Toward the Automatic Updating of Land-Cover Maps by a Domain-Adaptation SVM Classifier and a Circular Validation Strategy

Lorenzo Bruzzone, *Senior Member, IEEE*, and Mattia Marconcini, *Member, IEEE*

***Abstract*—In this paper, we address automatic updating of land-cover maps by using remote-sensing images periodically acquired over the same investigated area under the hypothesis that a reliable ground truth is not available for all the considered acquisitions. The problem is modeled in the domain-adaptation framework by introducing a novel method designed for land-cover map updating, which is based on a domain-adaptation support vector machine technique. In addition, a novel circular accuracy assessment strategy is proposed for the validation of the results obtained by domain-adaptation classifiers when no ground-truth labels for the considered image are available. Experimental results obtained on a multitemporal and multispectral data set confirmed the effectiveness and the reliability of the proposed system.**

***Index Terms*—Domain adaptation, kernel methods, partially unsupervised classification, semisupervised classification, support vector machines (SVMs), transfer learning, updating land-cover maps, validation strategy.**

I. Introduction

In THE LAST few years, the advances in remote-sensing technology have led to a growing interest in the use of space-borne data for large-scale mapping applications. In this framework, satellite images periodically acquired over the same geographical area have demonstrated to be particularly effective for providing and updating land-cover information in a timely and cost-effective manner. Thus, they make it possible to develop monitoring systems based on supervised classifiers that can map the information classes that characterize a specific geographical area on a regular basis. From an operational viewpoint, the implementation of such kind of systems requires the availability of adequate and reliable ground-truth labels for each new image to be categorized. In fact, although the considered images refer to the same area, it is reasonable to expect that there might occur relevant changes between the information class distributions that characterize each specific acquisition date due to several possible reasons (e.g., dissimilar illumination conditions, different acquisition system state, alterations in the phenologic state of vegetation, changes occurred on the ground, etc.). When ground-truth labels are available for all the items of the temporal series, the use of supervised classification approaches can be particularly effective [1], [2]. Nevertheless, gathering reliable ground truth for each specific acquisition date is not realistic and is generally very expensive both in terms of time and economic cost. Thus, in real applications, such constraint is rarely satisfied, and in several cases, it is not possible to rely on training data as frequently as required to ensure an efficient monitoring of the investigated site. For this reason, the process of temporal updating of land-cover maps results in a very complex and challenging problem.

Recently, in the remote-sensing community, great attention has been devoted to address ill-posed classification problems characterized by a small amount of training samples. In such situations, transductive and semisupervised[1] learning methods proved capable to improve the performances with respect to supervised classifiers by exploiting, in addition to the small-size available training data, unlabeled samples taken from the image being classified. In several applications, these approaches proved to be particularly effective and resulted in a relevant increase of the classification accuracy [3]. Due to its intrinsic high complexity, less attention has been devoted to solving the problem of land-cover map updating when ground truth is available only for one image of a temporal series. For addressing this challenging problem and improving the discrimination capability with respect to supervised classifiers, it is necessary to design different kinds of classifiers, which should be able to jointly exploit labeled and unlabeled patterns that refer to different images. In this framework, the authors have defined and developed in previous works partially unsupervised[2] classification techniques that aim at classifying an image for which no ground truth is available, by exploiting labeled training patterns of another image acquired over the same investigated area [4]–[7]. Nevertheless, even if these transfer learning approaches proved to be effective, they exhibited some limitations. The main drawback is due to the parametric nature of the proposed classifiers, which prevents their employment

Manuscript received May 15, 2008; revised August 1, 2008 and September 25, 2008. First published March 6, 2009; current version published March 27, 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).

Digital Object Identifier 10.1109/TGRS.2008.2007741

[1]A classifier which jointly exploits labeled and unlabeled data that refer to the same image is said to be as follows: 1) *transductive*, if it is specifically designed to suite only the unlabeled samples used in the learning process and cannot handle unlabeled unseen data, or 2) *semisupervised*, if it is designed to handle any unlabeled data of the considered image.

[2]The term "partially unsupervised" has been used to point out that, on the one hand, no ground-truth information is assumed to be available for a given investigated image; on the other hand, there exists a training set related to an image of the same geographical area acquired before the one to be classified.

in the cases where it is not possible to model explicitly the kind of distribution that governs the investigated classification problem. Aside from the intrinsic complexity of the considered problem, another important limitation in developing operational classification methods for addressing land-cover map updating is related to the lack of validation strategies that permit to assess the effectiveness of the classification results. In fact, under the assumption that no ground-truth labels are available for the image(s) being classified, standard statistical validation methods cannot be employed.

The problem of updating land-cover maps by classifying temporal series of images when only training samples collected at one time are available can be modeled in a more general framework, which is known in the pattern recognition and machine learning community as *domain adaptation* [8]–[11]. In domain-adaptation classification problems, unlabeled test patterns are drawn from a "target-domain" distribution different from the "source-domain" distribution of training samples. In such context, in this paper, we propose a novel domain-adaptation support vector machine (DASVM) technique that extends support vector machines (SVMs) to the domain-adaptation problem and permits the addressing of land-cover map updating in real operational cases where no ground truth is available for new images to be classified. Starting from a standard supervised learning, the proposed DASVM technique iteratively selects and labels the (unlabeled) patterns of the new image to be categorized that are most likely to be correctly classified. At the same time, original (labeled) samples are gradually erased, as they refer to a reference image different from the one being classified. In addition, as a second original contribution of this paper, we present a circular validation strategy that permits to automatically identify solutions that are consistent with the investigated problem and, thus, to assess the correctness of a land-cover map obtained with the DASVM technique when no ground-truth samples are available for the image being classified. In particular, this strategy (which can be used for validating the classification results of any domain-adaptation method) indirectly analyzes the robustness of the classification map obtained for the new image of a time series at the end of a circular (i.e., forward and backward) domain-adaptation process.

Several experimental results obtained on a multitemporal and multispectral data set related to the Lake Mulargia (Sardinia Island, Italy) confirmed the effectiveness and the reliability of the proposed system.

This paper is organized into seven sections. In Section II, a survey on both semisupervised and partially unsupervised methods is reported. Sections III and IV present the proposed DASVM technique and the circular validation strategy devised for addressing land-cover map updating, respectively. In Section V, experimental results are reported. Finally, Section VI draws the conclusions of this paper.

II. Literature Survey

In this section, we analyze the main contributions present in the literature on semisupervised and domain-adaptation methods.

A. Semisupervised Methods

In many remote-sensing applications, due to the practical impossibility to obtain a sufficient number of representative training samples for a reliable estimation of classifier parameters (particularly when many information classes are considered), supervised classification approaches may result in poor accuracies. Accordingly, in the last few years, the scientific community has devoted a growing interest to the definition of semisupervised classification techniques, which exploit both labeled and unlabeled patterns taken from the remote-sensing image being classified.

A possible approach to address this kind of problems is to use the expectation-maximization (EM) algorithm [12] for a maximum-likelihood (ML) estimation of the parameters that characterize the distributions of the considered information classes. In [13], Shahshahani and Landgrebe proved that additional unlabeled samples are helpful for semisupervised classification of hyperspectral remote-sensing images in the context of a Gaussian ML classifier under a zero-bias assumption. By assuming a Gaussian mixture model (GMM), the EM algorithm is employed to estimate model parameters with both labeled and unlabeled samples to better estimate the parameters of the GMM. In order to limit the risk of a possible negative influence of semilabeled samples (which are originally unlabeled samples that obtain labels during the learning process), in [14], a weighting strategy is introduced. Full weights are assigned to training samples, whereas reduced weights are defined for semilabeled samples during the estimation phase of the EM algorithm. Nevertheless, when only very few labeled patterns are available, the covariance matrices are generally highly variable. To overcome this problem, in [15], Jackson and Landgrebe proposed an adaptive covariance estimator. In the adaptive quadratic process, semilabeled samples are incorporated in the training set to estimate regularized covariance matrices so that the variance of these matrices can be smaller compared to the conventional counterparts estimated with labeled samples alone [14].

Recently, in the machine learning community, a growing attention has been focused on semisupervised approaches implemented under the cluster assumption: Each cluster of samples is assumed to belong to one data class; thus, the decision boundary is defined between clusters, i.e., in low-density regions of the feature space. In this context, transductive SVMs (TSVMs) [16], [17] and semisupervised SVMs (S^3VMs) [18] proved particularly effective in several applications. In particular, TSVMs and S^3VMs exploit specific iterative algorithms based on SVMs which gradually search a reliable separation hyperplane (in the kernel space) through a learning 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, the authors first proposed in [19] a semisupervised classifier specifically designed for addressing ill-posed problems in the context of remote sensing. In particular, this technique has the following properties: 1) It is based on a transductive procedure that exploits a weighting strategy for unlabeled patterns on the basis of a time-dependent criterion; 2) it is able to mitigate the effects of suboptimal model selection (which is unavoidable in the presence of small-size training sets); and 3) it can address multiclass problems.

A similar method that exploits, in place of SVMs, a modified version of the Kernel Fisher's discriminant using labeled and unlabeled data has been presented in [20] by Dundar and Landgrebe. In particular, the proposed technique is obtained through an optimization of a quadratic programming problem that minimizes the total cost of misclassified labeled data while maximizing the Rayleigh coefficient in the kernel space.

Other effective approaches have been presented in [21], where the authors introduced in remote sensing two different S^3VM algorithms for the classification of hyperspectral data implemented and optimized in the primal formulation. In this case, as proposed in [22], the constraints of the labeled and unlabeled samples are directly included in the cost function in order to obtain an unconstrained optimization problem. The first presented primal S^3VM optimizes the unconstrained objective function by the gradient descent technique, leading to the formulation of ∇S^3VMs. The second algorithm combines ∇S^3VMs with a graph-based kernel matrix.

A different class of promising methods includes graph-based semisupervised algorithms, which define a graph where the nodes are labeled and unlabeled patterns and edges reflect the similarity of samples. Each sample spreads its label information to its neighbors until a global stable state is achieved on the whole data set. Graph-based approaches aim at estimating an objective function on the graph which generally consists of a loss term and a regularizer. In this context, an interesting approach has been proposed in [23] where Camps-Valls *et al.* presented a graph-based classifier for hyperspectral images based on the algorithm described in [24], which takes advantage of both the high number of unlabeled samples present in the image and the integration of contextual information.

Another technique has been recently proposed by Gómez *et al.* in [25]. In particular, the authors extended to the remote-sensing domain the Laplacian SVM technique proposed in [26], which introduces an additional regularization term on the geometry of both labeled and unlabeled samples by using the graph Laplacian [27]. This method follows a noniterative optimization procedure in contrast to most transductive learning methods and provides out-of-sample predictions in contrast to graph-based approaches.

In order to increase the reliability of the semisupervised learning process, systems based on ensemble methods have also been devised. As an example, in [28] and [29], the authors proposed the employment of semilabeled-sample-driven bagging techniques.

