CHAPTER 3.2

APPROACHES BASED ON SUPPORT VECTOR MACHINES TO CLASSIFICATION OF REMOTE SENSING DATA

Lorenzo BRUZZONE and Claudio PERSELLO

Dept. of Information Engineering and Computer Science,
University of Trento,Via Sommarive, 14, I-38050 Trento, Italy;
e-mail: lorenzo.bruzzone@ing.unitn.it; claudio.persello@disi.unitn.it.

This chapter presents an extensive and critical review on the use of kernel methods and in particular of support vector machines (SVMs) in the classification of remote-sensing (RS) data. The chapter recalls the mathematical formulation and the main theoretical concepts related to SVMs, and discusses the motivations at the basis of the use of SVMs in remote sensing. A review on the main applications of SVMs in classification of remote sensing is given, presenting a literature survey on the use of SVMs for the analysis of different kinds of RS images. In addition, the most recent methodological developments related to SVM-based classification techniques in RS are illustrated by focusing on semisupervised, domain adaptation, and context sensitive approaches. Finally, the most promising research directions on SVM in RS are identified and discussed.

1. Introduction

In the last two decades there have been significant improvements both in the technology associated with the development of the sensors used in remote sensing (RS) to acquire signals and images for earth observation and in the analysis techniques adopted for extracting information from these data useful for operational applications. The modern technology resulted in the definition of different kinds of sensors for Earth observation based on different principles and with different properties. We have observed the development of passive multispectral and hyperspectral scanners, and active instruments like Synthetic Aperture Radar (SAR) and Lidar, as well as other instruments devoted to specific applications. Looking at this scenario from an historical viewpoint, we are passed from multispectral data having relatively low spatial and spectral resolution (like the MSS of the Landsat satellites) to a new generation of sensors characterized by very high geometrical resolution (VHR); (e.g., Ikonos, Quickbird, Geoeye-1, etc.), which can acquire images with a metric or sub-metric resolution. Hyperspectral sensors (on airborne or satellite platforms) can acquire images characterized by very high spectral resolution, with hundreds of channels having less than 2 nm of bandwith. The acquisition of very high resolution SAR images from satellite platforms has also become possible thanks to the recent TerraSar-X and Cosmo-Skymed missions. In this context, the challenging properties of new generation of sensors require the definition of novel data analysis methods.

In this chapter we focus our attention on RS image classification methodologies, which are devoted to translate the features that represent the information present in the data in thematic maps representing land cover types according to the solution of a pattern recognition problem. In particular, we concentrate our attention on supervised classification algorithms, which require the availability of labeled samples for the training of the classification model. In this context, the availability of last generation RS images allowed the development of new applications that require the mapping of the Earth surface with high geometric precision and a high level of thematic details. However, the huge amount of data associated with these images requires the development of sophisticated automatic classification techniques capable to obtain accurate land-cover maps in a reasonable processing time.

In the last decades, a great effort has been devoted to exploit machine learning methods for classification of remote sensing images. This has been done by introducing the use of neural networks (NN) in remote sensing (with the pioneering work presented in [1]) for solving many different classification tasks. Several different paradigms and models of NN have been used in recent years for addressing remote sensing image classification problems, ranging from standard Multilayer Perceptron (MLP) network [1-3], to Radial Basis Functions (RBF) neural network [4, 5], structured neural networks [6] and hybrid architectures. Also more complex and structured architecture have been exploited for solving specific problems, like compound classification of multitemporal data [7], multiple classification systems made up of neural algorithms [8, 9], etc. All these methods share as common property the idea to perform the learning of the classification algorithm according to the minimization of the empirical risk, defined in different ways. However, the last frontiers of machine learning classifiers in RS are represented by methods based on the structural risk minimization principle (which allows one to effectively tune the tradeoff between empirical risk and generalization capability) rather than on the empirical risk minimization. The related statistical learning theory (formulated from Vapnik [10]) is at the basis of the support vector machine (SVM) classification approach. SVM is a classification technique based on kernel methods that has been proved very effective in solving complex classification problems in many different application domains. In the last few years, SVM gained a significant credit also in remote sensing applications. The pioneering work of Gualtieri in 1998 [11] related to the use of SVM for classification of hyperspectral images has been followed from several different experiences of other researchers that analyzed the theoretical properties and the empirical performances of SVM applied to different kinds of classification problems [12-28]. The investigations include classification of hyperspectral data [11-18], multispectral images [19-26], VHR images [27], as well as multisource and multisensor classification scenarios [28-30]. SVMs revealed to be very effective classifiers and currently they are among the most adequate techniques for the analysis of last generation of RS data.

