A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples

Lorenzo Bruzzone, *Senior Member, IEEE*, and Claudio Persello, *Student Member, IEEE*

***Abstract*—This paper presents a novel context-sensitive semisupervised support vector machine (CS^{4}VM) classifier, which is aimed at addressing classification problems where the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns). Unlike standard context-sensitive methods, the proposed CS^{4}VM classifier exploits the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase to improve the robustness to possible mislabeled training patterns. This is achieved according to both the design of a semisupervised procedure and the definition of a novel contextual term in the cost function associated with the learning of the classifier. In order to assess the effectiveness of the proposed CS^{4}VM and to understand the impact of the addressed problem in real applications, we also present an extensive experimental analysis carried out on training sets that include different percentages of mislabeled patterns having different distributions on the classes. In the analysis, we also study the robustness to mislabeled training patterns of some widely used supervised and semisupervised classification algorithms (i.e., conventional support vector machine (SVM), progressive semisupervised SVM, maximum likelihood, and k-nearest neighbor). Results obtained on a very high resolution image and on a medium resolution image confirm both the robustness and the effectiveness of the proposed CS^{4}VM with respect to standard classification algorithms and allow us to derive interesting conclusions on the effects of mislabeled patterns on different classifiers.**

***Index Terms*—Context-sensitive classification, image classification, mislabeled training patterns, noisy training set, remote sensing, semisupervised classification, support vector machines (SVMs).**

I. Introduction

THE CLASSIFICATION of remote sensing images is often performed by using supervised classification algorithms, which require the availability of labeled samples for the training of the classification model. All these algorithms are sharply affected from the quality of the labeled samples used for training the classifier, whose reliability is of fundamental importance for an adequate learning of the properties of the investigated scene (and, thus, for obtaining accurate classification maps). In supervised classification, the implicit assumption is that all labels associated with training patterns are correct. Unfortunately, in many real cases, this assumption does not hold, and small amounts of training samples are associated with a wrong information class due to errors occurred in the phase of collection of labeled samples. Labeled samples can be derived by the following: 1) *in situ* ground truth surveys; 2) analysis of reliable reference maps; or 3) image photointerpretation. In all these cases, mislabeling errors are possible. During the ground truth surveys, mislabeling errors may occur due to imprecise geolocalization of the positioning system; this leads to the association of the identified land-cover label with a wrong geographic coordinate and, thus, with the wrong pixel (or region of interest) in the remotely sensed image. Similar errors may occur if the image to be classified is not precisely georeferenced. When reference maps are used for extracting label information, possible errors present in the maps propagate to the training set. The case of image photointerpretation is also critical, as errors of the human operator may occur, leading to a mislabeling of the corresponding pixels or regions.

Mislabeled patterns bring distort (wrong) information to the classifier (in this paper, we call them *noisy* patterns). The effect of noisy patterns in the learning phase of a supervised classifier is to introduce a bias in the definition of the decision regions, thus decreasing the accuracy of the final classification map. We can expect two different situations with respect to the distribution of noisy samples in the training set: 1) mislabeled samples may be uniformly distributed over all considered classes, or 2) mislabeled patterns can specifically affect one or a subset of the classes of the considered classification problem. The two different situations result in a different impact on the learning phase of the classification algorithms. Let us analyze the problem according to the Bayes decision theory and to the related estimates of class conditional densities (likelihood) and class prior probabilities (priors) [1]. If noisy samples are uniformly distributed over classes, the estimations of class conditional densities result corrupted, while the estimations of prior probabilities are not affected from the presence of mislabeled patterns. On the contrary, if noisy samples are not uniformly distributed over classes, both the estimations of prior probabilities and of class conditional densities are biased from mislabeled patterns. Therefore, we expect that supervised algorithms, which (explicitly or implicitly) consider the prior probabilities for the classification of a generic input pattern (e.g., Bayesian classifier and k-nearest neighbor (k-NN) [1]–[3]), are more sensitive to unbalanced noisy sample distributions over classes than other algorithms that take into account only the class conditional densities (e.g., maximum likelihood (ML) [1], [2]).

In this paper, we address the aforementioned problems by the following: 1) presenting a novel context-sensitive

Manuscript received February 15, 2008; revised July 25, 2008 and October 2, 2008. First published May 12, 2009; current version published June 19, 2009.

The authors are with the Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it; claudio.persello@disi.unitn.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2008.2011983

semisupervised support vector machine (CS^{4}VM) classification algorithm, which is robust to noisy training sets, and 2) analyzing the effect of noisy training patterns and of their distribution on the classification accuracy of widely used supervised and semisupervised classifiers.

The choice of developing a support vector machine (SVM)-based classifier is related to the important advantages that SVMs exhibit over other standard supervised algorithms [4]–[8]: 1) relatively high empirical accuracy and excellent generalization capabilities; 2) robustness to the Hughes phenomenon [9]; 3) convexity of the cost function used in the learning of the classifier; 4) sparsity of the solution; and 5) possibility to use the kernel trick for addressing nonlinear problems. In particular, the generalization capability of SVM (induced by the minimization of the structural risk) gives to SVM-based classifiers an intrinsic higher robustness to noisy training patterns than other standard algorithms that are based on the empirical risk minimization principle. In this framework, we propose an SVM-based technique for image classification particularly developed to improve the robustness of standard SVM to the presence of noisy samples in the training set. The main idea behind the proposed CS^{4}VM is to exploit the spatial context information provided by the pixel belonging to the neighborhood system of each training sample (which are called context patterns) in order to contrast the bias effect due to the possible presence of mislabeled training patterns. This is achieved by both a semisupervised procedure (aiming to obtain the semilabels for context patterns) and the definition of a novel contextual term in the cost function associated with the learning of the CS^{4}VM. It is worth noting that this use of the contextual information is completely different from that of traditional context-sensitive classifiers (see, e.g., [10]–[16]), where contextual information is exploited for regularizing classification maps in the decision phase.

Another important contribution of this paper is to present an extensive experimental analysis to investigate and compare the robustness to noisy training sets of the proposed CS^{4}VM and of other conventional classifiers. In greater detail, we considered the (Gaussian) ML classifier (which is based on a parametric estimation of the class conditional densities and does not consider the prior probabilities of the classes), the k-NN classifier (which is based on a distribution-free local estimation of posterior probabilities that implicitly considers the class prior probabilities), the standard SVM classifier, and the progressive semisupervised SVM (PS^{3}VM) [17]. The five considered classification algorithms were tested on two different data sets: 1) a very high resolution (VHR) multispectral image acquired by the Ikonos satellite and 2) a medium-resolution multispectral image acquired by the Landsat-5 Thematic Mapper (TM). The experimental analysis was carried out, considering training sets that include different amounts of noisy samples having different distributions over the considered classes.

This paper is organized into six sections. Section II presents the proposed CS^{4}VM technique. Section III describes the design of the experiments carried out with different classifiers. Sections IV and V illustrate the experimental results obtained on the Ikonos and Landsat data sets, respectively. Finally, Section VI, after discussion, draws the conclusion of this paper.

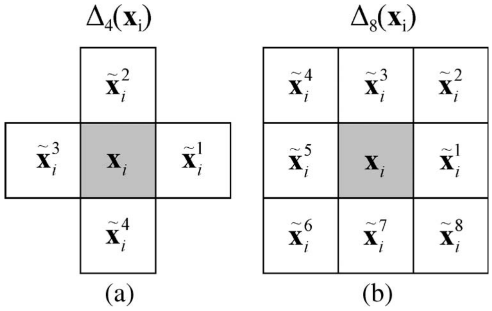

Fig. 1. Examples of neighborhood systems for the generic training pixel $\mathbf{x}_i$. (a) First-order system $\Delta_4(\mathbf{x}_i)$. (b) Second-order system $\Delta_8(\mathbf{x}_i)$.

II. Proposed CS^{4}VM

Let $\mathcal{I}$ denote a multispectral d-dimensional image of size $I \times J$ pixels. Let us assume that a training set $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ made up of N pairs $(\mathbf{x}_i, y_i)_{i=1}^N$ is available, where $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^N \subset \mathcal{I}$ is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_i\}_{i=1}^N$ is the corresponding set of labels. For the sake of simplicity, since SVMs are binary classifiers, we first focus the attention on the two-class case (the general multiclass case will be addressed later). Accordingly, let us assume that $y_i \in \{+1; -1\}$ is the binary label of the pattern $\mathbf{x}_i$. We also assume that a restricted amount δ of training samples $\mathbf{x}_i$ may be associated with wrong labels y_i, i.e., labels that do not correspond to the actual class of the considered pixel. Let $\Delta_M(\mathbf{x})$ represent a local neighborhood system (whose shape and size depend on the specific investigated image and application) of the generic pixel $\mathbf{x}$, where M indicates the number of pixels considered in the neighborhood. Generally, $\Delta_M(\mathbf{x})$ is a first- or second-order neighborhood system (see Fig. 1). Let $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_i^j | \tilde{\mathbf{x}}_i^j \in \Delta_M(\mathbf{x}_i), \forall \mathbf{x}_i \in \mathcal{X}, j = 1, \ldots, M\}$ be the set of (unlabeled) context patterns $\tilde{\mathbf{x}}_i^j$ made up of the pixels belonging to the neighborhood $\Delta_M(\mathbf{x}_i)$ of the generic training sample $\mathbf{x}_i$. It is worth noting that adjacent training pixels belong to both $\mathcal{X}$ and $\tilde{\mathcal{X}}$.