B. Domain-Adaptation Methods

Semisupervised approaches proved to be useful for improving the discrimination capability with respect to supervised classifiers when both available labeled and unlabeled data refer to the same domain. However, in the case of addressing the updating of land-cover maps, domain-adaptation methods are necessary as, unlike the semisupervised case, they permit the joint exploitation of labeled and unlabeled patterns that refer to different images (i.e., different domains). While, in the last few years, several techniques have been designed for handling remote-sensing semisupervised problems, at the present, only few methods have been proposed for tackling the challenging problem of land-cover map updating. In this context, partially unsupervised techniques that address domain adaptation in the framework of remote sensing have been previously proposed by the authors in [4]–[7]. In [4], a partially unsupervised approach is proposed, which can update the parameters of an already trained parametric ML classifier on the basis of the distribution of a new image for which no ground-truth information is available. In [5], in order to take into account the temporal correlation between images acquired over the same area at different times, the partially unsupervised ML classification approach is reformulated in the framework of the Bayesian rule for cascade classification (i.e., the classification process is performed by jointly considering information contained in all the items of a temporal series). The basic idea in both approaches consists in modeling the observed spaces by a mixture of distributions, whose components are estimated through the employment of unlabeled data according to a proper inference applied to training samples of the reference image. This is achieved by using a specific version of the EM algorithm with finite GMM [13]. In [6] and [7], partially unsupervised classification approaches based on a multiple-classifier system and a multiple-cascade-classifier system (MCCS) have been defined, respectively. In particular, in [7], the proposed MCCS architecture is made up of an ensemble of partially unsupervised classifiers integrated in a multiple-classifier architecture. Each classifier of the ensemble is developed in the framework of cascade classification. Both a parametric ML classification approach and a nonparametric radial basis function neural-network (RBF-NN) classification technique are used as basic classifiers. In addition, in order to increase both the effectiveness and the robustness of the ensemble, hybrid ML and RBF-NN cascade classifiers are defined. In this way, the resulting partially unsupervised MCCS is characterized by a higher reliability than single algorithms composing the ensemble.

However, in general, the updating of land-cover map is still a scarcely investigated problem that deserves to be further studied given the relevant impact that its solution can have on many application domains.

III. Proposed Domain-Adaptation Classifier

In this section, we describe the proposed DASVM classifier. Note that the rationale for developing a domain-adaptation technique in the framework of SVMs is due to the effectiveness of this classification methodology. Accordingly, in the following, we first briefly overview the main concepts of supervised SVMs; then, we discuss the considered assumptions, and finally, we present the formulation of the DASVM algorithm.

A. SVM: Background

The success of SVMs [16], [17] is mainly related to their desirable properties that can be summarized as follows: 1) high classification accuracies and very good generalization capabilities with respect to other traditional classifiers; 2) existence and uniqueness of an optimal solution; 3) possibility of representing the optimization problem in a dual formulation, which makes

SVMs scalable to large data sets and allows one to express the solution in terms of only a subset of the training samples; 4) capability to address classification problems in which no explicit parametric models on the distribution of information classes are assumed; and 5) possibility of defining nonlinear decision boundaries by implicitly mapping the available observations into a higher dimensional space.

Let $\mathcal{T} = \{\mathcal{X}, \mathcal{Y}\} = \{(\mathbf{x}_l, y_l)\}_{l=1}^{N}$ represent the available training set, where $\mathcal{X} = \{\mathbf{x}_l\}_{l=1}^{N}$ is a subset of N patterns drawn from the remote-sensing image being classified and $\mathcal{Y} = \{y_l\}_{l=1}^{N}$, $y_l = \pm 1$, is the set of associated true labels. SVMs aim at linearly separating data by the hyperplane $h : f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \mathbf{0}$, where $\mathbf{x}$ represents a generic sample, $\mathbf{w}$ is a vector normal to h, and b is a constant such that $b/\|\mathbf{w}\|^2$ represents the distance of h from the origin. The distance between the two hyperplanes $h_1 : \mathbf{w} \cdot \mathbf{x} + b = -1$ and $h_2 : \mathbf{w} \cdot \mathbf{x} + b = +1$ parallel to h is called *margin*. Note that the larger is the margin, the higher is expected to be the generalization capability of the classifier. Accordingly, since maximizing the margin is equivalent to minimizing the norm of $\mathbf{w}$, the objective of SVMs is to solve the following minimization problem:

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum\limits_{l=1}^{N} \xi_l \right\} \\ y_l(\mathbf{w} \cdot \mathbf{x}_l + b) \geq 1 - \xi_l \qquad \forall l = 1, \ldots, N \\ \xi_l > 0 \end{cases} \tag{1}$$

where ξ_n's are slack variables allowing for (permitted) errors and C is the associated *penalization parameter* (also called *regularization parameter*), which permits to tune the generalization capability. Since direct handling of inequality constraints is difficult, it is possible to exploit Lagrange theory which permits to obtain the corresponding equivalent dual representation

$$\begin{cases} \max\limits_{\alpha} \left\{ \sum\limits_{l=1}^{N} \alpha_l - \frac{1}{2} \sum\limits_{l=1}^{N} \sum\limits_{m=1}^{N} y_l y_m \alpha_l \alpha_m \mathbf{x}_l \cdot \mathbf{x}_m \right\} \\ \sum\limits_{l=1}^{N} y_l \alpha_l = 0 \\ 0 \leq \alpha_l \leq C \qquad \forall l = 1, \ldots, N \end{cases} \tag{2}$$

where the coefficients $\alpha_{l=1}^{N}$ are referred to as Lagrange multipliers. According to the Karush–Kuhn–Tucker conditions [30], it is possible to demonstrate that the solution is a linear combination of the only training patterns associated with nonzero Lagrange multipliers (i.e., either mislabeled training samples or correctly labeled training samples falling into the margin band $\mathcal{M} = \{\mathbf{x} | -1 \leq f(\mathbf{x}) \leq 1\}$), denoted as *support vectors*.

When the available data cannot be linearly separated directly in the input space, they can be projected into a higher dimensional feature space (e.g., a Hilbert space) with a nonlinear mapping function $\Phi(\cdot)$ defined in accordance with Cover's theorem [32]. As a consequence, the inner product between the two patterns $\mathbf{x}_l$ and $\mathbf{x}_m$ in (2) becomes $\Phi(\mathbf{x}_l) \cdot \Phi(\mathbf{x}_m)$. In order to avoid considering the $\Phi(\cdot)$ mapping explicitly, according to the Mercer's theorem, it is possible to exploit a *kernel function* $K(\mathbf{x}_l, \mathbf{x}_m) = \Phi(\mathbf{x}_l) \cdot \Phi(\mathbf{x}_m)$, which ensures that the objective function is convex. After obtaining the optimal values of the multipliers $\alpha_{l=1}^{N}$ (e.g., carrying out optimization in (2) with some quadratic optimization techniques), for any given sample $\mathbf{x}$, the predicted label becomes

$$\hat{y} = \operatorname{sgn}\left[f(\mathbf{x})\right] = \operatorname{sgn}\left[\sum_{l=1}^{N} y_l \alpha_l K(\mathbf{x}_l, \mathbf{x}) + b\right]. \tag{3}$$

Note that, for addressing multiclass problems, different strategies have been proposed so far in the literature (the reader is referred to [30] for greater details).

B. Proposed DASVM: Assumptions

Let $\mathcal{I}_1$ and $\mathcal{I}_2$ represent two remote-sensing images acquired over the same area at different times (t_1 and t_2, respectively). Let $\mathcal{X}_1 = \{\mathbf{x}_i^1 | \mathbf{x}_i^1 \in \mathcal{I}_1\}_{i=1}^{N}$ and $\mathcal{X}_2 = \{\mathbf{x}_i^2 | \mathbf{x}_i^2 \in \mathcal{I}_2\}_{i=1}^{M}$ denote two subsets of $\mathcal{I}_1$ and $\mathcal{I}_2$ composed of N and M patterns, respectively. Note that $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ represent the d-dimensional feature vectors associated with the ith sample of $\mathcal{I}_1$ and $\mathcal{I}_2$ (where d represents the dimensionality of the input space). In the formulation of the proposed DASVM technique, we make the following assumptions.

1) The same set of L information classes, $\Omega = \{\omega_i\}_{i=1}^{L}$, characterizes the two images $\mathcal{I}_1$ and $\mathcal{I}_2$.
2) A set of ground-truth labels $\mathcal{Y}_1 = \{y_i^1 | y_i^1 \in \Omega\}_{i=1}^{N}$ for $\mathcal{X}_1$ is available; thus, it is possible to define a training set $\mathcal{T}_1 = \{\mathcal{X}_1, \mathcal{Y}_1\} = \{(\mathbf{x}_l^1, y_l^1)\}_{l=1}^{N}$ for $\mathcal{I}_1$.
3) A set of ground-truth labels $\mathcal{Y}_2 = \{y_i^2 | y_i^2 \in \Omega\}_{i=1}^{M}$ for $\mathcal{X}_2$ is not known; thus, it is not possible to define a training set for $\mathcal{I}_2$.

Under the aforementioned hypothesis, our goal is to perform an effective domain adaptation from $\mathcal{I}_1$ to $\mathcal{I}_2$ and, thus, to obtain an accurate and robust classification of $\mathcal{I}_2$ by exploiting labeled training samples $\mathcal{T}_1$ from the reference image $\mathcal{I}_1$ and unlabeled samples $\mathcal{X}_2$ from the new image $\mathcal{I}_2$. Note that obtaining a good adaptation requires an adequate modeling of the relationship between $\mathcal{I}_1$ and $\mathcal{I}_2$. In this framework, there exist two bound conditions defined on the basis of the correlation between the distributions $P^1(\mathbf{x}, y)$ and $P^2(\mathbf{x}, y)$ that govern $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively: 1) If $P^1(\mathbf{x}, y)$ and $P^2(\mathbf{x}, y)$ are independent, then the available training set $\mathcal{T}_1$ for $\mathcal{I}_1$ is useless for building a model for $\mathcal{I}_2$, and 2) if $P^1(\mathbf{x}, y) \equiv P^2(\mathbf{x}, y)$, then adaptation is not necessary and standard supervised learning algorithms can be employed. Nevertheless, in real applications, the aforementioned distributions are generally neither independent nor identical. In these situations, it is reasonable to assume the existence of an intrinsic relationship between the two images that makes adaptation possible. We expect that the probability to succeed in the adaptation process is associated with the complexity of the problem, which depends on the similarity between $P^1(\mathbf{x}, y)$ and $P^2(\mathbf{x}, y)$.

