A Novel Technique for Sub-pixel Image Classification Based on Support Vector Machine

Francesca Bovolo, *IEEE Member*, Lorenzo Bruzzone, *IEEE Fellow Member*, Lorenzo Carlin

Dept. of Engineering and Computer Science, University of Trento, Via Sommarive, 14

I-38123, Povo, Trento, Italy, Phone: +39-0461-882056, Fax: +39-0461-882093

E-mail: lorenzo.bruzzone@ing.unitn.it

Abstract – **This paper presents a novel support vector machine classifier designed for sub-pixel image classification (pixel/spectral unmixing). The proposed classifier generalizes the properties of SVMs to the identification and modeling of the abundances of classes in mixed pixels by using fuzzy logic. This results in the definition of a fuzzy-input fuzzy-output support vector machine (F^2SVM) classifier that can: i) process fuzzy information given as input to the classification algorithm for modeling the sub-pixel information in the learning phase of the classifier, and ii) provide a fuzzy modeling of the classification results, allowing a relation many-to-one between classes and pixels. The presented binary F^2SVM can address multicategory problems according to two strategies: the fuzzy one-against-all (FOAA) and the fuzzy one-against-one strategies (FOAO). These strategies generalize to the fuzzy case techniques based on ensembles of binary classifiers used for addressing multicategory problems in crisp classification problems. The effectiveness of the proposed F^2SVM classifier is tested on three problems related to image classification in presence of mixed pixels having different characteristics. Experimental results confirm the validity of the proposed sub-pixel classification method.**

I. Introduction

Image classification is an important and challenging task in various application domains, including biomedical imaging, biometry, video-surveillance, industrial visual inspection, and remote sensing. The main objective of image classification is to assign to each pixel (or each object extracted from the image with a proper segmentation procedure) a semantic label associated with one of the information classes that characterize the analyzed scene. Usually image classification is addressed under the assumption that a given pixel can belong only to one class [1]. However, in some real problems, the geometrical resolution of the sensor is not sufficient to guarantee that the radiance measurement associated with a pixel is the contribution given from a single information class (object) in the scene. On the contrary, in many cases, the pixel measurement is given from a mixture of the reflectance of patterns which belong to different classes located in the same resolution cell of the sensor. This is the case of medium resolution remote sensing images [2], in which it is quite common that a pixel is associated with the radiometric response of more than one kind of land-cover class. Another example is related to biological applications, where multispectral fluorescence microscopy can be used for the identification of different co-localized fluorescent molecules that can be associated with the same resolution cell of the sensor [3]-[5]. From a slightly different perspective, the spectral unmixing problem has a high importance also in the analysis of hyperspectral images. The very high spectral resolution of this kind of data allows one a detailed characterization of the spectral signatures of the objects present in the investigated scene. This makes it possible to identify the abundances of constituents of a given material within the resolution cell. In these conditions conventional crisp (hard) classification methods preclude a proper analysis of the image as: i) it is not possible to model the sub-pixel abundances of each class in output from the classifier; and ii) the training phase of the classifier is affected from the use of mixed pixels (and not class endmembers) that provide unreliable information on the reflectance of the represented class. In this scenario, the image classification problem should be solved with a sub-pixel classification approach, where a pixel can be associated with multiple classes with different membership grades.

In the literature, two main kinds of approaches have been considered for solving sub-pixel classification problems: the ones based on linear models and the ones based on non-linear models [6]. Unmixing methods based on linear models assume that the radiance measured in presence of more than one information class in a resolution cell is a linear mixture. The linearity hypothesis typically holds when endmembers are spatially localized in specific areas within the resolution cell of the sensor and do not interfere among each other [6],[7]. These methods estimate the abundances of classes by deriving the parameters of the linear model for each analyzed pixel [6],[7]. Approaches based on nonlinear models assume that the radiance measured in presence of more than one information class in a resolution cell is a nonlinear mixture. The nonlinearity hypothesis typically holds when endmembers are scattered within the spatial resolution cell of the sensor and interfere among each other [7]. Non-linear methods (which are more complex but in many applications also more adequate to model the nature of the sub-pixel radiance [6]) can be based on the definition of parametric nonlinear models [6],[7] or on the use of distribution free machine learning techniques [8]. This last approach is effective when it is possible to rely on a training set that allows one exploiting the powerful properties of machine learning for extracting the model of the mixture directly from the observed data.

Linear and nonlinear techniques for sub-pixel image classification can be implemented according to supervised classification paradigms based on fuzzy sets. On the one hand, fuzzy classification models can employ fuzzy set theoretic principles to perform a soft partition of the input space where continuous class memberships (ranging from 0 to 1) may overlap with one another in the data space. On the other hand, in sub-pixel classification problems, fuzzy memberships can be used as a valuable methodological tool for modeling the membership grade of a pixel to a given class. In this way the fuzzy membership is not used for expressing uncertainty, but, according to the so-called probabilistic fuzzy-set theoretic framework, as soft information for modeling the membership of each pixel to different information classes. In this paper we focus our attention on nonlinear sub-pixel classification based on machine-learning techniques and fuzzy modeling of the class membership.

Looking at the machine learning literature, one of the most effective approaches to pattern classification is that based on support vector machines (SVMs). The SVM formulation (developed by Vapnik) is based on the *Structural Risk Minimization* principle, which is an inductive principle for model selection that aims at providing a tradeoff between hypothesis space complexity (the Vapnik-Chervonenkis dimension of approximating functions) and quality of fitting the training data (empirical error) [9]-[11]. Thanks to this formulation, the SVM approach has excellent properties, like: i) good generalization ability; ii) high effectiveness in hyperdimensional feature space (important when dealing for example with hyperspectral images); iii) learning phase associated with the minimization of a convex cost function that guarantees the uniqueness of the solution; iv) possibility to be implemented in a parallel architecture (thus reducing the overall computational time by an adequate parallel processing). Due to the aforementioned attractive properties and their good performances, SVMs are well accepted from the scientific community and have applied to many image classification and recognition fields, such as remote sensing [12]-[16], biomedical applications, spatial data analysis [17], character recognition [18], etc. However, a major limitation of standard SVM classifiers in image classification is that they produce a crisp output, i.e. they are based on the assumption that a given pixel can belong only to one information class. Thus, this theoretically elegant and powerful methodology cannot be used to address sub-pixel classification problems. In order to face this limitation, we present an approach that extends SVMs to manage sub-pixel (soft) information in image classification by using the concepts developed in the fuzzy set theory. In the literature, only relatively few researchers studied the general problem of extending SVMs to fuzzy problems [19]-[23]. Among the others, a pioneering work was proposed by Lin and Wang [19], who defined a *Fuzzy SVM*, i.e. a binary classifier capable to consider in the learning phase the uncertainty associated with each training pattern and to provide a crisp output like standard SVM. The basic idea is to weight the relevance of training patterns according to their uncertainty in the learning process. However, the Fuzzy SVM in [19] cannot fully exploit the fuzzy information present in the data as: i) it is able to manage the fuzzy information of an input pattern, but it cannot produce a soft output; and ii) each single pattern in the training set is considered with a weight that models the uncertainty and not a membership value to more than one class. In addition, only binary problems are considered and no discussion on possible generalization to multiclass fuzzy problems is reported. These limitations make Fuzzy SVM unsuitable to be applied to sub-pixel image classification.

In this paper we define a novel *Fuzzy-input Fuzzy-output Support Vector Machine* classifier (called F^2SVM) which is specifically designed for addressing image sub-pixel classification problems. F^2SVM is a classifier capable to learn the sub-pixel information present in a training set (fuzzy input) and to estimate the membership (abundance) of each unknown pixel in the analyzed image to the classes that describe the considered problem (fuzzy output). The novelties that F^2SVM presents with reference to standard SVM-based image classification methods are: i) a *sub-pixel learning procedure* (the membership grade of a pixel to a class is modeled by using a soft cost function in the training phase); ii) a *sub-pixel decision algorithm* (the output is not a crisp value, but a fuzzy membership grade that describes the abundances of each pixel toward each class); iii) the *generalization to the multicategory case* (two strategies, called *Fuzzy One Against All* (FOAA) and *Fuzzy One Against One* (FOAO), are proposed for combining the fuzzy outputs given by a set of binary F^2SVMs for addressing multicategory sub-pixel classification problems). Furthermore, the proposed approach simultaneously satisfies the critical sum-to-one and the non-negative abundance constraints [6]. It is worth noting that, although the presented F^2SVM technique has been developed in the probabilistic fuzzy-set framework for addressing sub-pixel image classification problems, it introduces general concepts that can be used in other fuzzy problems dealing with uncertainty modeling.

The proposed technique was tested on three different image classification problems. The first one is a simulated multispectral image. The second problem deals with the analysis of real multispectral images. The third problem concerns the sub-pixel classification in hyperspectral images. In all cases, the presented method increased the classification accuracy with respect to an effective machine-learning procedure based on fuzzy multilayer perceptron neural networks [8].

The paper is organized into six sections. The next section presents the background on supervised crisp SVM. Section III introduces the notation and describes the proposed F^2SVM in the binary case, by detailing the sub-pixel learning and the sub-pixel decision procedures. Section IV presents the proposed FOAA and FOAO strategies for the generalization of F^2SVM to multicategory problems. Section V addresses the design of experiments and illustrates the main concepts associated with the *Fuzzy Multi-Layer Perceptron* (*FMLP*) neural network used for comparisons. The three data sets used in the experiments and the related results are presented in Sections VI, VII and VIII. Concluding remarks are given in Section IX.

II. BACKGROUND: CRISP SUPPORT VECTOR MACHINE CLASSIFIER

In order to define the proposed F^2SVM algorithm, it is necessary to give an overview of standard crisp SVM. (Detailed discussions on crisp SVM can be found in [3]-[11],[24]).

Let $\boldsymbol{x} \in \Re^d$ be the pattern representing a pixel of a generic image in a d-dimensional feature space[1], and $\Omega=\{\omega_1,\omega_2\}$ the set of information classes that defines a binary classification problem. In the crisp formulation, $\boldsymbol{x}$ can belong only to one of the classes in Ω. Let the classes ω_1 and ω_2 be coded with "+1" and "-

[1] d is the number of attributes used for describing a generic pixel in the classification problem. For example, if a multispectral image is considered, d is equal to the number of available spectral channels.

1”, respectively. Let us assume that a training set L made up of N patterns[2] is available.

The SVM classifier attempts to separate samples belonging to the two considered classes by defining a maximum margin hyperplane in the original feature space (linear SVM) or in a transformed space where samples are mapped for obtaining linear separability according to a nonlinear mapping function $\varphi(\cdot)$ (non-linear SVM) [25]. In both cases the learning of the SVM is based on the combination of two criteria: i) empirical error minimization, and ii) control of model complexity. The former aims at optimizing the classification results in terms of accuracy on the training samples; the latter controls the capacity (or flexibility) of the function used for avoiding overfitting. These criteria are combined for defining the cost function to be minimized.

