A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability

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

***Abstract*—This paper presents a novel approach to feature selection for the classification of hyperspectral images. The proposed approach aims at selecting a subset of the original set of features that exhibits at the same time high capability to discriminate among the considered classes and high invariance in the spatial domain of the investigated scene. This approach results in a more robust classification system with improved generalization properties with respect to standard feature-selection methods. The feature selection is accomplished by defining a multiobjective criterion function made up of two terms: 1) a term that measures the class separability and 2) a term that evaluates the spatial invariance of the selected features. In order to assess the spatial invariance of the feature subset, we propose both a supervised method (which assumes that training samples acquired in two or more spatially disjoint areas are available) and a semisupervised method (which requires only a standard training set acquired in a single area of the scene and takes advantage of unlabeled samples selected in portions of the scene spatially disjoint from the training set). The choice for the supervised or semisupervised method depends on the available reference data. The multiobjective problem is solved by an evolutionary algorithm that estimates the set of Pareto-optimal solutions. Experiments carried out on a hyperspectral image acquired by the Hyperion sensor on a complex area confirmed the effectiveness of the proposed approach.**

***Index Terms*—Expectation–maximization (EM) algorithm, feature selection, hyperspectral images, image classification, remote sensing, robust features, semisupervised feature selection, stationary features.**

I. Introduction

HYPERSPECTRAL remote sensing images, which are characterized by a dense sampling of the spectral signature of different land-cover types, represent a very rich source of information for the analysis and automatic recognition of land-cover classes. However, supervised classification of hyperspectral images is a very complex methodological problem due to many different issues [1]–[5]: 1) the small value of the ratio between the number of training samples and the number of available spectral channels (and thus of classifier parameters), which results in the Hughes phenomenon [6]; 2) the high correlation among training patterns taken from the same area, which violates the required assumption of independence of samples included in the training set (thus reducing the information conveyed to the classification algorithm by the considered samples); and 3) the nonstationary behavior of the spectral signatures of land-cover classes in the spatial domain of the scene, which is due to physical factors related to ground (e.g., different soil moisture or composition), vegetation, and atmospheric conditions. All the aforementioned issues result in decreasing the robustness, the generalization capability, and the overall accuracy of classification systems used to generate the land-cover maps.

In order to address the aforementioned problems, in the recent literature, different promising approaches have been proposed for hyperspectral image classification. Among the others, we recall the following: 1) the use of supervised kernel methods [and in particular of support vector machines (SVMs)], which are intrinsically robust to the Hughes phenomenon [1], [2]; 2) the use of semisupervised learning methods that take into account both labeled and unlabeled samples in the learning of the classifier [3]; and 3) the joint use of kernel methods and semisupervised techniques [4], [5]. On the one hand, SVMs are supervised classifiers that result in augmented generalization capability with respect to other classification methods thanks to the structural risk minimization principle, which allows one to effectively control the tradeoff between the empirical risk and the generalization property. On the other hand, semisupervised approaches can increase the capability of classification algorithms to derive discrimination rules that better fit with the nonstationary behavior of features in the hyperspectral image under investigation, by considering also the information of unlabeled samples. These classification methods proved to be quite effective in mitigating some of the aforementioned problems. Nevertheless, the problem of the spatial variability of the features can be addressed (together with the sample size problem) at a different and complementary level, i.e., in the feature extraction and/or feature-selection phase. To this purpose, the feature extraction phase should aim at deriving discriminative features that are also as stationary as possible in the spatial domain. The feature-selection phase should aim at selecting a subset of the available features that satisfies the following: 1) allows the classifier to effectively discriminate

Manuscript received November 18, 2008; revised January 30, 2009. First published July 10, 2009; current version published August 28, 2009.

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

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

Digital Object Identifier 10.1109/TGRS.2009.2019636

the considered classes and 2) contains features that have the most invariant as possible behavior in the spatial domain. In this paper, we focus on the development of a feature-selection approach to the identification of robust and spatially invariant features. It is worth noting that, although, in the literature, several feature-selection algorithms have been proposed for the analysis of hyperspectral data (e.g., [9]–[12]), to the authors' knowledge, little attention has been devoted to the aforementioned problem.

The feature-selection techniques that are most widely used in remote sensing generally require the definition of a criterion function and a search strategy. The criterion function is a measure of the effectiveness of the considered subset of features, and the search strategy is an algorithm that aims at efficiently finding a solution (i.e., a subset of features) that optimizes the adopted criterion function. In standard feature-selection methods [9]–[17], the criterion functions typically adopted are statistical measures that assess the separability of the different classes on a given training set but do not explicitly take into account the stationarity of the features (e.g., the variability of the spectral signature of the land-cover classes). This approach may result in selecting a subset of features that retains very good discrimination properties in the portion of the scene close to the training pixels (and therefore with similar behavior), but are not appropriate to model the class distributions in separate portions on the scene, which may present different spectral behavior. Considering the typical high spatial variability of the spectral signature of land-cover classes in hyperspectral images, this approach can lead to an *overfitting* phenomenon in the feature-selection phase, resulting in poor generalization capabilities of the classification system. Note that we use here the term *overfitting* with an extended meaning with respect to the conventional sense, which traditionally refers to the phenomenon that occurs when inductive algorithms model too closely the training data, losing generalization capability. In this paper, we observe that there is an intrinsic spatial variability of the spectral signature of classes in the hyperspectral image, and thus, we expect that the generalization ability of the system is strongly affected by this property of hyperspectral data, which is much more critical than in standard multispectral images.

In this paper, we address the aforementioned problem by proposing a novel approach to feature selection that aims at identifying a subset of features that exhibits both high discrimination ability among the considered classes and high invariance in the spatial domain of the investigated scene. This approach is implemented by defining a novel criterion function that is based on the evaluation of two terms: 1) a standard separability measure and 2) a novel invariance measure that assesses the stationarity of features in the spatial domain. The search algorithm, adopted for deriving the subsets of features that jointly optimize the two terms, is based on the optimization of a multiobjective problem for the estimation of the Pareto-optimal solutions. For the assessment of the two terms of the criterion function, we propose both a supervised and a semisupervised method that can be adopted according to the amount of available reference data. The proposed approach can be integrated in the design of any system for hyperspectral image classification (e.g., based on parametric or distribution-free supervised algorithms, kernel methods, and semisupervised classification techniques) for increasing the robustness and the generalization capability of the classifier.

This paper is organized into six sections. The next section presents the background and a brief overview on existing feature-selection algorithms for the classification of hyperspectral data. Section III presents the proposed novel approach to the selection of features for the classification of hyperspectral images, and two possible methods to implement it according to the available reference data. Section IV describes the adopted hyperspectral data set and the design of the experimental analysis carried out for assessing the effectiveness of the proposed approach. Section V presents the obtained experimental results on the considered data set. Section VI draws the conclusions of this paper.

II. Background on Feature Selection in Hyperspectral Images

The process of feature selection aims at reducing the dimensionality of the original feature space by selecting an effective subset of the original features while discarding the remaining measures. Note that this approach is different from feature transformation (extraction), which consists in projecting the original feature space onto a different (usually lower dimensional) feature space [9], [14], [18], [19]. In this paper, we focus our attention on feature selection, which has the important advantage to preserve the physical meaning of the selected features. Moreover, feature selection results in a more general approach than feature transformation alone by considering that the features given as input to the feature-selection module can be associated with the original spectral channels of the hyperspectral image and/or with measures that extract information from the original channels and from the spatial context of each single pixel [20], [21] (e.g., texture, wavelets, average of groups of contiguous bands, derivatives of the spectral signature, etc.).