C. Proposed DASVM: Formulation

In the following, for simplicity, we describe the proposed DASVM technique in the case of a two-class problem. DASVMs directly take into account that unlabeled samples $\mathcal{X}_2$

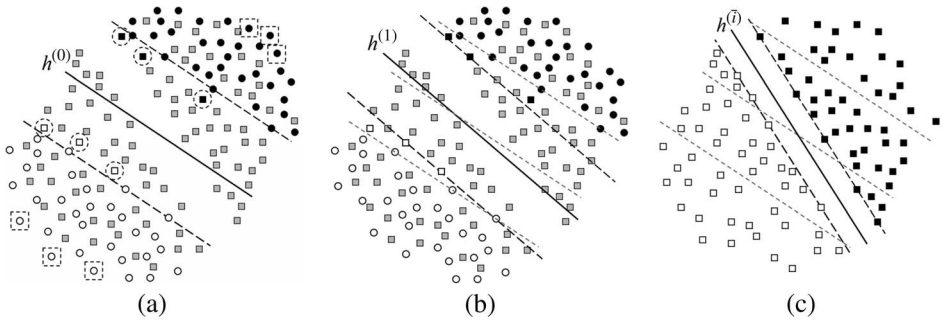


Fig. 1. (Solid line) Separation hyperplane and (dashed lines) margin bounds at different stages of the DASVM algorithm for a toy data set. Original patterns at time t_1 are shown as white and black circles. Semilabeled patterns are shown as white and black squares. Remaining unlabeled patterns at time t_2 are represented as gray squares. Feature space structure obtained: (a) At the first iteration (the dashed circles and squares highlight the current semilabeled patterns and the current original training patterns to delete, respectively; in the example, $\rho = 3$), (b) at the second iteration, and (c) at the last iteration, respectively, in an ideal situation (the dashed gray lines represent both the separation hyperplane and the margin bounds at the beginning of the learning process).

are drawn from a new image $\mathcal{I}_2$ different from the reference image $\mathcal{I}_1$ of training samples. Accordingly, we assume that labeled samples of $\mathcal{I}_1$ should be considered only for initially constraining the learning problem for $\mathcal{I}_2$ (i.e., they are the only labeled patterns at the beginning of the domain-adaptation learning process). It seems reasonable that, at the end of the training process, the final decision function should be defined on the basis of samples at t_2 alone (i.e., the semilabeled samples iteratively selected in the learning process), as they are the only ones referring to the target domain. In the light of this reasoning, in the proposed DASVM technique, original training samples at time t_1 are gradually erased in order to obtain a final discriminant function ruled only by semilabeled samples at time t_2.

The proposed DASVM algorithm is made up of three main phases: 1) initialization (only $\mathcal{T}_1$ is used for initializing the discriminant function); 2) iterative domain adaptation ($\mathcal{T}_1$ and $\mathcal{X}_2$ are jointly used for gradually adapting the discriminant function to $\mathcal{I}_2$); and 3) convergence (only samples of $\mathcal{X}_2$ are used for defining the final discriminant function). In the following, we will denote $\mathcal{T}^{(i)}$ and $\mathcal{X}_2^{(i)}$ as the training set and the unlabeled set (i.e., the set containing the unlabeled samples that have not been inserted into the training set $\mathcal{T}^{(i)}$) at the current iteration i, respectively. These phases are described into details in the following.

1) Phase 1—Initialization: At the beginning of the learning process, an initial separation hyperplane is determined on the basis of training data available for $\mathcal{I}_1$. We have that $\mathcal{T}^{(0)} \equiv \mathcal{T}_1 = \{(\mathbf{x}_l^1, y_l^1)\}_{l=1}^N$ and $\mathcal{X}_2^{(0)} = \{\mathbf{x}_u^2\}_{u=1}^M$. As for standard supervised SVMs, the bound cost function to minimize is the following:

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} \left\{ \frac{1}{2} \left\| \mathbf{w}^{(0)} \right\|^2 + C \sum\limits_{l=1}^{N} \xi_l^1 \right\} \\ y_l^1 \left(\mathbf{w}^{(0)} \cdot \mathbf{x}_l^1 + b^{(0)} \right) \geq 1 - \xi_l^1 \qquad \forall l = 1, \ldots, N \\ \xi_l^1 \geq 0. \end{cases} \tag{4}$$

2) Phase 2—Iterative Domain Adaptation: The second iterative phase represents the core of the proposed algorithm. At the generic iteration i, all the original unlabeled samples of $\mathcal{X}_2$ are associated with an estimated label $\hat{y}_u^{2(i)} = \operatorname{sgn}[f^{(i)}(\mathbf{x}_u^2)]$, determined according to the current decision function $f^{(i)}(\mathbf{x}_u^2) = \mathbf{w}^{(i)} \cdot \mathbf{x}_u^2 + b^{(i)}$.

In order to seize the classification problem at time t_2, a subset of the (remaining) unlabeled samples $\mathcal{X}_2^{(i)}$ is iteratively selected and moved, together with the corresponding estimated labels, into the training set $\mathcal{T}^{(i+1)}$. On the one hand, the higher the distance from the separation hyperplane $h^{(i)} : \mathbf{w}^{(i)} \cdot \mathbf{x} + b^{(i)} = 0$, the higher the chance for an unlabeled sample to be correctly classified. On the other hand, the current unlabeled samples falling into the margin band $\mathcal{M}^{(i)} = \{\mathbf{x} | -1 \leq f^{(i)}(\mathbf{x}) \leq 1\}$ are those with the highest probability to be associated with nonzero Lagrange multipliers (and, thus, to affect the position of $h^{(i+1)}$) once inserted in the training set with their current estimated label (patterns falling outside the margin band, in fact, are more likely to be associated with null multipliers). In light of these two observations, at each iteration, we progressively take into account unlabeled samples falling into $\mathcal{M}^{(i)}$ closest to the margin bounds (i.e., they are the ones further from the decision boundary $h^{(i)}$). Let us define the two following subsets:

$$\mathcal{H}_{\text{up}}^{(i)} = \left\{ \left(\mathbf{x}_u^2, \hat{y}_u^{2(i)} = +1 \right) | \mathbf{x}_u^2 \in \mathcal{X}_2^{(i)}, \right. \\ \left. 1 \geq f^{(i)}\left(\mathbf{x}_u^2\right) \geq f^{(i)}\left(\mathbf{x}_{u+1}^2\right) \geq 0 \right\} \tag{5}$$

$$\mathcal{H}_{\text{low}}^{(i)} = \left\{ \left(\mathbf{x}_u^2, \hat{y}_u^{2(i)} = -1 \right) | \mathbf{x}_u^2 \in \mathcal{X}_2^{(i)}, \right. \\ \left. -1 \leq f^{(i)}\left(\mathbf{x}_u^2\right) \leq f^{(i)}\left(\mathbf{x}_{u+1}^2\right) < 0 \right\}. \tag{6}$$

The current unlabeled samples in the upper and lower sides of the margin band $\mathcal{M}^{(i)}$ are inserted with their corresponding estimated labels into $\mathcal{H}_{\text{up}}^{(i)}$ (i.e., $\hat{y}_u^{2(i)} = +1$) and $\mathcal{H}_{\text{low}}^{(i)}$ (i.e., $\hat{y}_u^{2(i)} = -1$), respectively. In particular, samples of $\mathcal{H}_{\text{up}}^{(i)}$ and $\mathcal{H}_{\text{low}}^{(i)}$ are sorted in ascending order with respect to their distance from the upper and lower bounds of $\mathcal{M}^{(i)}$, respectively. At each iteration, the first ρ unlabeled patterns in both the two aforementioned subsets (where ρ is a strictly positive free parameter defined *a priori* by the user) are selected and moved to the training set $\mathcal{T}^{(i)}$ [see Fig. 1(a)]. Note that a similar strategy is proposed in [33], where only the two unlabeled patterns inside the margin band with the maximum and the minimum values of the decision function are iteratively considered; nevertheless, we assume that, in general, two patterns may not be sufficiently representative for properly tuning the position of the separation

hyperplane. As the cardinality of $\mathcal{H}_{\text{up}}^{(i)}$ and $\mathcal{H}_{\text{low}}^{(i)}$ may become lower than ρ, the set of unlabeled patterns selected at the generic iteration i becomes

$$\mathcal{H}^{(i)} = \left\{ \left(\mathbf{x}_u^2, \hat{y}_u^{2(i)} \right) \in \mathcal{H}_{\text{up}}^{(i)} \middle| \, 1 \le u \le \lambda^{(i)} \right\} \cup \left\{ \left(\mathbf{x}_u^2, \hat{y}_u^{2(i)} \right) \in \mathcal{H}_{\text{low}}^{(i)} \middle| \, 1 \le u \le \delta^{(i)} \right\} \quad (7)$$

where $\lambda^{(i)} = \min(\rho, |\mathcal{H}_{\text{up}}^{(i)}|)$ and $\delta^{(i)} = \min(\rho, |\mathcal{H}_{\text{low}}^{(i)}|)$. Patterns of $\mathcal{H}^{(i)}$ are then merged with $\mathcal{T}^{(i)}$.

Due to the fact that the training set iteratively changes, it is reasonable to expect that also the position of the separation hyperplane $h^{(i)}$ changes at each iteration [see Fig. 1(b)]. Accordingly, a dynamical adjustment is necessary for taking into account that a given semilabeled sample could be associated with different estimated labels between two successive iterations. Let

$$\mathcal{S}^{(i)} = \left\{ \left(\mathbf{x}_u^2, \hat{y}_u^{2(i-1)} \right) \in \mathcal{T}^{(i)} \middle| \, \hat{y}_u^{2(i)} \neq \hat{y}_u^{2(i-1)} \right\} \quad (8)$$

represent the set of semilabeled samples belonging to $\mathcal{T}^{(i)}$ whose labels at iteration i are different than those at iteration $i-1$. If the label of a semilabeled pattern at iteration i is different from the one at iteration $i-1$, it means that the system is no more confident on that sample. Accordingly, such a label is erased, and the "inconsistent" semilabeled pattern is reset to the unlabeled state and moved to $\mathcal{X}_2^{(i+1)}$. In this way, it is possible to reconsider this pattern at the following iterations of the learning procedure.

As it will be pointed out in the following, the proposed DASVM algorithm aims at gradually increasing the regularization parameter for the patterns that belong to the new image $\mathcal{I}_2$ according to a time-dependent criterion. Accordingly, the set $\mathcal{J}^{(i)}$ containing all the semilabeled patterns at the ith iteration is partitioned into a finite number of subsets $\gamma \in \mathbb{N}_0$, where γ is defined as the maximum number of iterations for which the user allows the regularization parameter for semilabeled samples to increase. In particular, we have that

$$\mathcal{J}^{(i)} = \mathcal{J}_1^{(i)} \cup \mathcal{J}_2^{(i)} \cup \cdots \cup \mathcal{J}_\gamma^{(i)}$$
$$\begin{cases} \mathcal{J}_1^{(i)} = \mathcal{H}^{(i)} \\ \mathcal{J}_k^{(i)} = \mathcal{J}_{k-1}^{(i-1)} - \mathcal{S}^{(i)} \qquad \forall k = 2, \ldots, \gamma - 1 \\ \mathcal{J}_\gamma^{(i)} = \left(\mathcal{J}_\gamma^{(i-1)} \cup \mathcal{J}_{\gamma-1}^{(i-1)} \right) - \mathcal{S}^{(i)} \end{cases} \quad (9)$$

where the generic kth subset includes all the semilabeled samples that have been labeled in the same way for k successive iterations. Each subset of $\mathcal{J}^{(i)}$ will be associated with a specific regularization parameter.

