated with the highest agreement among classifiers (most confident samples) are selected. On the contrary, multiview-based AL methods select the samples associated with the highest disagreement (most uncertain samples). The case of PS^{3}VM is more peculiar with respect to standard self-training or multiview methods, because the sample selection is based on a trade-off between label prediction confidence and expected information content. Very confident samples (very distant from the hyperplane) are associated to very low probability of becoming support vectors at the next iterations and their expected information content is therefore very low. On the basis of this observation, in PS^{3}VM, Q selects the samples that lie furthest from the discriminating hyperplane, but fall inside the margin (most certain among informative samples). A graphical representation of the above-mentioned query strategies is reported in Figure 2. For the aforementioned reasons, a semi-supervised approach usually requires several more iterations (and samples) than AL in order to reach the convergence.

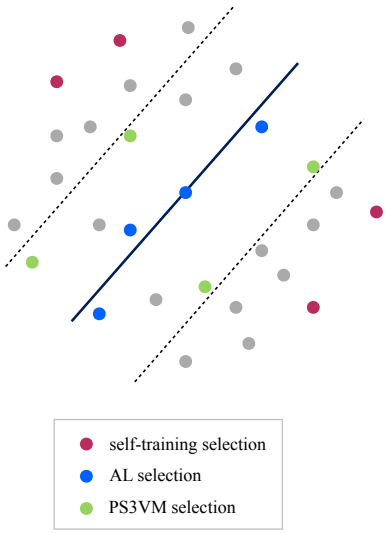

Fig. 2. Sample selection by the query function with different SSL and AL strategies.

Let us now point out the main general differences between the two learning paradigms. Differences mainly reside in their implicit assumptions, which is very important to be aware of before applying these approaches to real data. The main assumption of AL is that the supervisor is infallible, i.e., it is able to correctly label any selected sample. In real applications, it is important to correctly design the classification system in order to minimize possible ambiguities for the human annotator and give him the possibility to reject the samples that he is not able to annotate. However, adopting the standard assumption on the supervisor, AL methods can reach the convergence, without particular requirements on the initial training set and the underlying data distribution. This is generally not true in the case of SSL. The different SSL methods we described in Section III share common prerequisite in order to be effective, i.e., to obtain more accurate predictions by incorporating the unlabeled samples in the learning process with respect to standard supervised algorithms. Basically, the unlabeled data has to carry information that is useful in the inference of the right information class of unknown samples, i.e., $P(\mathbf{x})$ has to carry information about $P(y|\mathbf{x})$. In general, the basic assumption for using SSL is the so-called cluster assumption, which states that if patterns are in the same cluster, they are likely to be of the same class [15]. Note that the cluster assumption does not imply that each class forms a single, compact cluster: it only means that, usually, we do not observe objects of two (or more) distinct classes in the same cluster. This assumption can be equivalently formulated in the following way (also called low-density separation assumption): the decision boundaries among classes should lie in low-density regions of the feature space. The equivalence is easy to see: a decision boundary in a high-density region would cut a cluster into two different classes. Although the two formulations of the same assumption are conceptually equivalent, they inspired different algorithms (e.g., generative probabilistic models are inspired by the cluster assumption, whereas the TSVM implments the low-density separation assumption) [15]. If the cluster assumption does not hold, semi-supervised learning will not yield an improvement over supervised learning. It might even happen that using unlabeled samples, the SSL algorithm degrades the classification accuracy by misguiding the inference.

B. Incorporating AL concepts in SSL

On the basis of our analysis, we propose a novel SSL algorithm by incorporating in the PS^{3}VM concepts that are commonly adopted in batch-mode AL. In particular, we propose to integrate the *diversity* criterion and a *multiclass-based confidence* measure in the semi-labeled sample selection process of the PS^{3}VM algorithm.

The PS^{3}VM algorithm may require the inclusion of several semi-labeled samples in the training set before reaching the convergence. If many unlabeled samples are available, as it is usually the case in RS classification problems, this can result in a high number of iterations and therefore a slow training phase. In order to reduce the number of iterations to reach convergence, we introduce the use of a *diversity* criterion in the second phase of the algorithm for minimizing the redundancy among the selected semi-labeled samples. Moreover, we note that in original PS^{3}VM algorithm, multiclass problems are addressed by running the iterative learning process independently for each binary classifier of an OAA architecture and combining the results only at the end, when all binary PS^{3}VMs have reached the convergence. We propose a variant of the original algorithm that considers a multiclass confidence measure for semi-labeled sample selection, which is based on the MCLU AL algorithm. In this way, we run a single iterative learning process where an OAA architecture of binary SVMs is trained at every iteration considering both

labeled and semi-labeled samples. Semi-labeled samples are selected considering the multiclass confidence measure defined in equation (7) in combination with a diversity criterion. Let us define the set of samples H as follows:

$$H = \{\mathbf{x} | \mathbf{x} \in \mathcal{U}, 0 \leq c(\mathbf{x}) \leq 2)\}. \quad (8)$$

H is defined in this way such that in the binary case when using an OAA ensemble of two SVMs, we have that $H = H_{up} \cup H_{down}$, i.e., H contains all the samples inside the margin. We define the set of candidate semi-labeled samples J by taking γ samples from H with $c(\mathbf{x})$ closer to 2 (i.e., the ones closer to the margin). Then, we select $\rho < \gamma$ diverse samples among the γ candidates with an incremental procedure. In the first step, we initialize the set of selected samples $\mathcal{X}$ with the sample $\mathbf{x}_{up} = \arg\max_{\mathbf{x} \in J} c(\mathbf{x})$. Then we incrementally include in $\mathcal{X}$ the sample of J that minimizes the similarity with the closest sample already included in that set. The similarity between two samples is computed considering the kernel cosine-angular similarity [41]:

$$k^*(\mathbf{x}, \mathbf{x}_i) = \frac{k(\mathbf{x}, \mathbf{x}_i)}{\sqrt{k(\mathbf{x}, \mathbf{x}) k(\mathbf{x}_i, \mathbf{x}_i)}}, \quad (9)$$

where $k(\cdot, \cdot)$ is a positive semidefinite kernel function. The process is stopped when ρ samples are included in $\mathcal{X}$. The sample selection process is described in Algorithm 2. After the selection of the semi-labeled samples, the SVM is re-trained as in the original algorithm. In the reminder of the paper we refer to this SSL algorithm with diversity as PS^{3}VM-D.

Algorithm 2 Semi-Labeled Sample Selection in PS^{3}VM-D

1: $\mathbf{x}_{up} = \arg\max_{\mathbf{x} \in J} c(\mathbf{x})$
2: $\mathcal{X} = \{\mathbf{x}_{up}\}$
3: **repeat**
4: $\quad \mathbf{x}_{up} = \arg\min_{\mathbf{x} \in J \setminus \mathcal{X}} \{\max_{\mathbf{x}_i \in \mathcal{X}} k^*(\mathbf{x}, \mathbf{x}_i)\}$
5: $\quad \mathcal{X} = \mathcal{X} \cup \{\mathbf{x}_{up}\}$
6: **until** $|\mathcal{X}| = \rho$

C. Combining AL and SSL

AL and SSL methods can be combined in order to define a learning framework that exploits both labeled and semi-labeled samples for the training of the classifier and for selecting new samples to be labeled by the user. Few studies in this direction have been reported in the following articles [49]–[52]. In [49], the QBC AL algorithm is combined with EM SSL for assigning class labels to the samples that remain unlabeled. The method is applied to text classification problems. Muslea *et al.* extended this idea by using multiple views for both active and semi-supervised learning [50]. In [51], AL and SSL are combined on the basis of a confidence score obtained by a boosting algorithm. The method is applied to spoken language classification problem. In [52], at each iteration of the AL process new labeled samples are added to the training set together with pseudo-labeled samples. Many different ways for combining AL and SSL are actually possible. Here, we point out some strategies that can be effective in RS problems.

