A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images

Lorenzo Bruzzone, *Senior Member, IEEE*, Mingmin Chi, *Student Member, IEEE*, and Mattia Marconcini, *Student Member, IEEE*

***Abstract*—This paper introduces a semisupervised classification method that exploits both labeled and unlabeled samples for addressing ill-posed problems with support vector machines (SVMs). The method is based on recent developments in statistical learning theory concerning transductive inference and in particular transductive SVMs (TSVMs). TSVMs exploit specific iterative algorithms which gradually search a reliable separating hyperplane (in the kernel space) with a transductive process that incorporates both labeled and unlabeled samples in the training phase. Based on an analysis of the properties of the TSVMs presented in the literature, a novel modified TSVM classifier designed for addressing ill-posed remote-sensing problems is proposed. In particular, the proposed technique: 1) is based on a novel transductive procedure that exploits a weighting strategy for unlabeled patterns, based on a time-dependent criterion; 2) is able to mitigate the effects of suboptimal model selection (which is unavoidable in the presence of small-size training sets); and 3) can address multiclass cases. Experimental results confirm the effectiveness of the proposed method on a set of ill-posed remote-sensing classification problems representing different operative conditions.**

***Index Terms*—Ill-posed problems, labeled and unlabeled patterns, machine learning, remote sensing, semisupervised classification, support vector machines (SVMs), transductive inference.**

I. Introduction

ONE OF THE main critical issues in the application of supervised classification methods to the analysis of remote-sensing images is the definition of a proper training set for learning of the classification algorithm. In this context, two major problems should be considered: 1) a problem related to the quantity of the available training patterns and 2) a problem related to the quality of the available training samples.

As regards the quantity of the training patterns, often in practical applications, the number of available ground-truth samples is not sufficient to achieve a reliable estimate of the classifier parameters in the learning phase of the algorithm. In particular, if the number of training samples is relatively small compared to the number of features (and thus of classifier parameters to be estimated) [1], the well-known problem of the curse of dimensionality (i.e., the Hughes phenomenon [2]) arises. This results in the risk of overfitting the training data and may involve poor generalization capabilities in the classifier.

As regards the quality of the training data, two major issues are related to remote-sensing problems: 1) nonstationary nature of the spectral signatures of the land-cover classes in the spatial domain and 2) correlation among the training pixels taken from the same area. The nonstationary behavior of the spectral signatures in the spatial domain of the image depends on different physical factors related to both ground and atmospheric conditions. This behavior would require training samples representing each land-cover class to be collected from different portions of the scene, for a complete capturing of the information present in different realizations of spectral signatures of classes. However, since this is often unfeasible in real applications, the incomplete representation of statistical properties of the spectral signatures may affect the accuracy of the classification system. Another critical problem is that the training samples are often taken from the same site and appear as neighboring pixels in remote-sensing images. As the autocorrelation function of an image is not impulsive in the spatial domain, this violates the required assumption of independence among the samples included in the training set, thus reducing the information conveyed to the classification algorithm by the considered training patterns. Both the above problems result in unrepresentative training sets that affect the accuracy of the classification process.

The limited quantity and quality of the training samples involve the definition of ill-posed classification problems [1], [3], which cannot be solved properly with standard supervised classification techniques. This is critical in many remote-sensing applications, especially when considering hyperspectral images acquired by last-generation sensors [4]. Two major families of approaches have been investigated in the remote-sensing community for addressing ill-posed problems: 1) use of semisupervised learning methods that exploit both labeled and unlabeled samples [1], [3], [5], [6] and 2) use of supervised kernel-based methods, like inductive support vector machines (ISVMs).

Concerning semisupervised methods, in remote-sensing literature, ill-posed classification problems were mainly addressed by semisupervised classifiers based on parametric or semiparametric techniques that approximate the class distributions by a specific statistical model [4], [5], [7]. In [5], ill-posed problems were systematically investigated in remote-sensing applications for the first time. In terms of Fisher information, Shahshahani and Landgrebe have proved that the additional unlabeled samples are helpful for semisupervised

Manuscript received October 27, 2005; revised February 17, 2006.

L. Bruzzone and M. Marconcini are with the Department of Information and Communication Technologies, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

M. Chi is with the Department of Information and Communication Technologies, University of Trento, 38050 Trento, Italy and also with the Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China.

Digital Object Identifier 10.1109/TGRS.2006.877950

classification in the context of a Gaussian maximum-likelihood (GML) classifier, under a zero-bias assumption. In [7], by assuming a Gaussian mixture model (GMM), Tadjudin and Landgrebe used the iterative expectation–maximization (EM) algorithm to estimate model parameters from both labeled and unlabeled samples. In order to limit the negative influence of semilabeled samples (which are unlabeled samples that obtained the labels at the previous learning iteration [7]) to the estimation of the parameters of a GML classifier, a weighting strategy was introduced in [4], i.e., full weights were assigned to the training samples, while reduced weights were set for the semilabeled samples during the estimation phase of the EM algorithm. However, a major problem is that the covariance matrices are highly variable when the size of the training set is small. To overcome this problem, an adaptive covariance estimator was proposed in [4] to deal with ill-posed problems in the classification of hyperspectral data. In the adaptive quadratic process, the semilabeled samples are incorporated in the training set to estimate the regularized covariance matrices, so that the variance of these matrices can be smaller compared to the conventional counterparts [7]. However, these techniques cannot be used properly when reliable statistical models cannot be defined for the class distributions (e.g., in the case of the classification of multisource/multisensor images) [8]. Nevertheless, studies on semisupervised learning with nonparametric techniques are rarely reported in the classification of remote-sensing images. In previous works [3], [8], the authors addressed the ill-posed problem by a semisupervised learning with nonparametric techniques. In particular, the use of an ensemble of nonparametric classifiers with a semisupervised learning was proposed, where each classifier (except for the original one) was designed with both labeled and unlabeled samples in the context of an iterative procedure. Then, as in bagging [9], the results obtained from the members of the ensemble were combined. In [8], the ill-posed problem was addressed in the context of a nonparametric k-NN classifier, based on an estimation procedure local to the specific portion of the feature space of the analyzed pattern. Due to the diversity of the selected semilabeled samples at different iterations, the ensemble-driven k-NN approach proved successful in dealing with ill-posed problems. In [3], bagging driven by the semilabeled samples was proposed in the context of an another nonparametric algori thm based on multilayer perceptron neural networks (MLPNNs) [10]. In this case, the error backpropagation (EBP) learning algorithm optimized a cost function that jointly considered all the available training samples as well as the semilabeled patterns. To limit the possible negative effects of the semilabeled samples, a weighting scheme was exploited in the EBP algorithm to reduce the importance assigned to the semilabeled patterns.