The main purpose of the proposed technique is to define and solve a bound minimization problem with respect only to the samples at time t_2; therefore, a strategy for gradually deleting the original labeled patterns of the reference image at time t_1 has been developed. Intuitively, the higher is the distance of a training sample from the separation hyperplane, the lower is the chance that it can be misclassified after a new tuning of the hyperplane due to the insertion of semilabeled samples into the training set $\mathcal{T}^{(i+1)}$. For this reason, in the proposed DASVM technique, we iteratively delete remaining original training samples of the reference image $\mathcal{I}_1$ that are furthest from the current decision boundary $h^{(i)}$. Let us define the two following subsets:

$$\mathcal{Q}_{\text{up}}^{(i)} = \left\{ \left(\mathbf{x}_l^1, y_l^1 \right) \in \mathcal{T}^{(i)} \middle| \, f^{(i)}\left(\mathbf{x}_l^1 \right) \ge f^{(i)}\left(\mathbf{x}_{l+1}^1 \right) \ge 0 \right\} \quad (10)$$

$$\mathcal{Q}_{\text{low}}^{(i)} = \left\{ \left(\mathbf{x}_l^1, y_l^1 \right) \in \mathcal{T}^{(i)} \middle| \, f^{(i)}\left(\mathbf{x}_l^1 \right) \le f^{(i)}\left(\mathbf{x}_{l+1}^1 \right) < 0 \right\} \quad (11)$$

where $\mathcal{Q}_{\text{up}}^{(i)}$ and $\mathcal{Q}_{\text{low}}^{(i)}$ contain the remaining labeled patterns for $\mathcal{I}_1$ which lie above and under the separation hyperplane, respectively, sorted in descending order with respect to their distance from $h^{(i)}$. At each iteration, for balancing the contribution of the new semilabeled samples, the number of patterns to erase from $\mathcal{Q}_{\text{up}}^{(i)}$ and $\mathcal{Q}_{\text{low}}^{(i)}$ is set equal to $\lambda^{(i)}$ and $\delta^{(i)}$ (i.e., the number of semilabeled patterns selected from the upper and the lower side of the margin band), respectively. If none of the remaining unlabeled samples at time t_2 falls into either the upper or the lower side of the margin band (i.e., $\mathcal{H}^{(i)} = \varnothing$), the number of patterns to delete is set to ρ. As a consequence, we have

$$\mathcal{Q}^{(i)} = \left\{ \left(\mathbf{x}_l^1, y_l^1 \right) \in \mathcal{Q}_{\text{up}}^{(i)} \middle| \, 1 \le l \le \nu^{(i)} \right\} \cup \left\{ \left(\mathbf{x}_l^1, y_l^1 \right) \in \mathcal{Q}_{\text{low}}^{(i)} \middle| \, 1 \le l \le \kappa^{(i)} \right\} \quad (12)$$

where

$$\nu^{(i)} = \begin{cases} \min\left(\lambda^{(i)}, \left| \mathcal{Q}_{\text{up}}^{(i)} \right| \right) & \text{if } \mathcal{H}^{(i)} \neq \emptyset \\ \min\left(\rho, \left| \mathcal{Q}_{\text{up}}^{(i)} \right| \right) & \text{if } \mathcal{H}^{(i)} = \emptyset \end{cases}$$

$$\kappa^{(i)} = \begin{cases} \min\left(\delta^{(i)}, \left| \mathcal{Q}_{\text{low}}^{(i)} \right| \right) & \text{if } \mathcal{H}^{(i)} \neq \emptyset \\ \min\left(\rho, \left| \mathcal{Q}_{\text{low}}^{(i)} \right| \right) & \text{if } \mathcal{H}^{(i)} = \emptyset. \end{cases}$$

Let $\mu^{(i)}$ and $\eta^{(i)}$ represent the number of remaining original training samples at t_1 and semilabeled samples at t_2 in $\mathcal{T}^{(i)}$, respectively. For $i \ge 1$, the minimization problem can be written as

$$\begin{cases} \min\limits_{\mathbf{w}, b, \xi^1, \xi^2} \left\{ \frac{1}{2} \left\| \mathbf{w}^{(i)} \right\|^2 + C^{(i)} \sum\limits_{l=1}^{\mu^{(i)}} \xi_l^1 + \sum\limits_{u=1}^{\eta^{(i)}} C_u^* \xi_u^2 \right\} \\ y_l^1 \cdot \left(\mathbf{w}^{(i)} \cdot \mathbf{x}_l^1 + b^{(i)} \right) \ge 1 - \xi_l^1 \qquad \forall l = 1, \ldots, \mu^{(i)} \\ \hat{y}_u^{2(i-1)} \cdot \left(\mathbf{w}^{(i)} \cdot \mathbf{x}_u^2 + b^{(i)} \right) \ge 1 - \xi_u^2 \qquad \forall u = 1, \ldots, \eta^{(i)} \\ \xi_l^1, \xi_u^2 \ge 0. \end{cases} \quad (13)$$

It is possible to notice that, with respect to supervised SVMs, the training patterns of the image at time t_1 are associated with a regularization parameter $C^{(i)}$ which varies at each iteration. Moreover, as pointed out before, semilabeled samples are associated with a specific regularization parameter $C_u^* = C_u^*(k) \in \mathbb{R}^+$ that depends on the kth subset $\mathcal{J}_k^{(i-1)}$ they belong to at iteration $i-1$. The purpose of $C^{(i)}$ and C_u^* is to control the number of misclassified samples of the current training set $\mathcal{T}^{(i)}$ associated with $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. In particular, the larger their values, the higher is the influence of the associated samples in tuning the position of the separation

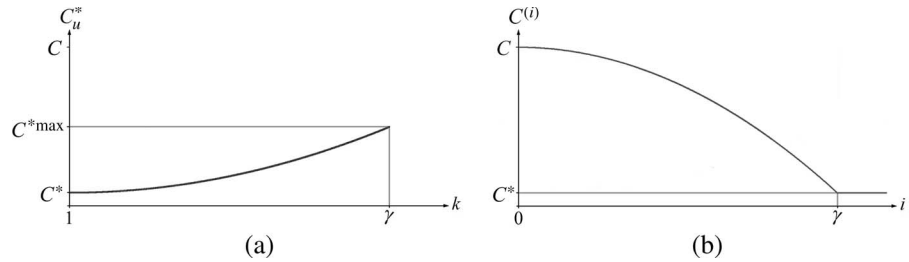


Fig. 2. (a) Behavior of C_u^*, regularization parameter for the semilabeled patterns belonging to $\mathcal{J}^{(i)} = \mathcal{J}_1^{(i)} \cup \mathcal{J}_2^{(i)} \cup \cdots \cup \mathcal{J}_\gamma^{(i)}$, versus k (index corresponding to the subset $\mathcal{J}_k^{(i)}$, which is related to the number of iterations in which a semilabeled pattern is associated with the same label). (b) Behavior of the regularization parameter for the original training patterns at time t_1, $C^{(i)}$, versus the number of iterations i.

hyperplane. It is reasonable to expect that the two distributions $P^1(\mathbf{x}, y)$ and $P^2(\mathbf{x}, y)$ could be rather different; thus, in order to avoid instabilities in the learning process, unlabeled samples are considered gradually. In particular, a weighting strategy based on a temporal criterion is adopted. As regards the regularization parameter for the semilabeled patterns, the proposed algorithm let it grow in a quadratic way, depending on the number of successive iterations they have been assigned the same estimated label [see Fig. 2(a)]. For each semilabeled sample $(\mathbf{x}_u^2, \hat{y}_u^{2(i)})$, we have

$$C_u^* = \frac{\tau \cdot C - C^*}{(\gamma - 1)^2}(k-1)^2 + C^* \Leftrightarrow \left(\mathbf{x}_u^2, \hat{y}_u^{2(i)}\right) \in \mathcal{J}_k^{(i)}, \quad k = 1, \ldots, \gamma \tag{14}$$

where C^* is a free parameter that represents the initial regularization value for semilabeled samples and τ, $0 < \tau \leq 1$, tunes the maximum cost value of semilabeled samples (i.e., a reasonable choice has proved to be $\tau = 0.5$). For balancing the growing importance of semilabeled samples in affecting the position of the current separation hyperplane, the algorithm makes the cost factor $C^{(i)}$ for the original labeled samples at time t_1 decrease in a quadratic way too [see Fig. 2(b)]

$$C^{(i)} = \max\left(\frac{C^* - C}{\gamma^2} i^2 + C, C^*\right). \tag{15}$$

In this way, while, at the beginning, the position of the separation hyperplane strongly depends on labeled patterns available for $\mathcal{I}_1$, then their influence gets always lower as the number of iteration increases (until $i = \gamma$).

It is worth noting that the strategy adopted for deleting patterns of $\mathcal{I}_1$ might result also in the removal of mislabeled original training samples, which affect the position of the hyperplane as they are associated with nonzero Lagrange multipliers. Nevertheless, this aspect does not seem to be critical as the following are observed.

1) If mislabeled original training samples fall far away from the hyperplane (i.e., $\xi \gg 1$), it is reasonable to assume that they are outliers (e.g., they could have been associated with a wrong label while defining the training set at time t_1). In practice, we expect that only few original training samples may exhibit such a behavior. Accordingly, even if they are still associated with a nonnegligible value of the regularization parameter, their deletion would not significantly affect the position of the hyperplane.
2) If mislabeled original training samples fall into the wrong side of the margin band, or close to the wrong margin bound, they are expected to be deleted when the system is close to convergence or, in general, when the number of current semilabeled samples is high. In such a scenario, by taking into account the adopted weighting criterion, it is reasonable to assume that the position of the hyperplane strongly depends on semilabeled samples. Thus, the deletion of these original training samples is not critical (note that, as this happens after several iterations of the learning process, training patterns would be associated with a small value of the regularization parameter).

The second phase ends when the convergence criteria described hereinafter are satisfied.

3) Phase 3—Convergence: As the proposed DASVMs aim at defining a discriminant function on the basis of samples at time t_2 alone, it is reasonable to assume that convergence can be obtained only if all the original labeled samples at t_1 have been completely erased, i.e., $\mathcal{Q}^{(i)} = \varnothing$. Under this assumption, from a theoretical viewpoint, it can be assumed that convergence is reached if none of the remaining unlabeled samples of $\mathcal{X}_2$ lies into the margin band $\mathcal{M}^{(i)}$, i.e., $\mathcal{H}^{(i)} = \varnothing$ [see Fig. 1(c)]. Nevertheless, such a choice might result in a high computational load. Moreover, it should be considered also that even when the margin band is empty, the number of inconsistent semilabeled patterns cannot be negligible. For these reasons, the following empirical stopping criteria have been defined:

$$\begin{cases} \mathcal{Q}^{(i)} = \varnothing \\ \left|\mathcal{H}^{(i)}\right| \leq \lceil \beta \cdot M \rceil \\ \left|\mathcal{S}^{(i)}\right| \leq \lceil \beta \cdot M \rceil \end{cases} \tag{16}$$

where M is the number of original unlabeled samples taken from $\mathcal{I}_2$ and β is a user-defined parameter that permits to tune the sensitivity of the learning process. This means that convergence is reached if the number of both mislabeled and remaining unlabeled patterns lying into $\mathcal{M}^{(i)}$ at the current iteration is lower than or equal to $\lceil \beta \cdot M \rceil$. Accordingly, the final bound minimization problem at the last iteration $\bar{i}$ is defined only on the basis of the semilabeled samples

$$\begin{cases} \min\limits_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \sum\limits_{u=1}^{\eta^{(\bar{i})}} C_u^* \xi_u \right\} \\ \hat{y}_u^{2(\bar{i}-1)} \cdot \left(\mathbf{w} \cdot \mathbf{x}_u^2 + b\right) \geq 1 - \xi_u \qquad \forall u = 1, \ldots, \eta^{(\bar{i})} \\ \xi_u \geq 0. \end{cases} \tag{17}$$

At the end of the learning process, all the patterns $\mathbf{x}_u^2 \in \mathcal{I}_2$ are labeled according to the resulting separation hyperplane, i.e., $\hat{y}_u^2 = \text{sgn}[\mathbf{w} \cdot \mathbf{x}_u^2 + b]$.