Let us formalize a general feature-selection problem for the classification of a hyperspectral image $\mathcal{I}$, where each pixel, described by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in an n-dimensional feature space, is to be assigned to one of C different classes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$. The set Υ is made up of the n features in input to the feature-selection process (which can be the original channels and/or measures extracted from them). Let $P(\omega_i)$, where $\omega_i \in \Omega$, be the *a priori* probabilities of the land-cover classes in the considered scene, and let $p(\mathbf{x} \mid \omega_i)$ be the conditional probability density functions for the feature vector $\mathbf{x}$, given the class $\omega_i \in \Omega$. Let us further assume that a training set $T = \{\mathcal{X}, \mathcal{Y}\}$ made up of l pairs $(\mathbf{x}_i, y_i)$ is available, where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $\forall i = 1, 2, \ldots, l$, is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_l\}$, $y_i \in \Omega$, $\forall i = 1, 2, \ldots, l$, is the corresponding set of class labels. The aim of the feature-selection process is to select the most effective subset $\boldsymbol{\theta}^* \subset \Upsilon$ of m features (with $m < n$), according to a criterion function and a search strategy. This can be obtained according to different algorithms that broadly fall into three categories [22]: 1) the *filter* model; 2) the *wrapper* model; and 3) the *hybrid* model. The filter model is based on the general characteristics of the considered data and filters out the most irrelevant

features without involving the classification algorithm. Usually, this is accomplished according to a measure that assesses the separability among classes. The wrapper model depends on a particular classification algorithm and exploits the classifier performance as the criterion function. It searches for a subset of features that optimizes the accuracy of the adopted inductive algorithm, but it is generally computationally more expensive than the filter model. The hybrid model takes advantage of the aforementioned two models by exploiting their different evaluation criteria in different search stages. It uses a criterion function that depends on the available data to identify the subset of candidate solutions for a given cardinality m and then exploits the classification algorithm to select the final best subset. In the next sections, we focus our literature analysis on the filter methods and only on the background concepts that are relevant for the developed technique.

A. Criterion Functions

In standard filter approaches to feature selection, the typically adopted criterion functions are based on statistical distance measures that assess the separability among class distributions $p(\mathbf{x}\,|\,\omega_i)$, $\forall \omega_i \in \Omega$, on the basis of the available training set T. Statistical distance measures are usually adopted as they represent practical criteria to easily approximate the Bayes error. The commonly adopted measures to evaluate the separability between the distributions of two classes ω_i and ω_j are [9], [14]

Divergence:

$$\mathrm{Div}_{ij}(\boldsymbol{\theta}) = \int_{\mathbf{x}} \{p(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_j)\} \ln \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} d\mathbf{x} \quad (1)$$

Bhattacharyya distance:

$$\mathrm{B}_{ij}(\boldsymbol{\theta}) = -\ln \left\{ \int_{\mathbf{x}} \sqrt{p(\mathbf{x}|\omega_i)p(\mathbf{x}|\omega_j)} d\mathbf{x} \right\} \quad (2)$$

Jeffries–Matusita (JM) distance:

$$\mathrm{JM}_{ij}(\boldsymbol{\theta}) = \left\{ \int_{\mathbf{x}} \left[\sqrt{p(\mathbf{x}|\omega_i)} - \sqrt{p(\mathbf{x}|\omega_j)} \right]^2 d\mathbf{x} \right\}^{1/2}. \quad (3)$$

The JM distance can be rewritten according to the Bhattacharyya distance B_{ij}

$$JM_{ij}(\boldsymbol{\theta}) = \sqrt{2\left\{1 - \exp\left[-B_{ij}(\boldsymbol{\theta})\right]\right\}}. \quad (4)$$

In multispectral and hyperspectral remote sensing images, the distributions of classes $p(\mathbf{x}\,|\,\omega_i)$, $\omega_i \in \Omega$ are usually modeled with Gaussian functions with mean vectors μ_i and covariance matrices $\boldsymbol{\Sigma}_i$. Under this assumption, we can write

$$\mathrm{Div}_{ij}(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{Tr}\left\{(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)\left(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1}\right)\right\} + \frac{1}{2}\mathrm{Tr}\left\{\left(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_j^{-1}\right)(\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}}\right\} \quad (5)$$

$$\mathrm{B}_{ij}(\boldsymbol{\theta}) = \frac{1}{8}(\mu_i - \mu_j)^{\mathrm{T}} \left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2}\ln\left(\frac{\frac{1}{2}|\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j|}{\sqrt{|\boldsymbol{\Sigma}_i|\,|\boldsymbol{\Sigma}_j|}}\right) \quad (6)$$

where $\mathrm{Tr}\{\cdot\}$ is the trace of a matrix. An important drawback of the divergence is that its value quadratically increases with respect to the separation between the mean vectors of the class distributions. This behavior does not reflect the classification accuracy behavior, which asymptotically tends to one when the class distributions are perfectly separated. On the contrary, the JM distance exhibits a behavior that saturates when the separability between the two considered classes increases. For this reason, the JM distance is generally preferred to either the divergence or the Bhattacharyya distance.

The previously described measures evaluate the statistical distance between a pair of class distributions. In order to extend the separability measures to multiclass problems, a usually adopted separability indicator is obtained by computing the average distance among all pairwise distances. Thus, a multiclass separability measure can be defined as

$$\Delta(\boldsymbol{\theta}) = \sum_{i=1}^{C} \sum_{j>i}^{C} P(\omega_i)P(\omega_j)S_{ij}(\boldsymbol{\theta}) \quad (7)$$

where $S_{ij}(\boldsymbol{\theta})$ is a statistical distance measure (e.g., Bhattacharyya distance, divergence, and JM distance) between the distributions $p(\mathbf{x}\,|\,\omega_i)$ and $p(\mathbf{x}\,|\,\omega_j)$ of the two classes ω_i and ω_j, and $P(\omega_i)$ and $P(\omega_j)$ are the prior probabilities of the classes ω_i and ω_j in the considered scene, respectively.

Other measures adopted for feature selection are based on scatter matrices that allow one to characterize the variance within classes and between classes [14]. Using these measures, the canonical analysis aims at maximizing the ratio between among-class variance and within-class variance, resulting in the selection of features that simultaneously exhibit both requirements, i.e., high among-class variance and low within-class variance. Another example of indicator that can be adopted as criterion function is the mutual information, which measures the mutual dependence of two random variables. In the context of feature selection, the mutual information can be used to assess the capability of the considered feature vector $\mathbf{x}_i \in \boldsymbol{\theta}$ to predict the correct class label $y_i \in \Omega\ \forall i = 1, 2, \ldots, l$. To this purpose, a definition of the mutual information that considers the discrete nature of y should be adopted (for deeper insight on feature selection based on mutual information, we refer the reader to [23] and [24]).