In the case of linear SVM, the discriminat function $f(\boldsymbol{x})$ can be written as:

$$f(x)=\sum_{i=1}^{N} w\left\langle x, x_i \right\rangle + b \tag{1}$$

where $\boldsymbol{w}$ is a vector normal to the hyperplane and b is a constant such that $b/\|\boldsymbol{w}\|^2$ represents the distance of the hyperplane from the origin (Figure 1 shows an example of how a crisp SVM classifier works). If the data in the input space cannot be linearly separated, they can be projected into a higher dimensional feature space (i.e. a Hilbert space $\mathcal{H}$) with a nonlinear mapping function $\varphi(\cdot)$ defined in accordance with the Cover’s theorem [26],[27]. As a consequence, the inner product between two mapped feature vectors becomes:

$$f(x)=\sum_{i=1}^{N} w\left\langle \varphi(x), \varphi(x_i) \right\rangle + b \tag{2}$$

The discriminant function $f(\boldsymbol{x})$ can be derived by minimizing the following cost function, which expresses the above-mentioned tradeoff between empirical error minimization and solution complexity:

$$\psi(\boldsymbol{w},\xi_i) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad subject\ to \begin{cases} \omega_i\left[\boldsymbol{w}\cdot\varphi(\boldsymbol{x}_i)+b\right]\geq 1-\xi_i, & i=1,2,...,N \\ \xi_i \geq 0 \quad and \quad C>0 \end{cases} \tag{3}$$

where C is the *regularization parameter*, ξ_i are non-negative *slack variables* necessary to deal with noisy and nonlinearly separable data (a nonzero ξ_i indicates that the pixel $\boldsymbol{x}_i$ is misclassified because it is on the wrong side of the hyperplane), ω_i is the label of the training pattern $\boldsymbol{x}_i$, and N is the total number of training samples. The final crisp decision function can be written as:

$$\hat{\omega} = sign\left[f(\boldsymbol{x})\right] \tag{4}$$

The primal minimization problem in (3) can be solved according to the Lagrange theory obtaining a dual problem in which the following convex objective function should be maximized:

[2] In this paper the generic pattern is defined with $\boldsymbol{x}$ and patterns used to train the classifier (pixels that belong to the training set) are indicated with $\boldsymbol{x}_i$, $i = 1,\ldots, N$.

$$W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\omega_i\omega_j\left\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\right\rangle \quad subject\ to\ \begin{cases}\sum_{i=1}^{N}\omega_i\alpha_i=0\\ 0\leq\alpha_i\leq C\ \ and\ \ C>0 \qquad i=1,2,\ldots,N\end{cases} \tag{5}$$

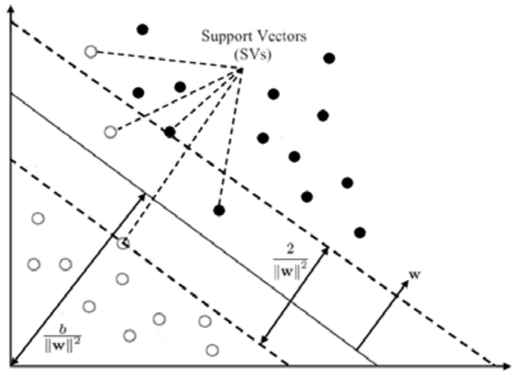


Figure 1. Illustration of a crisp SVM binary classifier: separation hyperplane (solid line) and margin bounds (dashed lines).

The Lagrangian $W(\boldsymbol{\alpha})$ should be maximized with respect to Lagrange multipliers α_i (which are associated with training points $\boldsymbol{x}_i$). This problem can be solved according to Quadratic Programming (QP) methods [11]. Patterns associated to nonzero Lagrange multipliers are called support vectors: the ones corresponding to $0<\alpha_i<C$ are called *non-bound support vectors* and fall inside the margin, while the ones corresponding to $\alpha_i = C$ are called *bound support vectors* and fall on the margin. These samples can be regarded as errors because they are associated to a nonzero ξ_i. Support vectors are the only patterns in the training set that determine the optimal hyperplane position.

Since in non-linear SVM we do not have any knowledge on functions $\varphi(\cdot)$, the QP problem solution is not possible using (5). Due to the Mercer's theorem [24],[28], by replacing the inner product with a positive defined kernel function $K(\cdot,\cdot)$, it is possible to avoid representing the feature vector explicitly, i.e.

$$\left\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\right\rangle=K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{6}$$

Accordingly, it is possible to prove that the discriminant function can be rewritten in the dual formulation as [11],[29]:

$$f(x)=\sum_{i=1}^{N}\alpha_i\omega_i K(x,x_i)+b \tag{7}$$

where b is calculated using the primal-dual relationship [29], and only samples with nonzero Lagrange multipliers α_i affect the solution. Thus, the decision function is obtained by applying (4) to (7).

The most widely used positive definite kernels that satisfy Merecer's conditions are the following.

- *Linear kernel*:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle \tag{8}$$

- *Polynomial Kernel*:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\left\langle\boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle+1)^d, \quad d\in\Re^+ \tag{9}$$

- *Radial Basis Function (RBF) Kernel*:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2\right),\quad \sigma\in\Re^+ \tag{10}$$

Unlike in other classification techniques, such as Multi-Layer Perceptron Neural Networks, the Mercer kernel $K(\cdot,\cdot)$ ensures that the objective function is convex, thus there are no local maxima in the function to be optimized.

The standard SVM classifier is defined as a binary supervised classification algorithm, which can discriminate two different classes. In the literature, many approaches for handling multiclass problems (R>2) have been proposed. Among the others we recall: the One-Against-All (OAA) and One-Against-One (OAO) strategies [30]. Let $\Omega=\{\omega_1,..., \omega_R\}$ be the set of R classes to be identified. In the OAA architecture, R different binary SVMs are trained. Each binary classifier is aimed at distinguishing the samples of a generic class $\omega_i\in\Omega$ from the samples of all the remaining classes $\Omega-\omega_i$. A given pattern is labeled according to the class of the classifier that results in the highest output value. In the OAO architecture, one classifier for each pair of classes ω_i and ω_j (with $i\neq j$) is considered. On the whole, we have $R(R-1)/2$ classifiers. A given pattern is classified according to a simple majority voting algorithm [31]. We refer the reader to [30] for greater details on multiclass strategies.

III. Proposed F²SVM for Sub-Pixel Classification: Binary Problems

A. Notation

According to the fuzzy framework, $\boldsymbol{x}\in\Re^d$ can belong to different classes with given membership values. In greater detail, a pixel $\boldsymbol{x}$ belongs to a generic class $\omega_k\in\Omega$ with a membership grade specified by $M_k(\boldsymbol{x})$, where $M_k(\boldsymbol{x})$ is a component of the memberships vector $\boldsymbol{M}(\boldsymbol{x})=[M_1(\boldsymbol{x}),\ldots,M_R(\boldsymbol{x})]$, with $0\leq M_k(\boldsymbol{x})\leq 1$. As we use fuzzy concepts for representing the membership of a pixel to different classes, we develop the proposed F²SVM in the probabilistic fuzzy-set theoretic framework by imposing the following constraint[3]:

$$\sum_{k=1}^{R} M_k(\boldsymbol{x})=1 \tag{11}$$

Given this fuzzy modeling of the problem, it is always possible to assign a crisp label to a pixel by hardening the soft classification solution, i.e. by assigning the pixel to the class having the maximum membership value.

In this paper we define a pixel belonging to more than one class as *mixed pixel*. The membership vector associated to a mixed pixel has more than one element different from zero. These patterns play an important role in the learning of F²SVM because they allow one deriving the model that describes the sub-pixel (soft) information in the considered data set.

The training of the proposed binary F²SVM is divided into two stages: the *learning of the input* and the *learning of the output* (a preliminary version of this procedure is presented in [32]). At the end of the *learning of the input*, we obtain a classifier that computes the optimal separating hyperplane by considering the position of the training pixels in the kernel space and their fuzzy membership vectors $\boldsymbol{M}(\boldsymbol{x}_i)$. At the end of the *learning of the output*, the classifier estimates the fuzzy membership vector $\boldsymbol{m}(\boldsymbol{x})$ for each unknown pattern.

In the next sub-sections we present the proposed learning and decision procedures in the binary case.

B. Fuzzy Learning of the Input

By extending and developing concepts previously presented in [19], we introduce the sub-pixel information (fuzzy memberships) of training patterns in hyperplane computation. The goal of the proposed fuzzy learning method is to obtain an SVM able to learn the fuzzy information inherent in the training set and to manage pixels belonging to different classes with different memberships.

Similarly to the crisp case, let us first consider a binary classification problem (R=2), where the classes ω_1 and ω_2 are coded with "+1" and "-1", respectively. Let us assume that a training set L composed of N patterns is available. For each training pattern the vector of memberships $\boldsymbol{M}(\boldsymbol{x}_i)$ is defined as follows:

$$\boldsymbol{M}(\boldsymbol{x}_i)=\{M_{+1}(\boldsymbol{x}_i), M_{-1}(\boldsymbol{x}_i)\}, \qquad M_{+1}(\boldsymbol{x}_i)+M_{-1}(\boldsymbol{x}_i)=1 \qquad i=1,2,...,N \tag{12}$$

where $M_{+1}(\boldsymbol{x}_i)$ and $M_{-1}(\boldsymbol{x}_i)$ are the abundances of the i-th pixel toward classes ω_1 and ω_2, respectively. If the components of $\boldsymbol{M}(\boldsymbol{x}_i)$ are both nonzero, $\boldsymbol{x}_i$ is a *mixed pixel*.

Let N_{mixed} be the number of mixed pixels in L. In the learning stage of F^2SVM, each mixed training pattern in L should contribute to both ω_1 and ω_2 proportionally to its memberships to the two classes. However, in the notation used in Section II, each pattern in the training set should belong to a single class. Since mixed pixels in L do not satisfy this requirement (they have two nonzero membership values), we have to manipulate L to obtain a new training set L_f. The manipulation is a cloning operation that consists in duplicating each mixed pixel in L in two new patterns with the same feature vector as the original sample. The first new pattern belongs to class ω_1 with membership $\mu_j = M_{+1}(\boldsymbol{x}_i)$, while the second new pattern belongs to class ω_2 with membership $\mu_{j+1} = M_{-1}(\boldsymbol{x}_i)$. The unmixed pixels in L remain unchanged in L_f and are labeled with ω_1 or ω_2. L_f is made up of $N_f = N+N_{mixed}$ patterns identified by ($\boldsymbol{z}_j$, ω_j, μ_j). $\boldsymbol{z}_j$ is the feature vector characterizing the *j-th* pattern in training set L_f. Here $\boldsymbol{z}_j$ is used instead of $\boldsymbol{x}_j$ to distinguish patterns belonging to training set L from those belonging to training set L_f, which are implicitly sorted in a different way.

In order to include the fuzzy memberships μ_j in the hyperplane computation, the cost function to minimize in the computation of the fuzzy hyperplane becomes the following:

[3] This assumption implicitly means that we consider classification problems in which we have an exhaustive representation of classes (i.e. all classes present in the scene are modeled in the training set). It is worth noting that this constraint can be removed in the solution of more general problems in which fuzzy memberships are used for modeling uncertainty, without affecting the validity of the proposed method.

$$\psi(\boldsymbol{w},\xi_i) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{j=1}^{N_f}\mu_j\xi_j \quad subject\ to \begin{cases} \omega_j\left[\boldsymbol{w}\cdot\varphi(\boldsymbol{z}_j)+b\right] \geq 1-\xi_j, & j=1,2,\ldots,N_f \\ \xi_j \geq 0 \quad and \quad C>0 \end{cases} \tag{13}$$

In this cost function, a μ_j value smaller than 1 (which is peculiar of mixed patterns) reduces the effect of the corresponding slack variable ξ_j such that the importance of pattern $\boldsymbol{z}_j$ is decreased. It is worth noting that in our technique a pixel can play a role for both classes in the learning, involving an intrinsic training of the hyperplane position on the basis of the available sub-pixel information. In greater detail, a mixed pixel attends two times in the QP problem because of its duplication in L_f. The first time it appears as a pattern belonging to class ω_1 (i.e., "+1") weighted by the fuzzy membership $M_{+1}(\boldsymbol{x}_i)$; the second time as a pattern belonging to class ω_2 (i.e., "-1") weighted by the fuzzy membership $M_{-1}(\boldsymbol{x}_i)$. From a theoretical and conceptual point of view, this is an important difference with respect to the learning of the fuzzy SVM proposed in [19]. We can observe that the learning of a traditional crisp SVM can be seen as a limit case of the proposed F^{2}SVM learning in the case in which there are no mixed pixels. On the contrary, if we consider a fuzzy training set L with $N_{mixed} \neq 0$, the solution obtained by the F^{2}SVM can be significantly different from that yielded by a standard crisp SVM. Figure 2 shows a qualitative example in which the proposed fuzzy learning method is compared with the crisp SVM learning algorithm.