In all these cases the success of SVMs is due to the important properties of this approach, which integrated with the effectiveness of the classification procedure and the elegance of the theoretical developments, result in a very solid classification methodology in many different RS data classification domains. As it will be explained in the following section, this mainly depends on the fact that SVMs implement a classification strategy that exploits a margin-based "geometrical" criterion rather than a purely "statistical" criterion. In other words, SVMs do not require an estimation of the statistical distributions of classes to carry out the classification task, but they define the classification model by exploiting the concept of margin maximization.

The main properties that make SVM particularly attractive in RS applications can be summarized as follows [31-33]:

- their intrinsic effectiveness with respect to traditional classifiers thanks to the structural risk minimization principle, which results in high classification accuracies and very good generalization capabilities (especially in classification problems defined in high dimensional feature spaces and with few training samples, which it is a typical situation in the classification of last generation of RS images);
- the possibility to exploit the kernel trick to solve non-linear separable classification problems by projecting the data into a high dimensional feature space and separating the data with a simple linear function;
- the convexity of the objective function used in the learning of the classifier, which results in the possibility to solve the learning process according to linearly constrained quadratic programming (QP) characterized from a unique solution (i.e., the system cannot fall into sub-optimal solutions associated with local minima);

- the possibility of representing the convex optimization problem in a dual formulation, where only non-zero Lagrange multipliers are necessary for defining the separation hyperplane (which is a very important advantage in the case of large datasets). This is related to property of sparseness of the solution;

Moreover, SVMs exhibit important advantages with respect to NN approaches. Among the others we recall: i) higher generalization capability and robustness to the Hughes phenomenon; ii) lower effort required for the model selection in the learning phase (i.e., they involve less control parameters and thus computational time for their optimum values selection) and the implicit automatic architecture definition; iii) optimality of the solution obtained by the learning algorithm.

The objective of this chapter is to review the state of the art of SVM for the classification of RS data. In particular, Section II recalls the basic principles of SVM for pattern classification. Section III presents a literature survey about the most relevant papers that report studies about the application of SVM to the classification of different kinds of RS images and papers that propose advanced systems based on the SVM approach for the analysis of RS data. Along with this state-of-the-art review, we discuss about the operative adoption of SVM for the analysis of RS images and the direction of the future research on this topic. Finally, section IV draws the conclusion of the chapter.

2. Support Vector Machines Classifiers

Let us consider the problem of supervised classification of a generic *d*-dimensional image $\mathcal{I}$ of size $I \times J$ pixel. Let us assume that a training set $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ made up of N pairs $(\mathbf{x}_i, y_i)_{i=1}^{N}$ is available, where $\mathcal{X} = \{\mathbf{x}_i \,|\, \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^{N} \subset \mathcal{I}$ is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ is the corresponding set of labels. For the sake of simplicity, since SVMs are binary classifiers, we first focus the attention on the two-class case (the general multiclass case will be addressed later). Accordingly, let us assume that $y_i \in \{+1; -1\}$ is the binary label of the pattern $\mathbf{x}_i$. The goal of the binary SVM is to divide the *d*-dimensional feature space in two subspaces, one for each class, through a separating hyperplane $H: y = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$. The final decision rule used to find the membership of a test sample is based on the sign of the discrimination function $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ associated to the hyperplane. Therefore, a generic pattern $\mathbf{x}$ will be labeled according to the following rule:

$$f(\mathbf{x}) > 0 \quad \Rightarrow \quad \mathbf{x} \in \mathit{class}\ +1$$
$$f(\mathbf{x}) \le 0 \quad \Rightarrow \quad \mathbf{x} \in \mathit{class}\ -1 \tag{1}$$