At each iteration of the proposed DASVM technique, the objective function to minimize is convex. This means that, independently from the number of considered semilabeled patterns, a unique solution for the considered minimization problem exists; therefore, the system is always able to reach convergence. However, note that it is not possible to guarantee the convergence toward a final solution that is satisfactory for the investigated task. In fact, this depends on the definition of the unlabeled samples considered and, implicitly, on the "similarity" between source and target domains.

The earlier described algorithm is defined for two-class problems. When a multiclass problem has to be investigated, a one-against-all (OAA) strategy [30], [34] can be employed.

IV. Proposed Circular Validation Strategy

One of the main limitations in developing methods for addressing updating of land-cover maps when no ground-truth information is available for the new image being classified is the lack of validation strategies that permit to assess the effectiveness of the classification results. To this end, in this section, we present a novel circular strategy for validating the solutions obtained with classifiers designed for handling domain-adaptation problems. The joint use of this accuracy assessment strategy and of the DASVM technique described in Section III permits to address operational land-cover map updating problems. It is worth noting that the proposed circular validation strategy is general and can be used with any domain-adaptation classifier.

A. Proposed Circular Validation Strategy: Rationale

The proposed strategy is based on the two following observations.

1) *Observation 1:* The only ground-truth information is that available for the reference image $\mathcal{I}_1$. Therefore, for assessing the accuracy of the classification of the new image $\mathcal{I}_2$ at time t_2, an indirect procedure that exploits the training set $\mathcal{T}_1$ at time t_1 must be defined.
2) *Observation 2:* Given the reference image $\mathcal{I}_1$ and the new image $\mathcal{I}_2$ acquired over the same geographical area at times t_1 and t_2, respectively, it is reasonable to expect that, in real applications, the related distributions $P^1(\mathbf{x}, y)$ and $P^2(\mathbf{x}, y)$ are neither independent nor identical. In this hypothesis, there exists a direct relationship between $\mathcal{I}_1$ and $\mathcal{I}_2$, which depends on the similarity between the distributions that govern the two images.

On the basis of these observations, the proposed strategy relies on the following rationale. Let us consider that, starting from an acceptable accuracy for $\mathcal{I}_1$, the considered algorithm results in a satisfactory solution for $\mathcal{I}_2$. In such a situation, the domain-adaptation process is able to seize the relationship between $\mathcal{I}_1$ and $\mathcal{I}_2$, and thus the structure of the problem at time t_2; therefore, we expect that by applying again the same learning algorithm in the reverse sense (using the classification labels in place of the missing ground-truth labels for $\mathcal{I}_2$, keeping the same learning parameters, and considering $\mathcal{I}_1$ as the new image to be categorized), it is possible to obtain again a good discrimination capability at time t_1. On the contrary, if the domain-adaptation classifier does not identify an acceptable solution for $\mathcal{I}_2$, this means that it does not capture the relationship between the two images but converges to a solution which is not related to the investigated problem. In this condition, at the end of the backward process from time t_2 to time t_1, it seems impossible to recover a reliable solution for $\mathcal{I}_1$, as, exploiting an unreliable training set for $\mathcal{I}_2$, the domain-adaptation classifier has almost no probability to correctly reseize the classification problem for the reference image. On the basis of these expected properties, we can use the accuracy evaluated on the original training samples $\mathcal{T}_1$ for validating the solution obtained for the unlabeled instances of the new image after a circular (forward and backward) application of the considered domain-adaptation algorithm.

In the following, we formally define the proposed circular strategy in the specific case of DASVM classifiers.

B. Proposed Circular Validation Strategy: Formulation

Let us define $\Lambda(\mathcal{Y}_j, \hat{\mathcal{Y}}_j)$ as a classification accuracy measure (e.g., the overall accuracy, the kappa coefficient of accuracy) that evaluates the similarity between a set of estimated labels $\hat{\mathcal{Y}}_j$ (i.e., a solution) predicted by a generic classifier and the corresponding set of true labels $\mathcal{Y}_j$. If $\Lambda(\mathcal{Y}_j, \hat{\mathcal{Y}}_j) \geq \Lambda_{\text{th}}$, where Λ_{th} represents a threshold for Λ, we assume that the solution $\hat{\mathcal{Y}}_j$ is consistent with $\mathcal{Y}_j$ (i.e., it is acceptable for the problem under investigation). Accordingly, let us define the following four sets.

1) $\mathcal{A}$: It contains all the solutions consistent with the reference image at time t_1 (i.e., $\Lambda(\mathcal{Y}_1, \hat{\mathcal{Y}}_1) \geq \Lambda_{\text{th}}$).
2) $\mathcal{B}$: It contains all the solutions consistent with the new image at time t_2 (i.e., $\Lambda(\mathcal{Y}_2, \hat{\mathcal{Y}}_2) \geq \Lambda_{\text{th}}$).
3) $\mathcal{C}$: It contains all the solutions nonconsistent with the reference image at time t_1 (i.e., $\Lambda(\mathcal{Y}_1, \hat{\mathcal{Y}}_1) < \Lambda_{\text{th}}$).
4) $\mathcal{D}$: It contains all the solutions nonconsistent with the new image at time t_2 (i.e., $\Lambda(\mathcal{Y}_2, \hat{\mathcal{Y}}_2) < \Lambda_{\text{th}}$).

Let us train a supervised SVM using the training samples available for $\mathcal{I}_1$, and select a solution consistent at time t_1 that belongs to the set $\mathcal{A}$ (i.e., the system is in the state $\overline{A}$). Successively, we train a DASVM using $\mathcal{T}_1$ as training set and $\mathcal{X}_2$ as unlabeled set. For both the kernel function variables and the regularization parameter C, we keep the same values adopted for the aforementioned supervised learning. With a proper choice of the values for the domain-adaptation parameters (i.e., C^*, γ, and ρ), the solution at time t_2 is expected to belong to $\mathcal{B}$ (i.e., the system moves to the state $\overline{B}$); on the contrary, if the choice of the parameters is not adequate (i.e., $P^1(\mathbf{x}, y)$ is too different from $P^2(\mathbf{x}, y)$), the solution belongs to $\mathcal{D}$ (i.e., the system moves to the state $\overline{D}$).

Let us now consider the solution obtained for $\mathcal{X}_2$. We can address the reverse domain-adaptation problem (from time t_2

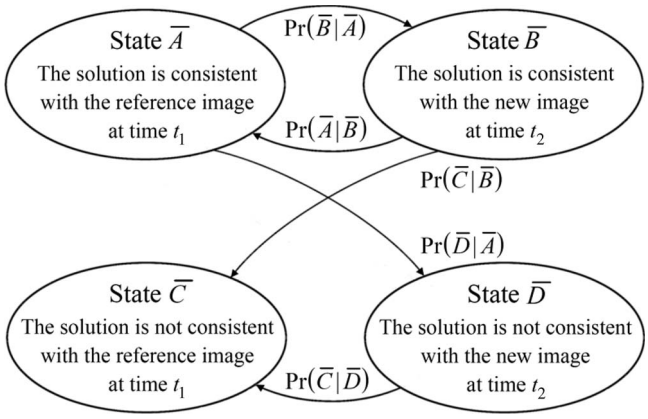


Fig. 3. Diagram of all the possible state transitions of solutions exploited from the proposed circular validation strategy.

to time t_1) with the same DASVM classifier by keeping the same learning parameters and jointly exploiting the estimated training set $\hat{\mathcal{T}}_2 = \{\mathcal{X}_2, \hat{\mathcal{Y}}_2\}$ (i.e., $\hat{\mathcal{Y}}_2$ represents the set of estimated labels for $\mathcal{X}_2$) and the subset of unlabeled patterns $\mathcal{X}_1$ at time t_1 (considered without the corresponding ground-truth labels $\mathcal{Y}_1$). As the ground-truth labels $\mathcal{Y}_1$ are known, we can compute the value for Λ associated with the results obtained after the circular learning process. If $\Lambda(\mathcal{Y}_1, \hat{\mathcal{Y}}_1) < \Lambda_{\text{th}}$, the classification accuracy at time t_1 is considered nonacceptable, then the solution belongs to $\mathcal{C}$ (i.e., the system moves to the state $\overline{C}$). On the contrary, if $\Lambda(\mathcal{Y}_1, \hat{\mathcal{Y}}_1) \geq \Lambda_{\text{th}}$, the resulting solution is consistent (i.e., it belongs to $\mathcal{A}$) and the system moves back to the state $\overline{A}$. Fig. 3 shows all the possible transitions of the system.

Our assumption is that, when the system starting from the state $\overline{A}$ is able to return again into the state $\overline{A}$, the classification accuracy at time t_2 is satisfactory. Accordingly, the land-cover map obtained by the corresponding DASVM for $\mathcal{I}_2$ represents a reasonable updating of the land-cover map for the reference image $\mathcal{I}_1$. This aspect is crucial because it means that, in such situations, we are able to assess that the estimated map at time t_2 has an acceptable accuracy even if no prior ground-truth information is available. Let $\Pr(\overline{Y}|\overline{X})$ denote the probability that the system joins the generic state $\overline{Y}$ starting from the generic state $\overline{X}$ by applying the DASVM algorithm.

The two main hypotheses under which the proposed validation technique is assumed to be effective are the following.

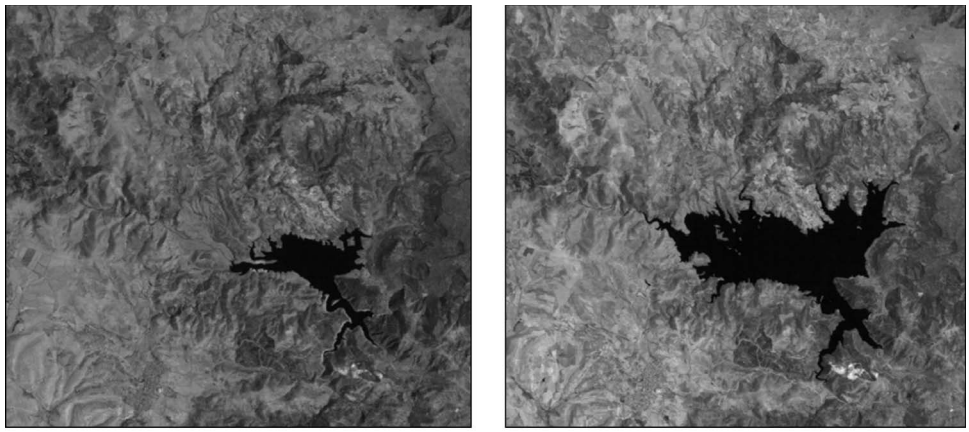
1) *Starting from the state $\overline{D}$, the system must never move back to the state $\overline{A}$.* If the solution obtained in the forward sense (from time t_1 to time t_2) for unlabeled instances $\mathcal{X}_2$ is not satisfactory, by applying the considered domain-adaptation algorithm in the backward sense (from time t_2 to time t_1), it must never be possible to obtain an acceptable solution at time t_1 (i.e., $\Pr(\overline{A}|\overline{D}) = 0$), but rather to obtain a solution that is not consistent with the reference image $\mathcal{I}_1$ (i.e., $\Pr(\overline{C}|\overline{D}) = 1$). This hypothesis is very reasonable because it seems almost impossible to recover a correct solution at time t_1 starting from a completely wrong solution at t_2. This has some analogies with the definition of specific trajectories that model transitions between different states in chaotic systems.
2) *Starting from the state $\overline{B}$, the system can return to the state $\overline{A}$.* If there exists a set of satisfactory solutions obtained in the forward sense (from time t_1 to time t_2) for unlabeled instances $\mathcal{X}_2$ (i.e., $\mathcal{B} \neq \varnothing$), by applying the considered domain-adaptation algorithm in the backward sense (from time t_2 to time t_1), it must be possible to obtain for at least one of them a solution acceptable for $\mathcal{I}_1$ (i.e., $\Pr(\overline{A}|\overline{B}) > 0$).