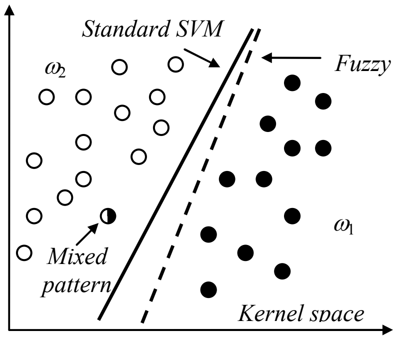


Figure 2. Qualitative example of Fuzzy SVM learning vs. standard SVM learning. The training set includes only one mixed pixel belonging to classes ω_1 (white) and ω_2 (black) with memberships 0.6 and 0.4, respectively. In the hardened (crisp) version of the training set used to train crisp SVM, this pattern is assigned to class ω_1. The hyperplane computed by the F^{2}SVM is closer to class ω_1 than the one obtained with standard SVM, as the importance of the mixed pattern in the QP problem is weighted by a membership lower than 1 to class ω_1.

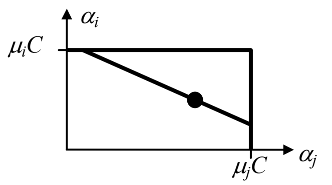
It is possible to prove that the minimization of (13) is equivalent to the maximization of the following dual formulation obtained with the Lagrange theory [19]:

$$W(\alpha) = \sum_{j=1}^{N_f}\alpha_i - \frac{1}{2}\sum_{j=1}^{N_f}\sum_{i=1}^{N_f}\alpha_j\alpha_i\omega_j\omega_i K(\boldsymbol{z}_j,\boldsymbol{z}_i) \quad subject\ to \begin{cases} \sum_{j=1}^{N_f}\omega_j\alpha_j = 0 \\ 0 \leq \alpha_j \leq \mu_j C \quad and \quad C>0 & j=1,2,\ldots,N_f \end{cases} \tag{14}$$

where fuzzy memberships μ_j multiply directly the regularization parameter C. $W(\alpha)$ has to be maximized with respect to Lagrange multipliers α_i, also. This problem can be solved according to Quadratic Programming methods [11]. To this end, we use a modified *Sequential Minimal Optimization* (SMO) algorithm [33],[34], which is an iterative procedure that decomposes the overall QP problem into QP subproblems using Osuna's Theorem to ensure convergence [35]. At each step, SMO: i) chooses two Lagrange multipliers; ii) finds the optimal values for these multipliers; and iii) updates the SVM to reflect the new optimal values. According to the Karush-Khun-Tucker (KKT) theorem, the Lagrange multipliers α_i that solve (14) must satisfy the following conditions [24],[36]:

$$KKT\ conditions\ \begin{cases} \alpha_j\left[\omega_j\left(\boldsymbol{w}\cdot\varphi(z_j)+b\right)-1+\xi_j\right]=0 \\ \mu_j C>0\ \ and\ \ \xi_j\geq 0 \qquad\qquad j=1,2,\ldots,N_f \\ \left(\mu_j C-\alpha_j\right)\xi_j=0 \end{cases} \tag{15}$$

where μ_j must be positive to obtain a correct interpretation of the KKT conditions (the condition $0\leq M_k(\boldsymbol{x})\leq 1$ ensures that $\mu_j > 0$). It is possible to show that in the crisp SVM, these conditions cause α_j and α_i to lie on a diagonal line in a *squared* box of side C [33]. According to (15), in the F[2]SVM we have that the support vectors α_j and α_i are bounded from $\mu_j C$ and $\mu_i C$, respectively. Therefore, in the SMO we change the constraints of the problem on the basis of the fuzzy memberships: α_j and α_i have to lie on a diagonal line in a *rectangular* box of sides $\mu_j C$ and $\mu_i C$, as shown in Figure 3.



$\mu_j = 0.75 \quad \Rightarrow\ 0 \leq \alpha_j \leq \mu_j C = 0.75\cdot C$

$\mu_i = 0.5 \quad \Rightarrow\ 0 \leq \alpha_i \leq \mu_i C = 0.5\cdot C$

Figure 3. Qualitative scheme of the SMO for fuzzy learning.

C. Fuzzy Learning of the Output

The fuzzy learning of the input allows the SVM classifier modeling the sub-pixel information present in the available soft training set in the definition of the hyperplane. However, this does not allow the SVM to provide the membership of a sample in output from the classifier. In order to obtain a soft (fuzzy) output we should properly consider the distance of the pattern from the hyperplane given from $f(\boldsymbol{x})$. However $f(\boldsymbol{x})$ is an uncalibrated output. Thus, to estimate fuzzy memberships for an unknown pattern, the output has to be normalized in order to take into account the learning set fuzziness. To this end, we propose to analyze the properties of the training patterns in L with respect to the separating hyperplane in the kernel space. We can construct a diagram (see Figure 4) that plots the distances of training pixels from the hyperplane $f(\boldsymbol{x}_i)$, with i=1,…, N, (on the abscissa axis) versus the membership $M_{+1}(\boldsymbol{x}_i)$ (on the axis of ordinates). As mixed pixels in the training set L are closer to the separating hyperplane than the pure pixels, we can model the training set sub-pixel information by inspiring to the idea proposed in [37]. In particular, we propose to model the fuzzy membership referred to class ω_1 according to a *sigmoid* function[4] defined as:

$$o_{+1}(\boldsymbol{x})=\left(1-e^{\left[A\cdot f(\boldsymbol{x})+B\right]}\right)^{-1} \tag{16}$$

where $o_{+1}(\boldsymbol{x})$, is the sigmoid value referred to the unknown pattern with feature $\boldsymbol{x}$ toward class ω_1. The sigmoid shape is defined by parameters A and B: the former tunes the curve spreading (it represents the slope of the tangent to the sigmoid for membership equal to 0.5), the latter indicates the horizontal offset. Figure 5 shows a sigmoid (solid line) with A=-1 and B=0.

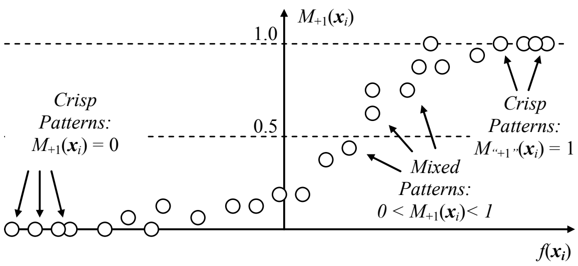


Figure 4. Membership of training patterns to class ω_1 (i.e.,"+1") versus their distance from the hyperplane $f(\boldsymbol{x}_i)$.

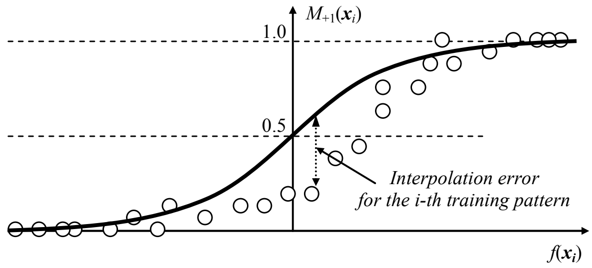


Figure 5. Interpolating sigmoid and interpolation errors e_i for the generic training pattern $\boldsymbol{x}_i$.

The values of parameters A and B that define the sigmoid that better fits the fuzzy membership behaviors are computed by a simple iterative algorithm that jointly optimize A and B. The algorithm minimizes the overall root mean square error (RMSE) between the known fuzzy memberships of the patterns in the training set and the ones obtained according to (16). The algorithm stops when the error difference between two consecutive steps is lower than a properly defined threshold ε. By tuning ε we can control the precision of the algorithm and the quality of the interpolating sigmoid. At convergence, the algorithm finds a sigmoid (see the example in Figure 6) that models the fuzzy membership behavior toward class ω_1 according to the information present in the training set, and allows us to estimate the membership grades to this class of unknown pixels. Due to the sum-to-one assumption of pixel memberships (see (11)), we can estimate the membership degree of unknown patterns toward class ω_2 by computing the curve symmetric to that for class ω_1 (the dotted sigmoid in Figure 6) as[5]:

$$o_{-1}(\boldsymbol{x}_i)=1-o_{+1}(\boldsymbol{x}_i)=1-\left(1-e^{\left[A\cdot f(\boldsymbol{x}_i)+B\right]}\right)^{-1}=e^{\left[A\cdot f(\boldsymbol{x}_i)+B\right]}\Big/\left(e^{\left[A\cdot f(\boldsymbol{x}_i)+B\right]}+1\right) \tag{17}$$

Combining the fuzzy learning SVM with the sigmoid applied to the decision phase we obtain the desired F^2SVM. The pair [$o_{-1}(\boldsymbol{x})$; $o_{+1}(\boldsymbol{x})$] is the fuzzy output of the F^2SVM, i.e. the estimated abundances (memberships) of an unknown pixel $\boldsymbol{x}$ to the two classes defined in the binary problem.

It is worth noting that a hardened output can be obtained by assigning a generic unknown pattern to the class with the highest membership grade. However, the hardened output of F^2SVM does not correspond to the output of crisp SVM, except in the particular case in which B=0. In fact, only in this situation the sigmoid has value 0.5 when $f(\boldsymbol{x})$=0. Thus, membership estimation is not only a way to describe more deeply the classification results, but it also plays an important role in the decision process of F^2SVM.

[4] It is worth noting that other fuzzy membership functions could be considered.

[5] In the general case in which (11) does not hold, a second sigmoid should be defined for estimating the membership grades of patterns to ω_2 according to (16).

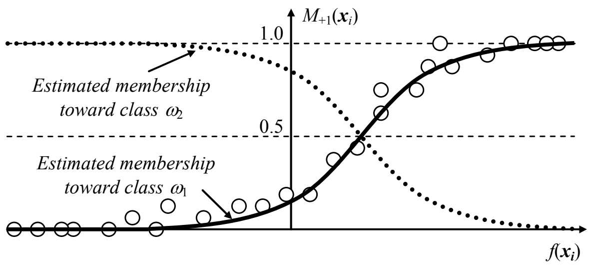


Figure 6. Sigmoid adapted to training set sub-pixel (soft) information.

IV. Proposed F²SVM for Sub-Pixel Classification: Multiclass Problems

Let us now consider a multiclass problem made up of R information classes $\Omega=\{\omega_1,..., \omega_R\}$. As in the binary case, we have to convert the original training set L in a new set L_f, in which each mixed pixel is replicated as many times as the number of nonzero components of its membership vector $\boldsymbol{M}(\boldsymbol{x}_i)$. If h ($1\leq h\leq R$) is the number of nonzero elements in $\boldsymbol{M}(\boldsymbol{x}_i)$, the mixed pattern $\boldsymbol{x}_i$ is mapped into h new patterns, which are characterized by the feature vector of $\boldsymbol{x}_i$, a label $\omega_k\in\Omega$ and a single nonzero membership value $\mu_k\in\boldsymbol{M}(\boldsymbol{x}_i)$. As mentioned in Sec. II, a multiclass problem can be faced by training a set of binary F^2SVM classifiers and combining their decisions. We propose two strategies that generalize to the fuzzy case architectures and decision rules developed for crisp classifiers. These strategies are described in the following sub-sections.