The training of an SVM consists in finding the position of the hyperplane *H*, estimating the values of the vector $\mathbf{w}$ and the scalar *b*, according to the solution of an optimization problem. From a geometrical point of view, $\mathbf{w}$ is a vector perpendicular to the hyperplane *H* and thus defines its orientation. The distance of the *H* to the origin is $b/\|\mathbf{w}\|$, while the distance of a sample $\mathbf{x}$ to the hyperplane is $f(\mathbf{x})/\|\mathbf{w}\|$. Let us define the *functional margin* $F = \min\{y_i f(\mathbf{x}_i)\}$, $i = 1,...,N$ and the *geometric margin* $G = F/\|\mathbf{w}\|$. The geometric margin represents the minimum Euclidean distance between the available training samples and the hyperplane.

A. Training of Linear SVM - Maximal Margin algorithm.

In the case of a linearly separable problems, the learning of an SVM can be performed with the maximal margin algorithm, which consists in finding the hyperplane *H* that maximizes the geometric margin *G*. Rescaling the hyperplane parameters $\mathbf{w}$ and *b* such that the functional margin $F=1$, it turns out that the optimal hyperplane can be determined as the solution of the following convex quadratic programming problem:

$$\begin{cases} \min_{\mathbf{w},b}: \frac{1}{2}\|\mathbf{w}\|^2 \\ y_i \cdot [\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b] \ge 1, \quad \forall i = 1,\ldots,N \end{cases} \tag{2}$$

Let H_1 and H_2 be two hyperplane parallel to the separating hyperplane *H* and equidistant from it:

$$H_1 : f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = +1$$
$$H_2 : f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = -1 \tag{3}$$

The goal of the training phase is to find the values of $\mathbf{w}$ and *b* such that the geometric distance between H_1 and H_2 is maximized with the condition that there is no sample between them. Since direct handling of inequality constraints is difficult, Lagrange theory is usually exploited by introducing Lagrange multipliers $\alpha_{i=1}^{N}$ for the quadratic optimization problem. This leads to an alternative dual representation:

$$\begin{cases} \max\limits_{\alpha} : \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i , \mathbf{x}_j \rangle \right\} \\ \sum_{i=1}^{N} y_i \alpha_i = 0, \;\; \alpha_i \geq 0, \quad 1 \leq i \leq N \end{cases} \tag{4}$$

The Karush–Kuhn–Tucker (KKT) complementarity conditions provide useful information about the structure of the solution. They state that the optimal solution $\boldsymbol{\alpha}^*$, $(\mathbf{w}^*, b^*)$ should satisfy:

$$\alpha_i^* \left[y_i (\langle \mathbf{w}^* \cdot \mathbf{x}_i \rangle + b^*) - 1 \right] = 0, \qquad i = 1, \ldots., N \tag{5}$$

This implies that only input samples $\mathbf{x}_i$ for which the functional margin is one (and that therefore lie closest to the hyperplane, i.e., lie on H_1 or H_2) are associated to Lagrange multipliers $\alpha_i > 0$. All the other multipliers α_i^* are zero. Hence, only these samples are involved in the expression for the weight vector. It is for this reason that they are called *support vectors* (SV). Thus we can write that $\mathbf{w}^* = \sum_{i=1}^{N} y_i \alpha_i^* \mathbf{x}_i = \sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i$. It is worth noting that the term b does not appear in the dual problem, and should be calculated making use of the primal constraints:

$$b = -\frac{\max_{y_i=-1}(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle) + \min_{y_i=+1}(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle)}{2} \tag{6}$$

Once the values for $\mathbf{w}$ and b are determined by solving the optimization problem, one generic test sample is classified on the basis of the sign of the discriminant function, that can be expressed as:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \left(\sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i \right) \cdot \mathbf{x} + b = \sum_{i \in SV} y_i \alpha_i^* \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b\,. \tag{7}$$

Note that the training samples appear only in the form of dot product. This property of the dual form will be exploited later to extend the formulation to nonlinear problems.