The idea behind the proposed methodology is to exploit the information of the context patterns $\tilde{\mathcal{X}}$ to reduce the bias effect of the δ mislabeled training patterns on the definition of the discriminating hyperplane of the SVM classifier, thus decreasing the sensitivity of the learning algorithm to unreliable training samples. This is accomplished by explicitly including the samples belonging to the neighborhood system of each training pattern in the definition of the cost function used for the learning of the classifier. These samples are considered by exploiting the labels derived through a semisupervised classification process (for this reason, they are called semilabeled samples) [18]–[20]. The semilabeled context patterns have the effect to mitigate the bias introduced by noisy patterns by adjusting the position of the hyperplane. This strategy is defined according to a learning procedure for the proposed CS^{4}VM that is based on two main steps: 1) supervised learning with original training samples and classification of the (unlabeled) context patterns and 2) contextual semisupervised learning based on both original labeled patterns and semilabeled context patterns according to a novel cost function. These two steps are described in detail in the following sections.

A. Step 1—Supervised Learning and Classification of Context Patterns

In the first step, a standard supervised SVM is trained by using the original training set $\mathcal{D}$ in order to classify the patterns belonging to the neighborhood system of each training pixels. The learning is performed according to the soft-margin SVM algorithm, which results in the following constrained minimization problem:

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{N}\xi_i\right\} \\ y_i\cdot[\mathbf{w}\cdot\Phi(\mathbf{x}_i)+b] \geq 1-\xi_i \quad \forall i=1,\ldots,N \\ \xi_i \geq 0 \end{cases} \tag{1}$$

where $\mathbf{w}$ is a vector normal to the separation hyperplane, b is a constant such that $b/\|\mathbf{w}\|$ represents the distance of the hyperplane from the origin, $\Phi(\cdot)$ is a nonlinear mapping function, ξ_i's are slack variables that control the empirical risk (i.e., the number of training errors), and $C \in \mathbb{R}_0^+$ is a regularization parameter that tunes the tradeoff between the empirical error and the complexity term (i.e., the generalization capability). The aforementioned minimization problem can be rewritten in the dual formulation by using the Lagrange optimization theory, which leads to the following dual representation:

$$\begin{cases} \max\limits_{\alpha}\left\{\sum\limits_{i=1}^{N}\alpha_i - \frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)\right\} \\ \sum\limits_{i=1}^{N} y_i\alpha_i = 0 \qquad \forall i=1,\ldots,N \\ 0 \leq \alpha_i \leq C \end{cases} \tag{2}$$

where α_i's are the Lagrange multipliers associated with the original training patterns $\mathbf{x}_i \in \mathcal{X}$, and $k(\cdot,\cdot)$ is a kernel function such that $k(\cdot,\cdot) = \Phi(\cdot)\Phi(\cdot)$. The kernel function is used for implicitly mapping the input data into a high-dimensional feature space without knowing the function $\Phi(\cdot)$ and still maintaining the convexity of the objective function [6]. Once α_i $(i = 1,\ldots,N)$ are determined, each context pattern $\tilde{\mathbf{x}}_i^j$ in the neighborhood system $\Delta_M(\mathbf{x}_i)$ of the training pattern $\mathbf{x}_i$ is associated with a semilabel $\tilde{y}_i^j$ according to

$$\hat{\tilde{y}}_i = \operatorname{sgn}\left[\sum_{n=1}^{N} y_n\alpha_n k\left(\mathbf{x}_n, \tilde{\mathbf{x}}_i^j\right) + b\right] \ \forall \mathbf{x}_n \in \mathcal{X},\ \forall \tilde{\mathbf{x}}_i^j \in \tilde{\mathcal{X}} \tag{3}$$

where, given $f(\mathbf{x}) = \sum_{i=1}^{N} y_i\alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$, b is chosen so that $y_i f(\mathbf{x}_i) = 1$ for any i with $0 < \alpha_i < C$.

B. Step 2—Context-Sensitive Semisupervised Learning

Taking into account the semilabels (i.e., the labels obtained in the previous step) of the context patterns belonging to $\tilde{\mathcal{X}}$, we define the following novel context-sensitive cost function for the learning of the classifier:

$$\Psi(\mathbf{w},\xi,\psi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\sum_{j=1}^{M}\kappa_i^j\psi_i^j \tag{4}$$

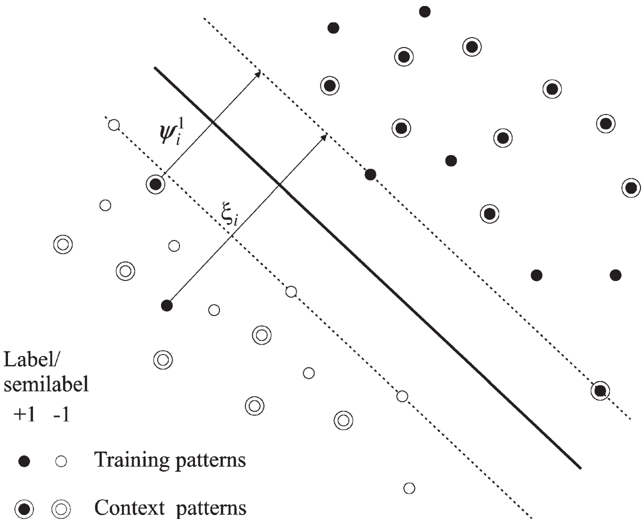

Fig. 2. Example of training and related context patterns in the kernel-induced feature space.

where ψ_i^j's are context slack variables and $\kappa_i^j \in \mathbb{R}_0^+$ are parameters that permit one to weight the importance of context patterns (see Fig. 2). The resulting constrained minimization problem associated with the learning of the CS^{4}VM is the following:

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi,\psi} \Psi(\mathbf{w},\xi,\psi) \\ y_i\cdot[\mathbf{w}\cdot\Phi(\mathbf{x}_i)+b] \geq 1-\xi_i \\ \qquad\qquad\qquad\qquad \forall i=1,\ldots,N \\ \tilde{y}_i^j\cdot\left[\mathbf{w}\cdot\Phi\left(\tilde{\mathbf{x}}_i^j\right)+b\right] \geq 1-\psi_i^j \\ \qquad\qquad\qquad\qquad \forall j=1,\ldots,M \\ \psi_i^j, \xi_i \geq 0. \end{cases} \tag{5}$$

The cost function in (4) contains a novel contextual term (made up of $N \cdot M$ elements) whose aim is to regularize the learning process with respect to the behavior of the context patterns in the neighborhood of the training pattern under consideration. The rationale of this term is to balance the contribution of possibly mislabeled training samples according to the semilabeled pixels of the neighborhood. The context slack variables $\psi_i^j = \psi_i^j(\tilde{\mathbf{x}}_i^j, \tilde{y}_i^j, \mathbf{w}, b)$ depend on $\tilde{\mathbf{x}}_i^j \in \Delta_M(\mathbf{x}_i)$ and, accordingly, permit one to directly take into account the contextual information in the learning phase. They are defined as

$$\psi_i^j = \max\left\{0, 1-\tilde{y}_i^j\cdot\left[\mathbf{w}\cdot\Phi\left(\tilde{\mathbf{x}}_i^j\right)+b\right]\right\}$$
$$\forall i=1,\ldots,N, \quad \forall j=1,\ldots,M. \tag{6}$$

The parameters $\kappa_i^j \in \mathbb{R}_0^+$ weight the context patterns $\tilde{\mathbf{x}}_i^j$ depending on the agreement of their semilabels $\tilde{y}_i^j$ with that of the related training sample y_i. The hypothesis at the basis of the weighting system of the context patterns is that the pixels in the same neighborhood system have high probability to be associated to the same information class (i.e., the labels of the pixels are characterized by high spatial correlation). In particular, κ_i^j's are defined as follows:

$$\kappa_i^j = \begin{cases} \kappa_1 & \text{if} \quad y_i = \tilde{y}_i^j \\ \kappa_2 & \text{if} \quad y_i \neq \tilde{y}_i^j \end{cases} \tag{7}$$

where κ_1 and κ_2 are chosen from the user. The role of κ_1 and κ_2 is to define the importance of the context patterns. In particular, it is very important to define the ratios C/κ_i, $i = 1, 2$, which tune the weight of context patterns with respect to the patterns of the original training set. According to our hypothesis, in order to adequately penalize the mislabeled training patterns, it is suggested to fix $\kappa_1 \geq \kappa_2$ as, in general, contextual patterns whose semilabels are in agreement with the label of the related training pattern should be considered more reliable than those whose semilabels are different. The selection of κ_1 and κ_2 can be simplified by fixing *a priori* the ratio $\kappa_1/\kappa_2 = K$, thus focusing the attention only on κ_1 or on the ratio C/κ_1.