B. Search Strategies

In order to select the final subset of features that optimizes the adopted criterion function, a search strategy is needed. The search strategy generates possible solutions of the feature-selection algorithm and compares them by applying the criterion function as a measure of the effectiveness of each solution. An exhaustive search for the optimal solution involves the evaluation and comparison of the criterion function for all

$\binom{n}{m}$ possible combinations of features. This is an intractable problem from a computational point of view, even for low numbers of features [17]. The *branch*-and-*bound* method proposed by Narendra and Fukunaga [14], [15] is a widely used approach to compute the globally optimum solution for monotonic criterion function without explicitly exploring all possible combinations of features. Nevertheless, the computational saving is not sufficient for treating problems with hundreds of features. Therefore, in the case of feature selection for hyperspectral data classification, suboptimal approaches should be adopted. Several suboptimal search strategies have been proposed in the literature. The simplest suboptimal search strategies are the *sequential forward selection* (SFS) and the *sequential backward selection* (SBS) techniques [16], [17]. A serious drawback of both algorithms is that they do not allow backtracking. In the case of the SFS algorithm, once the features have been selected, they cannot be discarded. Similarly, in the case of the SBS search technique, once the features have been discarded, they cannot be added again to the subset of selected features. Two effective sequential search methods are those proposed by Pudil *et al.* [16], namely, the *sequential forward floating selection* (SFFS) method and the *sequential backward floating selection* (SBFS) method. They improve the standard SFS and SBS techniques by dynamically changing the number of features included (SFFS) or removed (SBFS) to the subset of selected features at each step, thus allowing the reconsideration of the features included or removed at the previous steps. Other effective strategies are those proposed in [12], where two search algorithms are presented (i.e., the *steepest ascent* and the *fast constrained search*), which are based on the formalization of the feature-selection problem in the framework of a discrete optimization problem in an adequately defined binary multidimensional space.

An alternative approach to the exploration of the feature space that is relevant to this paper is that based on genetic algorithms (GAs), whose application to feature-selection problems was proposed in [25]. Genetic algorithms exploit an analogy with biology, in which a group of solutions, encoded as *chromosomes*, evolve via natural selection [26]. A standard GA starts by randomly creating an initial population (with a predefined size). Solutions are then combined via a crossover operator to produce offspring, thus expanding the current population. The individuals in the population are evaluated according to the criterion function, and the individuals that less fit such a function are discarded to return the population to its original size. A mutation operator is generally applied in order to increase individuals' variations. The processes of crossover, evaluation, and selection are repeated for a predetermined number of generations (if no other stop criterion is met before) in order to reach a satisfactory solution. Several papers confirmed the effectiveness of GAs for standard feature-selection approaches (e.g., [27]–[29]), also for hyperdimensional feature space. Moreover, as it will be explained later, GAs become particularly relevant for this paper as they are effective when the criterion function involves multiple concurrent terms, and therefore, a multiobjective problem has to be optimized in order to estimate the Pareto-optimal solutions [30], [31].

III. Proposed Feature-Selection Approach

The main idea and novelty of the approach that we propose in this paper is to explicitly consider in the criterion function of the feature-selection process the spatial variability of the features (e.g., of the spectral signatures) on each land-cover class in the investigated scene, together with their discrimination capability. This results in the possibility to select a subset of features that exhibits both high capability to discriminate among different classes and high invariance in the spatial domain. The resulting subset of selected features implicitly improves the generalization capability in the classification process, which results in augmented robustness and accuracy in the classification of hyperspectral images with respect to feature subsets selected with standard methods. This property is particularly relevant when the considered scene is extended over large geographical areas and/or presents considerable intraclass variability of the spectral signatures.

From a formal viewpoint, the aim of the proposed approach is to select the subset $\boldsymbol{\theta}^* \subset \Upsilon$ of m features (with $m < n$) that optimizes a novel criterion function made up of two measures that characterize the following: 1) the capability of the subset of features to discriminate among the considered classes in Ω and 2) the spatial invariance (stationary behavior) of the selected features. The first measure can be evaluated with standard statistical separability indices (as described in the previous section), whereas the spatial invariance property is evaluated according to a novel invariance measure that represents an important contribution of this paper. In particular, we propose two possible methods to evaluate the invariance of a subset of features: 1) a supervised method and 2) a semisupervised method. The supervised method relies on the assumption that the available training set T is made up of two subsets of labeled patterns T_1 and T_2 (such that $T_1 \cup T_2 = T$ and $T_1 \cap T_2 = \varnothing$) collected on disjoint (separate) areas on the ground. This property of the training set is exploited for assessing the spatial variability of the spectral signatures of the land-cover classes. We successively relax this hypothesis by proposing a semisupervised method that does not require the availability of a training subset T_2 spatially disjoint from T_1 (only a standard training set $T \equiv T_1$ acquired in a single area of the scene is needed) and takes advantage of unlabeled samples. This second method is based on an estimation of the distributions of classes in portions of the image separate from T, which is carried out by exploiting the information captured from unlabeled pixels. The final subset of features is selected by jointly optimizing the two concurrent terms of the criterion function. This is done by defining a proper search strategy based on the optimization of a multiobjective problem for deriving the subsets of features that exhibit the best tradeoff between the two concurrent objectives.

In the following sections, we present the proposed supervised and semisupervised methods for the evaluation of the criterion function. Then, we describe the proposed multiobjective search strategy for deriving the final subsets of features that exhibit both the aforementioned properties (which can be assessed with either the supervised or the semisupervised method, depending on the available reference data).

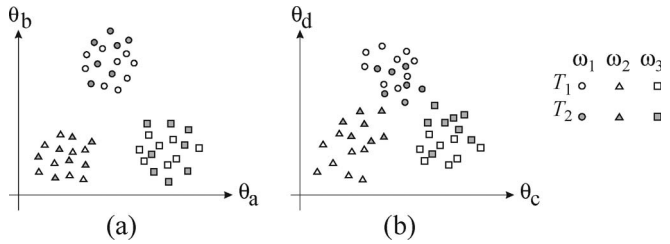

Fig. 1. Examples of feature subsets with different invariant (stationary) behaviors on two disjoint sets T_1 and T_2. (a) Feature subset that exhibits high separability and high invariance properties. (b) Feature subset with high separability on T_1 and high variability between T_1 and T_2.

A. Supervised Formulation of the Proposed Criterion Function

Let us first assume the availability of two subsets of labeled patterns T_1 and T_2 collected on disjoint areas on the ground (thus representing two different realizations of the class distributions). Under this assumption, we can define a novel criterion function that is based on two different terms: 1) a term that measures the class separability (discrimination term) and 2) a term that evaluates the spatial invariance of the investigated features (invariance term).

1) Discrimination Term Δ: This term is based on a standard feature-selection criterion function. In the proposed system, we adopt the definition given in (7), where the term $\Delta(\boldsymbol{\theta})$ evaluates the average measure of distance between all couples of class distributions $p(\mathbf{x}\,|\,\omega_i)$ and $p(\mathbf{x}\,|\,\omega_j)$ $\forall \omega_i, \omega_j \in \Omega$ and $i < j$. This term depends on the selected subset $\boldsymbol{\theta}$ of features, and the subset of m features $\boldsymbol{\theta}^*$ that maximizes this distance results in the best potential for discriminating land-cover classes in the area modeled by the training samples. It is important to note that the evaluation of the aforementioned term is usually performed by assuming Gaussian distributions of classes for calculating the statistical distance $S_{ij}(\boldsymbol{\theta})$. Under this assumption, also in the presence of two disjoint training sets, it is preferable to evaluate the discrimination term by considering only one subset of the training set (T_1 or T_2). This can be explained by considering that mixing up the two available training subsets T_1 and T_2 would result in mixing together two different realizations of the feature distributions, which, from a theoretical perspective, cannot be correctly modeled with Gaussian (monomodal) distributions.

2) Invariance Term P: In order to introduce the invariance term, let us first consider Fig. 1. This figure shows a qualitative example in a 2-D feature space of two subsets of features that exhibit different behavior of the samples extracted from different portions of a scene. The features of Fig. 1(a) present good capability to separate the class clusters and also exhibit high invariance on the two considered training sets. These properties allow the supervised algorithm to derive a robust classification rule, resulting in the capability to accurately classify samples that can be localized in both areas from which the samples of T_1 and T_2 are extracted. On the contrary, the features adopted in Fig. 1(b) exhibit good separability properties but low invariance. This feature subset leads the supervised learner to derive a classification rule that is not robust, resulting in poor classification accuracy in spatially disjoint areas.