- *Sequential application of AL and SSL.* A simple strategy to combine the two approaches is to execute them sequentially. AL is applied first for a number of iterations in order to include t additional labeled samples in the training set. Afterwards, a SSL method is executed. In this way, AL is used to build a sufficiently representative training set and then SSL is adopted for further increasing the classification accuracy by leveraging the information of unlabeled samples. Given that SSL has critical requirements on the initial training set in order to reach convergence, this approach has the advantage to adopt AL in order to set SSL in the right condition to effectively use the unlabeled samples in the second phase.
- *Interleaved approach.* This strategy generalizes the previous one by interleaving the application of AL and SSL. AL is first used for a given number of iterations in order to include t_1 labeled samples in the training set. Then SSL is adopted for selecting t_2 semi-labeled samples. The two phases of AL and SSL can be alternated multiple times. Given the characteristics of SSL, we expect that in general $t_2 >> t_1$.
- *SSL encapsulated in the AL process*. One can easily combine the two approaches by considering a SSL algorithm for the classifier G in the AL framework. This strategy is similar to the previous one by setting $t_1 = 1$ (or the batch size for batch-mode AL) and starting the process from the second phase (SSL phase). The method presented in [50] is a special case of this general strategy.
- *Collaborative AL and SSL.* At every iteration, new samples labeled by the user are considered together with semi-labeled samples for the training of the classifier. Different selection criteria for labeled and semi-labeled samples can be combined giving rise to different techniques. The techniques presented in [51], [52] are special cases of this general combination strategy.

VI. Experimental Analysis

The aim of our experimental analysis is to analyze and compare SSL and AL in different classification problems and to derive some insight about which approach is more appropriate according to the specific classification problem and the available initial training set. We considered both synthetic data and real multispectral and hyperspectral RS images associated with different classification problems. We used different initial training sets with different sizes and characterized by different types of sample selection bias. The experiments are carried out in order to emphasize the conditions where SSL can improve standard supervised methods and where this is not possible. In this latter case, in order to improve the classification accuracy, the manual labeling of new samples becomes necessary and AL can be adopted for selecting the most informative ones and guiding the user in the sample collection. We considered both binary and multiclass classification problems. As a baseline for our comparison we considered a standard supervised SVM classifier. As AL methods we considered MCLU and MCLU-ECBD methods. As SSL technique we adopted PS^{3}VM and the proposed PS^{3}VM-D algorithm. A

deeper analysis is performed on the considered hyperspectral data set by comparing different AL methods, considering also multiview-based methods, and providing results on the combination of AL with SSL using two different strategies. In all our experiments we run the AL and SSL techniques using ten different initial training sets and from them we derived averaged results. An RBF kernel function was adopted for all the experiments. In the experiments with AL methods, the model selection was performed at the first iteration using a grid-search for tuning the width of the RBF kernel and the regularization parameter. For the SSL techniques, the model selection was performed in two steps: the first step was used for tuning the parameters of the supervised SVM classifier, the second for tuning the parameters of the semi-supervised process. In all cases, we tuned the parameters of the classifiers for the optimization of the overall classification accuracy on the validation set.

A. Two-moon toy data set

We generated five different synthetic data sets in a bi-dimensional feature space varying the distance between the distributions of two intertwining moons (see Figure 3) and then carried out five different experiments accordingly. In this way, we analyze the capability of AL and SSL methods to cope with classification problems where the cluster assumption is becoming less and less acceptable. For each of these five experiments we generated 6000 samples (3000 for each class). 1000 samples were used in the test set. From the remaining 5000 samples, we derived ten different training sets of 20 samples and a validation set of 20 samples by randomly selecting samples in the box $0.3 < x < 0.3, -0.4 < y < 0.4$, i.e., in the region where the two moons are closer. In this way we generated training samples affected by sample selection bias. The remaining 4960 samples were used as pool without considering their labels in the experiments. We evaluated the distance between the samples of the two classes in the five experiments computing the Jensen-Shannon divergence [53] (see the second column of Table I). The obtained distances confirm the fact that the distributions of the two classes are becoming closer and closer in the various experiments (making the cluster assumption less and less true). Figures 4 and 5 report the curves of the overall accuracies (OAs) (averaged over ten trials) versus the number of semi-labeled samples in the five experiments obtained with PS^{3}VM and PS^{3}VM-D methods, respectively. Figure 6 reports the learning curves obtained with AL (MCLU-ECBD). From the obtained results we can easily observe that the AL method converged to very high classification accuracy (more than 99% of OA) in all five experiments after labeling about 120 samples. The complexity of the classification problem, i.e., the lower distance between the two classes in the five different experiments, affected only the number of samples needed by the algorithm to reach the convergence, but not the convergence capability. On the contrary, SSL methods are are much more sensitive to the distance between the distribution of the two classes. In particular, we observe that in experiment 1 the PS^{3}VM process could converge to high classification accuracy, significantly improving the performance of the supervised SVM. In experiments 2 and 3 the PS^{3}VM technique improved the classification accuracy of SVM by approximately 20%. In experiment 4 the convergence accuracy is reduced to about 64%, while in experiment 5 the PS^{3}VM decreased the classification accuracy obtained by the supervised SVM. The PS^{3}VM-D led to similar accuracies as the original PS^{3}VM algorithm. However, it significantly reduced the number of semi-labeled samples necessary to reach convergence. As expected, the standard supervised technique led to poor accuracy in these ill-posed classification problems. All averaged results are summarized in Table I. Figures 7 and 8 show the distribution of unlabeled and semi-labeled samples at different iterations of experiment 1 using the SSL algorithms PS^{3}VM and PS^{3}VM-D, respectively. From these figures, it is clear that the diversity-based SSL method better explores the real distribution of the classes and converges with a lower number of iterations. Figure 9 shows the distribution of unlabeled and labeled points at different iterations of experiment 1 using AL. In this case, very few iterations are sufficient for exploring the distributions of the two classes and reaching the convergence.

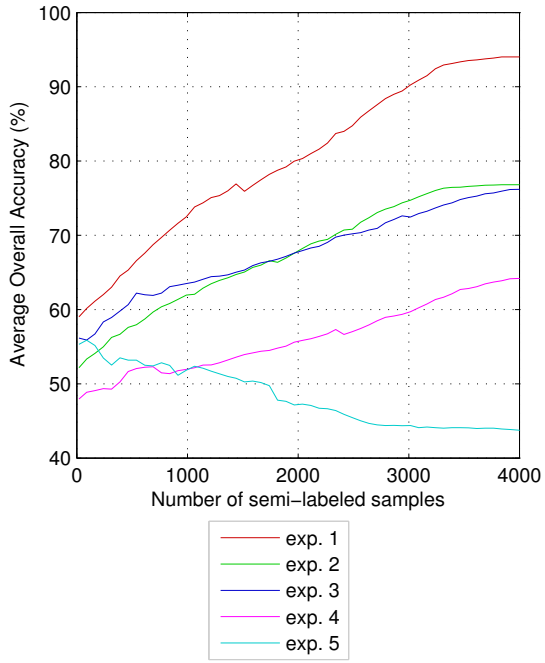

Fig. 4. Average (over ten trials) overall accuracy on the test set versus the number of semi-labeled training samples used by the classifier in the five experiments using PS^{3}VM (Two-moon toy data set).

B. Multiclass toy data set

We generated six different data sets in a two-dimensional feature space with four Gaussian-distributed classes. A pool and a test set are randomly generated by drawing 200 data points for each class from four Gaussian distributions with mean $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, and standard deviation σ varying from 0.01 to 0.15. Increasing values of σ determine increasing overlap between the distributions of the classes, making the cluster assumption less and less valid. Sets of

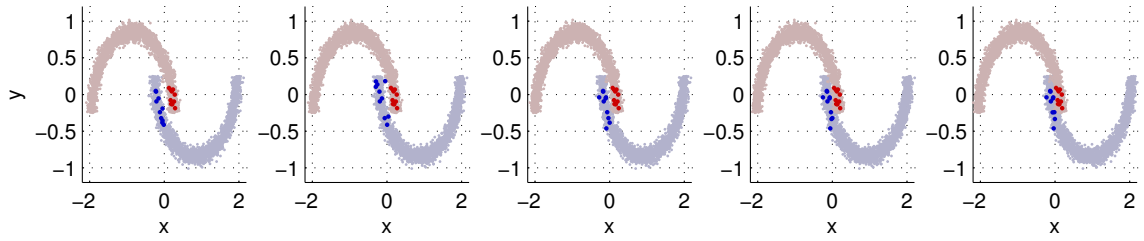

Fig. 3. Distribution of the synthetically generated samples. Dark points refer to the pool, while bright colored (red and blue) ones refer to the training set (Two-moon toy data set).