With regard to the kernel-induced methods, such as the ISVMs, recently, they have been applied promisingly to ill-posed remote-sensing classification problems [3], [6], [8], [11], [12]. The ISVMs are large-margin classifiers that define an optimal separating hyperplane in a properly chosen kernel-induced feature space. Due to Mercer's conditions, this is implemented by a convex optimization problem under nonlinear inequality constraints. Hence, compared to the other methodologies such as the MLPNNs [10], ISVMs have no local minima. Another important property of the ISVMs is that there is a sparse dual representation of the quadratic optimization problem due to Karush–Kuhn–Tucker (KKT) conditions [13], [14]. Hence, only a subset of the training samples (the ones associated with the nonzero Lagrange multipliers, which are called support vectors) contribute to the classification rule. In addition, for the nonlinear case, the data can be projected into a higher dimensional feature space with a nonlinear mapping function. The mapping can be represented implicitly by kernel functions in the sparse dual representation. As a consequence, the use of kernels makes it possible to train a linear machine in the kernel-induced feature space, potentially circumventing the high-dimensional feature problem inherent in ill-posed problems with the reduced nonzero (Lagrange) parameters [12], [15]. Moreover, for the nonseparable case, a loss function is exploited to penalize the training errors, thus reducing the risk of overfitting. However, for small-size training sets (i.e., in ill-posed problems), large deviations are possible for the empirical risk. In addition, the small sample size can force the overfitting or underfitting of the supervised learning. This may result in a low classification accuracy as well as in poor generalization capabilities. To address this problem, semisupervised SVMs (called transductive SVMs (TSVMs) in [6], [16]), which exploits both labeled and unlabeled samples, has been proposed in the machine learning community [17]. Bennett and Demiriz [11] used L1-norm linear SVMs to implement semisupervised SVMs showing little improvement when insufficient training information is available. In [18], the author solved the quadratic optimization problem for the implementation of the TSVMs with an application to text classification. The effectiveness of the TSVMs for text classification (in a high-dimensional feature space) was supported by theoretical and experimental findings. In the algorithm proposed by Joachims [18], an estimate of the ratio between the unlabeled positive and negative samples for the transductive learning should be known at the beginning. However, in real cases, prior knowledge is usually not available. Accordingly, in [19], a progressive TSVM algorithm was proposed to overcome the above drawback. In this algorithm, the positive and negative samples are labeled in a pairwise or pairwiselike way. From experimental results, Joachims [18] and Chen *et al.* [19] showed substantial improvement in the accuracy obtained with the TSVMs over that obtained with the ISVMs, especially for small-size training sets (i.e., in ill-posed classification problems). An important issue that should be highlighted is that, like ISVMs, TSVMs are binary classifiers. Nevertheless, at present, their generalization to the multiclass problems has not yet been investigated. Although there is still some debate about whether the transductive inference[1] can be successful in the semisupervised classification [20], it has been proved both empirically and theoretically [6], [11], [16], [18]

[1]It is worth noting that, from a theoretical viewpoint, the concept of transductive inference pioneered by Vapnik [6], [17] is closely related to semisupervised learning. Semisupervised learning refers to the use of both labeled and unlabeled data for training in the way to obtain classifiers defined over the whole space. Transductive learning, instead, is in contrast to the inductive inference: no general decision rule is inferred, but only the labels of the unlabeled (or test) points are predicted.

that TSVMs can be effective in handling problems where few labeled data are available (small-size labeled dataset), while the unlabeled data are easy to obtain (e.g., text categorization, biological recognition, etc.).

In this paper, we propose a TSVM classifier specifically designed for the analysis of ill-posed remote-sensing problems, which merges the advantages of semisupervised methods with those of kernel-based methods like the ISVMs. Besides introducing the TSVMs in the context of the remote-sensing data classification, this contribution presents three main methodological novelties related to the proposed TSVM. In particular:

1) It is based on an original iterative transductive procedure that exploits a weighting strategy for the unlabeled patterns, based on a time-dependent criterion.
2) It is able to mitigate the effects of a suboptimal model selection (which is unavoidable in the presence of small-size training sets).
3) It is developed in the multiclass case.

In order to assess the effectiveness of the proposed technique, many ill-posed classification problems have been defined using a multispectral Landsat Thematic Mapper image acquired over the Trentino area (Italy). The experimental results carried out on ill-posed classification problems in different operative conditions confirmed the effectiveness of the proposed TSVM, which showed both an increased robustness and a higher classification accuracy compared to standard ISVMs.

The rest of the paper is organized as follows. The next section introduces the basis of the ISVMs and TSVMs, and presents the proposed technique. Section III describes the datasets used in the experiments and reports the experimental results. Finally, Section IV draws the conclusion of this paper and discusses the future developments.

II. TSVMs for Semisupervised Classification

A. Basis of Inductive and Transductive SVMs

Let the given training set $X = (\mathbf{x}_l)_{l=1}^n$, made up of n labeled samples and the associated labels $Y = (y_l)_{l=1}^n$, $y_l \in \{1, \ldots, s, \ldots, S\}$. The unlabeled set $X^* = (\mathbf{x}_u^*)_{u=1}^m$ consists of m unlabeled samples and the corresponding predicted labels $Y^* = (y_u^*)_{u=1}^m$, obtained according to the classification model after learning with the training set.

The standard ISVM is a linear inductive learning classifier, where the data in the input space are separated by the hyperplane

$$y(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} + b \tag{1}$$

with the maximal geometric margin

$$\varphi(\mathbf{w}) = \frac{2}{\mathbf{w}^T \cdot \mathbf{w}} \tag{2}$$

where $\mathbf{w}$ is a vector normal to the hyperplane, b is a constant such that $b/\|w\|$ represents the distance of the hyperplane from the origin, and T denotes the transpose of a vector [13]. Hence, the objective of the ISVM is to solve the quadratic optimization problem in (2) with proper inequality constraints, i.e.,

$$\varphi(\mathbf{w}) = \min_{\mathbf{w}} \left\{ \frac{1}{2}(\mathbf{w}^T \cdot \mathbf{w}) \right\}$$
$$\text{subject to } \forall_{l=1}^n : y_l(\mathbf{w}^T \cdot \mathbf{x}_l + b) \geq 1. \tag{3}$$

To allow some training errors for generalization, the slack variables ξ_l and the associated regularization parameter C (whose value is user defined) are introduced for the nonseparable cases

$$\varphi(\mathbf{w}) = \min_{\mathbf{w},\xi_l} \left\{ \frac{1}{2}\mathbf{w}^T \cdot \mathbf{w} + C \sum_{l=1}^n \xi_l \right\}$$
$$\text{subject to } \begin{cases} \forall_{l=1}^n : y_l(\mathbf{w}^T \cdot \mathbf{x}_l + b) \geq 1 - \xi_l \\ \forall_{l=1}^n : \xi_l > 0. \end{cases} \tag{4}$$

Since direct handling of the inequality constraints is difficult, the Lagrange theory is usually exploited by introducing Lagrange multipliers for the quadratic optimization problem, which leads to the following alternative dual representation:

$$L(\boldsymbol{\alpha}) = \sum_{l=1}^n \alpha_l - \frac{1}{2} \sum_{l=1}^n \sum_{i=1}^n y_l y_i \alpha_l \alpha_i \langle \mathbf{x}_l, \mathbf{x}_i \rangle$$
$$\text{subject to } \begin{cases} 0 \leq \alpha_l \leq C, \quad 1 \leq l \leq n \\ \sum_{l=1}^n y_l \alpha_l = 0 \end{cases} \tag{5}$$

where $\langle \mathbf{x}_l, \mathbf{x}_i \rangle$ is the inner product between the two feature vectors and $\alpha_{l=1}^n$ are the Lagrange multipliers. The training samples associated with the nonzero multipliers (called support vectors) contribute to defining the separating hyperplane. If the data in the input space cannot be linearly separated, they are projected into a higher dimensional feature space with a nonlinear mapping function $\phi(\mathbf{x})$, where the inner product between the two mapped feature vectors becomes $\langle \phi(\mathbf{x}_l), \phi(\mathbf{x}_i) \rangle$. In this case, if we replace the inner product in (5) with a kernel function, we can avoid representing the feature vectors explicitly. The number of operations required to compute the inner product by evaluating the kernel function is not necessarily proportional to the number of features [14]. Hence, the use of kernels in a sparse dual representation potentially circumvents the high-dimensional feature problem with the reduced nonzero (Lagrange) parameters. In addition, due to Mercer's conditions on the kernels, unlike in other machine learning techniques based on neural networks (e.g., MLPNNs [10]), the positive semidefinite kernel $k(\mathbf{x}_l, \mathbf{x}_i)$ ensures that the objective function is convex, and hence, there are no local minima in the cost function of the ISVMs.