B. Training of Linear SVM – Soft Margin algorithm.

The maximum margin training algorithm can not be used in many real world problems where the available training samples are not linearly separable because of noisy samples and outliers (this is very common in real RS classification problems). In these cases, the soft margin algorithm is used in order to handle nonlinear separable data. This is done by defining the so called slack variables as:

$$\xi[(\mathbf{x}_i, y_i), (\mathbf{w}, b)] = \xi_i = \max[0, 1 - y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b)] \tag{8}$$

Slack variables allow one to control the penalty associated with misclassified samples. In this way the learning algorithm is robust to both noise and outliers present in the training set, thus resulting in high generalization capability. The optimization problem can be formulated as follows:

$$\begin{cases} \min\limits_{\mathbf{w},b} \left\{ \dfrac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{N}\xi_i \right\} \\ y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) \geq 1-\xi_i,\ \xi_i \geq 0,\ \forall i = 1,\ldots,N \end{cases} \tag{9}$$

where $C \geq 0$ is the regularization parameter that allows one to control the penalty associated to errors (if $C = \infty$ we come back to the maximal margin algorithm), and thus to control the tradeoff between the number of allowed mislabeled training samples and the width of the margin. If the value of *C* is too small, many errors are permitted and the resulting discriminant function will poorly fit with the data; on the opposite, if *C* is too large, the classifier may overfit the data instances, thus resulting in low generalization ability. A precise definition of the value of the *C* parameter is crucial for the accuracy that can be obtained in the classification step and should be derived through an accurate model selection phase.

Similarly to the case of the maximal margin algorithm, the optimization problem (9) can be rewritten in an equivalent dual form:

$$\begin{cases} \max\limits_{\alpha}: \left\{ \sum\limits_{i=1}^{N}\alpha_i - \dfrac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} y_i y_j \alpha_i \alpha_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \right\} \\ \sum_{i=1}^{N} y_i\alpha_i = 0,\ 0 \leq \alpha_i \leq C,\ \ 1 \leq i \leq N \end{cases} \tag{10}$$

Note that the only difference between (10) and (4) is in the constraint on the multipliers $\{\alpha_i\}_{i=1}^{N}$ that for the soft margin algorithm are bounded by the parameter *C*. For this reason this problem is also known as box constrained problem. The KKT conditions become in this case:

$$\begin{cases} \alpha_i[y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) - 1 + \xi_i] = 0, & i = 1,\ldots,l \\ \xi_i(\alpha_i - C) = 0, & i = 1,\ldots,l \end{cases} \tag{11}$$

Varying the values of the multipliers $\{\alpha_i\}_{i=1}^{N}$ three cases can be distinguished:

(i) if $\alpha_i = 0 \Rightarrow \xi_i = 0$ and $y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) \geq 1$;

(ii) if $0 < \alpha_i < C$, we have that $y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) + \xi_i = 1$, but given that $\xi_i = 0$ we have that $y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) = 1$;

(iii) if α_i = C, $\Rightarrow$ $y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) + \xi_i = 1$, but given that $\xi_i \geq 0$ we have that $y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) \leq 1$.

The KKT conditions can therefore be rewritten as:

$$\begin{cases} \alpha_i = 0 \quad \Rightarrow y_i f(\mathbf{x}_i) \geq 1 \\ 0 < \alpha_i < C_i \Rightarrow y_i f(\mathbf{x}_i) = 1 \\ \alpha_i = C_i \quad \Rightarrow y_i f(\mathbf{x}_i) \leq 1 \end{cases} \tag{12}$$