Under the aforementioned assumptions, the system never accepts solutions that are nonconsistent with the new image $\mathcal{I}_2$. It is worth noting that this aspect represents the keypoint for a correct validation in real operational problems, because this guarantees to avoid validation of inaccurate classification maps. In some cases, it may happen that solutions consistent at time t_2 are actually rejected; nevertheless, this is due to the fact that the learning parameters are not optimized for the backward process. Therefore, it does not represent a critical issue (it is important that at least one accurate solution can be accepted).

V. Experimental Results

In order to assess the effectiveness of the proposed system, we carried out several experiments. In particular, we considered a data set made up of two coregistered multispectral images acquired by the Thematic Mapper (TM) sensor of the Landsat 5 Satellite in September 1995 ($\mathcal{I}_1$) and in July 1996 ($\mathcal{I}_2$). The investigated site relates to a section of 412 $\times$ 382 pixels (i.e., about 11.7 km $\times$ 10.8 km), including Lake Mulargia, located in the Southern part of the Sardinia Island, Italy (see Fig. 4). In our trials, we took into account the five information classes that mainly characterized the area of interest at both the times, i.e., forest, pasture, urban area, water, and vineyard. As commonly done in the literature, among the seven available TM spectral bands, we did not take into account the low-resolution band associated with the thermal infrared channel (i.e., band 6). Moreover, in order to exploit the distribution-free nature of SVMs and to characterize the texture properties of the investigated land-cover classes, in addition to the six remaining TM bands, we also considered five additional texture features based on the gray-level co-occurrence matrix (GLCM), i.e., correlation, sum average, sum variance, difference variance, and entropy. The GLCM was obtained by compressing the original dynamic of 256 levels to 32 levels using an equal-probability quantizing algorithm [31]. The window size was set to 7 $\times$ 7, and the interpixel distance was fixed to 1.

Available prior knowledge about the area of interest was exploited to define two spatially uncorrelated sets of labeled samples at time t_1 (i.e., September 1995) and time t_2 (i.e., July 1996), respectively (see Table I). However, it is worth noting that prior ground information related to $\mathcal{I}_2$ was considered only for an objective and quantitative assessment of the performances of the proposed system. In the following, we will refer to the set of labeled samples available for $\mathcal{I}_1$ as $\mathcal{T}_1 = \{\mathcal{X}_1, \mathcal{Y}_1\}$ (i.e., $\mathcal{X}_1$ and $\mathcal{Y}_1$ represent the instances and the corresponding true labels, respectively), whereas we will



(a) (b)

Fig. 4. Band 5 of the multispectral Landsat-5 TM images used in the experiments. (a) Image acquired in September 1995. (b) Image acquired in July 1996.

TABLE I
NUMBER OF LABELED PATTERNS AVAILABLE FOR THE SEPTEMBER 1995 ($\mathcal{I}_1$) AND THE JULY 1996 ($\mathcal{I}_2$) IMAGES AND JENSEN–SHANNON DIVERGENCE ($D_{\rm JS}$) VALUES

Information Classes	Number of available labeled patterns		$D_{\rm JS}$
	$\mathcal{I}_1$	$\mathcal{I}_2$	
Pasture	554 (24.63%)	589 (30.22%)	0.517
Forest	304 (13.52%)	274 (14.06%)	0.278
Urban Area	408 (18.14%)	418 (21.45%)	0.391
Water	804 (35.75%)	551 (28.27%)	0.423
Vineyard	179 (7.96%)	117 (6.00%)	0.567
Overall	2249	1949	0.391

denote $\mathcal{X}_2$ as the subset of instances at t_2 for which true labels were actually available but were not exploited in the learning phase.

At the beginning of our analysis, we assessed the complexity of the considered problem by estimating the "distance" between the pattern distributions $P^1(\mathbf{x})$ and $P^2(\mathbf{x})$ related to $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. This was accomplished by computing the Jensen–Shannon divergence $(D_{\rm JS})$ [35] defined as

$$D_{\rm JS}\left[P^1(\mathbf{x}), P^2(\mathbf{x})\right] = \alpha \cdot D_{\rm KL}\left[P^1(\mathbf{x}) \| P^3(\mathbf{x})\right] + \beta \cdot D_{\rm KL}\left[P^2(\mathbf{x}) \| P^3(\mathbf{x})\right] \quad (18)$$

where $D_{\rm KL}$ represents the Kullback–Leibler divergence[3] and $P^3(\mathbf{x}) = \alpha \cdot P^1(\mathbf{x}) + \beta \cdot P^2(\mathbf{x})$. In particular, we fixed $\alpha = \beta = 0.5$ (this case is referred in the literature as *specific* $D_{\rm JS}$) for which it holds that $D_{\rm JS}[P^1(\mathbf{x}), P^2(\mathbf{x})] \in [0; \log 2]$. The existence of both lower and upper bounds for $D_{\rm JS}$ is very important as it let us guess how different the two distributions are. If $D_{\rm JS}[P^1(\mathbf{x}), P^2(\mathbf{x})] = 0$, then $P^1(\mathbf{x})$ and $P^2(\mathbf{x})$ can be considered identical, whereas if $D_{\rm JS}[P^1(\mathbf{x}), P^2(\mathbf{x})] = \log 2 \simeq 0.693$, $P^1(\mathbf{x})$ and $P^2(\mathbf{x})$ can be considered independent.

Table I shows that, as the two investigated images were acquired in different periods of the year, the resulting overall $D_{\rm JS}$ between the distributions at time t_1 and t_2 was considerable (i.e., 0.391). Moreover, the complexity of the considered problem is increased by the distances evaluated between the corresponding conditional class distributions at the two dates $D_{\rm JS}[P^1(\mathbf{x}|\omega_i), P^2(\mathbf{x}|\omega_i)]$, which permit the estimation of how far apart the pattern distributions related to the same information class in the two images are. In particular, a considerable difference is noticed for both the class pasture (i.e., $D_{\rm JS} = 0.517$) and the class vineyard (i.e., $D_{\rm JS} = 0.567$).

In all the trials, we chose the percentage overall accuracy $OA\%$ (i.e., the percentage of correctly labeled samples over the whole number of considered samples) as reference classification accuracy measure Λ and fixed $\Lambda_{\rm th} = OA\%_{\rm th} = 85$ (this value can be considered as a reasonable lower bound for an acceptable solution in the investigated problem). Accordingly, for both the classification problems at time t_1 and time t_2, a solution was assumed to be consistent if $OA\% \geq 85$. We employed Gaussian kernel functions (ruled by the free parameter σ) in the learning phase of both DASVMs and SVMs (used for comparison purposes), as they generally proved effective in addressing classification of multispectral images. Note that one of the constraints imposed by the DASVM algorithm is to use the OAA multiclass architecture; therefore, the same multiclass strategy was adopted also for supervised SVMs. In all the experimental trials, we employed the sequential minimal optimization algorithm [37] for training both the supervised SVMs and, with proper modifications, the proposed DASVMs. As pointed out in Section III, we fixed $\tau = 0.5$ (it is reasonable to assign the semilabeled samples at most a regularization parameter equal to one half of the regularization parameter for original training samples). Concerning the convergence criterion, on the basis of the performances exhibited by the system on a variety of preliminary experimental trials on different toy data sets, a reasonable empirical choice proved to be $\beta = 3 \cdot 10^{-2}$.

At the beginning of our analysis, we trained several supervised SVMs at time t_1 by exploiting all the labeled samples $\mathcal{T}_1 = \{\mathcal{X}_1, \mathcal{Y}_1\}$ available for the reference image $\mathcal{I}_1$. In particular, in order to identify models with good generalization capabilities, according to the literature [30], we adopted in the learning phase a tenfold cross-validation (CV) strategy, which allows one to effectively evaluate the quality of a predicted model using the same data set exploited for building the model

[3] $D_{\rm JS}$ is a symmetrized and smoothed version of $D_{\rm KL}$, which is defined as $D_{\rm KL}[P^1(\mathbf{x}) \| P^2(\mathbf{x})] = \sum_n p_n^1 \log(p_n^1/p_n^2)$, where p_n^1 and p_n^2 are point probabilities of $P^1(\mathbf{x})$ and $P^2(\mathbf{x})$, respectively [36].

TABLE II
PERCENTAGE OVERALL ACCURACY ($OA\%$), PRODUCER'S ACCURACIES ($PA\%$), AND USER'S ACCURACIES ($UA\%$) OBTAINED FOR THE NEW IMAGE $\mathcal{I}_2$ BY THE FOLLOWING: 1) THE SUPERVISED SVM TRAINED ACCORDING TO THE TENFOLD CV STRATEGY WITH THE LABELED PATTERNS AVAILABLE FOR $\mathcal{I}_1$ WHICH PROVIDED THE HIGHEST $OA\%$ AT TIME t_1 ($\text{SVM}_{t_1}^{\text{CV}}$) AND 2) THE PROPOSED DASVM TECHNIQUE WITH OPTIMAL SELECTION OF DOMAIN-ADAPTATION PARAMETERS ($\text{DASVM}^{\text{best}}$). THE AVERAGE ACCURACY ASSOCIATED WITH THE CONSISTENT SOLUTIONS OBTAINED BY THE PROPOSED DASVM TECHNIQUE AND CORRECTLY IDENTIFIED BY THE CIRCULAR VALIDATION STRATEGY ($\text{DASVM}^{\text{ave}}$) IS ALSO GIVEN

	$OA\%$	Pasture		Forest		Urban Area		Water		Vineyard	
		$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$
$\text{SVM}_{t_1}^{CV}$	78.76	53.48	94.03	99.27	61.82	80.34	98.25	100	99.28	52.14	22.02
DASVM^{best}	96.67 (+17.91)	96.60 (+43.12)	96.61 (+2.58)	99.63 (+0.36)	94.46 (+32.64)	97.85 (+17.51)	99.51 (+1.26)	100 (+0.00)	100 (+0.72)	70.01 (+17.87)	75.23 (+53.21)
DASVM^{ave}	93.12 (+14.36)	89.81 (+36.33)	96.36 (+2.33)	99.27 (+0.00)	81.68 (+19.86)	93.54 (+13.20)	99.24 (+0.99)	100 (+0.00)	100 (+0.72)	61.54 (+9.40)	59.02 (+37.00)

itself. This process allowed us to identify the values of the supervised learning parameters (i.e., σ and C) that better fit the structure of the classification problem at t_1 and resulted in solutions consistent for $\mathcal{I}_1$. Successively, we trained a number of DASVMs using $\mathcal{T}_1$ and $\mathcal{X}_2$ as labeled and unlabeled sets, respectively. It is worth noting that, in the learning process, we associated both σ and C to pairs of values that resulted in consistent solutions at time t_1, while we applied a grid search strategy on the remaining domain-adaptation parameters (i.e., C^*, ρ, and γ).