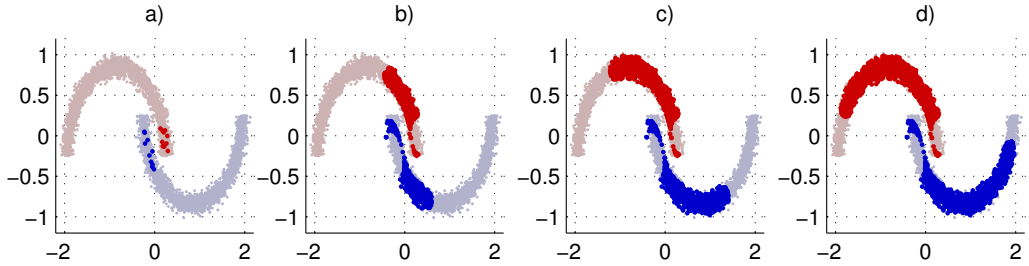

Fig. 7. Distribution of unlabeled and semi-labeled data at different iterations using the PS^{3}VM method for experiment 1: a) original training set, b) 600 semi-labeled points, c) 1200 semi-labeled points, and d) 1800 semi-labeled points.

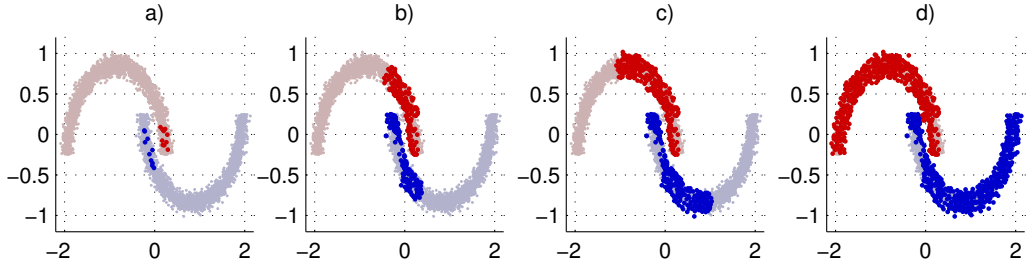

Fig. 8. Distribution of unlabeled and semi-labeled data at different iterations using the PS^{3}VM-D method for experiment 1: a) original training set, b) 300 semi-labeled points, c) 500 semi-labeled points, and d) 1000 semi-labeled points.

TABLE I
OVERALL ACCURACIES AVERAGED OVER TEN TRIALS OBTAINED BY THE CONSIDERED SUPERVISED, AL AND SSL TECHNIQUES IN THE FIVE EXPERIMENTS. THE SECOND COLUMN REPORTS THE JENSEN-SHANNON DIVERGENCE BETWEEN TRAINING SAMPLES OF THE TWO CLASSES. (TWO MOON TOY DATA SET).

Exp.	JS-Divergence	SVM	PS^{3}VM	PS^{3}VM-D	AL
1	0.371	59.25%	94.02%	96.24%	99.95%
2	0.361	51.37%	76.80%	78.55%	99.95%
3	0.341	55.87%	76.29%	66.67%	99.87%
4	0.314	47.85%	64.24%	56.82%	98.57%
5	0.305	55.97%	43.77%	30.76%	98.57%

samples of different sizes are selected with bias from the pool to define different training and validation sets. The selection bias in the definition of the training sets affects the estimation of the probability $P^{tr}(\mathbf{x}|y)$. For each of the six experiments, we obtained different training sets made up of 1, 2, 4, 8, 16 samples per class. Figure 10 shows the distribution of the pool and training samples in the case of 8 samples per class. We compared the standard supervised SVM with the MCLU and PS^{3}VM methods in the different classification problems defined by the values of σ and the number of training examples. Given the small number of samples and the limited complexity of the data set, diversity in not considered in these experiments.

The obtained results are reported in Table II. The classification accuracies obtained by supervised SVM strongly depend on both the complexity of the classification problem (here represented by the value of σ) and the number of available training samples. Very high accuracies are obtained with small values of σ, because the classification problem is not very complex. The accuracies decrease significantly by increasing the values of σ. The SSL method was effective in problems characterized by low/moderate overlap among the classes and sufficient number of training samples, leading to gain in accuracy up to 4.4% w.r.t. SVM. For complex problems with highly overlapping classes and few training samples, the SSL can also decrease the performance of supervised learning reducing the OA up to 3.3%. As expected, the AL method

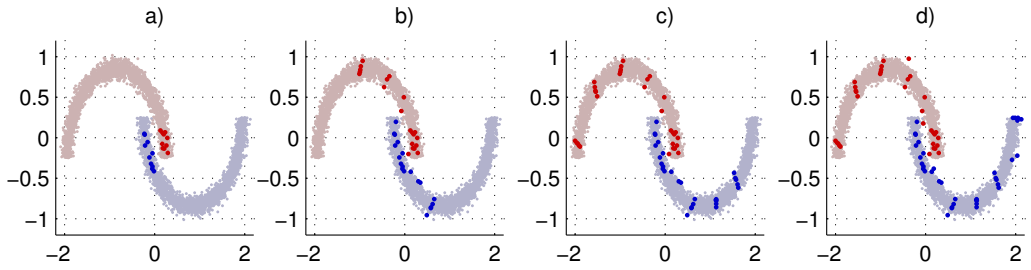

Fig. 9. Distribution of labeled data at different iterations of the AL method (MCLU-ECBD) for experiment 1: a) original training set, b) 40 labeled points, c) 60 labeled points, and d) 70 labeled points.

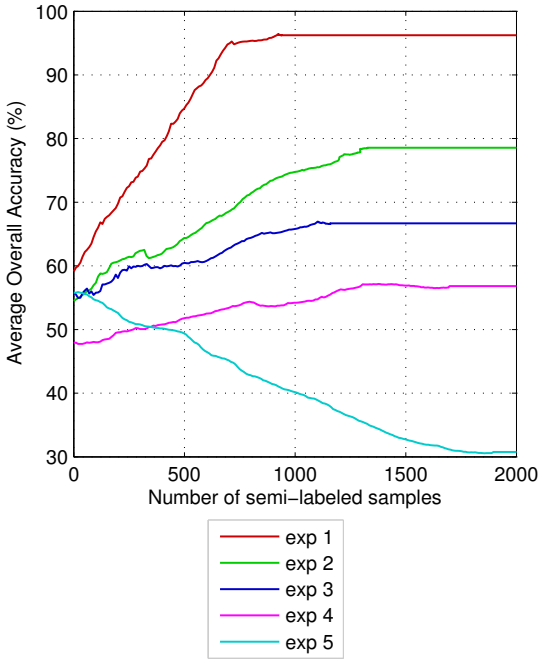

Fig. 5. Average (over ten trials) overall accuracy on the test set versus the number of semi-labeled training samples used by the classifier in the five experiments using PS^{3}VM-D (Two-moon toy data set).

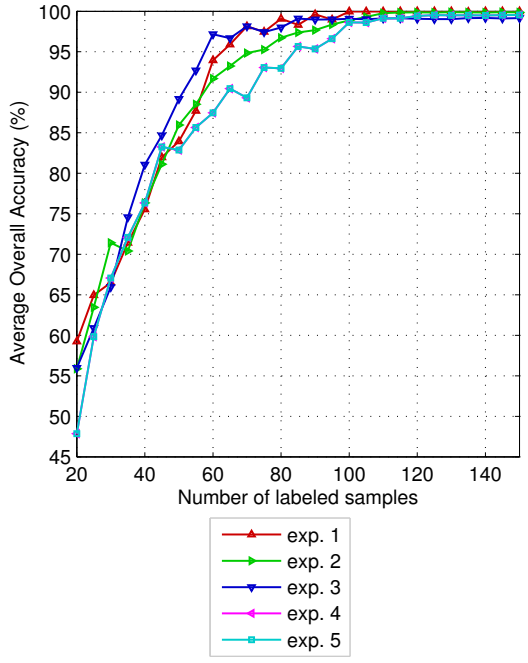

Fig. 6. Average (over ten trials) overall accuracy on the test set versus the number of training samples used by the classifier in the five experiments using AL (Two-moon toy data set).

reached classification accuracies (computed after labeling 100 samples) that are almost independent from the initial training set. They only depend on the overlap of the distribution of the classes.