A. Fuzzy OAA (FOAA) Strategy

The Fuzzy OAA strategy is conceptually very similar to the crisp OAA method. It requires R binary F^2SVMs to estimate the membership vector $\boldsymbol{m}(\boldsymbol{x})$ (i.e. the abundances) of the pixel $\boldsymbol{x}$ to the considered R classes (see Figure 7). The generic $F^2SVM_{k,\Omega\text{-}k}$ estimates the fuzzy memberships of the input pattern to classes ω_k and $\omega_{\Omega\text{-}k}$ (for simplicity Ω - k denotes the meta-class that groups all the information classes but ω_k, i.e. $\Omega - k \equiv \Omega - \omega_k$). Each $F^2SVM_{k,\Omega\text{-}k}$ is trained using all the N_f samples in the training set L_f, which are split into the set k (made up of the training samples that belong to the class ω_k) and the set Ω - k (made up of all the training samples in L_f that do not belong to class ω_k). The $F^2SVM_{k,\Omega\text{-}k}$ is trained to separate the information class ω_k from the meta-class $\Omega - \omega_k$ according to the algorithm presented in Section III.B. Once the learning stage has been completed, we can fit $sigmoid_{k,\Omega\text{-}k}$ to the fuzzy membership of the related training samples toward the class ω_k, according to their distance from the hyperplane using the iterative algorithm proposed in section III.C. With the $sigmoid_{k,\Omega\text{-}k}$ we can estimate $o_{k,\Omega\text{-}k}(\boldsymbol{x})$, which is the membership grade of an unknown pattern $\boldsymbol{x}$ toward the information class ω_k. It is worth noting that since we are not interested in the pattern membership to the meta-class Ω - k, we can use only the sigmoid referred to class ω_k.

Following this procedure for all the R $F^2SVM_{k,\Omega\text{-}k}$ we can train independently each binary classifier to estimate $o_{k,\Omega\text{-}k}(\boldsymbol{x})$, for k=1,2,...,R. At the end, when all the $F^2SVM_{k,\Omega\text{-}k}$ are trained and able to compute fuzzy outputs $o_{k,\Omega\text{-}k}(\boldsymbol{x})$, we obtain the membership vector $\boldsymbol{m}(\boldsymbol{x}) = \{m_1(\boldsymbol{x}), m_2(\boldsymbol{x}), \ldots, m_R(\boldsymbol{x})\}$ for the pattern $\boldsymbol{x}$. It is worth noting that in order to guarantee that the sum-to-one constraint is satisfied we must normalize the values of the outputs.

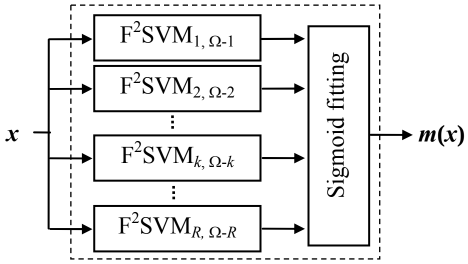


Figure 7. Architecture of the FOAA strategy

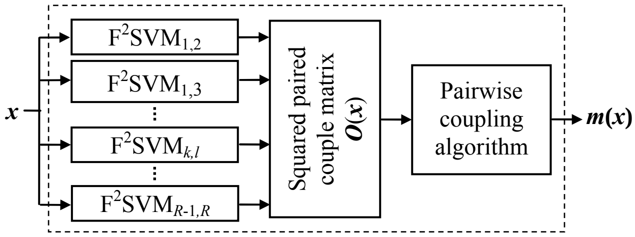


Figure 8. Architecture of the FOAO strategy.

B. Fuzzy OAO (FOAO) Strategy

In the fuzzy OAO strategy we define an architecture made up of $R(R\text{-}1)/2$ binary F^2SVM to estimate the memberships toward the R categories described in the classification problem (Figure 8). The generic $F^2SVM_{k,l}$ estimates the pair of values $[o_{k,l}(\boldsymbol{x}); o_{l,k}(\boldsymbol{x})]$: $o_{k,l}(\boldsymbol{x})$ is the fuzzy membership of the input pixel $\boldsymbol{x}$ to class ω_k against class ω_l, while $o_{k,l}(\boldsymbol{x})$ is the fuzzy membership of the input pattern $\boldsymbol{x}$ to class ω_l against class ω_k. Each $F^2SVM_{k,l}$ is trained using only the training samples in L_f belonging to classes ω_k and ω_l with nonzero memberships. Once the learning stage of the $F^2SVM_{k,l}$ has been completed and the related optimal separating hyperplane has been defined (see section III.B), it is possible to represent the membership μ_j of the training patterns versus their distance from the hyperplane. Figure 9.a shows an example of this process (white circles indicate memberships of patterns belonging to class ω_k, while black circles indicate memberships of patterns belonging to class ω_l; dotted lines join membership values referred to the same pixel in L). The example points out a critical issue of this multiclass architecture, which is related to the fact that some patterns in the specific binary sub-problem considered involving class ω_k and ω_l may not satisfy the sum-to-one assumption, i.e., the considered binary problem is not exhaustive. In fact, in the FOAO architecture we consider only two information classes for each classifier, thus a mixed training pixel belonging to more than two classes (or to class ω_k and to class $\omega_p \neq \omega_l$) presents $M_k(\boldsymbol{x}_i)+M_l(\boldsymbol{x}_i)\leq 1$ (see patterns highlighted with a dotted circle in Figure 9.a). For this reason, unlike in the binary F^2SVMs included in the FOAA architecture, the behavior of memberships to classes ω_k and ω_l are not symmetric. This makes it necessary to fit two different sigmoids: one to estimate membership $o_{k,l}(\boldsymbol{x})$ to class ω_k and the other to estimate membership $o_{l,k}(\boldsymbol{x})$ to class ω_l. The two sigmoids can be fitted using the same iterative algorithm described in Section III.C. Figure 9.b shows the two sigmoids computed for the qualitative example reported in Figure 9.a.

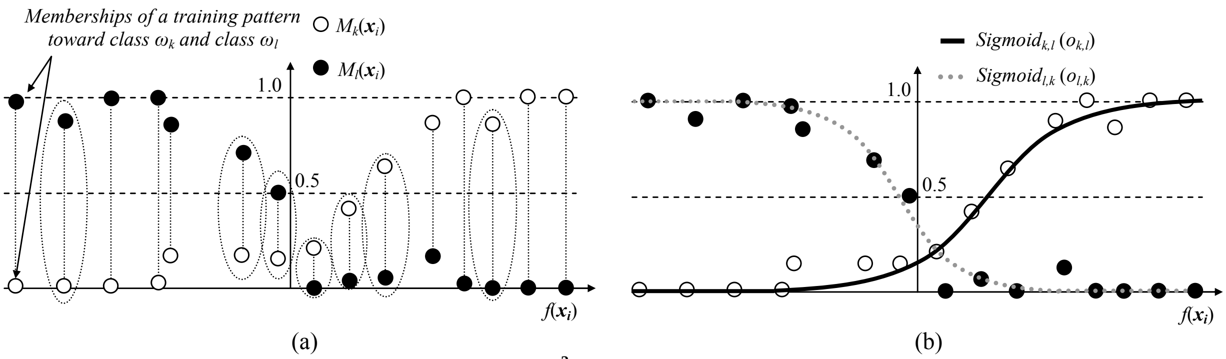


Figure 9. Qualitative example related to a generic $F^2SVM_{k,l}$ in the Fuzzy OAO strategy: a) memberships of training pixels to classes ω_k and ω_l; b) sigmoids defined for deriving the output of the $F^2SVM_{k,l}$.

When all the $F^2SVM_{k,l}$ have been trained, we can use the set of $R(R-1)/2$ binary $F^2SVM_{k,l}$ to estimate all the pairs $[o_{k,l}(\boldsymbol{x}); o_{l,k}(\boldsymbol{x})]$ for an unknown pattern $\boldsymbol{x}$. Unlike the conventional crisp OAO strategy (which assigns $\boldsymbol{x}$ to the class that wins the most pairwise comparisons), we have to relate the estimated pairwise memberships to the class memberships by adequately combining the outputs of all the binary classifiers. To this purpose, first the outputs $[o_{k,l}(\boldsymbol{x}); o_{l,k}(\boldsymbol{x})]$ of each $F^2SVM_{k,l}$ are normalized to obtain:

$$o_{k,l}(\boldsymbol{x})+o_{l,k}(\boldsymbol{x})=1.0 \quad for \quad k,l=1,2,...,R \tag{18}$$

Then, the normalized memberships are represented in a squared matrix $\boldsymbol{O}(\boldsymbol{x})$ defined as:

$$\boldsymbol{O}(\boldsymbol{x})=\begin{pmatrix} . & o_{1,2}(\boldsymbol{x}) & o_{1,3}(\boldsymbol{x}) & \cdots & \cdots & o_{1,R} \\ o_{2,1}(\boldsymbol{x}) & . & o_{2,3}(\boldsymbol{x}) & \cdots & \cdots & o_{2,R}(\boldsymbol{x}) \\ o_{3,1}(\boldsymbol{x}) & o_{3,2}(\boldsymbol{x}) & . & \cdots & \cdots & o_{3,R}(\boldsymbol{x}) \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & o_{R-1,R}(\boldsymbol{x}) \\ o_{R,1}(\boldsymbol{x}) & o_{R,2}(\boldsymbol{x}) & o_{R,3}(\boldsymbol{x}) & \cdots & o_{R,R-1}(\boldsymbol{x}) & . \end{pmatrix} \tag{19}$$

The main diagonal of the matrix is not defined for obvious reasons. Starting from the $R(R-1)$ values in $\boldsymbol{O}(\boldsymbol{x})$, we should derive a vector $\boldsymbol{m}(\boldsymbol{x})$ that describes the membership of an unknown pattern $\boldsymbol{x}$ toward all the R information classes of the problem. This should be accomplished by jointly using the pairwise memberships estimated from all binary F^2SVM classifiers for estimating the memberships $m_i(\boldsymbol{x})$ of the pattern $\boldsymbol{x}$ to class ω_i. We compute $\boldsymbol{m}(\boldsymbol{x})$ by using the following iterative *pairwise coupling* algorithm:

- *Initialization Step* $\{0\}$

$$m_k^{\{0\}}(\boldsymbol{x})=2\sum_{l\neq k} o_{k,l}(\boldsymbol{x})\Big/R(R-1) \quad for \quad k=1,2,...,R \tag{20}$$

$$v_{k,l}^{\{0\}}(\boldsymbol{x})=m_k^{\{0\}}(\boldsymbol{x})\Big/\left(m_k^{\{0\}}(\boldsymbol{x})+m_l^{\{0\}}(\boldsymbol{x})\right) \quad for \quad k,l=1,2,...,R \quad and \quad k\neq l \tag{21}$$

$m_k^{\{0\}}(\boldsymbol{x})$ is the summation of the values on line k of matrix $\boldsymbol{O}(\boldsymbol{x})$ normalized with $R(R-1)/2$. $v_{k,l}(\boldsymbol{x})$ is an approximation of $o_{k,l}(\boldsymbol{x})$ according to the Bradley-Terry model for paired comparisons [38], which relates the memberships estimated by the set of binary classifiers ($o_{k,l}(\boldsymbol{x})$ and their approximations $v_{k,l}(\boldsymbol{x})$) to the overall membership $m_k(\boldsymbol{x})$ toward the class ω_k.

- *Optimization and Updating Step* $\{t\}$

At the t-th iteration, we apply sequentially the following equations:

$$m_k^{\{t+1\}}(\boldsymbol{x}) \leftarrow m_k^{\{t\}}(\boldsymbol{x}) = \sum_{l\neq k} o_{k,l}(\boldsymbol{x}) \Big/ \sum_{l\neq k} \nu_{k,l}^{\{t\}}(\boldsymbol{x}) \quad \textit{for} \quad k=1,2,\ldots,R \tag{22}$$

$$m_k^{\{t+1\}}(\boldsymbol{x}) \leftarrow m_k^{\{t+1\}}(\boldsymbol{x}) \Big/ \sum_{l=1}^{R} m_l^{\{t+1\}}(\boldsymbol{x}) \quad \textit{for} \quad k=1,2,\ldots,R \tag{23}$$

$$\nu_{k,l}^{\{t+1\}} = m_k^{\{t+1\}}(\boldsymbol{x}) \Big/ \left(m_k^{\{t+1\}}(\boldsymbol{x}) + m_l^{\{t+1\}}(\boldsymbol{x})^{\{t+1\}} \right) \quad \textit{for} \quad k,l=1,2,\ldots,R \quad \textit{and} \quad k \neq l \tag{24}$$

Equation (23) normalizes $m_k(\boldsymbol{x})$ according to the sum-to-one assumption on membership values, while equation (24) updates $\nu_{k,l}(\boldsymbol{x})$ using the new $m_k(\boldsymbol{x})$ computed in the optimization step.