In all our trials, the choice of the C^* value was not critical. In particular, for very small values of C^* (i.e., in the range of 0.5–1), it was always possible to obtain good classification accuracies. As concerns γ and ρ, their behaviors proved to be correlated. Note the following: 1) The higher is ρ, the lower is expected to be the number of total iterations, and 2) to let the final regularization parameters for the most of semilabeled samples achieve sensible values, γ must be lower than the number of total iterations. Accordingly, there exists an intrinsic inverse proportion between ρ and γ. In our experiments, a reasonable range for ρ was 5–15. For $\rho = 5$, we obtained consistent solutions when $\gamma \in 45{-}55$, whereas for $\rho = 15$, we obtained consistent solutions when $\gamma \in 30{-}40$.

In order to assess the effectiveness of the circular validation strategy presented in Section IV, for each of the aforementioned DASVMs, we applied the proposed domain-adaptation algorithm in the reverse sense. In each case, we exploited the set of estimated labels $\hat{\mathcal{Y}}_2$ predicted at time t_2 for defining an estimated training set $\hat{\mathcal{T}}_2 = \{\mathcal{X}_2, \hat{\mathcal{Y}}_2\}$ for the new image $\mathcal{I}_2$. Then, we trained the correspondent backward DASVM using $\hat{\mathcal{T}}_2$ as labeled set and the set of instances $\mathcal{X}_1$ at time t_1 (considered without the associated true labels) as unlabeled set. Note that we kept the same values for the learning parameters employed in the forward learning. Finally, by exploiting the real ground-truth labels $\mathcal{Y}_1$ available for $\mathcal{I}_1$, we were able to determine whether the estimated final solution $\hat{\mathcal{Y}}_1$ was consistent or not with the reference image and, thus, to infer about the correctness of the solution $\hat{\mathcal{Y}}_2$ related to the new image $\mathcal{I}_2$.

At the end of our analysis, the proposed system proved to be particularly promising and exhibited very good classification performances. Table II shows the results obtained in terms of $OA\%$ and both percentage producer's and user's accuracies (i.e., $PA\%$ and $UA\%$, respectively) for each information class at time t_2 by the following: 1) the supervised SVM trained according to the tenfold CV strategy with the labeled patterns available for $\mathcal{I}_1$ which provided the highest $OA\%$ at time t_1 ($\text{SVM}_{t_1}^{\text{CV}}$) and 2) the proposed DASVM technique with optimal selection of domain-adaptation parameters ($\text{DASVM}^{\text{best}}$). Moreover, also the average accuracies associated with the consistent solutions obtained by the presented domain-adaptation algorithm and correctly identified by the circular validation strategy ($\text{DASVM}^{\text{ave}}$) are reported[4] (in order to obtain significant estimations, 400 backward DASVMs have been trained both starting from consistent and nonconsistent solutions at time t_2). It is worth noting that these values are particularly important as they represent an average measure for the quality of the solutions that were identified as consistent without exploiting any prior ground information about the new image $\mathcal{I}_2$.

From the table, it is noticed that the DASVM technique was able to sharply increase the accuracies with respect to the standard supervised approach. In particular, the average improvement in the $OA\%$ exhibited by the solutions automatically identified as consistent was remarkable (i.e., $+14.36$). Without any prior true label for $\mathcal{I}_2$, we were able to obtain a mean $OA\%$ equal to 93.12, which can be considered a very good result in the light of the complexity of the investigated problem. Moreover, when the domain-adaptation parameters were selected in an optimal (i.e., supervised) way, the proposed method proved capable to increase the $OA\%$ even up to 96.67 (i.e., $+17.91$ with respect to $\text{SVM}_{t_1}^{\text{CV}}$).

As shown by the values of D_{JS} reported in Table I, the distributions of pasture and vineyard experienced the most relevant changes between the two considered dates. Accordingly, it seems reasonable that $\text{SVM}_{t_1}^{\text{CV}}$ exhibited low performances on these two information classes at time t_2.

In general, the behavior of the DASVMs was significantly different with respect to the standard supervised approach [see, for instance, the confusion matrices reported in Table III(a) and (b)]. $\text{SVM}_{t_1}^{\text{CV}}$ often misclassified several pasture patterns as forest areas at time t_2, whereas the proposed domain-adaptation algorithm rarely incurred in such kind of errors, thus exhibiting a huge increase in the corresponding $PA\%$ (i.e., $+36.33$ on

[4]Note that also the solution obtained at time t_2 by $\text{DASVM}^{\text{best}}$ was correctly identified as consistent by the proposed circular validation strategy.

TABLE III
CONFUSION MATRICES THAT RESULTED FROM THE CLASSIFICATION OF THE NEW IMAGE $\mathcal{I}_2$ BY USING THE FOLLOWING: (a) THE SUPERVISED SVM TRAINED ACCORDING TO THE TENFOLD CV STRATEGY WITH THE LABELED PATTERNS AVAILABLE FOR $\mathcal{I}_1$ WHICH PROVIDED THE HIGHEST $OA\%$ AT TIME t_1 ($\text{SVM}_{t_1}^{\text{CV}}$) AND (b) THE PROPOSED DASVM TECHNIQUE WITH OPTIMAL SELECTION OF DOMAIN-ADAPTATION PARAMETERS ($\text{DASVM}^{\text{best}}$)

	Pasture	Forest	Urban Area	Water	Vineyard
Pasture	315	0	2	0	18
Forest	59	272	80	0	29
Urban Area	0	1	336	0	5
Water	0	0	0	551	4
Vineyard	215	1	0	0	61

(a)

	Pasture	Forest	Urban Area	Water	Vineyard
Pasture	569	0	2	0	18
Forest	1	273	0	0	15
Urban Area	0	0	409	0	2
Water	0	0	0	551	0
Vineyard	19	1	7	0	82

(b)

TABLE IV
PERCENTAGE OVERALL ACCURACY ($OA\%$), PRODUCER'S ACCURACIES ($PA\%$), AND USER'S ACCURACIES ($UA\%$) OBTAINED FOR THE NEW IMAGE $\mathcal{I}_2$ BY THE FOLLOWING: 1) THE SUPERVISED SVM TRAINED ACCORDING TO THE TENFOLD CV STRATEGY AT TIME t_2 EXPLOITING TRUE LABELS FOR THE PATTERNS OF $\mathcal{X}_2$ ($\text{SVM}_{t_2}^{\text{CV}}$); 2) THE PARTIALLY SUPERVISED RETRAINING TECHNIQUE FOR ML CLASSIFIERS ($\text{ML}_{\text{retrain}}$) PRESENTED IN [4]; AND 3) THE PARTIALLY UNSUPERVISED ML CASCADED CLASSIFIER ($\text{ML}_{\text{cascade}}$) PROPOSED IN [5]

	$OA\%$	Pasture		Forest		Urban Area		Water		Vineyard	
		$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$	$PA\%$	$UA\%$
$\text{SVM}_{t_2}^{CV}$	99.69	100	99.49	100	100	99.28	100	100	100	97.44	97.44
$\text{ML}_{\text{retrain}}$	92.76	94.06	88.50	87.22	88.19	93.06	97.49	100	100	64.10	73.53
$\text{ML}_{\text{cascade}}$	91.48	94.25	83.53	90.51	97.45	80.48	95.69	100	100	86.90	62.39

average, +43.12 in the best case). Urban areas exhibited a similar behavior, as they were frequently confused with forest by the supervised approach, whereas the presented DASVMs were able to avoid these errors, thus providing a noteworthy improvement both in the $PA\%$ of urban area (i.e., +13.20 on average, +17.51 in the best case) and in the $UA\%$ of forest (i.e., +19.86 on average, +32.64 in the best case). Even if the related D_{JS} was not negligible, $\text{SVM}_{t_1}^{\text{CV}}$ exhibited good accuracies for the class water also at time t_2. This is due to the fact that the spectral signature of this class is rather different with respect to those of the other information classes. The proposed system resulted effective also on the vineyard class, which proved to be the most critical information class (as confirmed by the very low accuracies provided by $\text{SVM}_{t_1}^{\text{CV}}$) mainly due to the fact that it has the lowest prior probability. Moreover, in this case, even if the accuracies are smaller than those obtained for the other classes, the gain of the DASVMs with respect to the supervised approach is significant both in terms of $PA\%$ (i.e., +9.40 on average, +17.87 in the best case) and $UA\%$ (i.e., +37.00 on average, +53.21 in the best case).

For the sake of comparison, in Table IV, we also evaluated the performances obtained by the following: 1) a supervised SVM trained according to a tenfold CV strategy with the samples belonging to $\mathcal{X}_2$ by exploiting the corresponding true labels ($\text{SVM}_{t_2}^{\text{CV}}$); 2) the partially unsupervised retraining technique for ML classifiers (denoted as $\text{ML}_{\text{retrain}}$) presented in [4]; and 3) the partially unsupervised ML cascade classifier (denoted as $\text{ML}_{\text{cascade}}$) proposed in [5].

As expected, $\text{SVM}_{t_2}^{\text{CV}}$ resulted in almost perfect separation between the considered information classes at time t_2. This is mainly due to the fact that the model selection for the free parameters has been optimized to seize the problem modeled on $\mathcal{I}_2$ according to the CV strategy. For a suboptimal selection of the values for the learning parameters, the supervised approach provided results comparable with those exhibited by the DASVM technique. DASVMs also proved capable to obtain higher accuracies than the $\text{ML}_{\text{retrain}}$ and $\text{ML}_{\text{cascade}}$ techniques, thus confirming once again the effectiveness of the presented method (see Table IV).

Let us now focus the attention on the circular validation procedure. Fig. 5 shows the empirical cumulative distribution function (cdf) of the $OA\%$ [denoted as $\hat{P}(OA\%)$] estimated from the solutions obtained for the reference image $\mathcal{I}_1$ at the end of the backward learning process when the system started from both the state $\overline{B}$ (i.e., solutions consistent with the new image $\mathcal{I}_2$, black line) and the state $\overline{D}$ (solutions nonconsistent with the new image $\mathcal{I}_2$, gray line).