C. Multispectral remote sensing data set

We considered a Quickbird image acquired in 2006 over a portion of the urban area of the city of Trento, Italy (see Figure 11). The four multispectral bands have been transformed to the same spatial resolution of 0.7 m as the panchromatic band using a GramSchmidt pan-sharpening procedure [54]. The size of the image is 2000 $\times$ 2000 pixel. From the original spectral bands three textural features have been extracted and stacked to the feature vector with the four multispectral bands and the panchromatic band. We defined a binary classification problem aimed to distinguish the two classes "urban area" and "rest" (mainly vegetation). The available labeled samples consist of ten initial training sets containing 40 samples for each experiment, a validation set with 64 samples, a pool with 5986 samples, and a test set with 14435 samples. We performed four different experiments varying the distribution of the initial training samples, thus simulating four different levels of sampling bias. In the first experiment, the initial training samples were randomly selected with uniform distribution from all the available samples in order to model the real distribution of the classes (no sampling bias). In the other experiments we selected training samples that are less and less representative of the real distribution of the classes, creating the conditions of a sample selection bias problem. This is obtained by selecting samples that belong only to one specific type of building, i.e., red roof buildings, for representing the class "urban area", and one specific type of agriculture field to represent the class "rest". Figure 12 reports the distribution of the training set and the pool (considering

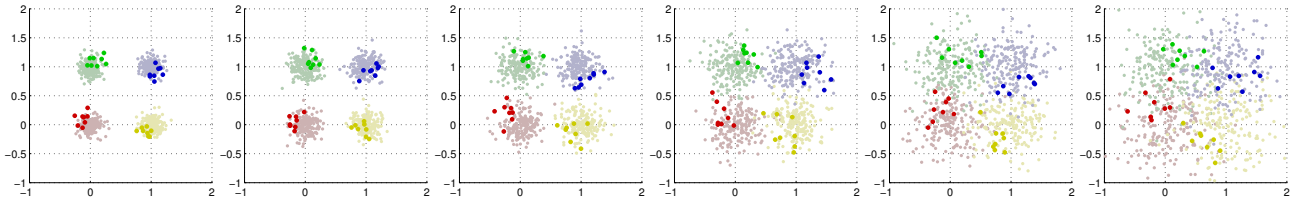

Fig. 10. Distribution of the synthetically generated samples. Dark points refer to the pool, while bright colored ones refer to the training set (Multiclass toy data set).

TABLE II
OVERALL ACCURACIES OBTAINED BY SUPERVISED SVM IN THE DIFFERENT CLASSIFICATION PROBLEMS (MULTICLASS TOY DATA SET).

SVM	Number of training samples per class				
σ	1	2	4	8	16
0.01	99.1% ±2.4	99.1% ±1.8	99.9% ±0.1	100.0% ±0.0	100.0% ±0.0
0.02	96.3% ±4.6	97.6% ±2.5	99.2% ±0.7	99.8% ±0.2	99.8% ±0.2
0.03	90.3% ±7.1	88.5% ±7.3	95.0% ±2.6	96.4% ±0.9	97.6% ±0.8
0.05	81.9% ±7.1	83.6% ±4.7	87.1% ±4.6	90.2% ±2.0	88.9% ±1.5
0.09	73.0% ±6.8	72.1% ±4.7	76.2% ±4.2	77.9% ±2.6	79.0% ±3.2
0.15	64.0% ±6.1	63.4% ±4.0	66.6% ±3.7	68.0% ±2.7	66.7% ±2.1

TABLE III
OVERALL ACCURACIES OBTAINED BY THE SSL METHOD IN THE DIFFERENT CLASSIFICATION PROBLEMS (MULTICLASS TOY DATA SET).

PS^{3}VM	Number of training samples per class				
σ	1	2	4	8	16
0.01	99.5% ±0.8	100.0% ±0.1	99.6% ±0.6	100.0% ±0.0	100.0% ±0.0
0.02	98.2% ±1.4	98.7 % ±1.3	99.3% ±1.4	100.0% ±0.0	100.0% ±0.1
0.03	91.9% ±6.5	91.2 % ±7.1	97.8% ±0.7	98.0% ±0.7	98.8% ±0.3
0.05	84.8% ±5.1	81.3 % ±6.1	91.5% ±2.5	92.5% ±1.8	92.4% ±1.2
0.09	72.9% ±7.7	68.8 % ±6.9	79.8% ±2.2	80.0% ±2.5	81.1% ±3.2
0.15	63.8% ±6.1	61.6 % ±4.2	68.3% ±2.5	68.6% ±2.9	67.4% ±2.0

TABLE IV
OVERALL ACCURACIES OBTAINED BY THE AL METHOD IN THE DIFFERENT CLASSIFICATION PROBLEMS AFTER LABELING 100 SAMPLES (MULTICLASS TOY DATA SET).

MCLU	Number of training samples per class				
σ	1	2	4	8	16
0.01	100.0% ±0.0	100.0% ±0.0	100.0% ±0.0	100.0% ±0.0	100.0% ±0.0
0.02	100.0% ±0.1	100.0% ±0.0	100.0% ±0.0	100.0% ±0.0	100.0% ±0.1
0.03	99.7 % ±0.1	99.6 % ±0.3	99.7 % ±0.2	99.5 % ±0.3	99.7% ±0.2
0.05	96.5 % ±1.1	96.6 % ±0.7	97.1 % ±0.8	97.0 % ±0.7	96.7% ±0.5
0.09	88.5 % ±1.7	88.9 % ±1.8	88.7 % ±1.6	89.0 % ±1.1	89.7% ±1.0
0.15	77.2 % ±5.2	78.8 % ±1.4	78.1 % ±1.9	79.7 % ±1.3	78.7% ±1.3

bands 3 and 4 of the multispectral image). Column two of Table V shows the Jensen-Shannon Divergence in order to evaluate the distance between the distribution of the pool and training samples (averaged over both the ten trials and the two information classes). There distances give us information about the capability of the training set to model the real distribution of the classes $P(\mathbf{x}|y)$. Note that in this case we evaluated the distances between the distributions of the same class considering the samples of the pool and the training set (not the distance between the two classes as done for the two-moon data set). The computed distance values confirm that the training set samples become less and less representative of the real distribution of the classes increasing the reference number of the experiment.

Figures 13 and 14 show the behavior of the averaged OA versus the number of semi-labeled samples added to the training set obtained in the different experiments using the PS^{3}VM and PS^{3}VM-D methods, respectively. From these graphs we can observe that in the first experiment, where the original training set can well represent the distribution of the classes, the SSL algorithms could slightly increase the accuracy of the supervised SVM classifier. In experiments 2 and 3, where the original training set does not properly model the real distribution of the data, the SSL techniques performed very well by increasing the accuracy of the standard SVM by about 20%. In experiment 4, where the distribution of the