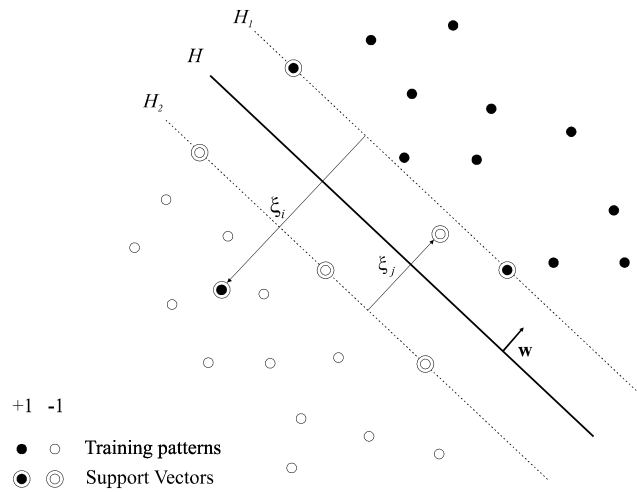

Figure 1. Qualitative example of a separating hyperplane in the case of a non linear separable classification problem.

The *support vectors* with multiplier $\alpha_i = C$ are called *bound support vectors* (BSV) and are associated to slack variables $\xi_i \geq 0$; the ones with $0 < \alpha_i < C_i$ are called *non bound support vectors* (NBSV) and lie on the margin hyperplane H_1 or H_2 ($y_i f(\mathbf{x}_i) = 1$).

C. Training of Non Linear SVM – Kernel Trick.

An important improvement to the above-described methods consists in considering non linear discriminant functions for separating the two information classes. This can be obtained by transforming the input data into a high dimension (Hilbert) feature space $\Phi(\mathbf{x}) \in \mathbb{R}^{d'}$ ($d' > d$) where the transformed samples can be better separated by a hyperplane. The main problem is to explicitly choose and calculate the function $\Phi(\mathbf{x}) \in \mathbb{R}^{d'}$ for each training samples. But given that the input points in dual formulation [see (10)] appear in the form of inner products, we can do this mapping in an implicit way by exploiting the so called kernel trick. Kernel methods provide an elegant and effective way of dealing with this problem by replacing the inner product in the input space with a kernel function such that:

$$K(x_i, x_j) = \langle(\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)\rangle \qquad i, j = 1, \ldots, N \tag{13}$$

implicitly calculating the inner product in the transformed space.

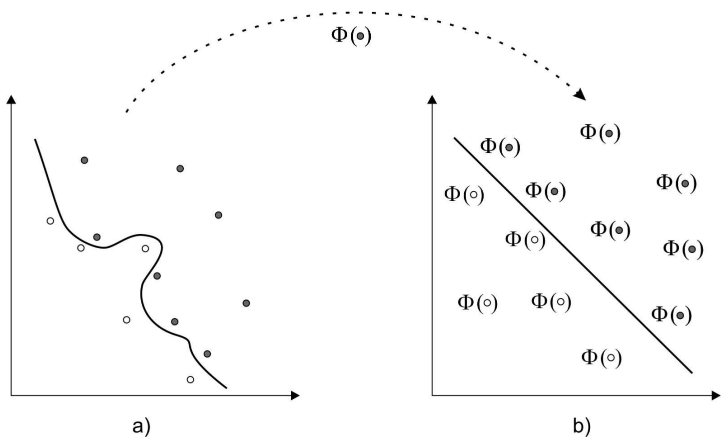

Figure 2. Transformation of the input data by means of a kernel function into a high dimension feature space. a) Input feature space; b) kernel induced high dimensional feature space.