For each point of the considered cdfs, the ordinate (referred to as *quantile*) represents the probability that the final $OA\%$ obtained after the circular learning process on the labeled samples available at time t_1 is lower or equal to the $OA\%$ value of the corresponding abscissa. Accordingly, it is noticed that

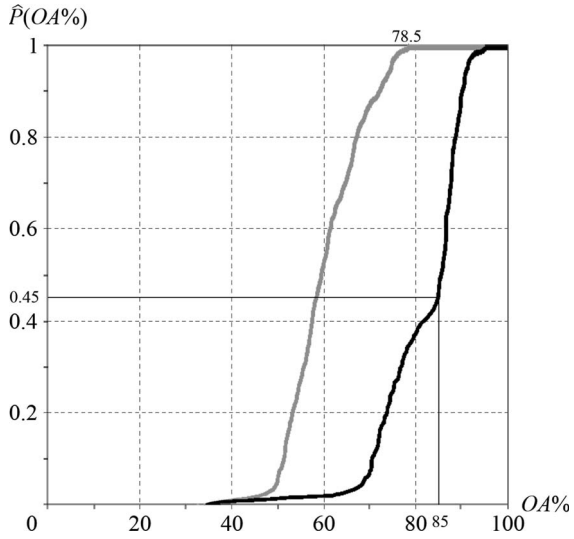


Fig. 5. Empirical cdf of the percentage overall accuracy ($OA\%$) estimated from the solutions obtained at the end of the backward learning process for $\mathcal{I}_1$ when the system started from both state $\overline{B}$ (i.e., solutions consistent with the new image $\mathcal{I}_2$, black line) and state $\overline{D}$ (i.e., solutions nonconsistent with the new image $\mathcal{I}_2$, gray line). If $OA\% < 85$, the system moved to the state $\overline{C}$ (i.e., solutions nonconsistent with the reference image $\mathcal{I}_1$); if $OA\% \geq 85$, the system moved to the state $\overline{A}$ (i.e., solutions consistent with the reference image $\mathcal{I}_1$).

the proposed validation strategy was always able to correctly reject solutions that were not acceptable at time t_2, since the distribution corresponding to solutions not consistent with the new image (gray line) saturates to 1 for $OA\% \approx 78.5$, which is lower than $OA\%_{\text{th}} = 85$. Indeed, when the DASVM started from a solution that did not adequately model the classification problem for $\mathcal{I}_2$, the system could not recover a solution consistent with the reference image $\mathcal{I}_1$, thus satisfying the most critical requirement for the operational employment of the proposed strategy (i.e., $\Pr(\overline{A}|\overline{D}) = 0$).

As concerns the distribution related to the solutions consistent with the new image (black line), the quantile corresponding to $OA\%_{\text{th}} = 85$ [i.e., $q_{0.85}(OA\%)$] is equal to 0.45. Accordingly, as the systems move to the state $\overline{C}$ if $OA\% < OA\%_{\text{th}}$, we have that $\Pr(\overline{C}|\overline{B}) = q_{0.85}(OA\%) = 0.45$. Therefore, $\Pr(\overline{A}|\overline{B}) = 1 - \Pr(\overline{C}|\overline{B}) = 0.55$, which means that the classifier moved back to the state $\overline{A}$ in the 55% of the cases. This result is very important, as it means that without considering any prior ground information for $\mathcal{I}_2$, it was possible to correctly identify more than a half of the correct solutions. Moreover, as discussed before, the average quality of these solutions is comparable with that obtained with optimal selection of learning parameters, thus confirming the effectiveness of the proposed circular validation strategy.

VI. Discussion and Conclusion

In this paper, we have presented a novel domain-adaptation classifier based on SVMs (DASVM) for land-cover map updating, which can be employed in real operational situations when ground-truth labels are available only for a reference image acquired over the same geographical area before the one being classified. In addition, we have proposed a circular validation strategy for the accuracy assessment of the classification results without labeled samples available for the image to be categorized.

The proposed DASVM technique extends the principles of SVMs to the domain-adaptation framework, by taking into account that unlabeled "test" samples are drawn from a new image $\mathcal{I}_2$ different from the reference image $\mathcal{I}_1$ of training samples. Starting from a standard supervised learning on labeled samples available for $\mathcal{I}_1$ (which determines an initial discriminant function for the new image), the DASVM technique iteratively selects and labels the unlabeled patterns of $\mathcal{I}_2$ that are most likely to be correctly classified. At the same time, original training patterns of $\mathcal{I}_1$ are gradually erased, as they refer to an image different from the one being classified; therefore, the final solution is ruled only by patterns of $\mathcal{I}_2$. In order to improve the robustness and to better control the behavior of the classifier, an adaptive weighting strategy for the regularization parameters based on a temporal criterion has been defined. This allows one to tune the influence of unlabeled patterns and, in general, prevents the system from both converging to improper solutions and providing unreliable results.

The presented circular validation strategy overcomes the problem of the lack of procedures for the validation of classification results in the context of land-cover map updating, when no prior ground information for the new image being classified is available. The proposed strategy has been developed under the assumption that, although different, the class distributions of images acquired over the same geographical area at different times are intimately related. In particular, we assume that a solution for $\mathcal{I}_2$ obtained with a domain-adaptation learning algorithm is consistent if the solution obtained by applying the same algorithm in the reverse sense (i.e., using the classification labels in place of missing true labels for the new image and considering the training patterns of the reference image as unlabeled) is consistent (i.e., sufficiently accurate) with $\mathcal{I}_1$ (this can be evaluated due to the availability of true labels for $\mathcal{I}_1$).

The experimental results obtained on a multitemporal data set made up of two multispectral images acquired by the TM sensor of the Landsat-5 satellite confirmed the effectiveness and the robustness of the proposed methods. On the one hand, the presented DASVM technique exhibited a very good discrimination capability and proved capable to outperform standard SVMs, resulting in high and satisfactory classification accuracies. On the other hand, the circular validation strategy allowed us to correctly identify solutions consistent with the new image even in critical conditions. In addition, it proved always able to reject solutions that were not consistent with the investigated classification problem (which is the most critical requirement for the operational employment of the proposed strategy).

It is worth noting that, if most of the unlabeled samples of the new image are incorrectly classified at the beginning of the learning process, it becomes difficult for the proposed system to recover a reliable land-cover map. Nevertheless, this is due to the fact that the convergence to a consistent solution is related to the intrinsic similarity between the considered images, which can be estimated by computing proper statistical distance measures (e.g., the Jensen–Shannon divergence).

Thus, as for semisupervised and TSVMs, when the two images are considerably different, it is not possible to guarantee for obtaining a reliable solution to the investigated classification problem. However, such critical situation is detected by the proposed circular validation strategy, which, in this case, is able to reject all the solutions.

As concerns the computational load of the proposed classifier, it is worth noting that each iteration of the DASVM algorithm requires a time equivalent to that necessary for training a supervised SVM. Indeed, while the number of semilabeled patterns drawn from the new image increases, at the same time, the number of labeled patterns of the reference image decreases. Accordingly, the cardinality of the training set does not increase with the number of iterations; therefore, the computational load grows almost linearly with the number of iterations. In our experiments, carried out on a PC mounting an Intel Core2 Duo processor at 2.6 GHz and a 4-Gb DDR2 RAM, the training phase of a supervised SVM took about 10 s. Concerning the proposed DASVM, it came out that, on average, some tens of iterations were necessary. Therefore, the average learning time resulted of about 9 min. Accordingly, by taking into account the huge increase of the classification accuracy with respect to the supervised approach, it is reasonable to consider the computational cost of DASVMs acceptable. In order to speed up further the learning process, similarly to what it is commonly done in the active learning framework, as a future development of this work, we are studying an adaptive version of the proposed method that iteratively takes into account only the patterns of the current training set identified as support vectors and the semilabeled samples selected at the current iteration.

REFERENCES

[1] J. A. Richards, *Remote Sensing Digital Image Analysis*, 2nd ed. New York: Springer-Verlag, 1993.

[2] P. H. Swain, "Bayesian classification in a time-varying environment," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 12, pp. 880–883, Dec. 1978.

[3] X. Zhu, "Semi-supervised learning literature survey," Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, TR-1530, 2005.

[4] L. Bruzzone and D. Fernàndez Prieto, "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 456–460, Feb. 2001.

[5] L. Bruzzone and D. Fernàndez Prieto, "A partially unsupervised approach to the automatic classification of multitemporal remote-sensing images," *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1063–1071, 2002.

[6] L. Bruzzone, R. Cossu, and G. Vernazza, "Combining parametric and non-parametric algorithms for a partially unsupervised classification of multitemporal remote-sensing images," *Inf. Fusion*, vol. 3, no. 4, pp. 289–297, Dec. 2002.

[7] L. Bruzzone and R. Cossu, "A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 9, pp. 1984–1996, Sep. 2002.

[8] H. Daumè, III and D. Marcu, "Domain adaptation for statistical classifiers," *J. Artif. Intell. Res.*, vol. 26, pp. 101–126, 2006.

[9] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in NLP," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, Prague, Czech Republic, 2007, pp. 264–271.

[10] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representation for domain adaptation," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 19. Vancouver, BC, Canada, 2006.

[11] S. Satpal and S. Sarawagi, "Domain adaptation of conditional probability models via feature subsetting," in *Proc. 11th Eur. Conf. Principles Practice Knowl. Discov. Databases*, Warsaw, Poland, 2007, pp. 224–235.

[12] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," in *J. R. Stat. Soc., B*, vol. 39, 1977, pp. 1–38.

[13] 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.

[14] 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.

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

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

[17] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Berlin, Germany: Springer-Verlag, 1999.

[18] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proc. Adv. Neural Inform. Process. Syst.*, 1998, vol. 10, pp. 368–374.

[19] 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.

[20] 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.

[21] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.

[22] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, May 2007.

[23] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[24] D. Zhou, J. Huang, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing System*, vol. 16, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, pp. 321–328.

[25] L. Gómez, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe-Maravilla, "Semisupervised image classification with Laplacian support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 336–340, Jul. 2008.

[26] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[27] M. I. Jordan, *Learning in Graphical Models*, 1st ed. Cambridge, MA: MIT Press, 1999.

[28] 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.

[29] 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.

[30] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[31] R. M. Haralick, K. Shanmugan, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[32] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[33] 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.

[34] L. Bruzzone, M. Chi, and M. Marconcini, "Semisupervised support vector machines for classification of hyperspectral remote sensing images," in *Hyperspectral Data Exploitation Theory and Applications*. New York: Wiley-Interscience, 2007, pp. 275–311.

[35] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[36] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[37] 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, pp. 185–208.

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 is currently a Full Professor of telecommunications. He teaches remote sensing, pattern recognition, and electrical communications. He is also currently the Head of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. 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 7 book chapters. 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). He conducts and supervises research on these topics within the frameworks of several national and international projects.

Dr. Bruzzone is a member of the International Association for Pattern Recognition and of the Italian Association for Remote Sensing. He 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 a Referee for many international journals and has served on the scientific committees of several international conferences. He was the General Chair and Cochair of the first and second IEEE International Workshop on the Analysis of Multi-temporal Remote-Sensing Images (MultiTemp). He is also 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. He is currently 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).

Mattia Marconcini (S'06–M'09) received the "Laurea" (B.S.) and the "Laurea Specialistica" (M.S.) degrees in telecommunication engineering (*summa cum laude*) and the Ph.D. degree in communication and information technologies from the University of Trento, Trento, Italy, in 2002, 2004, and 2008, respectively.

He is currently with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. His current research activities are in the areas of machine learning, pattern recognition, and remote sensing. In particular, his interests are related to transfer learning and domain-adaptation classification and to image segmentation problems. He conducts research on these topics within the frameworks of several national and international projects.

Dr. Marconcini was a finalist of the Student Prize Paper Competition of the 2007 IEEE International Geoscience and Remote Sensing Symposium (Barcelona, July 2007).