- *Conditional Step:* if the variation of each $m_k(\boldsymbol{x})$ is lower than a threshold defined by the user, the algorithm stops and current $\boldsymbol{m}(\boldsymbol{x})$ is the pairwise coupling result. Otherwise the algorithm repeats the *Optimization and Updating Step* and the *Conditional Step* until convergence.

At convergence, the set of $\nu_{k,l}(\boldsymbol{x})$ is the final approximation of $o_{k,l}(\boldsymbol{x})$ in matrix $\boldsymbol{O}(\boldsymbol{x})$. Thus, vector $\boldsymbol{m}(\boldsymbol{x})$ obtained by using this iterative algorithm is the membership vector that expresses the fuzzy information included in matrix $\boldsymbol{O}(\boldsymbol{x})$. It is possible to show that the elements $m_k(\boldsymbol{x})$ after the *Initialization Step* are in the same order as at the elements of $\boldsymbol{m}(\boldsymbol{x})$ at convergence [39]. However, values estimated during the *Initialization Step* tend to underestimate differences between memberships of the unknown pixel toward different classes. The pairwise coupling algorithm stretches the values of $m_k(\boldsymbol{x})$ to obtain better $o_{k,l}(\boldsymbol{x})$ approximations according to (21).

C. Discussion

The proposed FOAA and FOAO strategies have the same goal: estimating a fuzzy membership vector $\boldsymbol{m}(\boldsymbol{x})$ for all the unknown pixels in a multiclass problem. However, they generally do not reach the same numerical results and do not achieve the same overall accuracy in classification. For this reason, it is important to point out the properties and the characteristics of the two strategies:

- The FOAA strategy requires only R binary F^2SVM (each one with a single output sigmoid), while the FOAO architecture needs $R(R$-1)/2 binary F^2SVMs (each one with two output sigmoids).
- In the FOAA strategy, each $F^2SVM_{k,\Omega\text{-}k}$ is trained with all the N_f samples composing the training set L_f. In the FOAO strategy, the learning of each $F^2SVM_{k,l}$ is carried out by considering the subset of L_f that contains only patterns belonging to classes ω_k and ω_l. Thus, like in the standard crisp OAA architecture, the *learning time* required by an $F^2SVM_{k,l}$ is generally shorter than the time taken from an $F^2SVM_{k,\Omega\text{-}k}$.
- The FOAA strategy produces directly the fuzzy membership estimation; the FOAO technique, instead, exploits a pairwise coupling procedure that, starting from the outputs of the $F^2SVM_{k,l}$, estimates the membership vector of each unknown pattern. Hence the *classification time* in the FOAO case is longer than in the FOAA architecture.
- If the learning of the F^2SVMs is accurate, we expect that the FOAA strategy can result in a precise

modeling of fuzzy memberships, due to the direct estimation of the output soft information from training data. However if a normalization is not applied to the output the sum-to-one constraint is not guaranteed. The application of the normalization can introduce a bias on the estimated membership grades.

- The FOAO strategy estimates the class memberships of a pixel by exploiting the outputs of all the binary $F^2SVM_{k,l}$. Thus, on the one hand, it has the advantage to jointly consider all the pairwise contributions in the estimation of the membership grades. On the other hand, binary classifiers associated with pair of classes to which the analyzed pixel does not belong contribute to the fuzzy modeling of the output. This may affect the accuracy of the estimate, as these classifiers are unreliable on such pixels[6]. This can become particularly critical when a high number of classes R is considered.

Given a specific pixel unmixing problem, it is difficult to conclude on the best possible multiclass strategy. Such a choice depends on the number of the classes R, the number of available training samples, the distribution of classes in the feature space and the behavior of the abundances of pixels in the considered image.

V. Design of Experiments

A. Definition of Experiments

The first step of our experimental analysis was to test the effectiveness of the proposed F^2SVM on a simulated classification problem. To this end an image with 2 channels including 4 different classes and several fuzzy samples was generated. The goal of this first experiment was to test the effectiveness of F^2SVM in a controlled environment were a high number of labeled samples is available and uniformly distributed among classes. In the second step a possible application domain of F^2SVM for sub-pixel analysis was considered, i.e., remote sensing image classification. To validate the proposed technique in this domain we used two images acquired by two remote sensing sensors having different properties. The first one is a very high spatial resolution multispectral image; the second one has similar spatial properties but is an hyperspectral image (characterized by hundreds of channels that are associated with different portions of the electromagnetic spectrum). For both images a fuzzy training set (with N samples) and 2 uncorrelated fuzzy test sets were defined. The sub-pixel (fuzzy) information for each sample in the training and test sets was collected by a ground truth survey or by a proper photo-interpretation of the scene under investigation. The details about these images and the related data sets are described in the next subsections.

For all datasets in order to assess the effectiveness of the proposed F^2SVMs, we analyzed the accuracy from both the fuzzy and the crisp perspectives. With regard to sub-pixel properties, we defined the following *fuzzy* (*soft*) *accuracy measure*:

[6] This issue could be addressed by considering more complex pairwise coupling algorithms that apply a preliminary thresholding to the pairwise output of binary $F^2SVM_{k,l}$ [40].

$$a_f = 1 - \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{R}\left|M_k(\boldsymbol{x}_i) - m_k(\boldsymbol{x}_i)\right| \Big/ \sum_{p=1}^{R} M_p(\boldsymbol{x}_i) + \sum_{q=1}^{R} m_q(\boldsymbol{x}_i)\right) \tag{25}$$

where $M_k(\boldsymbol{x}_i)$ is the known membership degree for the i-th pixel toward class ω_k ($M_k(\boldsymbol{x}_i) \in \boldsymbol{M}(\boldsymbol{x}_i)$), while $m_k(\boldsymbol{x}_i)$ is the estimated membership value (fuzzy output) produced by the classifier. a_f can assume values in the range between 0 and 1. It has value 1 (fuzzy classification accuracy equal to 100%) only if all $m_k(\boldsymbol{x}_i)$ are equal to the correspondent $M_k(\boldsymbol{x}_i)$, whereas it has value 0 (fuzzy classification accuracy equal to 0%) when the estimated memberships are completely different from the given abundances. With regard to the crisp accuracy, we computed the *crisp overall accuracy* derived by considering the hardened output of the F^2SVM (given a pixel we can always convert fuzzy information in crisp information selecting the class with the maximum membership grade).

As in standard crisp SVMs, the model selection of F^2SVM requires to define the kernel function (and to estimate its parameters) and the regularization parameter C. In our experiments, for all images, we used a Gaussian Radial Basis Function (RBF) kernel function (which requires only the tuning of the Gaussian width σ) [see (10)] which is a universal kernel that includes other valid kernels as a particular case [41]. In all data sets input features were normalized between 0 and 1, and the spread of the kernel functions were fixed to be the same for all kernels. We derived the optimum parameter values (C and σ) according to an empirical grid-search model selection carried out with exponentially increasing sequences of values in the following ranges: $C \in [10^{-1}, 50]$ and $\sigma \in [10^{-4}, 10^{-2}]$. It is worth noting that different trials were carried out considering in the multiclass architectures binary F^2SVMs having: i) C and σ values optimized separately; ii) the same C and σ values. These trials resulted in similar accuracies; thus, for simplicity, in the paper we report the results obtained using the same values for all binary F^2SVM.

For the simulated data set the effectiveness of the F^2SVM was evaluated according to a 3-fold cross validation (CV) approach. The best parameter values that maximize the average overall accuracy on the 3 folds alternately used as test set were selected. Concerning remote sensing data sets, model selection was performed according to a 2-fold CV on test sets. In this case, since the number of available labeled fuzzy samples was small, we defined a training set with a sufficient number of samples for a reliable learning of the classifier, and two additional test sets with a smaller number of samples for validation. The classifier was first trained on the training set, and the optimal parameters were selected as those that maximizes the average accuracy on the test sets. Other approaches to model selection can be considered like leave-one-out, radius-margin bounds, span bounds [42],[43], etc.

B. Reference technique: Fuzzy Multi-Layer Perceptron Neural Networks

In order to understand the validity of the proposed F^2SVM, we compared the accuracies provided by this technique with those yielded by a *Fuzzy Multi-Layer Perceptron* (*FMLP*) neural network applied to pixel unmixing [8]. We selected this classifier because it is a widely used neuro-fuzzy inductive learning algorithm, in which fuzzy set theoretic concepts are combined with a mechanism of learning from data. FMLPs have been applied with success to many fields, included problems of spectral unmixing (sub-pixel classification) in the analysis of remote sensing images [8]. For this reason they represent a valuable reference for the proposed F^2SVM on the considered data set. Many different implementation of FMLP neural networks have reported in the literature, which are characterized by different complexity, efficiency, and computational speed. The FMLP technique we considered incorporates fuzzy set theoretic concepts in both input and output stages: the input vector consists of membership values to information classes, while the output vector is defined in terms of fuzzy class membership values (i.e. soft output values).

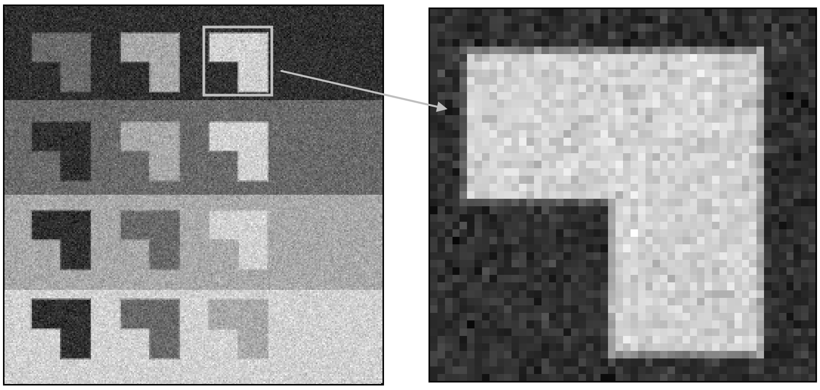
We considered a fully connected feedforward neural network architecture composed of an input layer, one (or more) hidden layers, and one output layer. In our experiments we used as many neurons in the input layer as the number of features that characterize each pixel; the output layer consisted of a number of neurons equal to the number of classes. Input neurons just propagate input features to the next layer. As activation function of the neurons in the hidden layers and in the output layer, we used the sigmoid function. It is worth noting that the adopted architecture models the soft output of the FMLP classifier according to a sigmoid function and no additional transformation are considered. This is motivated from the choice to carry out experiments in which the soft outputs of the proposed F^2SVM classifier and of the reference one are modeled by the same function. The learning of the FMLP algorithm was carried out according to an error backpropagation algorithm applied to a cost function based on the MSE error. The error on each training pattern to each class was properly weighted according to the corresponding fuzzy memberships. An adaptive learning rate was considered in the error backpropagation algorithm [8].

Different FMLP architectures were analyzed in our experiments on the two considered data sets. The numbers of hidden layers and neurons in the hidden layers were determined according to a tradeoff between complexity of representation and generalization ability of the net, according to standard empirical rules [8],[44]. We analyzed architectures with one or two hidden layers, and for each architecture we carried out three trials with different values of the initial weights. Finally, for each data set, we selected the architecture and the trial that resulted in the highest fuzzy accuracy on the test set.