In the above framework, to alleviate the problem of the small-size training set further, a transductive SVM approach has been proposed in [17]. The TSVM technique is based on an iterative algorithm [18], [19]. At the initial iteration, the standard ISVMs are used to obtain an initial separating hyperplane based on the training data alone $(\mathbf{x}_l)_{l=1}^n$. Then, "pseudo" labels are given to the unlabeled samples $(\mathbf{x}_u^*)_{u=1}^m$, which are thus called semilabeled data. After that, transductive samples chosen from the semilabeled patterns according to a given criterion are used

to define a hybrid training set made up of these same samples and the original training samples in X. The resulting hybrid training set is used at the following iterations to find a more reliable discriminant hyperplane with respect to the distribution of all the patterns of the image. This hyperplane separates (X, Y) and (X^*, Y^*) with the maximal margin and is derived as follows:

$$\varphi(\mathbf{w}(y_1^*,\ldots,y_d^*)) = \min_{\mathbf{w},\xi_l,\xi_u^*}\left\{\frac{1}{2}\mathbf{w}^T\cdot\mathbf{w}+C\sum_{l=1}^{n}\xi_l+C^*\sum_{u=1}^{d}\xi_u^*\right\}$$

$$\text{subject to}\begin{cases}\forall_{l=1}^{n}: y_l\left[\mathbf{w}^T\cdot\phi(\mathbf{x}_l)+b\right]\geq 1-\xi_l, \xi_l>0\\ \forall_{u=1}^{d}: y_u^*\left[\mathbf{w}^T\cdot\phi(\mathbf{x}_u^*)+b\right]\geq 1-\xi_u^*, \xi_u^*>0.\end{cases} \quad (6)$$

In order to handle the nonseparable training and transductive data, similarly to the ISVMs, the slack variables ξ_l and ξ_u^* and the associated penalty values C and C^* of both the training and transductive samples are introduced. d $(d \leq m)$ is the number of selected unlabeled samples for transductive learning (if $d = m$, all the unlabeled samples are used for the transductive learning, as in [18] and [19]). For ease of computation on the quadratic optimization problem, the Lagrange theory is applied to (6). Finally, similar to the ISVMs, the optimization problem can be solved by maximizing the following function:

$$\begin{aligned}L(\boldsymbol{\alpha},\boldsymbol{\alpha}^*,y^*) = &\sum_{l=1}^{n}\alpha_l+\sum_{u=1}^{d}\alpha_u^*\\ &-\frac{1}{2}\left(\sum_{l=1}^{n}\sum_{i=1}^{n}G_{li}y_l y_i\alpha_l\alpha_i\right.\\ &\quad+2\sum_{l=1}^{n}\sum_{u=1}^{d}G_{lu}^*y_l y_u^*\alpha_l\alpha_u^*\\ &\quad\left.+\sum_{u=1}^{d}\sum_{j=1}^{d}G_{uj}^*y_u^*y_j^*\alpha_u^*\alpha_j^*\right)\end{aligned}$$

$$\text{subject to}\begin{cases}0\leq\alpha_l\leq C, \quad 1\leq l\leq n\\ 0\leq\alpha_u^*\leq C, \quad 1\leq u\leq d\\ \sum_{l=1}^{n}y_l\alpha_l+\sum_{u=1}^{d}y_u^*\alpha_u^*=0\end{cases} \quad (7)$$

where $G_{li} = k(\mathbf{x}_l, \mathbf{x}_i)$, $G_{lu}^* = k(\mathbf{x}_l, \mathbf{x}_u^*)$, and $G_{uj}^* = k(\mathbf{x}_u^*, \mathbf{x}_j^*)$ are the Gram matrices.

Finally, after the Lagrange multipliers α_l and α_u^* in the transductive process are fixed, the output of the TSVMs becomes

$$f(\mathbf{x}) = \sum_{l=1}^{n}\alpha_l y_l k(\mathbf{x},\mathbf{x}_l)+\sum_{u=1}^{d}\alpha_u^*y_u^*k(\mathbf{x},\mathbf{x}_u^*)+b \quad (8)$$

and the decision function can be obtained as follows:

$$y(\mathbf{x}) = \operatorname{sgn}\left[f(\mathbf{x})\right]. \quad (9)$$

B. Proposed TSVM

The proposed TSVM exploits the standard theoretical approach of the TSVMs presented in the previous subsection. However, in designing the proposed TSVM, we address several important issues in ill-posed remote-sensing problems that have not been considered in the literature. The addressed issues are: 1) definition of a transductive procedure for the TSVMs characterized by high stability; 2) analysis of the model-selection problem without test/validation sets; and 3) solution of multiclass problems in the transductive framework.

1) Proposed TSVM—Novel Transductive Procedure: Two critical issues should be taken into account in the transductive algorithm: a) definition of the procedure for the selection of the transductive samples for the relearning of the classifier and b) design of a function for tuning the regularization values of the transductive samples. These issues are discussed in the following.

a) Selection of transductive samples: As regards the selection of transductive patterns, two points should be considered: 1) select the samples with an expected accurate labeling and 2) choose the "informative" samples. Due to the fact that support vectors bear the richest information (i.e., they are the only patterns that affect the position of the separating hyperplane) among the "informative" samples (i.e., the ones in the margin band), the unlabeled patterns closest to the margin bounds have the highest probability to be correctly classified. Therefore, in the algorithm, we design a selection procedure considering a balanced number of transductive and labeled samples based on the above observation. As it will be pointed out in the following, given a multiclass problem, an architecture made up of S different binary TSVMs should be used. Let us define $N_s^{\pm}$ to be the number of positive and negative margin support vectors (i.e., the ones that lie on the margin bounds), respectively, for the sth binary subproblem after the inductive learning phase. At each of the following iterations, the A transductive samples closest to each margin bound are chosen to define the positive/negative transductive candidate sets $\Psi^{\pm}$. A reasonable balancing between the labeled patterns and the unlabeled patterns introduced at the considered iteration is needed. From a theoretical point of view, only the unlabeled sample furthest from the hyperplane could be selected from each side of the margin band at each iteration. However, such a choice can make the whole process too slow and, hence, impractical for real applications. For speeding up the learning process without any significant loss of information, in the proposed algorithm, we fix $A = \min\{N_s^+, N_s^-\}$. Thus, at each iteration, the number of considered transductive patterns is comparable to the number of original support vectors. Moreover, to select the transductive samples in a small solution space, a threshold criterion is exploited in order to consider the density of the selected area. Here, the mean of the predicted values $y(\mathbf{x}_u^*)(1 \leq u \leq A)$ of the transductive samples in both candidate sets can be used to define the threshold values $T_{\text{hr}}^{\pm}$ for the positive and negative sets. In order to alleviate the problem of the unbalanced classes, a pairwise labeling strategy is adopted in the selection of the transductive samples. After fixing the threshold values, the candidate sets can be trimmed.