It is worth noting that the novel cost function defined in (4) maintains the important property of convexity of the cost function of the standard SVM. This allows us to solve the problem according to quadratic programming algorithms. By properly adjusting the Karush–Kuhn–Tucker conditions [i.e., the necessary and sufficient conditions for solving (5)], we derived the following dual bound maximization problem:

$$\begin{cases} \max\limits_{\alpha,\beta} \left\{ \sum\limits_{i=1}^{N} \left(\alpha_i + \sum\limits_{j=1}^{M} \beta_i^j \right) \right. \\ \qquad \left. -\frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{h=1}^{N} \begin{bmatrix} y_i y_h \alpha_i \alpha_h k\left(\mathbf{x}_i, \mathbf{x}_h\right) \\ +2 y_i \alpha_i \sum\limits_{j=1}^{M} \tilde{y}_h^i \beta_h^i k\left(\mathbf{x}_i, \tilde{\mathbf{x}}_h^j\right) \\ + \sum\limits_{q=1}^{M} \sum\limits_{j=1}^{M} \tilde{y}_i^q \tilde{y}_h^j \beta_i^q \beta_h^j k\left(\tilde{\mathbf{x}}_i^q, \tilde{\mathbf{x}}_h^j\right) \end{bmatrix} \right\} \\ \sum\limits_{i=1}^{N} \left(y_i \alpha_i + \sum\limits_{j=1}^{M} \tilde{y}_i^j \beta_i^j \right) = 0 \\ \qquad\qquad\qquad\qquad \forall i = 1, \ldots, N \\ \qquad 0 \leq \alpha_i \leq C \\ \qquad\qquad\qquad\qquad \forall j = 1, \ldots, M \\ \qquad 0 \leq \beta_i^j \leq \kappa_i^j \end{cases} \tag{8}$$

where α_i and r_i are the Lagrange multipliers associated with original training patterns, while β_i^j and s_i^j are the Lagrange multipliers associated with contextual patterns. The Lagrange multipliers α_i associated with the original labeled patterns are superiorly bounded by C (they all have the same importance). The upper bound for the Lagrange multipliers β_i^j associated with context patterns is κ_i^j, as it comes from (7). Once α_i and β_i^j $(i = 1, \ldots, N, j = 1, \ldots, M)$ are determined, the generic pattern $\mathbf{x}$ belonging to the investigated image $\mathcal{I}$ can be classified according to the following decision function:

$$\hat{y} = \operatorname{sgn} \left\{ \sum_{i=1}^{N} \left[y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{M} \tilde{y}_i^j \beta_i^j k\left(\tilde{\mathbf{x}}_i^j, \mathbf{x}\right) \right] + b \right\} \qquad \forall \mathbf{x}_i \in \mathcal{X}, \forall \tilde{\mathbf{x}}_i^j \in \tilde{\mathcal{X}} \tag{9}$$

where, given $f(\mathbf{x}) = \sum_{i=1}^{N} [y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{M} \tilde{y}_i^j \beta_i^j k(\tilde{\mathbf{x}}_i^j, \mathbf{x})] + b$, b is chosen so that $y_i f(\mathbf{x}_i) = 1$ for any i with $0 < \alpha_i < C$ and $\tilde{y}_i^j f(\tilde{\mathbf{x}}_i^j) = 1$ for any i and j with $0 < \beta_i^j < \kappa_i^j$.

It is worth noting that the proposed formulation could be empirically defined by considering different analytical forms for the kernels associated with the original training samples and the context patterns (composite kernel approach). From a general perspective, this would increase the flexibility of the method. However, as the training patterns and the context patterns are represented by the same feature vectors, the use of composite kernels (which would result in a further increase of the number of free parameters to set in the leaning of the classifier and, thus, in an increase of the computational cost required from the model-selection phase) does not seem useful.

C. Multiclass Architecture

Let us extend the binary CS^{4}VM to the solution of multiclass problems. Let $\Omega = \{\omega_1, \ldots, \omega_L\}$ be the set of L information classes that characterize the considered problem. As for the conventional SVM, the multiclass problem should be addressed with a structured architecture made up of binary classifiers. However, the properties of CS^{4}VM lead to an important difference with respect to the standard supervised SVM. This difference is related to the *step 2* of the learning of the CS^{4}VM. In this step, we assume to be able to give a reliable label to all patterns in the neighborhood system of each training pattern. In order to satisfy this constraint, we should define binary classification problems for each CS^{4}VM included in the multiclass architecture characterized from an exhaustive representation of classes.

Let each CS^{4}VM of the multiclass architecture solve a binary subproblem, where each pattern should belong to one of the two classes Ω_A or Ω_B, defined as proper subsets of the original set of labels Ω. The contextual semisupervised approach requires that, for each binary CS^{4}VM of the multiclass architecture, there must be an exhaustive representation of all possible labels, i.e.,

$$\Omega_A \cup \Omega_B = \Omega. \tag{10}$$

If (10) is not satisfied, some semilabels of context patterns $\tilde{\mathbf{x}}_i^j$ may not be represented in the binary subproblem and the context-sensitive semisupervised learning cannot be performed. According to this constraint, we propose to adopt a one-against-all (OAA) multiclass architecture, which is made up of L parallel CS^{4}VM, as shown in Fig. 3.

The lth CS^{4}VM solves a binary problem defined by the information class $\{\omega_l\} \in \Omega$ against all the others $\Omega - \{\omega_l\}$. In this manner, all the binary subproblems of multiclass architecture satisfy (10). The "winner-takes-all" rule is used for taking the final decision, i.e.,

$$\hat{\omega} = \underset{i=1,\ldots,L}{\arg\max} \{f_i(\mathbf{x})\} \tag{11}$$

where $f_i(\mathbf{x})$ represent the output of the ith CS^{4}VM.

It is worth noting that other multiclass strategies that are commonly adopted with standard SVM (such as the one-against-one) [21] cannot be used with the proposed CS^{4}VM as they do not satisfy (10). Nevertheless, other multiclass architectures could be specifically developed for the CS^{4}VM approach, which should satisfy the constraint defined in (10).

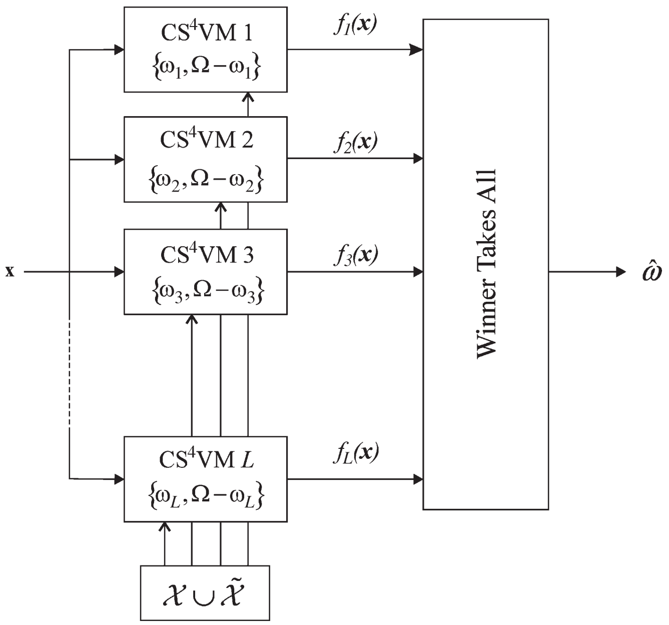
CS^{4}VM 1
$\{\omega_1, \Omega-\omega_1\}$
CS^{4}VM 2
$\{\omega_2, \Omega-\omega_2\}$
CS^{4}VM 3
$\{\omega_3, \Omega-\omega_3\}$
CS^{4}VM L
$\{\omega_L, \Omega-\omega_L\}$
$f_1(\mathbf{x})$
$f_2(\mathbf{x})$
$f_3(\mathbf{x})$
$f_L(\mathbf{x})$
$\mathbf{x}$
Winner Takes All
$\hat{\omega}$
$\mathcal{X} \cup \tilde{\mathcal{X}}$

Fig. 3. OAA architecture for addressing the multiclass problem with the proposed CS^{4}VM.