VI. Experimental Results: Simulated Dataset

The simulated dataset is an image of 256×256 pixels with 2 channels and represents a 4-classes classification problem (R=4). The spatial distribution of classes in the image was designed such that there exist boundaries between all possible pairs of classes. Along these boundaries a set of fuzzy samples were introduced that simulate the gradual transition from a class to the other. From this image two different fuzzy problems were generated that show a different noise level. The first problem is represented by the simulated 2-channel image with the addition of a Gaussian noise characterized by a standard deviation σ_N

= 15, while the second one was obtained by adding a Gaussian noise with $\sigma_N = 25$. Figure 10.a shows the first channel of the most noisy image and Figure 10.b a detail of it, where the mixed-pixel region between class 1 and 4 is visible.



(a) (b)

Figure 10. Simulated 256×256 pixels image corrupted by a Gaussian noise with σ_N = 25: (a) first channel, (b) zoom of the detail in white square.

The effectiveness of the proposed method was tested on 7008 randomly selected labeled pixels. Among them there are 5088 *pure* pixels (patterns that belong only to one information class) equally distributed among classes, and 1920 *mixed* pixels (patterns that have memberships to more than one class). Mixed pixels have membership values different from zero for two classes, with values in the set {0.25, 0.5, 0.75}. They are almost uniformly distributed among all the possible pairs of abundances. For each class there are 960 *mixed* pixels (each mixed pattern is considered one time for each class it belongs to). The set of labeled patterns was used for model selection according to a 3-fold CV strategy. The three folds are made of 2110, 2458 and 2440 pixels and were built preserving the relative frequency that pure and mixed pixels show in the whole set.

First of all we carried out the model selection for both the proposed F^{2}SVM (FOAA and FOAO architectures) and the FMLP neural network on both simulated data sets. Table I summarizes the optimum parameter values for the three classifiers, i.e. the ones that resulted in the highest average accuracy on the 3 folds when used as test set. The same parameter values were used for all the binary F^{2}SVMs making up each multiclass architecture. With regard to the FMLP neural network, the learning was carried out with the error back-propagation algorithm with learning rate equal to 0.001.

TABLE I
OPTIMUM PARAMETER VALUES OBTAINED WITH THE 3-FOLD CV (SIMULATED DATA SET)

Classification technique	Parameter	σ_N= 15	σ_N= 25
F^{2}SVM (FOAA)	C	0.01	0.01
	σ	0.5	0.3
F^{2}SVM (FOAO)	C	2.51	3.98
	σ	0.1	0.1
FMLP	# neurons	16	20

TABLE II
OVERALL CLASSIFICATION ACCURACIES PROVIDED BY THE PROPOSED F^2SVM (WITH THE FOAA AND FOAO ARCHITECTURES) AND BY THE FMLP CLASSIFIER WITH THE 3-FOLD CV (SIMULATED DATA SET).

Classification technique	σ_N= 15		σ_N= 25	
	Overall *fuzzy* accuracy (%)	**Overall *crisp* accuracy (%)**	**Overall *fuzzy* accuracy (%)**	**Overall *crisp* accuracy (%)**
F^2SVM (FOAA)	82.94	85.71	80.52	83.92
F^2SVM (FOAO)	80.82	86.00	77.55	84.01
FMLP	80.47	83.79	77.21	82.08

Table II reports for all the classifiers the fuzzy and crisp accuracies obtained in average on the 3 folds with the parameter values reported in Table I. From an analysis of Table II, it is possible to observe that the F^2SVM approach exhibited a higher overall fuzzy accuracy than the FMLP neural network. The overall fuzzy accuracy improvement achieved using the FOAA strategy is of about 2.5% for the simulated data set with σ_N= 15, while it is higher than 3% for the image with σ_N= 25 (in both cases it is of about 0.5% with the FOAO strategy). The improvement in the overall crisp accuracy is also around 2%. In this classification problem the overall fuzzy accuracy achieved using the FOAA strategy is higher than the one obtained with the FOAO strategy (we obtained a difference of about 1-3%).

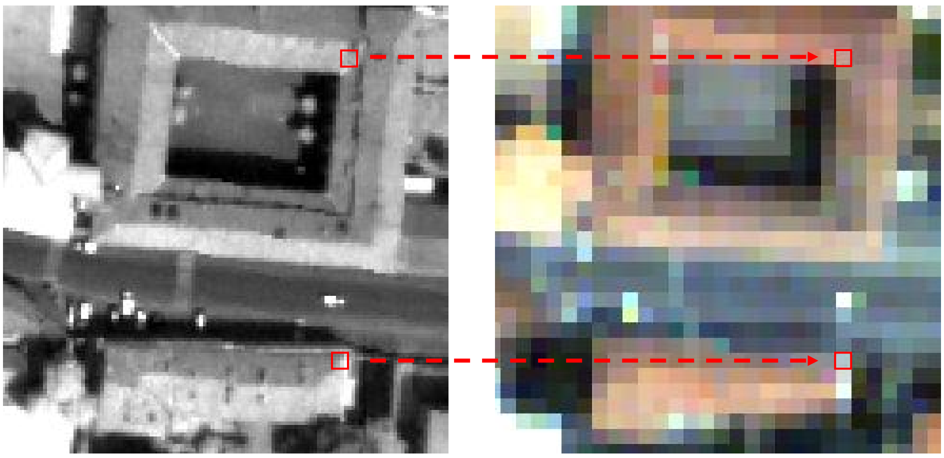
It is worth noting that the aforementioned results are interesting as they point out the superiority of the proposed F^2SVM on the fuzzy MLP in a quite simple problem with few classes and a proper number of fuzzy samples. In other words, this is a set up where the main properties of SVM are not fully exploited.

VII. EXPERIMENTAL RESULTS: MULTISPECTRAL IMAGE

The first remote sensing data set used in the experiments is associated with a very high geometrical resolution image acquired by the QuickBird satellite on the city of Pavia (Italy) (see Figure 11). This satellite takes a panchromatic high resolution image (60cm) and, simultaneously, a multispectral image made up of four spectral bands, with lower geometric resolution (2.4m). This means that one pixel in the multispectral image is mapped into sixteen pixels in the panchromatic image, as shown in Figure 12. We used the high spatial resolution of the panchromatic image for manually defining the fuzzy training set (which is given as input to the classifier in the learning phase) and the 2 test sets (which are used for accuracy assessment according to a 2-fold validation strategy) with sub-pixel information for the multispectral image. In greater detail, for each pixel in the multispectral image, we defined a ground truth membership vector analyzing the labels of the corresponding sixteen pixels in the panchromatic image. For example, a pixel in the multispectral image that corresponds to twelve pixels of road and to four pixels of red roof in the panchromatic image was assigned in the training set to the class “road” with membership 0.75 and to the class “red roof” with membership 0.25. This time consuming process for definition of reference data was carried out with high precision in order to obtain a reliable benchmark for assessing the effectiveness of the proposed sub-pixel classification technique. On this data set we defined R=11 information classes, a training set of N=376 samples, and two test sets of 159 and 228 samples, respectively. Patterns belonging to the three sets were extracted from different portions of the image in order to obtain as more uncorrelated as possible data sets. We assumed an exhaustive knowledge and representation of information classes present on the ground. It is worth noting that the problem complexity is very high because the ratio between R and N is small and training patterns are not uniformly distributed among classes.

Figure 11. Panchromatic image (1024x1024 pixels) acquired by the QuickBird satellite on the city of Pavia, Italy (multispectral data set).

Table III shows the number of *pure* pixels and *mixed* pixels included in the training and test sets (each mixed pattern is considered one time for each class it belongs to). Table IV presents the membership grades of all the mixed pixels for which reference data are available. From an analysis of the tables it is possible to observe that mixed pixels are distributed among 5 classes (i.e. road, red roof, trees. shadow, grey roof) with different abundances, whereas 6 classes contain exclusively pure pixels. This points out the complexity of the benchmark considered.



(a) (b)

Figure 12. Relationship between the geometrical resolution of (a) the panchromatic image and (b) the multispectral image (multispectral data set).

TABLE III
NUMBER OF PURE AND MIXED PIXELS INCLUDED IN THE TRAINING AND TEST SETS (MULTISPECTRAL DATA SET)

Class	Number of pixels					
	Training set		Test set 1		Test set 2	
	Pure	Mixed	Pure	Mixed	Pure	Mixed
Grass (ω_1)	51	0	20	0	38	0
Road (ω_2)	49	48	13	28	22	26
Water (ω_3)	13	0	5	0	2	0
Red roof (ω_4)	28	51	12	31	18	32
Trees (ω_5)	65	29	26	11	31	17
Shadow (ω_6)	16	79	8	36	21	54
White roof (ω_7)	14	0	4	0	4	0
Light grey roof (ω_8)	6	0	4	0	5	0
Grey roof (ω_9)	10	1	6	0	5	2
Dark roof (ω_{10})	12	0	6	0	13	0
Black roof (ω_{11})	8	0	2	0	3	0

TABLE IV
MEMBERSHIP GRADES OF ALL MIXED PIXELS FOR WHICH REFERENCE DATA ARE AVAILABLE (MULTISPECTRAL DATA SET)

Number of pixels	Membership grade		Number of pixels	Membership grade		Number of pixels	Membership grade		Number of pixels	Membership grade	
	ω_4	ω_6		ω_5	ω_4		ω_2	ω_6		ω_2	ω_4
5	12.50	87.50	5	6.25	93.75	6	18.75	81.25	5	12.50	87.50
5	31.25	68.75	5	18.75	81.25	8	31.25	68.75	8	31.25	68.75
3	37.50	62.50	3	25.00	75.00	6	43.75	56.25	6	43.75	56.25
3	43.75	56.25	4	31.25	68.75	3	50.00	50.00	8	50.00	50.00
9	50.00	50.00	3	37.50	62.50	4	56.25	43.75	5	56.25	43.75
5	56.25	43.75	3	43.75	56.25	6	62.50	37.50	6	68.75	31.25
5	62.50	37.50	5	50.00	50.00	4	68.75	31.25	4	75.00	25.00
7	68.75	31.25	5	56.25	43.75	4	75.00	25.00	4	81.25	18.75
10	75.00	25.00	3	62.50	37.50	3	81.25	18.75	4	87.50	12.50
4	81.25	18.75	5	68.75	31.25	5	87.50	12.50			
8	87.50	12.50	4	75.00	25.00		ω_6	ω_9		ω_2	ω_5
			3	81.25	18.75	3	12.50	87.50	3	50.00	50.00
			6	93.75	6.25						

First of all we carried out the model selection for both the proposed F^{2}SVM and the FMLP neural network. The optimum parameter values for the FOAA strategy resulted C=20 and $\sigma=5\cdot10^{-3}$, while for the FOAO strategy were C=28 and $\sigma=78\cdot10^{-3}$. The same values were used for all the binary F^{2}SVMs making up each multiclass architecture. With regard to the FMLP neural network, the classifier that resulted in the highest overall accuracy on the test set was made up of one hidden layer with 16 nodes (learning rate equal to 0.001). Table V reports the highest fuzzy and crisp accuracies obtained in average on the two test sets and on training set for all the classifiers. From an analysis of this table, it is possible to observe that the F^{2}SVM approach exhibited a sharply higher overall fuzzy accuracy than the FMLP neural network. In greater detail, considering the test sets, the overall fuzzy accuracy improvement achieved using the FOAA strategy is of about 16% (it is of about 11% with the FOAO strategy). The improvement in the overall crisp accuracy is smaller. This can be explained by the fact that on this data set, although the FMLP modeled the fuzzy membership values of pixels with a significantly lower accuracy than the F^{2}SVM, it preserved the proportions of the membership grades to the different classes in output from the classifier, thus obtaining relatively good crisp accuracies after hardening. In this classification problem the overall fuzzy accuracy achieved using the FOAA strategy is higher than the one obtained with the FOAO strategy (we obtained a difference of about 12% on the training set, and of about 5% on the average accuracy of test sets). This is probably due to the problem complexity involved from the high number of classes, which decreases the effectiveness of the Pairwise Coupling algorithm (as discussed in Sec. IV.C there are too many unreliable pairwise contributions in the computation of the membership grades of each pixel).