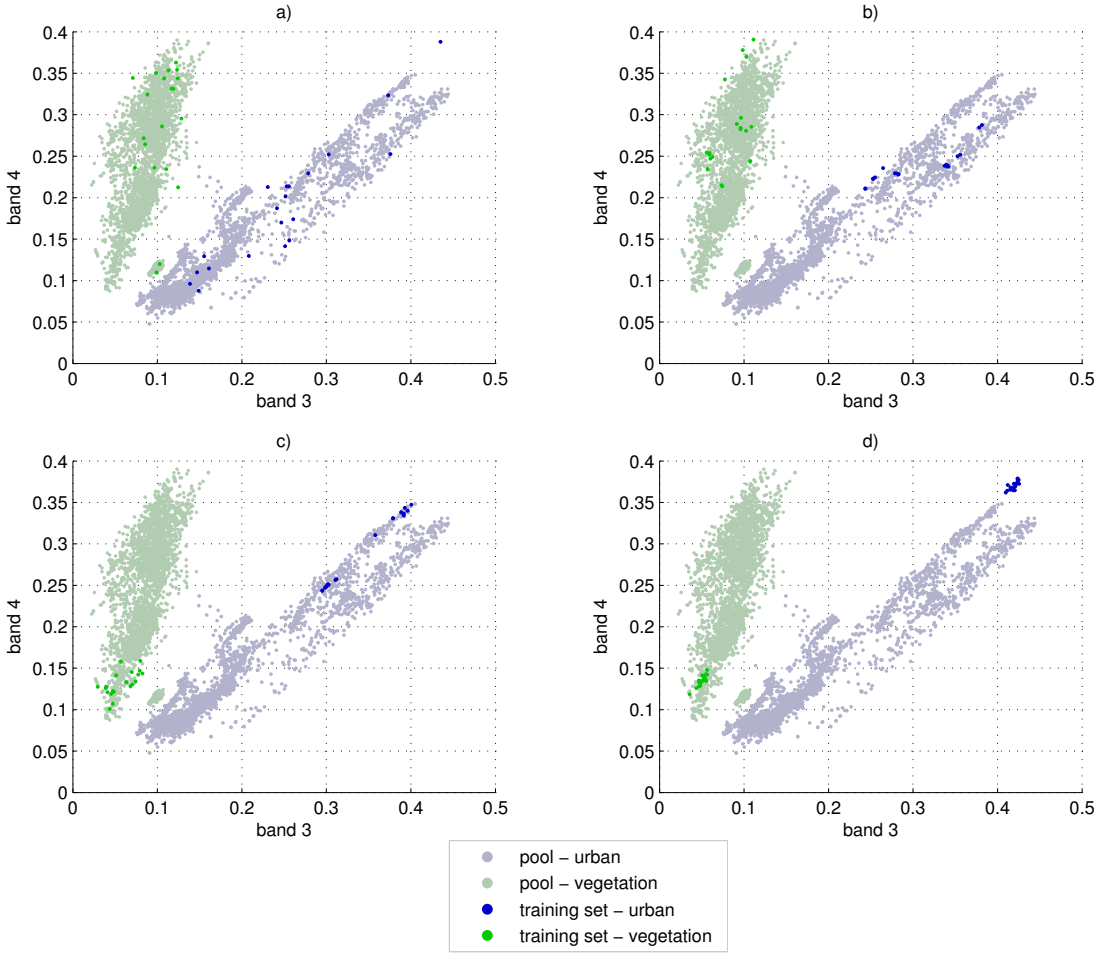

Fig. 12. Distribution of training and pool samples considering bands 3 and 4 of the multispectral image for: a) experiment 1, b) experiment 2, c) experiment 3, and d) experiment 4. (Multispectral RS data set).

training set is too far from the real one, PS^{3}VM is not able to effectively exploit the information of unlabeled samples and does not improve the accuracy of the SVM. Conversely, the PS^{3}VM-D could still improve the accuracy of standard SVM. We observe also on this data set that the diversity criterion can generally speed-up the convergence of the iterative algorithm, reducing the required training effort. In all experiments, the AL technique was able to significantly increase the classification accuracy of the standard SVM leading to an OA accuracy higher than 98% (see Figure 15). This is possible at the expense of additionally (manually) labeling about 80 samples from the pool. Table V reports the OAs obtained by the considered classifiers in the four experiments. As expected, also with this data set we can observe that AL was able to cope with different ill-posed classification problems. On the contrary, the performances of SSL techniques can decrease when the Jensen-Shannon Divergence is too high, i.e., the training set is less representative of the real distributions of the classes.

Figures 16 and 17 report the distribution of unlabeled and semi-labeled samples at different iterations of the PS^{3}VM learning process for experiment 2 and 4, respectively. In experiment 2, the SSL method can slowly explore the real distribution of the classes (including few mislabeled samples in the training set), converging to a good classification accuracy. In experiment 4, the PS^{3}VM starts considering several mislabeled samples after that 2000 samples are included in the training set, causing therefore a decrease in the OA. Figures 18 and 19 show the distribution of unlabeled and labeled samples at different iterations of the AL process for experiment 2 and 4, respectively. In such cases, the labeling of few samples close to the decision boundary is sufficient to converge to a reliable classification rule.

TABLE V
OVERALL ACCURACY (AVERAGED OVER TEN TRIALS) OBTAINED BY THE DIFFERENT CONSIDERED AL AND SSL TECHNIQUES IN THE FOUR EXPERIMENTS. (MULTISPECTRAL RS DATA SET).

Exp.	JS-Divergence	Supervised	PS^{3}VM	PS^{3}VM-D	AL
1	0.062	95.29%	95.82%	96.59%	98.26%
2	0.179	70.81%	91.18%	93.58%	98.07%
3	0.264	60.02%	82.85%	79.12%	98.62%
4	0.315	58.60%	52.46%	73.24%	98.57%

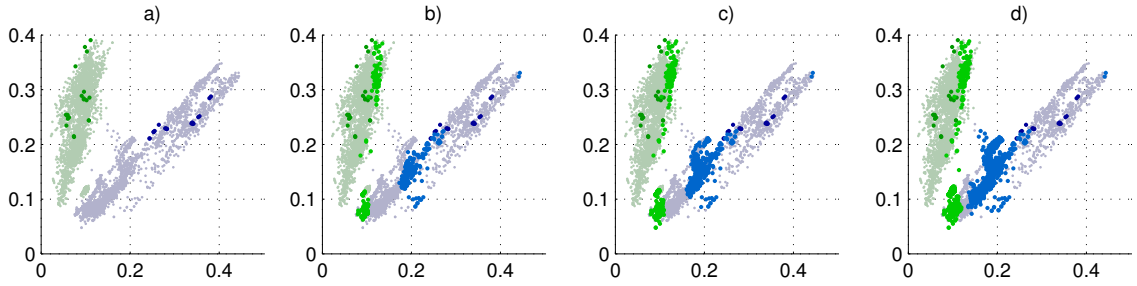

Fig. 16. Distribution of unlabeled and semi-labeled data at different iterations using the PS^{3}VM method for experiment 2 considering band 3 and 4. a) original training set, b) 500 semi-labeled points, c) 1000 semi-labeled points, and d) 2000 semi-labeled points (Multispectral RS data set).

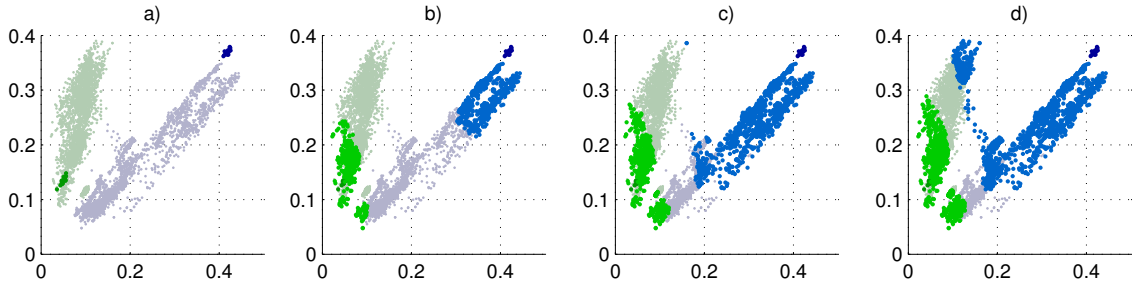

Fig. 17. Distribution of unlabeled and semi-labeled data at different iterations using the PS^{3}VM method for experiment 4 considering band 3 and 4. a) original training set, b) 1000 semi-labeled points, c) 2000 semi-labeled points, and d) 3000 semi-labeled points (Multispectral RS data set).

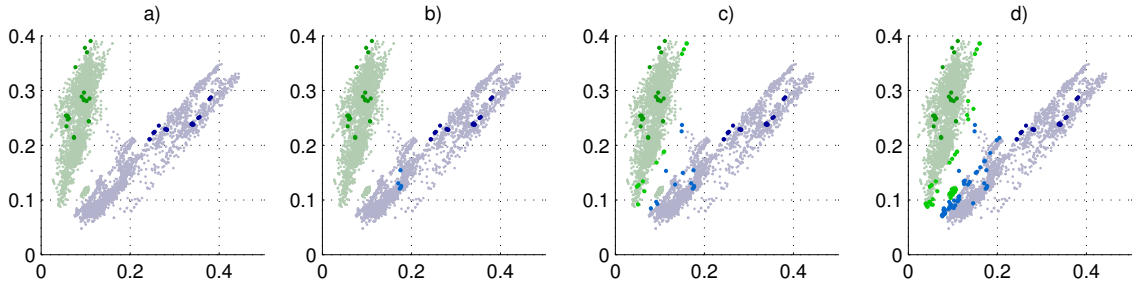

Fig. 18. Distribution of unlabeled and labeled data at different iterations of the AL process (MCLU-ECBD) for experiment 2 considering band 3 and 4. a) original training set, b) 5 added labeled points, c) 25 added labeled points, and d) 100 added labeled points (Multispectral RS data set).