The different behavior between the feature subsets in Fig. 1(a) and (b) can be modeled by considering the distance between the clusters that refer to the same land-cover class in the two disjoint training sets T_1 and T_2. Thus, we can introduce a novel term to explicitly measure the invariance (stationary behavior) of features on each class in the investigated image. It can be defined as

$$\mathrm{P}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{C} P^{T_1}(\omega_i) P^{T_2}(\omega_i) S_{ii}^{T_1 T_2}(\boldsymbol{\theta}) \tag{8}$$

where $S_{ii}^{T_1 T_2}$ is a statistical distance measure between the distributions $p^{T_r}(\mathbf{x}\,|\,\omega_i)$, $r = 1, 2$, of the class ω_i computed on T_1 and T_2, and $P^{T_r}(\omega_i)$ represents the prior probability of the class ω_i in T_r, $r = 1, 2$. This term evaluates the average distance between the distributions of the same class in different portions of the scene (i.e., on the two disjoint subsets of the training set). Unlike for $\Delta(\boldsymbol{\theta})$, we expect that a good (i.e., robust) subset of features should minimize the value of $\mathrm{P}(\boldsymbol{\theta})$. The computation of $\mathrm{P}(\boldsymbol{\theta})$ can be easily extended to more than two training subsets if labeled data collected on more than two disjoint regions are available. In the general case, when R spatially disjoint training sets are available, the invariance term can be defined as follows:

$$\mathrm{P}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{a=1}^{R}\sum_{b>a}^{R}\sum_{i=1}^{C} P^{T_a}(\omega_i) P^{T_b}(\omega_i) S_{ii}^{T_a T_b}(\boldsymbol{\theta}). \tag{9}$$

The process of selection of features that jointly optimize the discrimination term $\Delta(\boldsymbol{\theta})$ and the invariance term $\mathrm{P}(\boldsymbol{\theta})$ will be described in Section III-C.

B. Semisupervised Evaluation of the Criterion Function (Invariance Term Estimation)

The collection of labeled training samples on two (or more) spatially disjoint areas from the site under investigation can be difficult and/or very expensive. This may compromise the applicability of the proposed supervised method in some real classification applications. In order to overcome this possible problem, in this section, we propose a semisupervised technique to estimate the invariance term defined in (8), which does not require the availability of a disjoint training subset T_2. Here, we only assume that a training set T_1 is available, and we consider a set of unlabeled pixels $U = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathcal{I}$ (subset of the original image $\mathcal{I}$) that should satisfy two requirements: 1) U contains samples of all the considered classes, and 2) samples in U should be taken from portions of the scene separated from those on which the training samples T_1 are collected. The set U can be defined in either of the following ways: 1) by manually selecting clusters of pixels on a portion of the considered scene; 2) by randomly subsampling a set of pixels; or 3) by considering the whole image $\mathcal{I}$. It is worth noting that, in the proposed algorithm, the labels of classes are not required. We only assume that the unlabeled samples are collected according to a strategy that can implicitly consider all classes present in the scene.

The method is based on the semisupervised estimation of the terms $P^U(\omega_i)$ and $p^U(\mathbf{x}\,|\,\omega_i)$, $\omega_i \in \Omega$, which, in this case, characterize the prior probabilities and the conditional probability density functions in the disjoint area corresponding to the pixels in U, respectively. The distribution of the samples in U can be described by the following mixture model:

$$p^U(\mathbf{x}) = \sum_{i=1}^{C} P^U(\omega_i) p^U(\mathbf{x}\,|\,\omega_i). \tag{10}$$

We assume that $P^U(\omega_i)$ and $p^U(\mathbf{x}\,|\,\omega_i)$ are not known, while $p^U(\mathbf{x})$ is given from the data distribution. However, despite the expected variability, for each class $\omega_i \in \Omega$, the initial values of both the prior probability $P^U(\omega_i)$ and the conditional density function $p^U(\mathbf{x}\,|\,\omega_i)$ can be roughly approximated by the prior and the conditional density function in T_1, i.e.,

$$P^{U,0}(\omega_i) = P^{T_1}(\omega_i) \quad p^{U,0}(\mathbf{x}\,|\,\omega_i) = p^{T_1}(\mathbf{x}\,|\,\omega_i). \tag{11}$$

The problem can be addressed by estimating the parameter vector $\mathbf{J} = [P^U(\omega_i), \delta_i]_{i=1}^{C}$, where each component δ_i represents the vector of parameters that characterize the density function $p^U(\mathbf{x}\,|\,\omega_i)$, which, given its dependence from δ_i, can be rewritten as $p^U(\mathbf{x}\,|\,\omega_i, \delta_i)$. The components of $\mathbf{J}$ can be estimated by maximizing the pseudo log-likelihood function $L[p^U(\mathbf{x})]$ defined as

$$L\left[p^U(\mathbf{x})\,|\,\mathbf{J}\right] = \sum_{j=1}^{m} \log\left\{\sum_{i=1}^{C} P^U(\omega_i\,|\,\mathbf{J}) p^U(\mathbf{x}\,|\,\omega_i, \mathbf{J})\right\}. \tag{12}$$

The maximization of the log-likelihood function can be obtained with the expectation–maximization (EM) algorithm [32]. The EM algorithm consists of two main steps: an expectation step and a maximization step. The two steps are iterated, so that the value of the log-likelihood function $L[p^U(\mathbf{x})]$ increases at each iteration, until a local maximum is reached. For simplicity, let us consider that all the classes $\omega_i \in \Omega$ are Gaussian distributed. Under this assumption, the density function associated with each class ω_i can be completely described by the mean vector μ_i^U and the covariance matrix Σ_i^U, $i = 1, \ldots, C$. Therefore, the parameter vector to be estimated becomes

$$\mathbf{J} = \left[P^U(\omega_i), \mu_i^U, \Sigma_i^U\right]_{i=1}^{C}. \tag{13}$$

It can be proven that the equations to be used at iteration $s+1$ for estimating the statistical terms associated with a generic class ω_i are the following [3], [32], [33]:

$$P^{U,s+1}(\omega_i) = \frac{1}{m} \sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i) p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)} \tag{14}$$

$$\left[\mu_i^U\right]^{s+1} = \frac{\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i) p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)} \mathbf{x}_j}{\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i) p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}} \tag{15}$$

$$\left[\Sigma_i^U\right]^{s+1} = \frac{\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i) p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)} \left\{\mathbf{x}_j - \left[\mu_i^U\right]^{s+1}\right\}^2}{\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i) p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}} \tag{16}$$

where the superscripts s and $s+1$ refer to the values of the parameters at the sth and $s+1$th iterations, respectively. The estimates of the statistical parameters that describe the class distributions in the disjoint areas are obtained starting from the initial values of the parameters [see (11)] and iterating (14)–(16) up to convergence. An important aspect of the EM algorithm concerns its convergence properties. It is not possible to guarantee that the algorithm will converge to the global maximum of the log-likelihood function, although convergence to a local maximum can be ensured. A detailed description of the EM algorithm is beyond the scope of this paper, so we refer the reader to the literature for a more detailed analysis of such an algorithm and its properties [3], [32]. The final estimates obtained at convergence for each class $\omega_i \in \Omega$, i.e., $\hat{P}^U(\omega_i)$, and $\hat{p}^U(\mathbf{x}\,|\,\omega_i)$ (which depend on the estimated parameters $\hat{\mu}_i^U$ and $\hat{\Sigma}_i^U$) can be used in place of $P^{T_2}(\omega_i)$ and $p^{T_2}(\mathbf{x}\,|\,\omega_i)$ to estimate the invariance term $\hat{\mathrm{P}}(\boldsymbol{\theta})$ for each subset of features $\boldsymbol{\theta}$ considered. Thus, the semisupervised estimation of the invariance term becomes

$$\hat{\mathrm{P}}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{C} P^{T_1}(\omega_i) \hat{P}^U(\omega_i) \hat{S}_{ii}^{T_1 U}(\boldsymbol{\theta}). \tag{17}$$

The discrimination term $\Delta(\boldsymbol{\theta})$ can be calculated as in (7) with no difference with respect to the supervised method.