III. Design of Experiments

In this section, we describe the extensive experimental phase carried out to evaluate the robustness to the presence of noisy training samples of the proposed CS^{4}VM and of other standard supervised and semisupervised classification algorithms. In particular, we compare the accuracy (in terms of kappa coefficient [22]) obtained by the proposed CS^{4}VM with those yielded by other classification algorithms: the progressive semisupervised SVM (PS^{3}VM) [17], the standard supervised SVM, the ML, and the k-NNs. We carried out different kinds of experiments by training the classifiers: 1) with the original training samples (with their correct labels) and 2) with different synthetic training sets, where mislabeled patterns (i.e., patterns with wrong labels) were added to the original training set in different percentages (10%, 16%, 22%, and 28%) with respect to the total number of training samples. In the second kind of experiments, we manually introduced mislabeled training samples by considering the particular scene under investigation and simulating realistic mislabeling errors (e.g., caused by possible photointerpretation errors). The spatial location of wrong samples was distributed over the whole scene, by considering also clusters of pixels in the same neighborhood system. We analyzed the effects of noisy training sets on the classification accuracy in two different scenarios (which simulate different kinds of mislabeling errors): 1) Wrong samples are uniformly added to all the information classes (thus simulating the presence of mislabeling errors in the training points that does not depend on the land-cover type), and 2) wrong patterns are added to one specific class or to a subset of the considered classes (thus simulating a systematic error in the collection of ground truth samples for specific land-cover types).

In all the experiments, for the ML classifier, we adopted the Gaussian function as a model for the probability density functions of the classes. Concerning the k-NN classification algorithm, we carried out several trials, varying the value of k from 1 to 40 in order to identify the value that maximizes the kappa accuracy on the test set.

For the SVM-based classifiers (CS^{4}VM, PS^{3}VM, and standard SVM), we employed the sequential minimal optimization algorithm [23] (with proper modifications for the CS^{4}VM) and used Gaussian kernel functions (ruled by the free parameter σ that expresses the width of the Gaussian function). All the data were normalized to a range [0, 1], and the model selection for deriving the learning parameters was carried out according to a grid-search strategy on the basis of the kappa coefficient of accuracy obtained on the test set.

For the standard SVM, the value of $2\sigma^2$ was varied in the range [10^{-2}, 10], while the values of C were concentrated in the range [20, 200] after a first exploration in a wider range. For the model selection of both the CS^{4}VM and the PS^{3}VM, we considered the same values for C and $2\sigma^2$ as for the SVM in order to have comparable results. Moreover, for the proposed CS^{4}VM, we fixed the value of $K = \kappa_1/\kappa_2 = 2$ and used the following values for C/κ_1: 2, 4, 6, 8, 10, 12, and 14. For the definition of the context patterns, we considered a first-order neighborhood system. With regard to the PS^{3}VM, the value of $C^{*(0)}$ was varied in the range [0.1, 1], the one of γ was varied in the range [10, 100], and ρ was varied in the range [10, 100].

For simplicity, the model selection for all the SVM-based classifiers and the k-NN algorithm was carried out on the basis of the kappa coefficient of accuracy computed on the test set, which does not contain mislabeled samples. It is worth noting that this does not affect the relative results of the comparison, as the same approach was used for all the classifiers. It is important to observe that the proposed CS^{4}VM method does not rely on the assumption of noise-free samples in the test set for parameter settings. The use of context patterns is effective in mitigating the bias effect introduced by noisy patterns even if the selected model is optimized on a noisy test set. In this condition, we may have an absolute decrease of classification accuracy, but the capability to mitigate the effects of wrong samples on the final classification result does not change.

In the experiments, we considered two data sets: The first one is made up of a very high geometrical resolution multispectral image acquired by the Ikonos satellite over the city of Ypenburg (The Netherlands); the second one is made up of a medium-resolution multispectral image acquired by the sensor TM of Landsat-5 in the surroundings of the city of Trento (Italy). The results obtained on the two data sets are presented in the following two sections.

IV. Experimental Results: Ikonos Data Set

The first considered data set is made up of the first three bands (corresponding to visible wavelengths) of an Ikonos subscene of size 387 × 419 pixels (see Fig. 4). The 4-m spatial resolution spectral bands have been reported to a 1-m spatial resolution according to the Gram–Schmidt pansharpening procedure [24]. The available ground truth (which included

Fig. 4. Band 3 of the Ikonos image.

TABLE I
NUMBER OF PATTERNS IN THE TRAINING AND TEST SETS (IKONOS DATA SET)

Class		Number of patterns	
		Training Set	**Test Set**
Grass		63	537
Road		82	376
Building	Small-aligned	62	200
	White-roof	87	410
	Gray-roof	65	336
	Red-roof	19	92
Shadow		30	231

the information classes grass, road, shadow, small-aligned building, white-roof building, gray-roof building, and red-roof building) collected on two spatially disjoint areas was used to derive a training set and a test set for the considered image (see Table I). This setup allowed us to study the generalization capability of the systems by performing validation on areas spatially disjoint from those used in the learning of the classification algorithm. This is very important because of the nonstationary behavior of the spectral signatures of classes in the spatial domain. Starting from the original training set, several data sets were created by adding different percentages of mislabeled pixels in order to simulate noisy training sets as described in the previous section.

A. Results With Mislabeled Training Patterns Uniformly Added to All Classes

In the first set of experiments, different percentages (10%, 16%, 22%, and 28%) of mislabeled patterns (with respect to the total number of samples) were uniformly added to all classes of the training set. The accuracy yielded on the test set by all the considered classifiers versus the percentage of mislabeled patterns are reported in Table II and shown in Fig. 5. As one can see, with the original training set, the proposed CS^{4}VM exhibited a higher kappa coefficient of accuracy than the other classifiers. In greater detail, the kappa coefficient obtained with the CS^{4}VM is slightly higher than the ones obtained with the standard SVM and the PS^{3}VM (+1.6%) and sharply higher than those yielded by the k-NN (+6.6%) and the ML (+8%). This confirms that the semisupervised exploitation of contextual information of training patterns allows us to increase the classification accuracy (also if their labels are correct). In this condition, the PS^{3}VM classifier did not increase the classification accuracy of the standard SVM. When mislabeled samples were added to the original training set, the accuracies obtained with ML and k-NN classifiers sharply decreased, whereas SVM-based classifiers showed to be much more robust to "noise" (by increasing the number of mislabeled samples, the kappa accuracy decreased slowly). In greater detail, the kappa accuracy of the ML classifier decreased by 15.9% in the case of 10% of mislabeled samples with respect to the result obtained in the noise-free case, while the k-NN reduced its accuracy by 5.8% in the same condition. More generally, the k-NN classifier exhibited higher and more stable accuracies than the ML with all the considered amounts of noisy patterns. In all the considered trials, the proposed CS^{4}VM exhibited higher accuracy than the other classifiers. In addition, with moderate and large numbers of mislabeled patterns (16%, 22%, and 28%), it was more stable than the SVM and the PS^{3}VM. In the trials with noisy training sets, the PS^{3}VM classifier slightly increased the accuracy obtained by the standard SVM.

In order to better analyze the results of SVM and CS^{4}VM, we compared the average and the minimum kappa accuracies of the binary classifiers that made up the OAA multiclass architecture

TABLE II
KAPPA COEFFICIENT OF ACCURACY ON THE TEST SET WITH DIFFERENT PERCENTAGES OF MISLABELED PATTERNS ADDED UNIFORMLY TO THE TRAINING SET (IKONOS DATA SET)

% of mislabeled patterns	Kappa Accuracy				
	CS^{4}VM	**PS^{3}VM**	**SVM**	***k*-NN**	**ML**
0	0.927	0.907	0.907	0.861	0.847
10	0.919	0.910	0.907	0.803	0.688
16	0.921	0.869	0.866	0.787	0.801
22	0.893	0.862	0.861	0.781	0.727
28	0.905	0.874	0.860	0.763	0.675

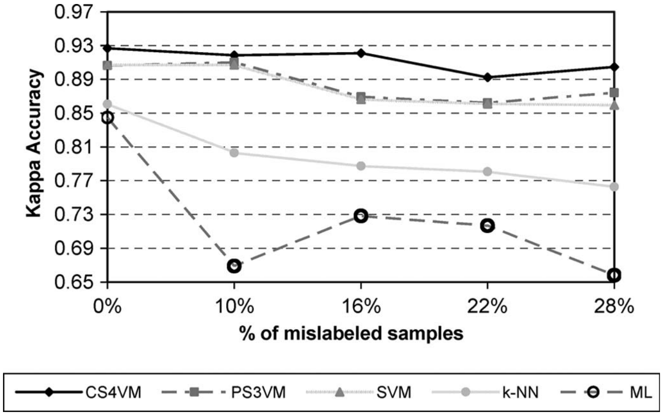

Fig. 5. Behavior of the kappa coefficient of accuracy on the test set versus the percentage of mislabeled training patterns uniformly distributed over all classes introduced in the training set (Ikonos data set).

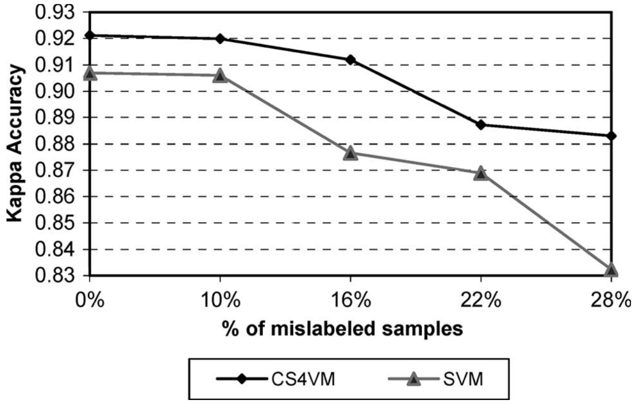

Fig. 6. Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns uniformly added to all classes (Ikonos data set).