TABLE V
OVERALL CLASSIFICATION ACCURACIES PROVIDED BY THE PROPOSED F^{2}SVM (WITH THE FOAA AND FOAO ARCHITECTURES) AND BY THE FMLP CLASSIFIER (MULTISPECTRAL DATA SET).

Classification technique	Overall *fuzzy* accuracy (%)		Overall *crisp* accuracy (%)	
	Training set	Average on Test sets	Training set	Average on Test sets
F^{2}SVM (FOAA)	85.59	78.81	91.09	82.86
F^{2}SVM (FOAO)	73.91	73.52	86.04	84.44
FMLP	63.41	62.81	83.76	82.89

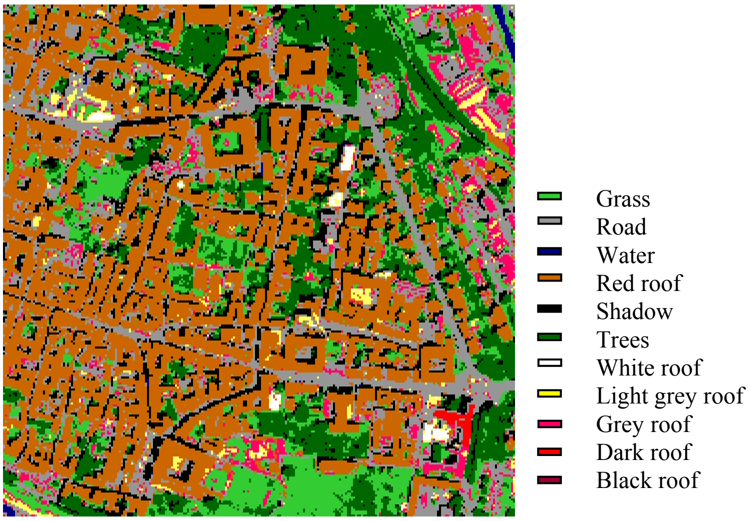


Figure 13. Classification map obtained by the hardened output of OAO-F^{2}SVM (multispectral data set).

The crisp classification map obtained by hardening the output of the F^{2}SVM classifier is reported in Figure 13. A comparison between this map and the one obtained by the FMLP neural network (not reported for space constraints) confirms the quantitative results and points out the good precision of the F^{2}SVM output. Similar conclusions can be drawn by a qualitative analysis of the abundances (fuzzy) maps (not reported for space constraints).

VIII. EXPERIMENTAL RESULTS: HYPERSPECTRAL IMAGE

The second remote sensing data set used for F^{2}SVM validation is made up of an hyperspectral image with 115 spectral channels acquired in different parts of the electromagnetic spectrum by the airborne ROSIS sensor. The image represents the San Felice lagoon area, near Venice (Italy), which is characterized from the presence of salt-marsh vegetation. Although the spatial resolution of each pixel is high also on this image (i.e., 1m), the species spatial variability results very high (in the scale of tens of centimeters) due to the particular kind of ecosystem. This peculiarity of salt-marsh vegetation makes this dataset particularly suitable for testing the robustness of the proposed sub-pixel classification algorithm [45]. The goal of this image classification problem is to describe the land-covers according to the identification of six information classes (R=6) associated with four different vegetation species [Spartina Maritima (ω_1), Liboneum Narbonese (ω_2), Juncus Maritimus (ω_3), Sarcocornia Fruticosa (ω_4)], Bare Soil (ω_5), and Water (ω_6). Figure 14 shows a false color composition of three spectral channels of the image. From the available 115 spectral bands, we selected the 17 more informative bands (D = 17) by excluding 35 noisy channels and then applying a feature-selection procedure based on the Jeffries-Matusita distance and the *Steepest Ascent* search method [46].

The training set and the two test sets were defined on the basis of a ground truth data collection procedure focused on the analysis of sub-pixel information. This data collection was carried out in the framework of the European project Hysens 2000 [45]. Several Region Of Interests (ROIs) were identified with a size of at least 3×3 pixels. The ROI boundaries were positioned according to the use of either differential GPS or laser theodolite. To assign membership grades to the ground truth points, an accurate analysis was carried out by independent operators according to the Braun-Blanquet visual method [47]. This analysis was supported by several high resolution (i.e., 2mm) digital photographs acquired within ROIs. Table VI reports the number of pure and mixed pixels included in the mentioned sets (each mixed pattern is considered one time for each class it belongs to). The fractional abundances of information classes within the defined ROIs are summarized in Table VII. From an analysis of the table, it is possible to observe that mixed pixels are distributed (with different abundances) among the four considered vegetation classes and the Bare Soil class. In greater detail, we can note that all the samples of classes Liboneum Narbonese (ω_2) and Juncus Maritimus (ω_3) are mixed samples. Class Water (ω_6), instead, does not share any pattern with the other classes. It is worth noting that from the viewpoint of the distribution of the soft information, this problem is less complex than that associated with the multispectral data set presented in the previous section. However, it is challenging as a relatively high number of features are provided as input to the classifier.

As in the previous data set, we derived the parameter values that resulted in the highest average accuracy on the test sets for all the classifiers. The best parameters of F^{2}SVM using FOAA strategy were σ=5·10^{-4} and C=13. The highest accuracy using the FOAO strategy was obtained with C=3.5 and σ=7·10^{-4}. Concerning the architecture of the FMLP neural network used for comparison, the highest overall fuzzy accuracy on the test set was obtained with one hidden layer made up of 16 nodes. The learning was carried out with the error back propagation algorithm with learning rate equal to 0.001.

Table VIII reports fuzzy and crisp accuracies obtained for training and test sets. By analyzing the table, one can observe that the proposed F^{2}SVM (with both the FOAA and the FOAO architectures) significantly increased both the fuzzy and crisp accuracies yielded by the FMLP neural network classifier.

Figure 14. False color composition of three spectral channels of the ROSIS image acquired on the Venice lagoon, Italy (hyperspectral data set).

TABLE VI
NUMBER OF PURE AND MIXED PIXELS INCLUDED IN THE TRAINING AND TEST SETS (HYPERSPECTRAL DATA SET)

Class	Number of pixels					
	Training set		Test set 1		Test set 2	
	Pure	Mixed	Pure	Mixed	Pure	Mixed
Spartina Maritima (ω_1)	27	38	9	18	16	24
Liboneum Narbonese (ω_2)	0	645	0	277	0	347
Juncus Maritimus (ω_3)	0	556	0	237	0	347
Sarcocornia Fruticosa (ω_4)	43	129	19	57	28	82
Bare Soil (ω_5)	79	154	35	65	51	154
Water (ω_6)	199	0	84	0	122	0

We obtained the best result using the FOAA multiclass strategy, which increased both the fuzzy overall accuracy and the crisp overall accuracy provided in average by the FMLP on the test sets of about 11% and 12%, respectively. This points out the effectiveness of the proposed approach, that provided a significantly better modeling of the sub-pixel information than the FMLP neural classifier. F^{2}SVM shows both good learning capabilities (we observed an improvement of about 17% in the fuzzy training accuracy and 16% in the crisp training accuracy) and good generalization capabilities (as proved by the accuracies on test sets).

The crisp classification map obtained by hardening the output of the F^{2}SVM classifier is reported in Figure 15. A comparison between this map and the one obtained by the FMLP neural network (not reported for space constraints) confirms the quantitative results and points out the high precision of the

F^2SVM output. Similar conclusions can be drawn by a qualitative analysis of the abundance (fuzzy) maps.

TABLE VII
MEMBERSHIP GRADES OF ALL MIXED PIXELS FOR WHICH REFERENCE DATA ARE AVAILABLE (HYPERSPECTRAL DATA SET)

Number of pixels	Membership grade		Number of pixels	Membership grade	
	ω_2	ω_5		ω_1	ω_2
77	0.50	0.50	80	0.90	0.10
	ω_2	ω_3		ω_2	ω_4
221	0.40	0.60	69	0.90	0.10
64	0.30	0.70	199	0.10	0.90
108	0.20	0.80		ω_5	ω_3
396	0.90	0.10	57	0.40	0.60
55	0.10	0.90	239	0.20	0.80

TABLE VIII
OVERALL CLASSIFICATION ACCURACIES PROVIDED BY THE PROPOSED F^2SVM WITH THE FOAA AND FOAO ARCHITECTURES AND BY THE FMLP CLASSIFIER (HYPERSPECTRAL DATA SET)

Classification technique	Overall *fuzzy* accuracy (%)		Overall *crisp* accuracy (%)	
	Training set	Average on Test sets	Training set	Average on Test sets
F^2SVM (FOAA)	88.67	82.02	95.31	89.37
F^2SVM (FOAO)	84.01	77.62	95.63	89.01
FMLP	72.07	71.07	79.32	78.43

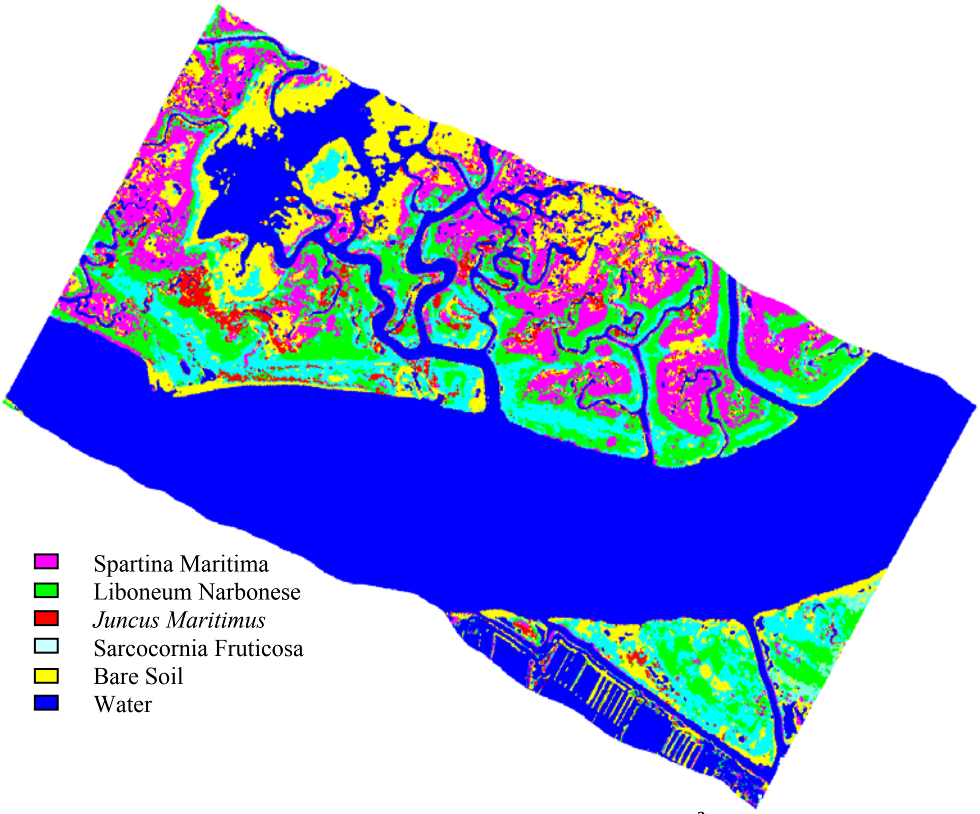


Figure 15. Classification map obtained by the hardened output of the F^2SVM (hyperspectral data set).