It is worth noting that, depending on the adopted set U of unlabeled pixels, the estimation of the prior probabilities and the class-conditional densities can reflect with different degree of accuracy the true values. In particular, the estimation of the elements of the covariance matrices $\hat{\Sigma}_i^U$, $i = 1, \ldots, C$, may become critical in some cases when the number of classes is high. Thus, in these cases, since small fluctuations in the accuracy of the estimation of the covariance terms $\hat{\Sigma}_i^U$, $i = 1, \ldots, C$, can strongly affect the invariance term values, the estimation of the invariance term can be simplified in the following ways: 1) by assuming that the covariance matrix is diagonal and 2) by considering only the first-order statistical moment (thus neglecting the second-order moments) for the evaluation of the statistical distance $\hat{S}_{ii}^{T_1 U}(\boldsymbol{\theta})$.

C. Proposed Multiobjective Search Strategy

Given the proposed criterion function that is made up of the discrimination term $\Delta(\boldsymbol{\theta})$ and the invariance term $\mathrm{P}(\boldsymbol{\theta})$ (which, depending on the available reference data, can be evaluated with the supervised or unsupervised methods, as described in the previous two sections), we address now the problem of defining a search strategy to select the subset (or the subsets) of features that (jointly) optimize(s) the two defined measures. To this purpose, one can define a global optimization function as

$$V(\boldsymbol{\theta}) = \Delta(\boldsymbol{\theta}) + K \cdot f\left[\mathrm{P}(\boldsymbol{\theta})\right] \tag{18}$$

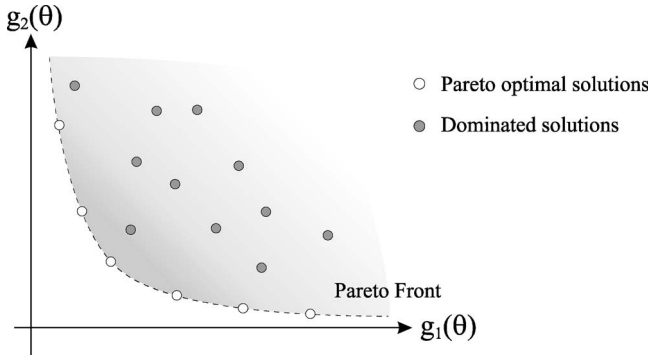

Fig. 2. Example of Pareto-optimal solutions and dominated solution in a two-objective search space.

where K tunes the tradeoff between discrimination ability and invariance of the selected subset of features and f is the monotonic decreasing function of $\mathrm{P}(\boldsymbol{\theta})$. The subset $\boldsymbol{\theta}^*$ of m features for which $V(\boldsymbol{\theta})$ has the maximum value represents the solution to the considered problem.

Nevertheless, the aforementioned formulation of the problem has two drawbacks: 1) the obtained criterion function is not monotonic (and thus, effective search algorithms based on this property cannot be used), and 2) the definition of f and K (which should be carried out empirically) affects significantly the final result. To overcome these drawbacks, we modeled this problem as a multiobjective minimization problem, where the multiobjective function $\mathbf{g}(\boldsymbol{\theta})$ is made up of two different (and possibly conflicting) objectives $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$, which express the discrimination ability $\Delta(\boldsymbol{\theta})$ among the considered classes and the spatial invariance $\mathrm{P}(\boldsymbol{\theta})$ of the subset of features $\boldsymbol{\theta}$, respectively. The multiobjective problem can therefore be formulated as follows:

$$\min_{|\boldsymbol{\theta}|=m} \{\mathbf{g}(\boldsymbol{\theta})\},$$
$$\text{where} \quad \mathbf{g}(\boldsymbol{\theta}) = [g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})] = [-\Delta(\boldsymbol{\theta}), \mathrm{P}(\boldsymbol{\theta})] \quad (19)$$

where $|\boldsymbol{\theta}|$ is the cardinality of the subset $\boldsymbol{\theta}$, i.e., the number of features m to be selected from the originally available n. This problem is solved in order to obtain a set of Pareto-optimal solutions O^* instead of a single optimal one. In greater detail, a solution $\boldsymbol{\theta}^*$ is said to be Pareto optimal if it is not dominated by any other solution in the search space, i.e., there is no other $\boldsymbol{\theta}$ such that $g_i(\boldsymbol{\theta}) \leq g_i(\boldsymbol{\theta}^*)$ $(\forall i = 1, 2)$ and $g_j(\boldsymbol{\theta}) < g_j(\boldsymbol{\theta}^*)$ for at least one j $(\forall j = 1, 2)$. This means that $\boldsymbol{\theta}^*$ is Pareto optimal if there exists no other subset of features $\boldsymbol{\theta}$ that would decrease an objective without simultaneously increasing the other one (Fig. 2 clarifies this concept with a graphical example). The set O^* of all optimal solutions is called Pareto-optimal set. The plot of the objective function of all solutions in the Pareto-optimal set is called Pareto front $\mathrm{PF}^* = \{\mathbf{g}(\boldsymbol{\theta}) | \boldsymbol{\theta} \in O^*\}$. Because of the complexity of the search space, an exhaustive search of the set of optimal solutions O^* is unfeasible. Thus, instead of identifying the true set of optimal solutions, we aim to estimate a set of nondominated solutions $\hat{O}^*$ with objective values as close as possible to the Pareto front. This estimation can be achieved with different multiobjective optimization algorithms (e.g., multiobjective evolutionary algorithms).

The main advantage of the multiobjective approach is that it avoids aggregation of metrics capturing multiple objectives into a single measure. On the contrary, it allows one to effectively identify different possible tradeoffs between the values of $\Delta(\boldsymbol{\theta})$ and $\mathrm{P}(\boldsymbol{\theta})$. This results in the possibility to evaluate in a more flexible way the tradeoffs between discrimination ability among classes and spatial invariance of each feature subset and to identify the subsets of features that simultaneously exhibit both properties. In particular, we expect that the most robust subsets of features (which will result in the best generalization capability of the classification system) are represented by the solutions that are localized close to the knee of the estimated Pareto front (or the solutions closest to the origin of the search space).