TABLE III
KAPPA COEFFICIENT OF ACCURACY EXHIBITED FROM THE BINARY CS^{4}VM AND SVM THAT RESULTED IN THE LOWEST ACCURACY AMONG ALL BINARY CLASSIFIERS INCLUDED IN THE MULTICLASS ARCHITECTURE VERSUS THE PERCENTAGES OF MISLABELED TRAINING PATTERNS UNIFORMLY ADDED TO ALL CLASSES (IKONOS DATA SET)

% of mislabeled patterns	Kappa Accuracy		
	CS^{4}VM	SVM	Δ(%)
0	0.783	0.756	2.7
10	0.784	0.767	1.8
16	0.757	0.738	1.9
22	0.751	0.691	6.0
28	0.755	0.509	24.6

(see Fig. 6 and Table III). It is possible to observe that the average kappa accuracy of the binary CS^{4}VMs was higher than that of the binary SVMs and exhibited a more stable behavior when the amount of noise increased. Moreover, the accuracy of the class most affected by the inclusion of mislabeled patterns in the training set was very stable with the proposed classification algorithm, whereas it sharply decreased with the standard SVM when large percentages of mislabeled patterns were included in the training set. This confirms the effectiveness of the proposed CS^{4}VM, which exploits the contributions of the contextual term (and, thus, of contextual patterns) for mitigating the effects introduced by the noisy samples.

B. Results With Mislabeled Training Patterns Concentrated on Specific Classes

In the second set of experiments, several samples of the class “grass” were added to the original training set with the wrong label “road” in order to reach 10% and 16% of noisy patterns. In addition, “white-roof building” patterns were included with label “gray-roof building” to reach 22% and 28% of noisy samples. The resulting classification problem proved quite critical, as confirmed by the significant decrease in the kappa accuracies yielded by the considered classification algorithms (see Fig. 7 and Table IV). Nevertheless, also in this case, the context-based training of the CS^{4}VM resulted in a significant increase of accuracy with respect to other classifiers. The kappa accuracy of the k-NN classifier dramatically decreased when the percentage of noisy patterns increased (in the specific case of 28% of mislabeled samples, the kappa accuracy decreased by 35.1% with respect to the original training set). The ML decreased its accuracy by 10.1%, with 10% of noisy patterns, but exhibited a more stable behavior with respect to the k-NN when the amount of noisy patterns was further increased. The standard SVM algorithm obtained accuracies higher than those yielded by the k-NN and ML classifiers, while the PS^{3}VM classifier, in general, slightly improved the accuracy of the standard SVM. However, with 28% of noisy patterns, the kappa accuracy sharply decreased to 0.629 (below the performance of ML). This behavior was strongly mitigated by the proposed CS^{4}VM (which exhibited kappa accuracy of 0.820 in the same conditions).

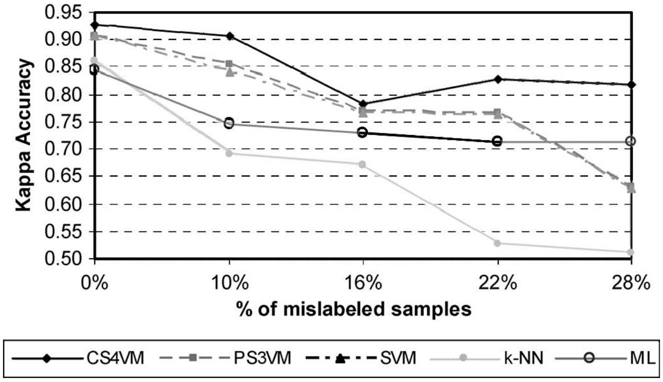

Fig. 7. Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns concentrated on specific classes of the training set (Ikonos data set).

TABLE IV
KAPPA COEFFICIENT OF ACCURACY ON THE TEST SET WITH DIFFERENT PERCENTAGES OF MISLABELED PATTERNS ADDED TO SPECIFIC CLASSES OF THE TRAINING SET (IKONOS DATA SET)

% of mislabeled patterns	Kappa Accuracy				
	CS^{4}VM	PS^{3}VM	SVM	k-NN	ML
0	0.927	0.907	0.907	0.861	0.847
10	0.906	0.855	0.841	0.690	0.746
16	0.781	0.769	0.765	0.672	0.734
22	0.828	0.767	0.762	0.525	0.722
28	0.820	0.632	0.629	0.510	0.721

Considering the behavior of the average kappa of the binary SVMs and CS^{4}VMs that made up the OAA multiclass architecture (see Fig. 8), it is possible to note that the CS^{4}VM always improved the accuracy of the standard SVM, and the gap between the two classifiers increased by increasing the amount of noisy samples. In the very critical case of 28% of mislabeled patterns, the context-based learning of CS^{4}VM improved the average kappa accuracy of binary SVMs by 9.2%. Moreover, the kappa coefficient of the class with the lowest accuracy with the proposed CS^{4}VM, even if small, was sharply higher than that of the standard SVM in all the considered trials (see Table V). This behavior shows that on this data set, the proposed method always improved the accuracy of the most critical binary classifier.

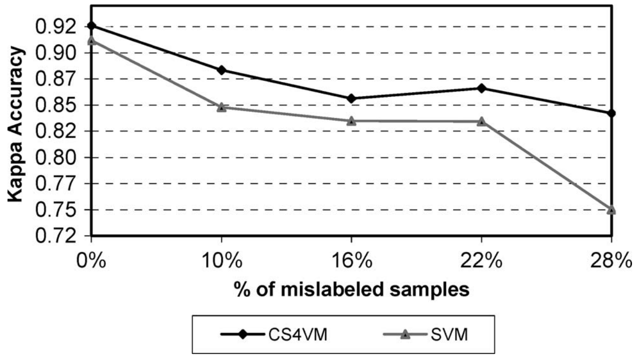

Fig. 8. Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns concentrated on specific classes (Ikonos data set).

TABLE V
KAPPA COEFFICIENT OF ACCURACY EXHIBITED FROM THE BINARY CS^{4}VM AND SVM THAT RESULTED IN THE LOWEST ACCURACY AMONG ALL BINARY CLASSIFIERS INCLUDED IN THE MULTICLASS ARCHITECTURE VERSUS THE PERCENTAGES OF MISLABELED TRAINING PATTERNS CONCENTRATED ON SPECIFIC CLASSES (IKONOS DATA SET)

% of mislabeled patterns	**Kappa Accuracy**		
	CS^{4}VM	**SVM**	**Δ(%)**
0	0.783	0.756	2.7
10	0.620	0.422	19.8
16	0.449	0.360	8.9
22	0.538	0.360	17.8
28	0.450	0.360	9.0

Fig. 9 shows the classification maps obtained by training the considered classifiers with 28% of mislabeled patterns added on specific classes (“roads” and “gray-roof buildings”) of the training set (the map obtained with the PS^{3}VM is not reported because it is very similar to the one yielded with the SVM classifier). As one can see, in the classification maps obtained with the SVM, the k-NN, and the ML algorithms, many pixels of the class grass are confused with the class road, while white-roof buildings are confused with the gray-roof buildings. This effect is induced by the presence of noisy training samples affecting the aforementioned classes. In greater detail, the SVM classifier was unable to correctly recognize the red-roof buildings, while the k-NN technique often misrecognized the shadows present in the scene as red-roof buildings and white-roof buildings as gray-roof buildings. Moreover, the thematic map obtained with the k-NN is very noisy and fragmented (as confirmed by the low kappa coefficient of accuracy). The thematic map obtained with the proposed CS^{4}VM clearly appears more accurate and less affected by the presence of mislabeled patterns.

V. EXPERIMENTAL RESULTS: LANDSAT DATA SET

The second data set consists of an image acquired by the Landsat-5 TM sensor with a GIFOV of 30 m. The considered image has size of 1110 × 874 pixels and was taken in the surrounding of the city of Trento (Italy) (see Fig. 10). A six-class classification problem (with forest, water, urban, rock, fields, and grass classes) was defined according to the available ground truth collected on two spatially disjoint areas and used to derive the training and test sets (see Table VI). As for the Ikonos data set, this setup allowed us to study the generalization capability of the algorithms by classifying areas spatially disjoint from those used in the learning of the classifier. The important difference between this data set and the previous one consists in the geometric resolution, which in this case is significantly smaller than in the previous case (30 m versus 1 m). Similar to the Ikonos data set, several noisy training sets were created by adding different amount of mislabeled pixels to the original data set: 1) with uniform distribution over the classes and 2) concentrated on a specific class.