D. Hyperspectral remote sensing data set

The last data set is made up of a hyperspectral image acquired by the Hyperion sensor of the EO-1 satellite in an area of the Okavango Delta, Botswana. The considered image has a spatial resolution of 30 m over a 7.7 km strip in 145 bands. For greater details on this data set, we refer the reader to [9]. Reference labeled samples for 14 land-cover classes are available for two spatially disjoint areas, which are referred in the following as Area 1 and Area 2, representing two different geographical areas with the same set of land-cover classes characterized by slightly different distributions. The labeled samples taken from Area 1 were randomly partitioned into two sets T_1 and VAL_1 and the samples of Area 2 were similarly partitioned into a training set T_2 and a test set TS_2, as in [55] (see Table VI for detailed information). Starting from T_1, we derived ten initial training sets by subsampling it at different rates from 5% to 50%. For each subsampling rate, ten different initial training sets are obtained. We used T_2 as pool of unlabeled samples and TS_2 as test set. In this way, we have derived biased training sets, where the bias is caused by a non-homogeneous sampling of the image in the spatial domain between the training and test sets, as we have described in Section II.

The mean OAs (averaged over ten trials for every experiment) obtained with the different learning paradigms are reported in Table VII. Both SSL methods improved the classification accuracies obtained by the supervised SVM in all experiments except the first two. This confirms, once again, the capability of SSL to improve the classification accuracies of standard supervised methods, when the training samples are few and biased. The experiments on this data set show the effectiveness of SSL, when the sampling bias is due to the

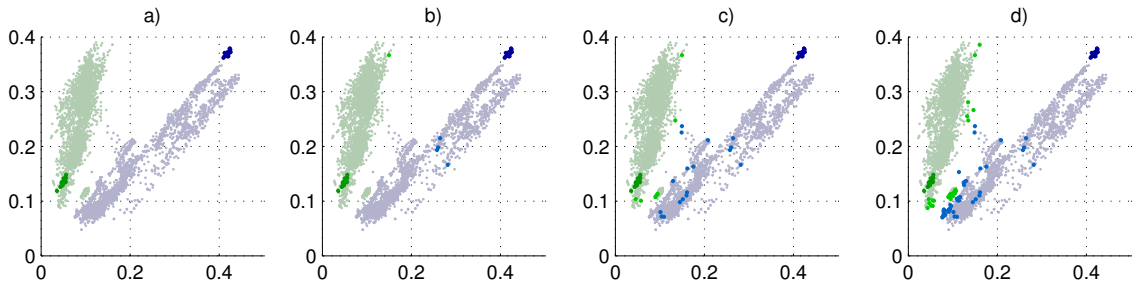

Fig. 19. Distribution of unlabeled and labeled data at different iterations of the AL process (MCLU-ECBD) for experiment 4 considering band 3 and 4. a) original training set, b) 5 added labeled points, c) 25 added labeled points, and d) 100 added labeled points (Multispectral RS data set).

Fig. 11. True color composition of the considered Quickbird image.

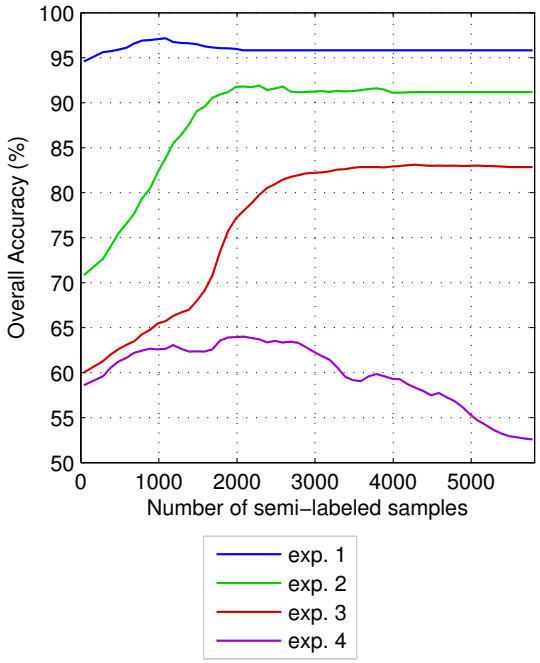

Fig. 13. Average (over ten trials) overall accuracy on the test set versus the number of semi-labeled samples considered by PS^{3}VM classifier in the four experiments. (Multispectral RS data set).

dependence of the selection variable s from the geographical location, which is a typical problem in the classification of RS images. We observe that PS^{3}VM-D performed slightly worse than PS^{3}VM in the first two experiments, but it reached higher accuracies in all other cases. The average OA obtained by the PS^{3}VM-D versus the number of semi-labeled samples added to the training set is shown in Figure 20 (for a significant subset of experiments). The MCLU-ECBD AL method converged to an accuracy above 96% after adding about 500 labeled samples to the training set in all considered experiments.

We performed additional experiments in order to compare different AL methods starting from the training sets of experiment 10. We compared the OAs obtained by the following methods: 1) random sampling (RAND), 2) MCLU, 3) MCLU-ECBD (with batch-size five), 4) Multiview-based AL using five views generated by correlation-partition-based clustering as reported in [47] (Multiview C), and 5) Multiview-based AL using five views randomly generated (Multiview R). All methods except MCLU-ECBD are applied to the selection of one sample per iteration. The obtained results (see Figure 21), show that the MCLU-ECBD technique led to the highest OAs compared to the other considered methods. The results confirm the effectiveness of multiclass confidence used in the MCLU. Moreover, the diversity approach adopted in MCLU-ECBD results in being very effective in both reducing the training effort and improving classification accuracies.

Finally, we performed other experiments in order to combine AL with SSL with two different strategies: 1) sequential strategy, and 2) SSL encapsulation in the AL process. In the sequential approach AL is first used for a number of iterations in order to include t additional labeled samples in the training set. Afterward a SSL method is executed. In our experiments, we first run the MCLU-ECBD and then the PS^{3}VM-D algorithm. In the second strategy, we encapsulated the PS^{3}VM-D algorithm (running 30 iterations) in the AL process using MCLU-ECBD. Figure 22 shows the learning curves obtained by the sequential strategy for different values of t and by the SSL-encapsulation strategy. We notice that the combination of the two approaches was effective in most

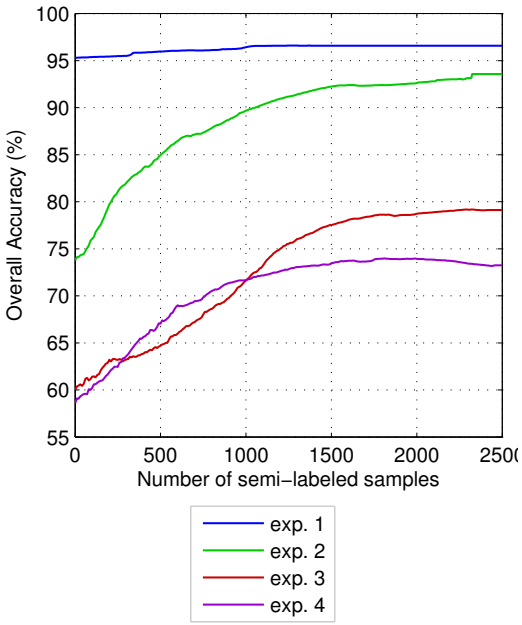

Fig. 14. Average (over ten trials) overall accuracy on the test set versus the number of semi-labeled samples considered by PS^{3}VM-D classifier in the four experiments. (Multispectral RS data set).

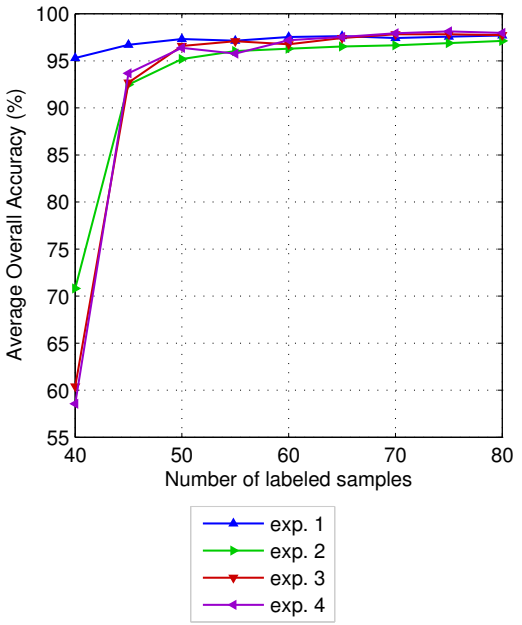

Fig. 15. Average (over ten trials) overall accuracy on the test set versus the number of training samples used by the classifier in the four experiments using AL (MCLU-ECBD). (Multispectral RS data set).

of our trials. For small values of t (i.e., $t = 10, 20, 30$), the application of SSL after AL could further increase the OA by approximately 2-3%, without any additional labeling cost for the user. For higher values of t (i.e, $t = 90, 140$), the accuracy obtained with AL was already high and SSL could scarcely improve it further. In those cases, only the selection of very uncertain samples could further increase the classification accuracy. The second strategy led to a small improvement with respect to the standard MCLU-ECBD AL method, without any additional labeling cost for the user.