IV. Data Set Description and Design of Experiments

In order to assess the effectiveness of the presented approach (with both the proposed supervised and semisupervised methods), we carried out several experiments on a hyperspectral image acquired over an extended geographical area. We considered a data set that is increasingly used as a benchmark in the literature and consists of data acquired by the Hyperion sensor of the Earth Observing 1 (EO-1) satellite in an area of the Okavango Delta, Botswana. The Hyperion sensor on EO-1 acquired the hyperspectral image with a spatial resolution of 30 m over a 7.7-km strip in 242 bands. Uncalibrated and noisy bands that cover water absorption range of the spectrum were removed, and the remaining 145 bands were given as input to the feature-selection technique. For more details on this data set, we refer the reader to [34]. The labeled reference samples were collected on two different and spatially disjoint areas (Area 1 and Area 2), thus representing possible spatial variabilities of the spectral signatures of classes. The samples taken on the first area were partitioned into a training set T_1 and a test set TS_1 by random sampling (these sets represent similar realizations of the spectral signatures of classes). The samples taken on the second area were used to derive a training set T_2 and a test set TS_2 according to the same procedure used for the samples of the first considered area (these two sets present possible variability in class distributions with respect to the first two sets). The numbers of labeled reference samples for each set and class are reported in Table I. After preliminary experiments were carried out in order to understand the size of the subset of features that led to the saturation of the classification accuracies, we performed different experiments (with both the supervised and semisupervised methods) by varying the size m of the selected subset of features in a range between 6 and 14 with step 2. The obtained subsets of features were used to perform the classification with a Gaussian maximum-likelihood (ML) classifier. The training of the ML classifier (estimation of Gaussian parameters for class-conditional densities) was carried out using the training set T_1. We compared the classification accuracies obtained on both test sets TS_1 and TS_2 performing the feature selection with the following: 1) the proposed approach with the supervised method for the estimation of the invariance term; 2) the proposed semisupervised method

TABLE I
NUMBER OF TRAINING (T_1 AND T_2) AND TEST (TS_1 AND TS_2) PATTERNS ACQUIRED IN THE TWO SPATIALLY DISJOINT AREAS

Class	Number of samples			
	Area 1		Area 2	
	T_1	TS_1	T_2	TS_2
Water	69	57	213	57
Hippo grass	81	81	83	18
Floodplain grasses1	83	75	199	52
Floodplain grasses2	74	91	169	46
Reeds1	80	88	219	50
Riparian	102	109	221	48
Firescar2	93	83	215	44
Island interior	77	77	166	37
Acacia woodlands	84	67	253	61
Acacia shrublands	101	89	202	46
Acacia grasslands	184	174	243	62
Short mopane	68	85	154	27
Mixed mopane	105	128	203	65
Exposed soil	41	48	81	14
Total	1242	1252	2621	627

for estimating the invariance term; and 3) a standard feature-selection technique that considers only the discrimination term.

The experiments with the supervised feature-selection method were carried out by considering the training set T_1 for the evaluation of the discrimination term $\Delta(\boldsymbol{\theta})$ and both T_1 and T_2 for the evaluation of the invariance term $\mathrm{P}(\boldsymbol{\theta})$. In our implementation, we adopted the JM distance (under the Gaussian assumption for the distribution of classes) as a statistical distance measure for both considered terms. The second set of experiments was carried out with the proposed semisupervised feature-selection method. In these experiments, we considered the training set T_1 for the evaluation of the discriminative term $\Delta(\boldsymbol{\theta})$, while the invariance term $\hat{\mathrm{P}}(\boldsymbol{\theta})$ was estimated from T_1 and the samples of T_2, which were used without their class label information as set U. For simplicity, we considered only the first-order moment to evaluate the statistical distance $\hat{S}_{ii}^{T_1 U}(\boldsymbol{\theta})$ (see the discussion in Section II-A). The standard feature selection was performed by selecting the subsets of features that maximize the JM distance on the training set T_1 with a (mono-objective) GA. Note that we did not mix up the two training sets T_1 and T_2 for both training the ML classifiers and evaluating the discrimination term, as the Gaussian approximation is no more reasonable for the two different Gaussian realizations of each class in T_1 and T_2 (see Section II-A).

In order to solve the defined two-objective minimization problem for the proposed methods (i.e., estimating the Pareto-optimal solutions), we implemented a modification of the "non-dominated sorting in genetic algorithm II" (NSGA-II) [31]. The original algorithm was modified in order to avoid solutions with multiple selections of the same feature. This has been accomplished by changing the random initialization of the chromosome population and by modifying the crossover and mutation operators. In all the experiments, the population size was set equal to 100, and the maximum number of generations was set equal to 50. The classification was carried out using all combinations of features $\hat{\boldsymbol{\theta}}^* \in \hat{O}^*$ that lie on the estimated Pareto front, and the subset $\hat{\boldsymbol{\theta}}^*$ that resulted in the highest accuracy on the disjoint test set TS_2 was finally selected. For the mono-objective GA, we adopted the same values for both the population size and the maximum number of generations as for the multiobjective GA.

V. EXPERIMENTAL RESULTS

A. Results With the Supervised Method for the Estimation of the Invariance Term

We first present the experimental results obtained with the proposed supervised method that allows us to derive important considerations about the validity of the proposed approach with respect to the standard one. In order to show the shortcomings of standard feature-selection algorithms for the classification of hyperspectral images, Fig. 3 shows the graphs of the accuracy obtained by the ML classifier on the adjoint (TS_1) and disjoint (TS_2) test sets versus the values of the discrimination term $\Delta(\boldsymbol{\theta})$ for different subsets of features. For the reported graphs, we used the solutions on the Pareto front estimated by the modified NSGA-II algorithm applied to the multiobjective minimization problem in (19), in the cases of six and eight features (these two cases are selected as examples; the other considered cases led to similar results). From this figure, it is possible to observe that the accuracy on TS_1 increases when the discrimination term increases, whereas the accuracy on TS_2 increases only until a certain value and then it decreases. Therefore, the simple maximization of the discrimination term (as standard approaches do) can lead to an overfitting phenomenon, which results in poor generalization capabilities, i.e., low capability to discriminate and correctly classify the land-cover classes in areas of the scene different from that associated with the collected training data. This confirms the significant variability of the spectral signature of classes in hyperspectral images.

The aim of the proposed approach is to overcome this problem. Let us now consider Fig. 4 that shows the Pareto fronts estimated by the proposed approach (employing the modified NSGA-II algorithm) in the cases of the selection of six and eight features. This figure represents the information of the kappa coefficient of accuracy, which is obtained by the classification of the test sets TS_1 and TS_2 with the considered subset of features $\hat{\boldsymbol{\theta}}^*$, as the color of the point, according to the reported color scale bar. The diagrams in Fig. 4(a)–(c) show that, for the classification of TS_1, the solutions with higher discrimination capability [lower values of $-\Delta(\boldsymbol{\theta})$] result in better accuracies. This behavior reveals (as expected) that only the discrimination term is important for selecting the most effective feature subset for the classification of pixels acquired in a similar area of pixels in T_1 (in these conditions, training and test patterns represent the same realization of the statistical distributions of classes). On the contrary, the diagrams in Fig. 4(b)–(d) show that the most accurate solutions for the classification of the spatially disjoint samples of TS_2 (which result in the highest kappa coefficient of accuracy) are located in a middle region, close to the knee of the estimated Pareto front. This confirms the importance of the invariance term, and that tradeoff solutions between the two competing objectives $\Delta(\boldsymbol{\theta})$ and $\mathrm{P}(\boldsymbol{\theta})$ should be identified in order to select the subset of features that leads to