A. Results With Mislabeled Training Patterns Uniformly Added to All Classes

Table VII shows the accuracies obtained in the first set of experiments, where mislabeled patterns were uniformly added to the information classes. Fig. 11 shows the behavior of the kappa accuracy versus the number of mislabeled patterns included in the training set for all the considered classifiers. It is possible to observe that with the noise-free training set, the proposed CS^{4}VM led to the highest accuracy, slightly improving the kappa coefficient of standard SVM by 0.8%. The ML classifier performed very well with the noise-free training set (the kappa accuracy was 0.923) but decreased its accuracy to 0.778 when only 10% of mislabeled patterns were introduced in the original training set, and its accuracy further decreased to 0.691 when the mislabeled samples reached 16%. The k-NN classifier led to a lower accuracy than the ML in the absence of noise but showed to be less sensitive to noisy patterns uniformly added to the training set, thus exhibiting a more stable behavior. On the contrary, the SVM-based classification algorithms proved to be robust to the presence of mislabeled training samples. Indeed, the excellent generalization capability of the SVM led to even slightly increase the classification accuracy when a small amount of mislabeled patterns was added to the training set. The PS^{3}VM algorithm resulted in a small improvement with respect to the SVM classifier in the trials where mislabeled samples were added to the training set. The kappa accuracy of the SVM classifier slightly decreased when the mislabeled samples exceeded 16%, reducing its accuracy by 3% with respect to the noise-free case. In these cases, the proposed CS^{4}VM further enhanced the robustness of SVM, leading to kappa accuracies that were always above 0.91.

This behavior is confirmed by the analysis of the average and minimum kappa computed on the binary classifiers (see Fig. 12 and Table VIII), which highlights that the CS^{4}VM significantly improved the accuracy with respect to the SVM. Such an improvement was more significant when increasing the amount of noise; thus, the CS^{4}VM resulted in a more stable value of the kappa coefficient with respect to the percentage of mislabeled patterns present in the training set. It is worth noting that on this data set, the proposed CS^{4}VM always improved the average kappa accuracy of the binary classifiers, even in cases where the global multiclass kappa coefficient of the CS^{4}VM was slightly smaller than the one obtained with the standard SVM. This can be explained by observing that the decision strategy associated with the OAA multiclass architecture in some cases

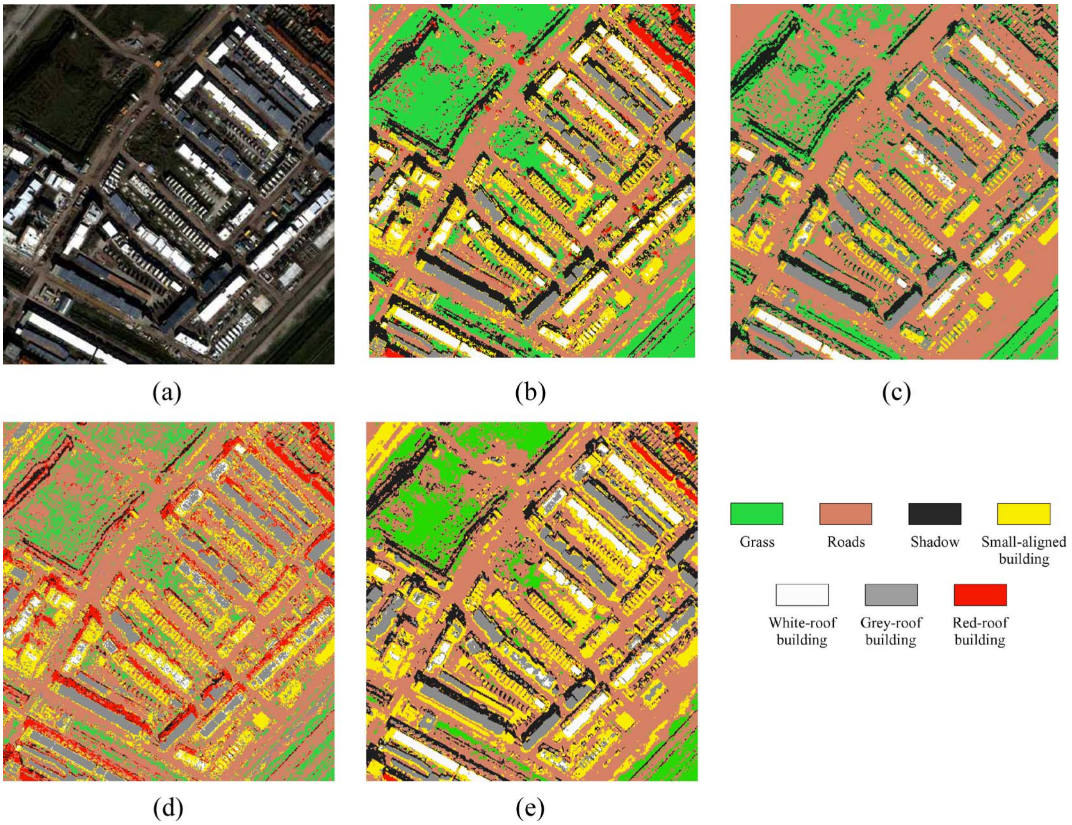

Fig. 9. (a) True color composition of the Ikonos image. Classification maps obtained by the different classifiers with the training set containing 28% of mislabeled patterns added on specific classes. (b) CS^{4}VM. (c) SVM. (d) k-NN. (e) ML.

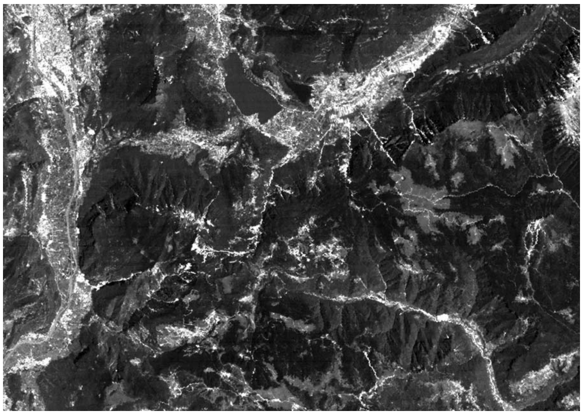

Fig. 10. Band 2 of the Landsat TM multispectral image.

TABLE VI
NUMBER OF PATTERNS IN THE TRAINING AND TEST SET (LANDSAT DATA SET)

Class	Number of patterns	
	Training Set	Test Set
Forest	128	538
Water	118	177
Urban	137	289
Rocks	45	51
Fields	93	140
Grass	99	227

TABLE VII
KAPPA COEFFICIENT OF ACCURACY ON TEST SET USING DIFFERENT PERCENTAGES OF MISLABELED PATTERNS ADDED UNIFORMLY TO THE TRAINING SET (LANDSAT DATA SET)

% of mislabeled patterns	Kappa Accuracy				
	CS^{4}VM	PS^{3}VM	SVM	k-NN	ML
0	0.927	0.915	0.919	0.912	0.923
10	0.930	0.935	0.931	0.905	0.778
16	0.935	0.932	0.930	0.893	0.691
22	0.921	0.891	0.886	0.868	0.686
28	0.916	0.886	0.886	0.840	0.681

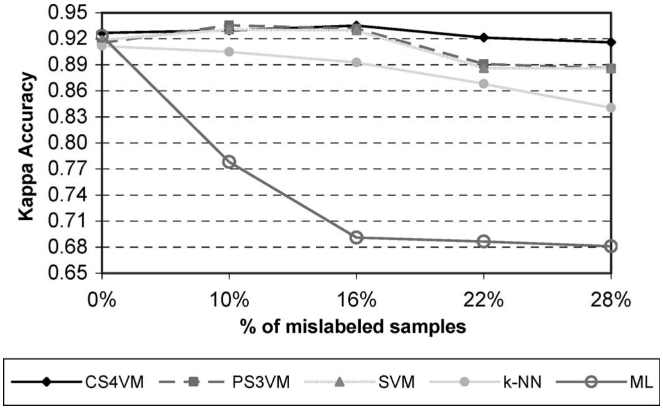

Fig. 11. Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns uniformly added to all classes (Landsat data set).

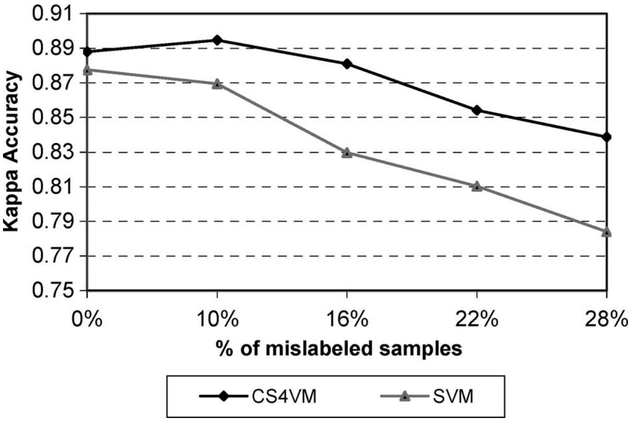

Fig. 12. Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns uniformly added to all classes (Landsat data set).