TABLE I
PROPOSED TRANSDUCTIVE PROCEDURE FOR A BINARY TSVM (THE SUBSCRIPT S IDENTIFIES A SPECIFIC BINARY TSVM IN THE CONTEXT OF A MULTICLASS ARCHITECTURE)

Initialization: $C_s^{*(0)} \leftarrow \frac{C_s}{(10 \cdot G)}$, $X^{(0)} \equiv X$

Begin

- Learning on training set $X^{(0)}$ (made up of original training samples)
- Fix $A = \min\{N_s^+, N_s^-\}$

do

- Select A transductive samples closest to both the margin bounds $\rightarrow$ define positive and negative candidate sets $\Psi^{\pm}$;
- Compute the positive and negative threshold values $T_{hr}^{\pm}$:

$$T_{hr}^{+} = D^{+} \cdot |f(\mathbf{x}_u)|_{\max}, \mathbf{x}_u \in \Psi^{+}, D^{+} = \frac{\sum_{u=1}^{A} f(\mathbf{x}_u)}{A}$$

$$T_{hr}^{-} = D^{-} \cdot |f(\mathbf{x}_j)|_{\max}, \mathbf{x}_j \in \Psi^{-}, D^{-} = \frac{\sum_{j=1}^{A} f(\mathbf{x}_j)}{A}$$

- Trim the transductive candidate sets:

$$\left.\begin{array}{l} N^{+} = \text{card}\{\mathbf{x}_u | \mathbf{x}_u \in \Psi^{+}, |f(\mathbf{x}_u)| \geq T_{hr}^{+}\} \\ N^{-} = \text{card}\{\mathbf{x}_j | \mathbf{x}_j \in \Psi^{-}, |f(\mathbf{x}_j)| \geq T_{hr}^{-}\} \end{array}\right\} \Rightarrow N = \min(N^{+}, N^{-})$$

if $(N = N^{+})$

$$\Psi_t^{+} = \Psi^{+}, \Psi_t^{-} = \{\mathbf{x}_1, \cdots, \mathbf{x}_u, \cdots, \mathbf{x}_N\}, \text{ where } f(\mathbf{x}_1) \geq \cdots \geq f(\mathbf{x}_u) \geq \cdots \geq f(\mathbf{x}_N)$$

else if $(N = N^{-})$

$$\Psi_t^{-} = \Psi^{-}, \Psi_t^{+} = \{\mathbf{x}_1, \cdots, \mathbf{x}_j, \cdots, \mathbf{x}_N\}, \text{ where } f(\mathbf{x}_1) \geq \cdots \geq f(\mathbf{x}_j) \geq \cdots \geq f(\mathbf{x}_N)$$

- Updating the training set: $X^{(i+1)} = X^{(i)} \cup (\Psi_t^{-} \cup \Psi_t^{+})$;
- Updating the unlabeled set: $X^{*(i+1)} = X^{*(i)} - (\Psi_t^{-} \cup \Psi_t^{+})$;
- Updating the regularization parameters of the transductive samples: $C_s^{*(i)} = \frac{C_s^{*max} - C_s^{*(0)}}{G^2} i^2 + C_s^{*(0)}$;
- $i = i + 1$;
- Learning on training set $X^{(i)}$ (made up of both original training samples and semilabeled samples);

while $(i < G)$

End

The transductive samples whose values are greater than or equal to the threshold values are kept in the candidate sets, and then, the minimum size between the positive and negative sets (i.e., $N = \min\{N^{+}, N^{-}\}$) is chosen as the size of the final candidate sets $\Psi_t^{\pm}$. As a consequence, the set whose cardinality is greater than N is appropriately pruned (see Table I). If i denotes the number of the current iteration, at the next iteration $(i+1)$, the new training set becomes

$$X^{(i+1)} = X^{(i)} \cup \left(\Psi_t^{-} \cup \Psi_t^{+}\right) \tag{10}$$

whereas the new unlabeled set is

$$X^{*(i+1)} = X^{*(i)} - \left(\Psi_t^{-} \cup \Psi_t^{+}\right). \tag{11}$$

The procedure is iterated until convergence (see in the following). Note that, if the label of a semilabeled pattern at iteration $(i+1)$ is different from the one at iteration (i) (label inconsistency), such a label is erased, and the pattern is reset to the unlabeled state. In the proposed procedure, it is possible to reconsider this pattern at the following iterations of the transductive learning procedure.

b) Tuning regularization values of transductive samples: In the learning process of the TSVMs, a proper choice for the regularization parameters C and C^{*} represents a crucial issue. The purpose of C and C^{*} is to control the number of misclassified samples that belong to the original training set and to the unlabeled set, respectively. On increasing their values, the penalty associated with the errors on the training and transductive samples increases. In other words, the larger the regularization parameter is, the higher is the influence of the associated samples on the selection of the discriminant hyperplane. As regards the transductive procedure, it has to be taken into account that the statistical distribution of the transductive patterns could be rather different compared to that of the original training data (i.e., in ill-posed problems, the available labeled samples are often not representative enough of the test data distribution). Thus, they should be considered gradually in the transductive process so as to avoid instabilities in the learning process. For this reason, we adopted a weighting strategy different from those proposed in [18] and [19]. For each sth

binary subproblem, we propose to increase the regularization parameter for the transductive patterns C_s^* in a quadratic way, depending on the number of the iterations, as follows:

$$C_s^{*(i)} = \frac{C_s^{*\max} - C_s^{*(0)}}{G^2} i^2 + C_s^{*(0)} \tag{12}$$

where i is the ith iteration of the transductive learning process, $C_s^{*(0)}$ is the initial value for the transductive samples, $C_s^{*\max}$ is the maximum cost value of the transductive samples (this is a user-defined parameter) and is related to that of the training patterns (e.g., $C_s^{*\max} = \rho \cdot C_s$, $\rho \leq 1$ being a constant), and G (another user-defined value) is the growth rate, which, together with the maximum regularization value, controls the asymptotic convergence of the algorithm. Based on (10), we can define an indexing table so as to identify the regularization values of the transductive samples easily according to the number of the iterations included in the training set. The learning phase stops at the Gth iteration (i.e., $i = G$); therefore, we have $C_s^{*(G)} = C_s^{*\max}$.

It is worth noting that, as in all semisupervised methods, also for the proposed TSVM it is not possible to guarantee an increase of accuracy with respect to the ISVM in all cases. If the initial accuracy is particularly low (i.e., most of the semilabeled samples are incorrectly classified), it is not possible to obtain good performances. In other words, the correct convergence of learning depends on the definition of unlabeled samples considered and, implicitly, on the "similarity" between the problems represented by the training patterns and the unlabeled samples. Nevertheless, this effect is common to all semisupervised classification approaches and is not peculiar of the proposed method (see, for example, [5]).

2) Proposed TSVM—Model Selection in Ill-Posed Problems: It is widely known that a key factor in ISVMs is the choice of the kernel function. When no prior knowledge is available (or prior information is not reliable, as in ill-posed problems), the best option seems to use spherical kernels [21], e.g., Gaussian radial basis function (RBF) kernels defined as

$$k(\mathbf{x}, \mathbf{x}_l) = \exp\left(- \sum_f \left[\frac{\|\mathbf{x}^f - \mathbf{x}_l^f\|^2}{2\sigma^2} \right] \right) \tag{13}$$

where f is a feature index. After choosing the kind of kernel function, the values of the kernel parameters (i.e., the spread σ in the RBF kernels) and of the regularization parameter penalizing the training errors should be estimated in the training phase. These parameters are called hyperparameters, and choosing their best values (i.e., those that minimize the expected test error) is called model selection. In order to improve the generalization capability of the ISVMs, the kernel parameters are set identically in the multiclass problems for small-size training sets, i.e., $\sigma_1 = \cdots = \sigma_s = \cdots = \sigma_S = \sigma$. Due to the fact that the regularization parameters work with a similar function as does the prior probabilities of classes in the Bayesian formulation [22], if we can incorporate prior knowledge in the selection of the regularization values, classification accuracy can be improved significantly [22]. Considering problems with unbalanced prior probabilities of classes, it could be worthwhile to set a specific regularization parameter C_s for each sth subproblem. As in the selection of the parameters for the Fisher kernel in [23], the regularization values C_s are set to be proportional to one minus the corresponding *a priori* probability P_s of the analyzed class s, i.e., $C_s = F(1 - P_s)$, where F is a scaling factor. Hence, the model selection in the initial phase of the proposed TSVM is reduced to the choice of two parameters: σ and F.