IX. DISCUSSION AND CONCLUSION

In this paper, a novel Fuzzy-input Fuzzy-output SVM (F^2SVM) technique for binary and multicategory pixel unmixing in image classification has been proposed. The proposed F^2SVM technique is able to learn the sub-pixel information inherent a fuzzy training set and to estimate the abundances (fuzzy memberships) of unknown pixels to different classes. The presented classifier explicitly manages in a nonlinear way sub-pixel information associated with each pixel, both in binary problems and in multicategory problems (thanks to the proposed multiclass strategies FOAA and FOAO). This is a very important property in image classification problems, as in many applications the geometrical resolution of the sensor is not sufficient for guaranteeing that pixels represent only the radiometric response of a single information class present in the investigated scene. In this critical situation, on the one hand, standard crisp classifiers do not allow to properly modeling the complexity of the signal associated with the images and thus provide unreliable outputs; on the other hand, the use of a crisp learning strategy for mixed pixels misleads the classifier on the true radiometric properties of classes during the training phase.

Besides the global architecture of the classifier and the idea to exploit the SVM approach to solve spectral unmixing problems, the main specific novelties of the proposed F^2SVM are the following:

i. The use of input fuzzy membership information to model the sub-pixel abundances of unknown patterns in the learning of SVM;
ii. The proposed fuzzy output estimation method (which is based on adapted sigmoid functions that relate pattern distances from the hyperplane with the estimated membership behavior);
iii. The multiclass FOAA and FOAO strategies (which generalize to the fuzzy case the standard OAA and OAO techniques).

It is worth noting that the proposed F^2SVM has all the desirable properties of the crisp supervised SVM approach, i.e.: i) convexity of the cost function used in the learning of the classifier; ii) robustness to the effects of the Hughes phenomenon when dealing with a high-dimensional feature space; iii) sparsity of the solution that results in very good generalization capabilities; iv) possibility to be implemented in parallel architectures.

Experimental results obtained on three data sets associated with images having different properties confirm the effectiveness of the proposed F^2SVM, which provided sharply higher fuzzy accuracies (especially in the case of real remote sensing image classification problems) than a FMLP neural network and satisfactory abundance maps. These results were expected due to the aforementioned properties of F^2SVM and point out that the proposed technique seems very promising for sub-pixel image classification.

With regard to the presented multiclass strategies, in all our experiments the highest fuzzy accuracies were obtained by the FOAA strategy, which outperformed the FOAO method. This is due to the fact that the FOAO architecture estimates the fuzzy memberships of a pixel by considering the outputs of all the pairwise classifiers, thus including in the estimation also binary classifiers associated with classes that have no relationships with the pixel. This results in the use of unreliable outputs in the final computation of the memberships, thus mitigating the potential advantage of the joint processing of the output of all binary classifiers in the computation of the class abundances.

The main drawback of the proposed method is the need of having as input to the classifier soft information about labeled samples for which the fuzzy memberships (abundances) to the different classes should be known. This information is available (or can be collected) in some application domains, whereas it is difficult to have in others. Another limitation of the proposed technique is associated with the relatively high computational load required from the learning of the classifier. As in standard supervised crisp SVM, this time is mainly due to the model selection phase, which requires to test many combinations of the values of the regularization parameter *C* and the kernel parameters for an adequate modeling of the considered problem. Nonetheless, this computational load is not higher than that required from other machine learning classifiers (e.g. the considered FMLP neural network).

Future developments of this work are devoted: i) to address the main drawback related to the FOAO strategy for multiclass problems by adaptively selecting for each pixel the relevant binary classifiers to include in the Pairwise Coupling processing; ii) to apply the F^2SVM technique to other image classification problems by considering different application domains, and iii) to include in the sub-pixel classification procedure also the use of the information present in the spatial neighborhood system of each pixel.

ACKNOWLEDGEMENTS

This work was supported by the Italian Ministry of Education, University and Research (MIUR). The ROSIS data set used in the experimental part of this paper has been acquired and defined in the context of the European project Hysens 2000 DLR-EU. The authors wish to thank Dr. Andrea Garzelli of the University of Siena (Italy) for providing the Quickbird image used in the experiments, and Mr. Benedetto Borasca and Mr. Michele Zusi for their contribution to a preliminary version of this work.

REFERENCES

[1] A. Baraldi, L. Bruzzone, P. Blonda, "A multiscale expectation-maximization semisupervised classifier suitable for badly posed image classification" *IEEE Trans. Image Proc.*, vol. 15, no. 8, pp. 2208-2225, 2006.

[2] L. Bruzzone, D. Fernández Prieto, "An adaptive semi-parametric and context-based approach to unsupervised change detection in multitemporal remote sensing images", *IEEE Trans. Image Proc.*, vol. 11, no. 4, pp. 452-466, 2002.

[3] L. Miao, H. Qi and H. Szu, "A maximum entropy approach to unsupervised mixed-pixel decomposition," *IEEE Trans. Image Proc.*, vol. 16, no. 4, pp. 1008-1021, Apr. 2007.

[4] H. Tsurui, H. Nishimura, S. Hattori, S. Hirose, K. Okumura and T. Shirai, "Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition," *J. Histochemi. Cytochem.*, vol. 48, no. 5, pp. 653-662, 2000.

[5] M. E. Dickinson, G. Bearman, S. Tille, R. Lansford and S. E. Fraser, "Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy," *BioImag.*, vol. 31, no 6, pp. 1272-1278, 2001.

[6] N. Keshava, "A survey of spectral unmixing algorithms," *Lincoln Lab. J.*, vol. 14, no. 1. pp. 55-78, 2003.

[7] K. J. Guilfoyle, M. L. Althouse and C.-I. Chang, "A quantitative and comparative analysis of linear and nonli-

near mixture models using radial basis function neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 8, pp. 2314-2318, 2001.

[8] A. Baraldi, E. Binaghi, P. Blonda, P. A. Brivio, A. Rampini, "Comparison of the multilayer perceptron with neuro-fuzzy techniques in the estimation of cover class mixture in remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 5, pp. 994-1005, 2001.

[9] V. Vapnik, "An Overview of Statistical Learning Theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, 1999.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, *Springer*, 1995.

[11] V. Vapnik, *Statistical Learning Theory*, *John Wiley*, 1998, NY, p. 732.

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

[13] 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, 2005.

[14] A. Baraldi, L. Bruzzone, P. Blonda and L. Carlin, "Badly posed classification of remotely sensed images-an experimental comparison of existing data labeling systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 1, pp 214-235, 2006.

[15] L. Bruzzone and L. Carlin, "A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587-2600, 2006.

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

[17] M. Kanevski, A. Pozdnukhov, S. Canu and M. Maignan, "Advanced Spatial Data Analysis and Modeling with Support Vector Machines", *Int. J. Fuzzy Systems*, vol. 4, no. 1, pp. 606-615, 2002.

[18] J.X. Dong, A. Krzyzak and C.Y. Suen, "Fast SVM training algorithm with decomposition on very large data sets," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 603–618, 2005.

[19] C. Lin and S. Wang, "Fuzzy Support Vector Machines", *IEEE Trans. Neural Networks*, vol. 13, no. 2, 2002.

[20] E. C. C. Tsang, D. S. Yeung and P. P. K. Chan, "Fuzzy Support Vector Machines for solving Two-Class problems," *Proc. Second Int. Conf. Machine Learning and Cybernetic*, Wan, 2-5 Nov. 2003.

[21] D. Tsujinishi and S. Abe, "Fuzzy least squares support vector machines for multiclass problems," *Neural Networks*, vol. 16, no. 5-6, pp. 785-792, 2003.

[22] Jayadeva, R. Khemchandani and S. Chandra, "Fast and robust learning through Fuzzy Linear Proximal Support Vector Machines," *Neurocomputing*, vol. 61, pp. 401-411, 2004.

[23] H. Huang and Y. Liu, "Fuzzy Support Vector Machines for Pattern Recognition and Data Mining," *Int. J. Fuzzy Systems*, vol. 4, no. 3, 2002.

[24] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[25] G. Wahba, "Support vector machines, reproducing kernel hilbert spaces, and randomized gacv," in *Advances in Kernel Methods: Support Vector Learning*, Eds. by B. Schoelkopf, C.J.C. Burges and A.J. Smola, MIT Press, ch. 6, pp. 69-87, 1998.

[26] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. 14, pp. 326–334, 1965.

[27] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition," in *Artificial Neural Networks: Concepts and Theory*, P. Mehra and B. Wah, Eds. Los Alamitos, CA: *IEEE Comput. Soc. Press*, 1992. Reprinted.

[28] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. Roy. Soc.,* London 1909,

[29] B. Schölkopf and A. Smola, Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, MA: MIT Press, 2001.

[30] C.W. Hsu and C.J. Lin, "A comparison of methods for multiclass support vector machines," *Technical report, Dept. of Computer Science and Information Engineering, National Taiwan University*, Taipei, Taiwan, 2001.

[31] J. Furnkranz, "Round Robin Classification," *J. Machine Learning Res.*, vol. 2, pp. 721-747, 2002.

[32] B. Borasca, L. Bruzzone, L. Carlin, M. Zusi, "A Fuzzy-input Fuzzy-output SVM Technique for Classification of Hyperspectral Remote Sensing Images", in Proc. NORSIG 2006, Reykjavík, Iceland, 7-9 June, 2006.

[33] J. C. Platt, "Fast training of Support Vector Machines using Sequential Minimal Optimization," *Advances in kernel methods - Support Vector Learning*, 1998.

[34] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, "Improvements to Platt's SMO Algo-

rithm for SVM Classifier Design," *Technical report, Dept of CSA, IISc*, Bangalore, India, 1999.
[35] E. Osuna, R. Freund and F. Girosi, "Improved Training Algorithm for Support Vector Machines," *Proc. IEEE NNSP*, 1997.
[36] C. Burges, "A tutorial on support vectormachines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.
[37] J. C. Platt, "Probabilistic Outputs for Support Vectors Machines and Comparisons to Regularized Likelihood Methods," In: Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D. (eds.): *Advances in Large Margin Classifiers*. Cambridge, MA, MIT Press (1999).
[38] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs I. The method of paired comparisons," *Biometrika,* vol. 39, pp. 324-345, 1952.
[39] T. Hastie and R. Tibshirani, "Classification by Pairwise Coupling," *The Annals of Statistics*, vol. 26, no. 2, 451-471, 1998.
[40] Z. Li; S. Tang, "Face recognition using improved pairwise coupling support vector machines," *Proc. 9th Int. Conf. Neural Information*, (ICONIP '02), Vol. 2, pp. 876 – 880, 2002.
[41] S. S. Keerthi, C. J. Lin, "Asymptotic behaviors od support vector machines with Gaussian kernel," Classifier Design," *Neural Computation,* vol. 15, no. 7, pp. 1667-1689, 2003.
[42] O. Chapelle and V. Vapnik, "Model selection for support vector machines," In *Advances in Neural Information Processing Systems* 12, S.Solla, T. Leen and K. Muller, eds., MIT PRESS, 1999.
[43] O. Chapelle, V. Vapnik, O. Bousquet and S. Mukherjee, "Choosing kernel parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131-159, 2002.
[44] S.B. Serpico, L. Bruzzone and F. Roli, "An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images," Pattern Recognition Letters, Vol. 17, pp. 1331-1341, 1996.
[45] E. Belluco, M. Camuffo, S. Ferrari, S. Modenese, S. Silvestri, A. Marani and M. Marani, "Mapping salt-marsh vegetation by multispectral and hyperspectral remote sensing," *Remote Sensing of Environment*, vol. 105, no. 1, pp. 54-67, 2006.
[46] L. Bruzzone and S. B. Serpico, "A technique for feature selection in multiclass cases," *Int. J. Remote Sen.*, vol. 21, pp. 549-563, 2000.
[47] S. Silvestri, M. Marani, J. Settle, F. Benvento and A. Marani, "Salt marsh vegetation radiometry. Data analysis and scaling," *Remote Sensing of Environment*, vol. 80, pp. 473-482, 2002.