TABLE VIII
KAPPA COEFFICIENT OF ACCURACY EXHIBITED FROM THE CS^{4}VM AND SVM THAT RESULTED IN THE LOWEST ACCURACY AMONG ALL BINARY CLASSIFIERS INCLUDED IN THE MULTICLASS ARCHITECTURE VERSUS THE PERCENTAGES OF MISLABELED TRAINING PATTERNS UNIFORMLY ADDED TO ALL CLASSES (LANDSAT DATA SET)

% of mislabeled patterns	Kappa Accuracy		
	CS^{4}VM	SVM	Δ(%)
0	0.701	0.701	0.0
10	0.701	0.701	0.0
16	0.650	0.627	2.3
22	0.650	0.579	7.1
28	0.641	0.498	14.3

could "recover" the errors of binary classifiers by assigning the correct label to a pattern when comparing the output of binary classifiers. Nevertheless, the increased average accuracy of the binary CS^{4}VMs is an important property because it involves more stable and reliable classification results.

Fig. 13 shows the classification maps obtained by training the classifiers with 28% of mislabeled patterns uniformly added to all the classes. It is possible to observe that the map generated by the proposed CS^{4}VM is the most accurate. In the maps yielded by the SVM, the k-NN, and the ML algorithms, several pixels are misclassified as water (the map obtained with the PS^{3}VM is not reported as very similar to the SVM map). In greater detail, the map obtained with the k-NN presents confusion between the classes water and urban, and the classes forest and water. In the map obtained by the ML, grass areas are often confused with forest.

B. Results With Mislabeled Training Patterns Concentrated on a Specific Class

In the second set of experiments, several samples of the class "forest" were added to the class "fields" to reach 10%, 16%, 22%, and 28% of mislabeled patterns with respect to the total number of training samples. Moreover, in this case, the presence of errors that systematically affected one class severely impacted the performance of the supervised classification algorithms. When a low percentage (10%) of noisy patterns was added to the original training set, all the considered classifiers decreased their kappa coefficient of accuracy by more than 12% (see Table IX and Fig. 14). In contrast to the first set of experiments, also the SVM algorithm suffered the presence of this type of noisy training set, reducing its accuracy by 18.4% (while the k-NN decreased its accuracy by 20.2% and the ML by 22.5%). The semisupervised approach based on the PS^{3}VM was not able to improve the accuracies of the standard SVM. The CS^{4}VM could partially recover the accuracy of standard SVM by increasing the kappa accuracy by 7.4%, thus limiting the effect of mislabeled patterns. When the amount of noisy patterns further increased, PS^{3}VM, SVM, ML, and k-NN classifiers did not further decrease significantly their kappa accuracies.

This behavior is confirmed from the average kappa coefficient of accuracy of the binary classifiers versus the percentage of mislabeled training patterns (see Fig. 15). In this case, we do not report the results of the binary classifiers exhibiting the lowest accuracy, because the complexity of the problem resulted in unreliable kappa values on this class (even if, also in this case, the CS^{4}VM outperformed the SVM).

VI. DISCUSSION AND CONCLUSION

In this paper, we have proposed a novel classification technique based on SVM that exploits the contextual information in order to render the learning of the classifier more robust to possible mislabeled patterns present in the training set. Moreover, we have analyzed the effects of mislabeled training samples on the classification accuracy of supervised algorithms, comparing the results obtained by the proposed CS^{4}VM with those yielded by a PS^{3}VM, a standard supervised SVM, a Gaussian ML, and a k-NN. This analysis was carried out, varying both the percentage of mislabeled patterns and their distribution on the information classes. The experimental results obtained on two different data sets (a VHR image acquired by the Ikonos satellite and a medium-resolution image acquired by the Landsat-5 satellite) confirm that the proposed CS^{4}VM approach exhibits augmented robustness to noisy training sets with respect to all the other classifiers. In greater detail, the proposed CS^{4}VM method always increased the average kappa coefficient of accuracy of the binary classifiers included in the OAA multiclass architecture with respect to the standard SVM classifier. Moreover, in many cases, the CS^{4}VM sharply increased the accuracy on the information class that was most affected by the mislabeled patterns introduced in the training set.

By analyzing the effects of the distribution of mislabeled patterns on the classes, it is possible to conclude that errors concentrated on a class (or on a subset of classes) are much more critical than errors uniformly distributed on all classes. In greater detail, when noisy patterns were added uniformly to all classes, we observed that the proposed CS^{4}VM resulted in higher and more stable accuracies than all the other classifiers. The supervised SVM and the PS^{3}VM exhibited relatively high accuracies when a moderate amount of noisy patterns was included in the training set, but they slowly decreased their accuracy when the percentage of mislabeled samples increased. On the contrary, both the ML and the k-NN classifiers are very sensitive even to the presence of a small amount of noisy

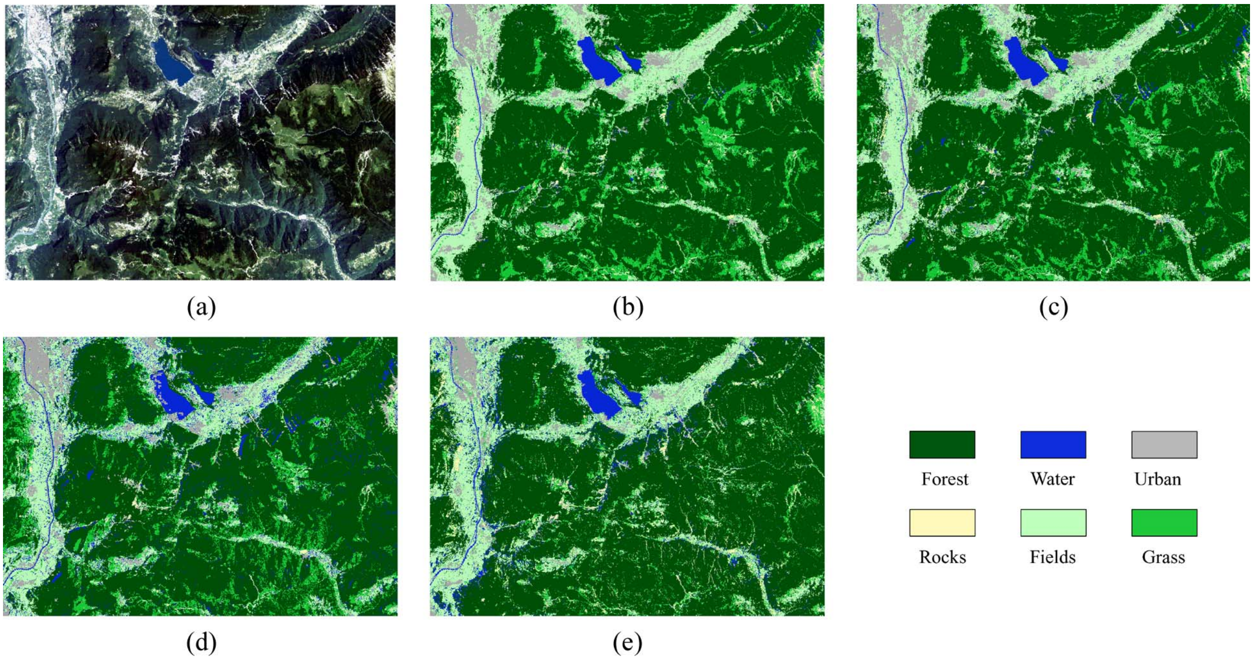

Fig. 13. (a) True color composition of Landsat image. Classification maps obtained by the different classifiers with the training set containing 28% of noisy patterns uniformly added to all classes. (b) CS^{4}VM. (c) SVM. (d) k-NN. (e) ML.

TABLE IX
KAPPA COEFFICIENT OF ACCURACY ON THE TEST SET USING TRAINING SETS WITH DIFFERENT PERCENTAGES OF MISLABELED PATTERNS ADDED TO A SPECIFIC CLASS (LANDSAT DATA SET)

% of mislabeled patterns	Kappa Accuracy				
	CS^{4}VM	PS^{3}VM	SVM	k-NN	ML
0	0.927	0.915	0.919	0.882	0.923
10	0.809	0.738	0.735	0.680	0.699
16	0.712	0.706	0.695	0.652	0.678
22	0.691	0.664	0.661	0.632	0.671
28	0.658	0.651	0.648	0.632	0.666

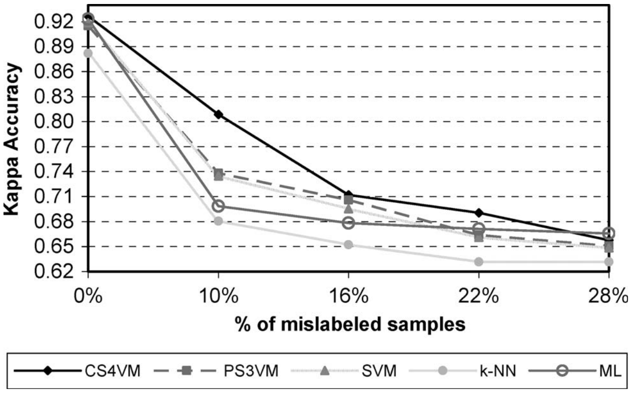

Fig. 14. Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns concentrated on a specific class (Landsat data set).