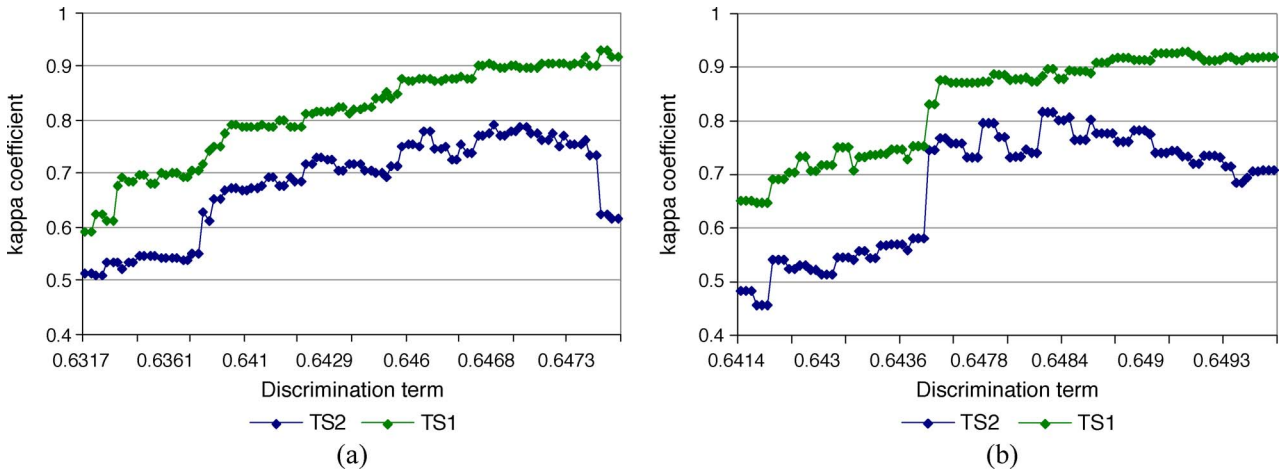

Fig. 3. Behaviors of the kappa coefficients of accuracy on the test sets TS_1 and TS_2 versus the values of the discrimination term $\Delta(\theta)$. Cases of (a) six and (b) eight features.

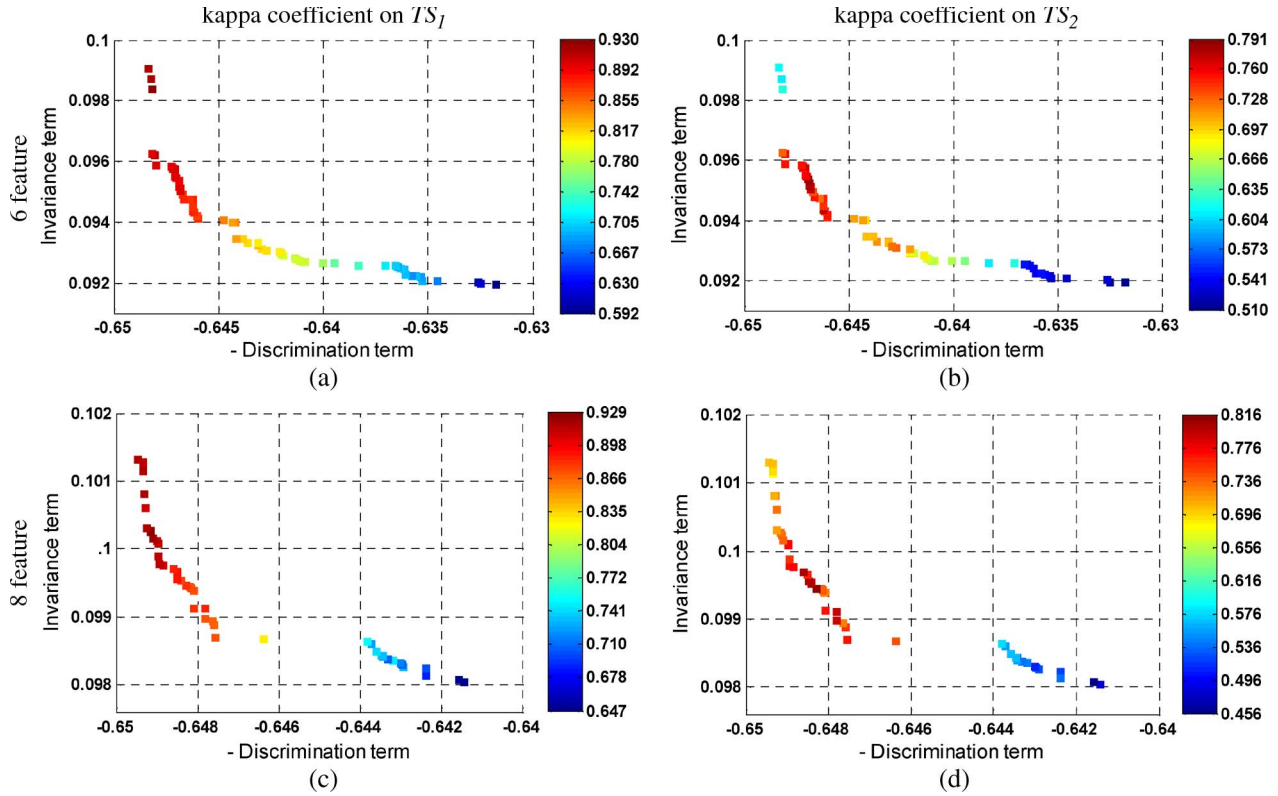

Fig. 4. Pareto fronts estimated by the proposed approach with the supervised method. (a) and (b) Six-feature case. (c) and (d) Eight-feature case. The color indicates the kappa coefficient of accuracy on (a)–(c) TS_1 and (b)–(d) TS_2 according to the reported color scale bar.

better generalization capabilities and, thus, higher classification accuracy in areas of the hyperspectral image different from the training one.

Table II reports the comparison of the classification accuracies obtained on TS_1 and TS_2 by selecting the subset of features with the proposed multiobjective supervised and semisupervised methods, as well as the standard method. From this table, it is possible to observe that the obtained accuracy on the disjoint test set TS_2 are, in general, significantly lower that those obtained on the adjoint test set TS_1, confirming the presence of consistent variability in the spatial domain of the spectral signatures of the classes. This phenomenon severely challenges the generalization capability of the classification system. Nevertheless, we can observe that, for all considered cases, the proposed multiobjective feature-selection methods allowed one to significantly increase the accuracy on the test set TS_2 with respect to the standard method, while the accuracy on the adjoint test set TS_1 only slightly decreased. On average, the proposed supervised method resulted in an increase of the classification accuracy on the disjoint test set of 21.3% with respect to the standard approach, while it slightly decreased by 4.2% the accuracy on the adjoint test set.

The obtained results clearly confirm that the proposed approach is effective in exploiting the information of the two

TABLE II
KAPPA COEFFICIENT OF ACCURACIES OBTAINED BY THE ML CLASSIFIER WITH THE FEATURES SELECTED BY THE PROPOSED SUPERVISED AND SEMISUPERVISED METHODS AND THE STANDARD APPROACH

Number of features	Kappa coefficient of Accuracy on (Test Set TS_2)			Kappa coefficient of Accuracy on (Test Set TS_1)		
	Proposed Semisup. Method	Proposed Supervised method	Standard method	Proposed Semisup. Method	Proposed Supervised method	Standard method
6	0.780	0.791	0.580	0.894	0.902	0.931
8	0.767	0.816	0.577	0.906	0.884	0.939
10	0.777	0.813	0.592	0.938	0.912	0.942
12	0.722	0.808	0.591	0.914	0.900	0.954
14	0.739	0.799	0.625	0.912	0.913	0.953
Average	0.757	0.805	0.593	0.913	0.902	0.944

distinct available training sets to select subsets of robust and invariant features, which can improve the generalization capabilities of the classification system. We further observe that very few spectral channels (6–14 bands out of the originally available 145) are sufficient for effectively representing and discriminating the considered information classes, thus significantly reducing the problems associated with the Hughes phenomenon. The computational cost of the proposed supervised method is comparable with that of the standard mono-objective algorithm. In our experiments, which were carried out on a personal computer mounting an Intel Pentium D processor at 3.4 GHz and a 2-GB DDR2 RAM, the feature selection with the supervised multiobjective method took an average time of about 4 min, while the standard method took about 3 min. This is due to the fact that the evaluation of the discrimination term $\Delta(\boldsymbol{\theta})$ (which has to be computed also with standard feature-selection methods) requires a computational cost that is proportional to $C(C-1)/2$, while the introduced invariance term $\mathrm{P}(\boldsymbol{\theta})$ has a computational cost that is proportional to C. Therefore, the additional cost due to the evaluation of the new term becomes lesser and lesser when the number of classes increases.