TABLE VI
NUMBER OF AVAILABLE LABELED SAMPLES FOR THE HYPERSPECTRAL REMOTE SENSING DATA SET.

Class	Number of Samples			
	Area 1		Area 2	
	T_1	VAL_1	T_2	TS_2
Water	69	57	213	57
Hippo Grass	81	81	83	18
Floodplain Grasses 1	83	75	199	52
Floodplain Grasses 2	74	91	169	46
Reeds	80	88	219	50
Riparian	102	109	221	48
Firescar	93	83	215	44
Island Interior	77	77	166	37
Acacia Woodlands	84	67	253	61
Acacia Shrublands	101	89	202	46
Acacia Grasslands	184	174	243	62
Short Mopane	68	85	154	27
Mixed Mopane	105	128	203	65
Exposed Soil	41	48	81	14
Total	1242	1252	2621	627

TABLE VII
MEAN AND STANDARD DEVIATION OF THE OVERALL ACCURACY OBTAINED BY THE DIFFERENT CONSIDERED LEARNING PARADIGMS IN THE TEN EXPERIMENTS USING INITIAL TRAINING SETS OF DIFFERENT SIZES $|T|$. (HYPERSPECTRAL RS DATA SET).

Exp.	$\|T\|$	SVM	PS^{3}VM	PS^{3}VM-D	MCLU-ECBD
1	58	70.3% $\pm$4.3	73.2% $\pm$3.6	69.1% $\pm$7.2	96.6% $\pm$0.5
2	116	74.1% $\pm$2.4	76.1% $\pm$2.2	74.7% $\pm$6.6	96.4% $\pm$0.6
3	174	74.7% $\pm$2.7	76.3% $\pm$3.0	78.0% $\pm$4.5	96.5% $\pm$0.8
4	232	74.2% $\pm$2.8	75.7% $\pm$2.2	76.0% $\pm$5.4	96.4% $\pm$0.6
5	290	74.7% $\pm$1.3	76.3% $\pm$2.0	78.4% $\pm$2.6	96.0% $\pm$0.6
6	348	74.9% $\pm$2.2	75.9% $\pm$2.2	78.7% $\pm$3.6	96.5% $\pm$0.7
7	406	75.8% $\pm$1.9	77.0% $\pm$1.8	78.9% $\pm$2.4	96.1% $\pm$0.6
8	464	75.7% $\pm$1.4	77.6% $\pm$2.0	80.9% $\pm$1.4	96.6% $\pm$0.4
9	522	75.7% $\pm$1.0	77.5% $\pm$0.9	81.0% $\pm$1.0	96.6% $\pm$0.4
10	580	76.3% $\pm$0.1	78.1% $\pm$0.1	81.1% $\pm$0.2	96.6% $\pm$0.3

VII. DISCUSSION

In this paper we have presented a comparative study in order to analyze AL and SSL in relation to the classification of RS images. We reviewed the main categories and methods of both learning paradigms. The two approaches have been theoretically compared in light of a conceptual framework to describe the workflow of AL and iterative SSL methods. We also investigated the combination of AL and SSL in order to jointly leverage the advantages of both approaches and proposed a novel SSL iterative method that is inspired by concepts that are usually considered in AL methods.

The two learning paradigms have been experimentally compared in the classification of different simulated and real RS data sets, addressing different problems characterized by *small* and *biased* training sets. On the basis of our analysis we derived the following general conclusions about these two approaches:

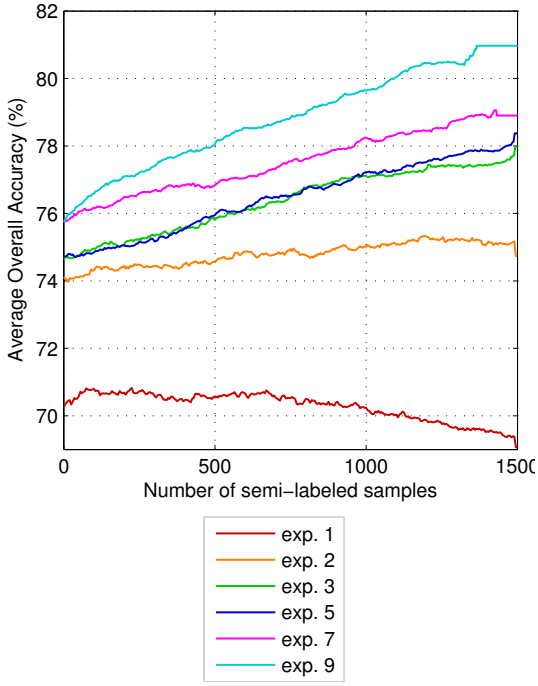

Fig. 20. Average (over ten trials) OA obtained by the PS^{3}VM-D with respect to the number of semi-labeled samples added to the training set for experiments 1, 2, 3, 5, 7, 9 (Hyperspectral RS data set).

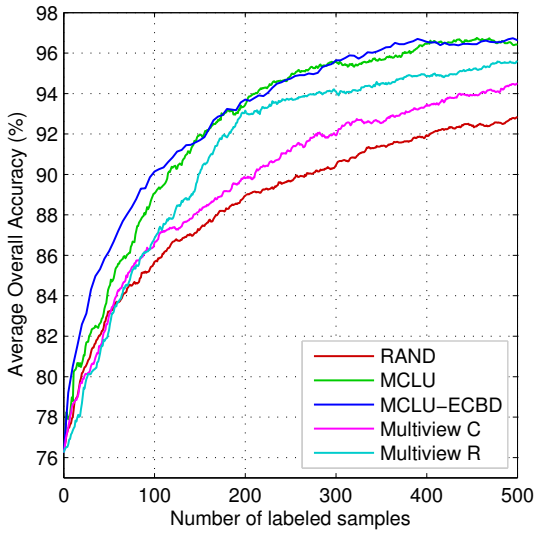

Fig. 21. Average (over ten trials) OA obtained versus the number of labeled samples using the following methods: 1) random sampling (RAND), 2) MCLU, 3) MCLU-ECBD, 4) Multiview-based AL using five views generated by correlation-partition-based clustering as reported in [47] (Multiview C), and 5) Multiview-based AL using five views randomly generated (Multiview R). (Hyperspectral RS data set)

1) SSL techniques can improve the classification accuracies obtained by supervised classifiers depending on the initial conditions and assuming that the cluster assumption holds. When the training samples fairly represent the real distribution of the classes, the SSL technique can improve the accuracy of the supervised classifier.

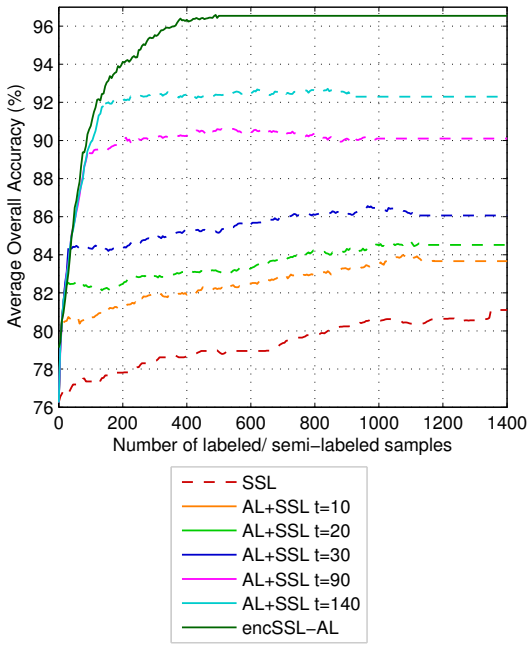

Fig. 22. Average (over ten trials) OA obtained versus the number of labeled/semi-labeled samples added to the training using different combination strategies for AL and SSL: 1) sequential strategy (AL+SSL), and 2) SSL encapsulation in the AL process (encSSL-AL). For the sequential strategy, solid and dashed lines correspond to AL and SSL phase, respectively (Hyperspectral RS data set).