The soft margin algorithm for nonlinear function can be represented by the following optimization problem:

$$\begin{cases} \max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \right\} \\ \sum_{i=1}^{N} y_i \alpha_i = 0, \ 0 \le \alpha_i \le C \text{ and } 1 \le i \le N \end{cases} \tag{14}$$

And the discrimination function becomes:

$$f(\mathbf{x}) = \sum_{i \in SV} y_i \alpha_i^* k(\mathbf{x}_i \cdot \mathbf{x}) + b \tag{15}$$

The condition for a function to be a valid kernel is given by the Mercer's theorem [32]. The most widely used non-linear kernel functions are the following [31]:

- homogeneous polynomial function: $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^p, \quad p \in \mathbb{Z}$
- inhomogeneous polynomial function: $k(\mathbf{x}_i, \mathbf{x}_j) = (c + (\mathbf{x}_i \cdot \mathbf{x}_j))^p, \quad p \in \mathbb{Z}, c \ge 0$
- Gaussian function: $k(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, \quad \sigma \in \mathbb{R}$

D. Multiclass architectures

As stated in the previous section, SVMs are binary classifiers. However, several strategies have been proposed to address multiclass problems with SVMs. Let $\Omega = \{\omega_1, ..., \omega_L\}$ be the set of L information classes associated with the different land cover types present in the study area. In order to define a multiclass architecture based on different binary classifiers, the general approach consists of: i) defining an ensemble of binary classifiers; and ii) combining them according to some decision rules. The definition of the ensemble of binary classifiers involves the definition of a set of two-class problems, each modeled with two groups Ω_A and Ω_B of classes. The selection of these subsets depends on the kind of approach adopted to combine the ensemble. In the following, we describe the two most widely adopted (parallel) multiclass strategies, i.e., the *One-Agains-All* (OAA) and *One-Against-One* (OAO) strategies.

1) *One-Against-All:* the one-against-all (OAA) strategy represents the earliest and one of the most common multiclass approach used for SVMs. It involves a parallel architecture made up of L SVMs, one for each class (Figure 3). Each SVM solves a two-class problem defined by one information class against all the others, i.e.,

$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \Omega - \omega_i \end{cases} \tag{16}$$

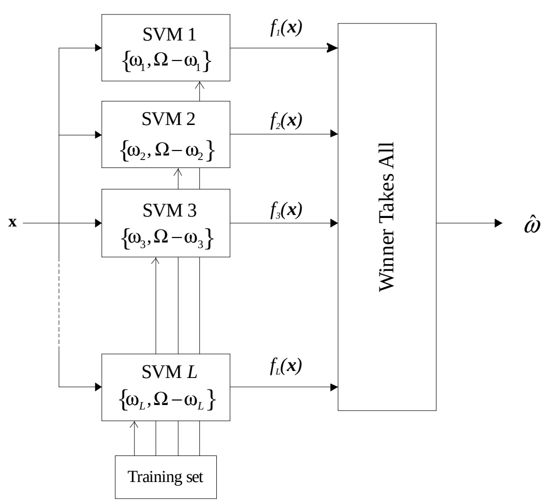

Figure 3 Block diagram of the *One-Against-All* multiclass architecture

The *winner-takes-all* rule is used for the final decision, i.e., the winning class is the one corresponding to the SVM with the highest output (discriminant function value).

2) *One-Against-One:* the main problem of the OAA strategy is that the discrimination between an information class and all the others often leads to the estimation of complex discriminant functions. In addition, a problem with strongly unbalanced prior probabilities should be solved by each SVM. The idea behind the *one-against-one* (OAO) strategy is that of a different reasoning, in which simple classification tasks are made possible thanks to a parallel architecture made up of a large number of SVMs. The OAO strategy involves $L(L-1)/2$ SVMs, which model all possible pairwise classifications. In this case, each SVM carries out a binary classification in which two information classes ω_i and ω_j are analyzed against each other by means of a discriminant function $f_{ij}(\mathbf{x})$. Consequently, the grouping becomes:

$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \omega_j \end{cases} \tag{17}$$

Before the decision process, it is necessary to compute for each class $\omega_i \in \Omega$ a score function $S_i(\mathbf{x})$, which sums the favorable and unfavorable votes expressed for the considered class

$$S_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^{L} \operatorname{sgn}[f_{ij}(\mathbf{x})] \tag{18}$$

The final decision in the OAO strategy is taken on the basis of the *winner-takes-all* rule, which corresponds to the following maximization:

$$\mathbf{x} \in \omega \Leftrightarrow \omega = \arg\max_{i=1,\dots,L} \{S_i(\mathbf{x})\} \tag{19}$$