B. Results With the Semisupervised Method for the Estimation of the Invariance Term

Often, in real applications, a disjoint training set T_2 is not available to the user, and the proposed supervised method cannot be used. In these cases, the semisupervised approach can be adopted. It is worth noting that, from the perspective of the semisupervised method, the supervised technique represents an upper bound of the accuracy and generalization ability that can be obtained (if the same samples with and without labels are considered). Thus, in this case, the results presented in the previous section can be seen as the best performances that can be obtained on the considered samples.

As expected, the semisupervised method led to accuracies that were slightly lower than that of the supervised method, but it still maintained a significant improvement with respect to the traditional approach. On average, the semisupervised method increased the classification accuracy on TS_2 by 16.4% with respect to the standard feature-selection method, while it decreased the accuracy on TS_1 by 3.1%. The small decrease in performances with respect to those obtained by the supervised method is due to the approximate estimation of the invariance term carried out with the EM algorithm, which cannot ensure convergence to the optimal solution. However, the semisupervised method has the very important advantage to considerably increase the generalization capabilities of the classification systems with respect to the traditional approach without requiring additional reference data. The computation cost of this method is slightly higher with respect to the standard method, because of the time required by the EM algorithm to perform the estimation necessary to evaluate the invariance term. In our experiments, the average time for the feature selection with the semisupervised approach was about 60 min (15 times more than that with the supervised method).

VI. CONCLUSION

In this paper, we presented a novel feature-selection approach to the classification of hyperspectral images. The proposed approach aimed at selecting subsets of features that exhibited, at the same time, high discrimination ability and high spatial invariance, improving the robustness and the generalization properties of the classification system with respect to standard techniques. The feature selection was accomplished by defining a multiobjective criterion function that considered the evaluation of both a standard separability measure and a novel term that measured the spatial invariance of the selected features. In order to assess the invariance in the scene of the feature subset, we proposed both a supervised method (assuming the availability of training samples acquired in two or more spatially disjoint areas) and a semisupervised method (which required only a standard training set acquired in a single area of the scene and which exploited the information of unlabeled pixels in portions of the scene spatially disjoint from the training areas). The multiobjective problem was solved by an evolutionary algorithm for the estimation of the set of Pareto-optimal solutions.

Experimental results showed that the proposed feature-selection approach selected subsets of the original features that sharply increased the classification accuracy on disjoint test samples, while it slightly decreased the accuracy on the adjoint test set with respect to standard methods. This behavior confirms that the proposed approach results in augmented generalization capability of the classification system. In this regard, we would like to stress the importance of evaluating the accuracy on a disjoint test set, because this allows one to estimate the

accuracy in the classification of the whole considered image. In particular, the proposed supervised method is effective in exploiting the information of the two available training sets, and the proposed semisupervised method can significantly increase the generalization capabilities of the classification system without requiring additional reference data with respect to traditional feature-selection algorithms. This can be achieved at the cost of an acceptable additional computational time.

It is important to note that the proposed approach is defined in a general way, thus allowing different possible implementations. For instance, the discrimination and invariance terms can be evaluated considering statistical distance measures that are different from those adopted in our experimental analysis, and other multiobjective optimization algorithms can be adopted as search strategy for estimating the Pareto-optimal solutions. This general definition of the approach results in the possibility of further developing the implementation that we adopted for our experimental analysis. As an example, as future developments of this paper, the proposed approach could be integrated with classification algorithms that are different from the adopted ML classifier, e.g., the SVM and/or other kernel-based classification techniques, for further improving the accuracy of the classification system. In addition, we think that the overall classification system can be further improved by jointly exploiting the proposed feature-selection approach and a semisupervised classification technique for a synergic and complete exploitation of the unlabeled-sample information.

ACKNOWLEDGMENT

The authors would like to thank Prof. M. Crawford (Purdue University, West Lafayette, IN) for kindly providing the data set used in the experimental part of this paper. They would also like to thank Dr. A. Boni and Dr. A. Marconato for the valuable discussion on multiobjective optimization.

REFERENCES

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

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

[3] 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.

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

[5] M. Chi and L. Bruzzone, "Semi-supervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pt. 2, pp. 1870–1880, Jun. 2007.

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

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

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

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

[10] P. W. Mausel, W. J. Kramber, and J. K. Lee, "Optimum band selection for supervised classification of multispectral data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 1, pp. 55–60, Jan. 1990.

[11] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.

[12] S. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[13] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1318–1321, Nov. 1995.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[15] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, no. 9, pp. 917–922, Sep. 1977.

[16] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods for feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.

[17] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[18] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4/5, pp. 411–430, May/Jun. 2000.

[19] S. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, Feb. 2007.

[20] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[21] M. N. Do and M. Vetteri, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[22] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[23] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[24] B. Guo, R. I. Damper, S. R. Gunn, and J. D. B. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classification," *Pattern Recognit.*, vol. 41, no. 5, pp. 1653–1662, May 2008.

[25] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognit. Lett.*, vol. 10, no. 5, pp. 335–347, Nov. 1989.

[26] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[27] F. Z. Brill, D. E. Brown, and W. N. Martin, "Fast genetic selection of features for neural network classifiers," *IEEE Trans. Neural Netw.*, vol. 3, no. 2, pp. 324–328, Mar. 1992.

[28] J. H. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 44–49, Mar./Apr. 1998.

[29] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, no. 2, pp. 164–171, Jul. 2000.

[30] H. C. Lac and D. A. Stacey, "Feature subset selection via multi-objective genetic algorithm," in *Proc. Int. Joint Conf. Neural Netw.*, Montreal, QC, Canada, Jul. 31–Aug. 4, 2005, pp. 1349–1354.

[31] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[33] L. Bruzzone and D. F. 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.

[34] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

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

From 1998 to 2000, he was a Postdoctoral Researcher 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 and the Head of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science. He teaches remote sensing, pattern recognition, radar, and electrical communications. His current research interests include 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. He is an Evaluator of project proposals for many different governments (including the European Commission) and scientific organizations. He is the author or coauthor of 74 scientific publications in referred international journals, more than 140 papers in conference proceedings, and seven book chapters.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was a recipient of the Recognition of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Best Reviewers in 1999 and was a Guest Editor of a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of multitemporal remote sensing images (November 2003). He was the General Chair and Cochair of the First and Second IEEE International Workshops on the Analysis of Multitemporal Remote Sensing Images (MultiTemp) and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. From 2004 to 2006, he was an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a Referee for many international journals and has served on the Scientific Committees of several international conferences. He is a member of the Managing Committee of the Italian Inter-University Consortium for Telecommunications and a member of the Scientific Committee of the India–Italy Center for Advanced Research. Since 2009, he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society. He is also a member of the International Association for Pattern Recognition and the Italian Association for Remote Sensing (AIT).

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

He is with the Remote Sensing Group, Department of Information Engineering and Computer Science, University of Trento. His current research interests include remote sensing, image classification, pattern recognition, and machine learning. He is a Referee for the *Canadian Journal of Remote Sensing* and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.