Under these conditions it represents an effective method for addressing ill-posed problem characterized by *few* and/or *biased* initial training samples without requiring an additional effort of the user for labeling samples. On the contrary, the performances of the SSL technique significantly decrease when the training samples are too less representative of the true underlying distribution and when the cluster assumption does not hold. However, it is worth noting that on real data sets it might be difficult to asses the validity of the cluster assumption and therefore it is not easy to understand in practice whether SSL may improve the accuracy of standard supervised methods or not. This makes the use of SSL techniques difficult for non-experts of machine learning, when a proper validation of the classification results is not possible.

2) AL techniques can converge to good classification accuracies starting from any initial training set without any assumption on the distribution of the classes (i.e., the cluster assumption is not necessary) at the expense of additional labeling effort. For this reason, AL represents a good alternative approach to address ill-posed problems when the SSL assumption does not hold. In such cases, AL results in a very useful tool for guiding the user in the collection of the most informative labeled samples. It is important to observe that the labeling costs substantially depends on the type of labeling procedure, i.e., 1) by photo-interpretation, or 2) by ground survey. Applications where the annotation by photo-

interpretation is possible, may benefit more from the adoption of AL. In such cases, the labeling time is reasonably short compared to the computational time required by SSL methods.

3) The model selection is more critical with SSL techniques due to the importance of the parameter values on the final results and the lack of labeled data for validation. This is critical because makes it difficult to assess the quality of the obtained solution at convergence.
4) AL techniques typically require much less iterations to reach convergence with respect to iterative SSL techniques (e.g., PS^{3}VM). However, the additional iterations of SSL techniques are automatic and do not involve the user, but just the time taken by the considered computation machine.
5) The proposed PS^{3}VM-D technique, which adopts a diversity criterion and a multiclass-based confidence measure, allows the iterative SSL approach to better explore the distribution of the classes and to reach converge in less iterations with respect to the standard algorithm.
6) From our analysis, we can conclude that AL techniques are effective and ready to be used in operational applications. SSL techniques still require additional work to be adopted in operational applications in order to relate their convergence properties to the considered problem and to the available training samples, as well as to further investigate validation procedures.
7) AL and SSL paradigms can be effectively combined in order to define learning algorithms that exploit both labeled and semi-labeled samples in the training phase and for the selection of new samples to be labeled by the user.

Possible future developments of this work are: 1) further investigating the integration of AL and SSL for the development of hybrid solutions, 2) the deeper investigation of validation methods for SSL techniques using strategies like the ones proposed in [56], [57]. Regarding this latter point, novel validation strategies should be able to asses model consistency after the inclusion of semi-labeled samples in the training set. These strategies should detect if the learning algorithm is converging to a good solution or moving toward an inconsistent solution (which will decrease the accuracy of the classifier) also in the critical conditions with few and biased labeled data in which cross-validation cannot be used in a reliable way.

ACKNOWLEDGMENT

The work of Dr. Claudio Persello has been supported by the Autonomous Province of Trento and the European Community in the framework of the project "Trentino - PCOFUND-GA-2008-226070 (call 3 - post-doc 2010 Outgoing)". The authors would like to thank Prof. M. Crawford for kindly providing the hyperspectral data set.

REFERENCES

[1] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778 – 1790, aug. 2004.

[3] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[4] S. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[5] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. John Wiley & Sons, 2009.

[6] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363 –3373, nov. 2006.

[7] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe, "Semisupervised image classification with laplacian support vector machines," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 3, pp. 336 –340, july 2008.

[8] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067 – 1074, 2004.

[9] S. Rajan, J. Ghosh, and M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231–1242, april 2008.

[10] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, March 2011.

[11] S. Patra and L. Bruzzone, "A fast cluster-assumption based active-learning technique for classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1617 –1626, may 2011.

[12] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–161, 1979.

[13] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[14] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press, 1999.

[15] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.

[16] X. Zhu, "Semi-supervised learning literature survey," University of Wisconsin – Madison, Tech. Rep., 2008.

[17] I. Scudder, H., "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363 – 371, july 1965.

[18] S. Fralick, "Learning to recognize patterns without a teacher," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 57 –64, january 1967.

[19] A. Agrawala, "Learning with a probabilistic teacher," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 373 – 379, july 1970.

[20] I. Dopido, J. Li, P. R. Marpu, A. Plaza, J. B. Dias, and J. A. Benediktsson, "Semisupervised self-learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 4032–4044, 2013.

[21] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Workshop on Computational Learning Theory*, 1998.

[22] V. R. de Sa, "Learning classification with unlabeled data," *Adv. Neural Inf. Process. Syst.*, vol. 6, no. 112-119, 1994.

[23] Q. Jackson and D. Landgrebe, "An adaptive method for combined covariance estimation and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 5, pp. 1082 –1087, may 2002.

[24] B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087 –1095, sep 1994.

[25] S. Tadjudin and D. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 439 –445, jan 2000.

[26] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 200–209. [Online]. Available: http://dl.acm.org/citation.cfm?id=645528.657646

[27] L. Bruzzone, M. Chi, and M. Marconcini, *Semisupervised support vector machines for classification of hyperspectral remote sensing images*, C.-I. Chang, Ed. Wiley, 2007, vol. Hyperspectral Data Exploitation: Theory and Applications.

[28] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by svms optimized in the primal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1870 –1880, june 2007.

[29] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Mach. Learn.*, vol. 56, no. 1-3, pp. 209–239, jun 2004.

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

[31] M. Li and I. Sethi, "Confidence-based active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251 –1261, aug. 2006.

[32] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994.

[33] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," *International Conference on Machine Learning*, pp. 839–846, 2000.

[34] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 111–118, 2000.

[35] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, March 2002.

[36] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 287–294.

[37] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," *Proceedings of the International Conference on Machine Learning*, pp. 150–157, 1995.

[38] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 1–9.

[39] 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*, vol. 6, p. 2005, 2004.

[40] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[41] K. Brinker, "Incorporating diversity in active learning with support vector machines," *Proceedings of the 20th International Conference on Machine Learning*, pp. 59–66, 2003.

[42] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *Proceedings of the 25th European Conference on Information Retrieval Research*, ser. ECIR '03, 2003.

[43] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 79–.

[44] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification." in *International Conference on Machine Learning*, 2006.

[45] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Neural Information Processing Systems*, 2007, pp. 593–600.

[46] J. Azimi, A. Fern, X. Z. Fern, G. Borradaile, and B. Heeringa, "Batch active learning via coordinated matching," in *International Conference on Machine Learning*, 2012.

[47] W. Di and M. M. Crawford, "View generation for multiview maximum disagreement based active learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1942–1954, 2012.

[48] ——, "Active learning via multi-view and local proximity co-regularization for hyperspectral image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 618–628, 2011.

[49] A. K. McCallum and K. Nigam, "Employing em and pool-based active learning for text classication," in *Proc. International Conference on Machine Learning*, 1998.

[50] U. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning=robust multi-view learning," in *Proceedings of the 19th International Conference on Machine Learning*, 2002, pp. 435–442.

[51] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, pp. 171–186, 2005.

[52] M. Li, R. Wang, and K. Tang, "Combining semi-supervised and active learning for hyperspectral image classification," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2013.

[53] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[54] B. Aiazzi, S. Baronti, M. Selva, and L. Alparone, "Enhanced gram-schmidt spectral sharpening based on multivariate regression of ms and pan data," in *IEEE Int. Geoscience and Remote Sensing Symposium*, 2006, pp. 3806–3809.

[55] C. Persello, "Interactive domain adaptation for the classification of remote sensing images using active learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 736 – 740, 2013.

[56] L. Bruzzone and M. Marconcini, "Domain adaptation problems: a dasvm classification technique and a circular validation strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 770–787, 2010.

[57] ——, "Toward an automatic updating of land-cover maps by a domain adapatation SVM classifier and a circular validation strategy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1108–1122, 2009.