Due to the small-size labeled dataset available in ill-posed problems, it is obvious that the holdout, bootstrap, and n-cross-validation methods cannot be used for model selection. Therefore, the only two reasonable choices are the leave-one-out and resubstitution methods. Leave-one-out is a special case of cross validation. If the available labeled samples are N, N models have to be built. In each case, a training subset, containing all of the available originally labeled patterns but one, is generated (i.e., the test set is represented by the only pattern that is not employed during the learning phase), thus making the procedure slow. Moreover, leave-one-out can become particularly critical in the presence of strongly minoritary classes, especially when the information classes are represented by very few patterns (in the limit case in which only one pattern is available for a class, it cannot be applied). For these reasons, in the considered problem, kernel parameters and the scaling factor should be selected on the basis of the training error (resubstitution error). From a theoretical viewpoint, this can lead to a poor generalization capability in inductive learning. Nevertheless, the proposed transductive process is able to mitigate the overfitting problems by a proper exploitation of the unlabeled patterns, which results in the selection of a reliable separating hyperplane. It is worth noting that, even if the distributions of the originally labeled dataset and of the semilabeled dataset are slightly different, it is reasonable to expect that they represent a similar problem (if this assumption is not satisfied, the reliability of the semisupervised approach decreases). Hence, even if the transductive approach cannot optimize the model on the unlabeled patterns (the model selection is carried out on the training samples), it is able to adapt the model to all the available data. In other words, in the transductive process, the support vectors that define the discriminant function change with respect to those identified in the inductive learning carried out on the training samples, thus fitting the model to all the available samples.

3) Proposed TSVM—Multiclass Problems: Let us consider a multiclass problem defined by a set $\Omega = \{\omega_1, \ldots, \omega_S\}$ made up of S information classes. As in standard ISVMs, the TSVMs inherit the multiclass problem. Therefore, the transductive process too has to be based on a structured architecture made up of binary classifiers. However, there is an important difference between ISVMs and TSVMs, which leads to a fundamental constraint when considering multiclass architectures. The constraint is that in the learning procedure of the binary TSVMs it must be possible to give a proper classification label to all the unlabeled samples. This is discussed in the following.

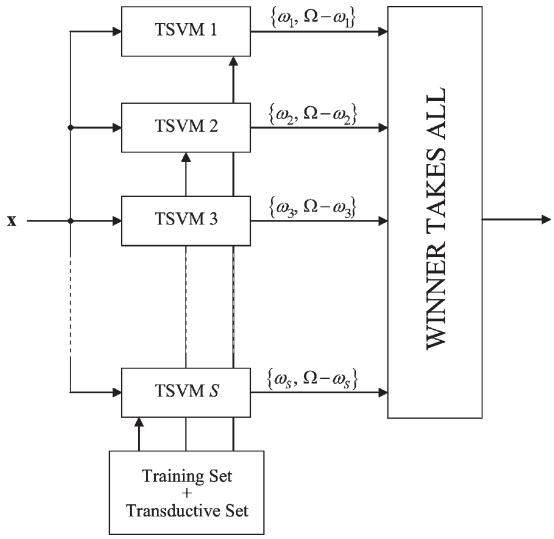


Fig. 1. OAA architecture for addressing multiclass problems with the proposed TSVM approach.

Let each binary TSVM of the multiclass architecture solve a subproblem, where each pattern must belong to one of two classes Ω_{A} and Ω_{B}, defined as proper subsets of the original set of labels Ω. The transductive approach imposes that, for each binary TSVM of the multiclass architecture, there must be an exhaustive representation of all possible labels. In other words, the following simple but important constraint must be fulfilled:

$$\Omega_{\mathrm{A}} \cup \Omega_{\mathrm{B}} = \Omega. \tag{14}$$

If (14) is not satisfied, it means that there are unlabeled patterns that the system is not capable of classifying correctly. In order to take this constraint into account, we propose to adopt a one-against-all (OAA) multiclass strategy that involves a parallel architecture made up of S different TSVMs (one for each class), as shown in Fig. 1. The sth TSVM solves a binary problem defined by one information class (e.g., $\{\omega_s\} \in \Omega$) against all the others (e.g., $\Omega - \{\omega_s\}$). In other words we have that

$$\Omega_{\mathrm{A}} = \{\omega_s\} \quad \Omega_{\mathrm{B}} = \Omega - \{\omega_s\}. \tag{15}$$

It is clear that, with this strategy, all the binary TSVMs of the multiclass architecture satisfy (14). The "winner-takes-all" (WTA) rule is used to make the final decision. For a generic pattern $\mathbf{x}$, the winning class is the one that corresponds to the TSVM with the highest output, i.e.,

$$\mathbf{x} \in \omega_s \Longleftrightarrow \omega_s = \arg \max_{i=1,\ldots,S} \{f_i(\mathbf{x})\} \tag{16}$$

where $f_i(\mathbf{x})$ represents the output of the ith TSVM [see (8)].

It is worth noting that, in the literature, there are also other multiclass combination schemes that are adopted with standard ISVMs. For example, the one-against-one (OAO) strategy is widely used and has proved to be more effective than the OAA strategy in many classification problems. However, the OAO scheme cannot be used with the TSVMs. In fact, this scheme involves $S \cdot (S-1)/2$ binary classifiers, which model all possible pairwise classification problems. Each element of the multiclass architecture carries out a classification in which two information classes $\omega_i \in \Omega$ and $\omega_j \in \Omega$ $(i \neq j)$ are analyzed against each other. Consequently, for the generic binary classifier, we have that

$$\Omega_{\mathrm{A}} = \{\omega_i\}, \qquad \Omega_{\mathrm{B}} = \{\omega_j\} \qquad j \neq i. \tag{17}$$

It is clear that all the members of this multiclass architecture violate the constraint in (14) (i.e., $\Omega_{\mathrm{A}} \cup \Omega_{\mathrm{B}} \neq \Omega$); therefore, the OAO strategy cannot be used in the transductive framework.

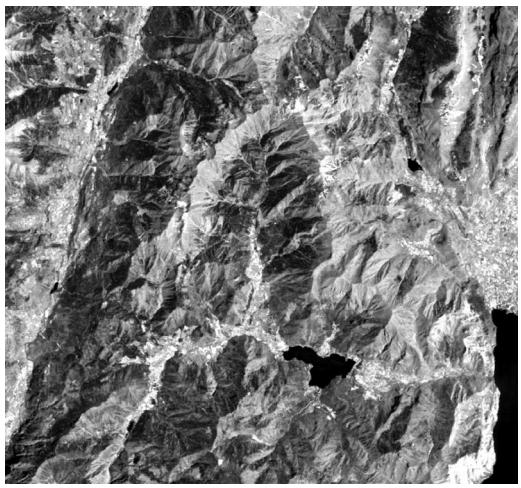


Fig. 2. Band 5 of the multispectral Landsat-5 Thematic Mapper image used in the experiments.