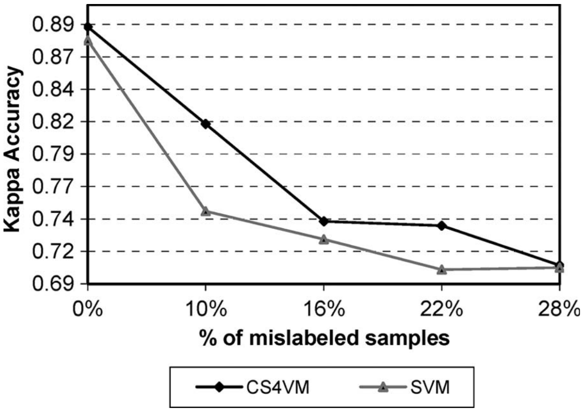

Fig. 15. Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns concentrated on a specific class (Landsat data set).

patterns and sharply decreased their accuracies by increasing the number of mislabeled samples. Nevertheless, the k-NN classifier resulted significantly more accurate than the ML classifier when mislabeled patterns equally affected the considered information classes. When noisy patterns were concentrated on a specific class of the training set, the accuracies of all the considered classifiers sharply decreased by increasing the amount of mislabeled training samples. Moreover, in this case, the proposed CS^{4}VM exhibited, in general, the highest and more stable accuracies. Nonetheless, when the number of mislabeled patterns increased over a given threshold, the classification problem became very critical and also the proposed technique significantly reduced its effectiveness. The standard SVM classifier still maintained higher accuracies than the ML and the k-NN techniques. The PS^{3}VM slightly increased the accuracies of the standard SVM. Unlike the previous case, the k-NN algorithm resulted in lower accuracies than the ML method. This is mainly due to the fact that mislabeled patterns concentrated on a single class (or on few classes) alter the prior probabilities, thus affecting more the k-NN classifier (which implicitly considers the prior probabilities in the decision rule) than the ML technique (which does not consider the prior probabilities of classes).

The proposed CS^{4}VM introduces some additional free parameters with respect to the standard supervised SVM, which should be tuned in the model-selection phase. The analysis on the effects of the values of these parameters on the classification results (carried out in the different simulations described in this paper) pointed out that the empirical selection of $K = \kappa_1/\kappa_2 = 2$ (which is reasonable, considering the physical meaning of this ratio) resulted in good accuracies on both data sets. This choice allows one to reduce the model-selection phase to tune the value of the ratio C/κ_1 in addition to the standard SVM parameters. Nonetheless, when possible, the inclusion of the choice of the κ_1/κ_2 value in the model selection would optimize the results achievable with the proposed approach. The optimal value for the ratio C/κ_1 depends on the considered data set and the type of mislabeling errors, but in general, we observed that higher weights for the context patterns (lower values for the ratio C/κ_1) can result in better classification accuracies when the percentage of mislabeled training patterns increases. This confirms the importance of the context term to increase the classification accuracy in the presence of noisy training sets.

It is worth noting that the considered PS^{3}VM classifier slightly improved the accuracy with respect to the standard SVM by exploiting the information of unlabeled samples, but it could not gain in accuracy when the amount of mislabeled patterns increased. Indeed, the PS^{3}VM is not developed to take into account the possible presence of mislabeled training patterns, which affect the first iteration of the learning phase propagating the errors to the semilabeled samples in the next iterations of the algorithm. On the contrary, the proposed CS^{4}VM is particularly developed to cope with "non fully reliable" training sets by exploiting the information of pixels in the neighborhood of the training points according to a specific weighting mechanism that penalizes less reliable training patterns. In addition, the proposed CS^{4}VM approach is computationally less demanding than the PS^{3}VM as it requires only two steps (this choice is done for limiting the computational complexity and is supported from empirical experiments that confirmed that increasing the number of iterations does not significantly change the classification results). On the contrary, the PS^{3}VM may require a large number of iterations before convergence.

The computational cost of the learning phase of the proposed CS^{4}VM method is slightly higher than that required from the standard supervised SVM. This depends on both the second step of the learning algorithm (which involves an increased number of samples, as semilabeled context patterns are considered in the process) and the setting of the additional parameters in the model-selection phase. In our experiments on the Ikonos data set, carried out on a PC mounting an Intel Pentium D processor at 3.4 GHz and a 2-Gb DDR2 RAM, the training phase of a supervised SVM took in average about 20 s, while the one of the proposed CS^{4}VM required about 3 min. It is important to point out that the additional cost of the proposed method concerns only the learning phase, whereas the computational time in the classification phase remains unchanged.

As a final remark, it is worth stressing that the proposed analysis points out the dramatic effects involved on the classification accuracy from a relatively small percentages of mislabeled training samples concentrated on a class (or on a subset of classes). This should be understood in order to define adequate strategies in the design of training data for avoiding this kind of errors.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[2] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 4th ed. Berlin, Germany: Springer-Verlag, 2006.

[3] 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.

[4] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2001.

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

[6] 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.

[7] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002. [Online]. Available: http://www.learning-with-kernels.org/

[8] 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.

[9] 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.

[10] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Proc. IEEE PReMI*, Kolkata, India, Dec. 2005, vol. 3776, pp. 260–265.

[11] A. A. Farag, R. M. Mohamed, and A. El-Baz, "A unified framework for MAP estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[12] F. Melgani and S. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003.

[13] G. Moser, S. Serpico, and F. Causa, "MRF model parameter estimation for contextual supervised classification of remote-sensing images," in *Proc. IEEE IGARSS*, Jul. 2005, pp. 308–311.

[14] P. Gamba, F. Dell'Acqua, G. Lisini, and G. Trianni, "Improved VHR urban area mapping exploiting object boundaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 8, pp. 2676–2682, Aug. 2007.

[15] M. Berthod, Z. Kato, S. Yu, and J. Zerubia, "Bayesian image classification using Markov random fields," *Image Vis. Comput.*, vol. 14, no. 4, pp. 285–295, May 1996.

[16] R. Nishii, "A Markov random field-based approach to decision-level fusion for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 10, pp. 2316–2319, Oct. 2003.

[17] L. Bruzzone, M. Chi, and M. Marconcini, "Semisupervised support vector machines for classification of hyperspectral remote sensing images," in *Hyperspectral Data Exploitation*, C.-I. Chang, Ed. Hoboken, NJ: Wiley, 2007, ch. 11, pp. 275–311.

[18] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[19] 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.

[20] 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.

[21] 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.

[22] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL: Lewis Publishers, 1999.

[23] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998, ch. 12, pp. 185–208.

[24] B. Aiazzi, S. Baronti, M. Selva, and L. Alparone, "Enhanced Gram-Schmidt spectral sharpening based on multivariate regression of MS and pan data," in *Proc. IEEE IGARSS*, 2006, pp. 3806–3809.

Lorenzo Bruzzone (S'95–M'98–SM'03) received the Laurea (M.S.) degree (*summa cum laude*) in electronic engineering 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 with the University of Genoa. Since 2000, he has been with the University of Trento, Trento, Italy, where he his currently a Full Professor of telecommunications. He teaches remote sensing, pattern recognition, and electrical communications. He is the Head of the Remote Sensing Laboratory in the Department of Information and Communication Technology, University of Trento. He conducts and supervises research on these topics within the frameworks of several national and international projects. He is an Evaluator of project proposals for many different governments (including European Commission) and scientific organizations. He is the author (or coauthor) of 60 scientific publications in referred international journals, more than 120 papers in conference proceedings, and seven book chapters. He is a Referee for many international journals and has served on the scientific committees of several international conferences. His current research interests are in the area of remote-sensing image processing and recognition (analysis of multitemporal data, feature extraction and selection, classification, regression and estimation, data fusion, and machine learning).

Dr. Bruzzone is a member of the Managing Committee of the Italian Inter-University Consortium on Telecommunications and 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. He was the General Chair and Cochair of the First and Second IEEE International Workshop on the Analysis of Multitemporal Remote-Sensing Images (MultiTemp) and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. From 2004 to 2006, he served as an Associated Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and currently, he is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He 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).

Claudio Persello (S'07) received the Laurea (B.S.) and Laurea Specialistica (M.S.) degrees in telecommunication engineering from the University of Trento, Trento, Italy, in 2003 and 2005, respectively, where he is currently working toward the Ph.D. degree in information and communication technologies.

He is working with the Remote Sensing Group at the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the area of remote sensing, image classification, pattern recognition, and machine learning.

Mr. Persello is a referee for the *Canadian Journal of Remote Sensing* and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.