TABLE II
NUMBER OF PATTERNS FOR THE DIFFERENT TRAINING SETS (AVERAGE COMPUTED OVER THE TEN REALIZATIONS) AND THE TEST SET USED IN THE EXPERIMENTS

Classes	*Training Set*							*Test Set*
	Original	*Size 100*	*Size 50*	*Size 40*	*Size 30*	*Size 20*	*Size 10*	
Conifers	1704	38	19	15	11	7	3	1155
Forest	1154	25	12	10	8	5	2	681
Grass	883	20	9	8	6	4	2	336
Water	140	3	2	1	1	1	1	84
Urban Area	234	5	3	2	2	1	1	104
Rocks	419	9	5	4	2	2	1	113
Overall	4549	100	50	40	30	20	10	2473

III. EXPERIMENTAL RESULTS

A. Dataset Description

This section reports on the experimental results obtained by the proposed TSVM in the semisupervised classification of small-size training sets. The considered initial dataset is made up of a Landsat 5 Thematic Mapper image. The selected test site is a section (657×613) of a scene showing Lake Ledro in the Trentino region (a mountain area in the north of Italy). Fig. 2 shows channel 5 of the investigated image. For this dataset, a large number of labeled samples were defined from ground-reference data. The labeled samples were divided into a training

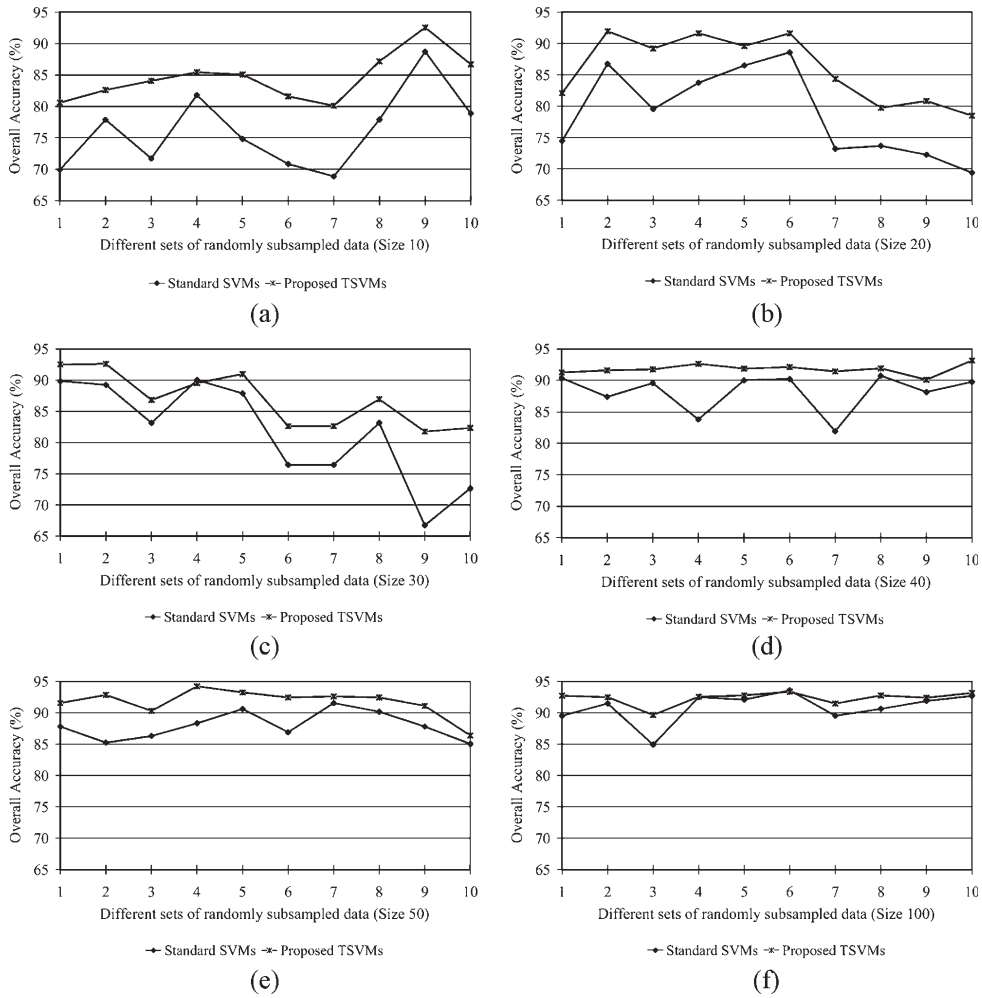


Fig. 3. Overall accuracies (in percent) provided by the standard ISVMs and the proposed TSVMs versus the size of the training sets. (a) Size 10, (b) size 20, (c) size 30, (d) size 40, (e) size 50, and (f) size 100.

and a test sets. In order to simulate ill-posed classification problems, a random subsampling strategy was applied to the training set (see Table II). In the greater detail, from 4549 original training patterns, experiments with 10, 20, 30, 40, 50, and 100 training samples were designed. For each size, ten different realizations of the training data were defined according to the application of a random procedure, with the assumption that there is at least one sample for each class. It is worth noting that, from a pattern recognition viewpoint, this is equivalent to evaluating the performances of the proposed method on 60 training sets made up of different samples and with different sizes. Therefore, this validation procedure is reliable and statistically stable. In all the datasets, seven features (i.e., all the spectral channels) and six land-cover classes (i.e., conifers, forest, grass, water, urban area, and rocks) were considered in the analysis. Hence, we are clearly in the presence of an ill-posed complex classification problem. In order to assess the effectiveness of the proposed TSVM approach, we considered the 2473 available test samples as unlabeled patterns, so that, after an initial classification, the transductive samples can be extracted from this set to be incorporated in the training set for transductive learning. Accuracy assessment was carried out on all 2473 samples. However, these samples have not been considered for model selection (the labels are assumed to be unavailable).

B. Model Selection

For all the experiments reported in this paper, we used Gaussian RBF kernels, as they proved effective in ill-posed classification problems [21]. However, the approach is general, and any other kernel function can be adopted. Datasets were normalized so that each feature is rescaled between zero and one. According to the analysis presented in Section II, in all experiments, we considered an OAA multiclass architecture [12], [24]. Model selection was carried out on the labeled datasets in the inductive learning phase, according to a grid-searching method. Due to the very small number of training samples, only a coarse grid-searching method was applied. The values considered for the kernel width were 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, and 64, while the scaling factors for the regularization parameters were 1, 2, 4, 8, 16, 32, 64, 128, and 256. In all trials, the transductive positive and negative samples were selected according to the process described in Table I. To gradually consider the influence of transductive samples, the initial values of their regularization parameters were set to

$C_s^{*(0)} = C_s/(10 \cdot G)$, where G is the growth rate in (10) and is user defined. At subsequent iterations, the values were automatically tuned in terms of the corresponding classification confidence and labeling consistency [see (10)]. We fixed $C_s^{*\max} = (1/2)C_s$. In our experiments, we observed that small values of G are sufficient to obtain high accuracy. In particular, the highest accuracies were obtained by setting $G = 5$ or $G = 10$, depending on the specific training set considered. It is worth noting that, from a computational viewpoint, each iteration requires a time equivalent to that of an inductive learning of the SVMs (this time slightly increases by increasing the number of semilabeled samples). In other words, small values of the growth-rate parameter considered in our experiments correspond to an increase of overall computational time that is approximately linearly proportional to G (the increased size of the training set does not significantly affect the learning time). As a consequence, the increase of the computational load with respect to standard ISVMs can be considered reasonable in light of the very complex problem considered. In order to compare the accuracies provided by the proposed TSVMs with those obtained by the standard ISVMs, in all the experiments, the parameter values for the training samples in the transductive process were kept the same as in the initial step.

C. Results

As described in Section III-B, the proposed algorithm was applied to all the small-size datasets, with a model selection carried out by the grid search on the training samples [25]. Fig. 3 shows the overall accuracy obtained by the proposed TSVM and the standard ISVM on the reference test set for all ten realizations of the training sets made up of 10, 20, 30, 40, 50, and 100 samples. On analyzing the diagrams, it can be observed that, on average, the proposed TSVM significantly increased overall classification accuracy with respect to the ISVM. Considering the best case [i.e., the set 9 defined by 30 training patterns, Fig. 3(b)], the overall accuracy obtained on the test set using only the initial training samples was 66.72%, while the accuracy provided by the proposed approach was 81.76% (the accuracy sharply increased by 15.04%). Other significant cases are sets 3 and 7 that are made up of ten training samples [see Fig. 3(a)], where the initial overall accuracies were 71.73% and 68.86%, while the accuracies provided by the proposed approach were 84.07% and 80.11%, respectively (the accuracies increased significantly by 12.33% and 11.24%, respectively). Table III shows the increase in the overall accuracy provided by the proposed TSVM compared to the standard ISVM for all ten realizations with different training set sizes. From an analysis of the table, it can be observed that the overall accuracy of almost all the datasets increased significantly (except for set 4, with 30 samples, and set 6, with 100 samples).

In order to analyze the effectiveness of the proposed TSVM further, Fig. 4 shows the statistical properties of the results provided by the proposed TSVM and the standard ISVM in terms of average overall accuracy and standard deviation over the ten realizations for each training size. In addition, the gap of the average accuracy and standard deviation between the proposed TSVM and standard ISVM is reported. It is worth noting that

TABLE III
AVERAGE INCREASE OF THE OVERALL ACCURACY PROVIDED BY THE PROPOSED TSVM, WITH RESPECT TO STANDARD ISVM FOR ALL TEN REALIZATIONS OF THE TRAINING SETS MADE UP OF 10, 20, 30, 40, 50, AND 100 SAMPLES. THE AVERAGE INCREASE IN THE ACCURACY (AVERAGE) AND THE STANDARD DEVIATION (STDDEV) FOR THE TEN REALIZATIONS OF EACH DIFFERENT SIZE ARE ALSO REPORTED

Datasets	*Size 10*	*Size 20*	*Size 30*	*Size 40*	*Size 50*	*Size 100*
1	10.68	7.64	2.67	0.89	3.76	3.19
2	4.73	5.18	3.36	4.21	7.60	1.01
3	12.33	9.66	3.68	2.18	4.0	4.73
4	3.64	7.89	-0.49	8.86	5.90	0.08
5	10.23	3.11	3.11	1.86	2.63	0.67
6	10.76	3.03	6.19	1.90	5.54	-0.20
7	11.24	11.16	6.19	9.50	1.05	1.94
8	9.26	6.07	3.80	1.17	2.26	2.14
9	3.84	8.57	15.04	1.94	3.32	0.53
10	7.84	9.10	9.66	3.40	1.33	0.49
AVERAGE	8.46	7.14	5.32	3.59	3.74	1.46
STDDEV	3.26	2.74	4.34	3.11	2.10	1.55

the overall accuracy obtained with the large-size training set (which consists of 4549 samples) is reported in Fig. 4(a) as the upper bound accuracy (up-bound) and is equal to 95.15%. The average overall accuracies obtained by the proposed TSVM increased those obtained by the standard ISVM significantly [see Fig. 4(a) and (b)]. Moreover, on analyzing Fig. 4(d), it can be seen that, with the training sets of different sizes, the standard deviations of the overall accuracies obtained over the ten realizations with the proposed TSVM (which are in a range between 0.82 and 5.4) were significantly lower than those obtained with the ISVM (which are in a range between 2.23 and 8.14). This confirms that the transductive inference improves the generalization capability of the classifier, besides increasing classification accuracy and stability. This can be explained by arguing that, when learning is done with very few training samples, it is not possible to find a good hyperplane to classify the entire image, while the use of both the transductive and the training samples can lead to the identification of a more reliable discriminant function. It is worth noting that the gap between the proposed approach and the ISVM decreases when increasing the number of training samples. This is due to the fact that, when the size of the training set increases, the representativeness of the training samples increases, resulting in a limited effect of transductive samples in defining the hyperplane that well represents the general classification problem. As a final remark, it is important to observe that in all the considered cases, the TSVMs cannot reach the upper bound of the test accuracy obtained with learning on the original 4549 training patterns. This is a reasonable result, since, by using only

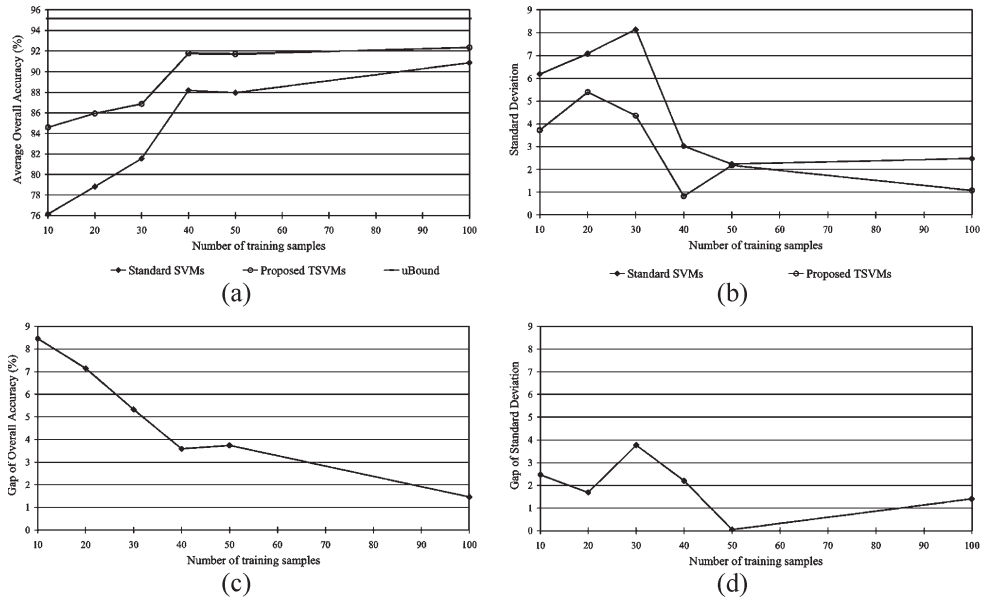


Fig. 4. (a) Average overall accuracies (in percent) provided by the standard ISVMs and the proposed TSVMs versus the size of the training sets (made up of 10, 20, 30, 40, 50, and 100 samples) on the ten realizations [the upper bound accuracy (up-bound) provided by the entire training set consisting of 4549 samples is also reported]. (b) Associated average standard deviations. (c) Average gap of the overall accuracies (in percent). (d) Average gap of the standard deviations.

unlabeled samples, the transductive process cannot recover all the information found in a complete representative training set.

IV. Discussion and Conclusion

In this paper, ill-posed classification problems have been addressed in the framework of SVMs and semisupervised learning techniques (which exploit both labeled and unlabeled samples in the definition of the discriminant function). In particular, we introduced TSVMs in the classification of remote-sensing images, which are based on specific learning algorithms that define the separating hyperplane according to a transductive process that integrates the unlabeled samples together with the training samples. With respect to the TSVMs presented in the literature, the proposed method also presents three main novelties: 1) design of a novel transductive procedure that exploits a time-dependent weighting strategy for unlabeled patterns; 2) mitigation of the effects of a suboptimal model selection (which is unavoidable in the presence of small-size training sets); and 3) generalization of binary TSVMs to the multiclass problems. It is worth noting that the proposed TSVM approach should be related to other semisupervised methods studied in the remote-sensing literature [4], [5], [7]. As regards the methods presented in [4], [5], and [7], the SVMs play the role of the GML to separate the positive and negative samples, while the transductive inference plays a role similar to that of the EM algorithm to progressively search more reliable discriminant functions with the additional unlabeled samples. However, unlike in the GML semisupervised classifier based on the EM algorithm, in the TSVM algorithm, the transductive samples are selected on the basis of a geometric analysis of the feature space, and only support-vector-like samples that contain the richest information are included in the training set. In addition, the regularization parameters of the semilabeled samples can be defined properly and incorporated directly in the transductive learning. This can better control the influence of the transductive samples in the semisupervised classification.

It is important to point out that, in order to address ill-posed classification problems, the free parameters to be estimated should be as few as possible. In our algorithm, the model selection is reduced to choosing four parameters, i.e., the kernel parameter, the regularization parameter for the originally labeled samples (C), the scaling factor for the regularization values of the training samples (F), and the growth rate (G) for the weights of the transductive patterns. However, as proved in the experimental results, the choice of G is not critical.

From the analysis of the experimental results obtained on the several simulated small-size datasets extracted from a real remote-sensing image, we can state that the proposed TSVM results both in high classification accuracy and very good stability.

As a future development of this work, we plan to apply the proposed TSVM to real hyperspectral remote-sensing images, which usually define ill-posed classification problems. This extension seems very promising, given the effectiveness of the ISVMs in the hyperspectral classification problems (already proved in previous works [12], [15]) and the advantages offered by semisupervised transductive learning described in this paper. An additional issue to be investigated is the problem that the use of the transductive samples (which could be misclassified) jointly with the original training samples for quadratic optimization may result in a nonconvex cost function made up of different "local minima" [26]. This problem could be alleviated by considering different optimization strategies.

Acknowledgment

The authors would like to thank the Istituto Sperimentale per l'Assestamento Forestale e per l'Alpicoltura (ISAFA), Trento, Italy, for providing both the Landsat image used in the

experiments and advice about the ground truth, and the anonymous referees for their constructive criticism.

REFERENCES

[1] A. Baraldi, L. Bruzzone, and P. Blonda, "Quality assessment of classification and cluster maps without ground truth knowledge," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 857–873, Apr. 2005.

[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[3] M. Chi and L. Bruzzone, "A semilabeled-sample-driven bagging technique for ill-posed classification problems," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 1, pp. 69–73, Jan. 2005.

[4] Q. Jackson and D. A. Landgrebe, "A adaptive method for combined covariance estimation and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1082–1087, May 2002.

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

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Berlin, Germany: Springer-Verlag, 1999, pp. 237–240, 263–265, and 291–299.

[7] S. Tadjudin and D. A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Trans. Geosci Remote Sens.*, vol. 38, no. 1, pp. 439–445, Jan. 2000.

[8] M. Chi and L. Bruzzone, "An ensemble-driven k-NN approach to ill-posed classification problems," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 301–307, Mar. 2006.

[9] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, Feb. 1996.

[10] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[11] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems*, vol. 10. Cambridge, MA: MIT Press, 1998, pp. 368–374.

[12] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[13] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[14] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[15] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[16] A. Gammerman, V. Vapnik, and V. Vowk, "Learning by transduction," in *Proc. Uncertainty Artif. Intell.*, Madison, WI, Jul. 1998, pp. 148–156.

[17] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998, pp. 339–371, 434–437, and 518–520.

[18] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. ICML*, 1999, pp. 200–209.

[19] Y. Chen, G. Wang, and S. Dong, "Learning with progressive transductive support vector machine," *Pattern Recognit. Lett.*, vol. 24, no. 12, pp. 1845–1855, Aug. 2003.

[20] T. Zhang and F. Oles, "A probability analysis on the value of unlabeled data for classification problems," in *Proc. Int. Joint Conf. Mach. Learn.*, Jun. 2000, pp. 1191–1198.

[21] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 131–159, 2002.

[22] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior knowledge in support vector kernels," in *Advances in Neural Information Processing Systems*, vol. 10. Cambridge, MA: MIT Press, 1998, pp. 640–646.

[23] M. M. Dundar and D. A. Landgrebe, "A cost-effective semisupervised classifier approach with kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[24] C. Hsu and C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

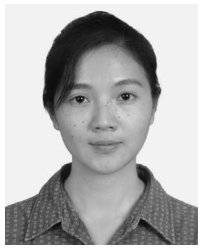
[25] C. Hsu, C. Chang, and C. Lin, *A practical guide to support vector classification*, Jul. 2003. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers.html

[26] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proc. 10th Int. Workshop Artif. Intell. and Statist.*, 2005, pp. 57–64.

Lorenzo Bruzzone (S'95–M'98–SM'03) received the laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

From 1998 to 2000, he was a Postdoctoral Researcher at the University of Genoa. From 2000 to 2001, he was an Assistant Professor at the University of Trento, Trento, Italy, and from 2001 to 2005, he was an Associate Professor at the same university. Since March 2005, he has been a Full Professor of telecommunications at the University of Trento, where he currently teaches remote sensing, pattern recognition, and electrical communications. He is currently the Head of the Remote Sensing Laboratory in the Department of Information and Communication Technology, University of Trento. His current research interests are in the area of remote-sensing image processing and recognition (analysis of multitemporal data, feature selection, classification, regression, data fusion, and machine learning). He conducts and supervises research on these topics within the frameworks of several national and international projects. Since 1999, he has been appointed Evaluator of project proposals for the European Commission. He is the author (or coauthor) of more than 150 scientific publications, including journals, book chapters, and conference proceedings. He is a Referee for many international journals and has served on the Scientific Committees of several international conferences.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was a recipient of the Recognition of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Best Reviewers in 1999 and was a Guest Editor of a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of multitemporal remote-sensing images (November 2003). He was the General Chair and Cochair of the First and Second IEEE International Workshop on the Analysis of Multi-temporal Remote-Sensing Images. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a member of the Scientific Committee of the India–Italy Center for Advanced Research. He is also a member of the International Association for Pattern Recognition and of the Italian Association for Remote Sensing (AIT).



Mingmin Chi (S'05) received the B.S. degree in electrical engineering from the Changchun University of Science and Technology, Changchun, China, in 1998, the M.S. degree in electrical engineering from Xiamen University, Xiamen, China, in 2002, and the Ph.D. degree in computer science on pattern recognition and remote sensing from the University of Trento, Trento, Italy, in 2006.

She is currently an Assistant Professor with the Department of Computer Science and Engineering, Fudan University, Shanghai, China. Her research interests include the design of supervised and semi-supervised pattern recognition and machine learning algorithms for signal/image processing and analysis, especially kernel-based methods and graphical models with applications to remote sensing and bioinformatics.

Mattia Marconcini (S'06) was born in Verona, Italy, in 1980. He received the laurea (B.S.) and laurea specialistica (M.S.) degrees in telecommunication engineering (*summa cum laude*) from the University of Trento, Trento, Italy, in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree in information and communication technologies at the University of Trento.

He is presently with the Pattern Recognition and Remote Sensing Group, Department of Information and Communication Technologies, University of Trento. His current research activities are in the area of machine learning and remote sensing. In particular, his interests are related to semisupervised, partially supervised, and context-sensitive classification problems. He conducts research on these topics within the frameworks of several national and international projects.