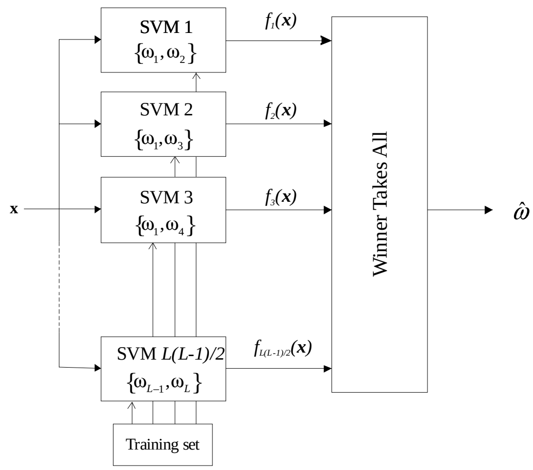

Figure 4 Block diagram of the *One-Against-One* multiclass architecture

Other multiclass architectures proposed in the literature are the Directed Acyclic Graph SVM (DAGSVM) [34] and different approaches based on binary hierarchical trees (BHT) [16,35].

3. SVM for the classification of RS Data

In the last decade many studies have been published in the RS literature on the application of SVM classifiers to the analysis of RS data. Table 1 (which is not exhaustive) presents some relevant papers about the applications of SVM to the classification of RS data, providing a short description of the study and the kind of data used for the experimental analysis. The SVM approach has been first applied to the classification of hyperspectral data [11], which require the classifier to operate in large dimensional feature spaces. Supervised classification of hyperspectral images is a very complex methodological problem due to many different issues, among which we recall the typical small value of the ratio between the number of training samples and the number of available spectral channels, which results in the so-called course of dimensionality (Hughes phenomenon) [36]. Thanks to the structural risk minimization principle and the margin-based approach, SVMs represent an effective choice for the classification of this specific kind of data. Several papers [11-18] confirm the effectiveness of SVMs in the classification of hyperspectral images, which outperform other classification algorithms both in terms of classification accuracy and generalization ability. In particular, in [16] it is found that SVMs are much more effective than other conventional nonparametric classifiers (i.e., the RBF neural networks and the *K*-NN classifier) in terms of classification accuracy, computational time, stability to parameter setting, and generalization ability. In [15], the SVM approach was compared with neural networks and fuzzy methods on six hyperspectral images acquired with the 128-band HyMap spectrometer. The authors of the study concluded that SVMs yield better outcomes than neural networks regarding accuracy, simplicity, and robustness. In [17], SVMs were compared with other kernel-based methods, i.e., with regularized radial basis function NN, kernel Fisher discriminant analysis, and regularized AdaBoost. The results obtained on an AVIRIS dataset show that SVMs are more beneficial, yielding better results than other kernel-based methods, ensuring sparsity and lower computational cost.

Nevertheless, SVMs revealed adequate for the analysis of many different kinds of RS data, i.e., multispectral imagery and SAR imagery (with different resolutions) and LIDAR data. Several papers present a comparison between SVM and other supervised algorithms applied to the classification of different kinds of RS images [20, 23, 25, 30]. In [20], for instance, the authors compared the accuracies obtained by the classification of a Landsat Thematic mapper (TM) scene with four different supervised classifiers, i.e., SVM, maximum likelihood (ML), MLP neural networks (NN), and decision tree classifier (DTC). The obtained results show that SVM was in general sharply more accurate than ML and DTC, and more accurate than NN in most of the cases. In [21], the SVM algorithm was applied to the classification of ASTER data acquired in an urban area of Beer Sheva, Israel. Field validations show that the classification is reliable for urban studies with high classification accuracy. In [23], the SVM classifier, as well as the well-known ML classifier and a context-based classifier based on Markov random fields, were applied to the automatic land cover classification of a Landsat TM image taken on the Tenerife Island. The authors found that SVM was more accurate than the other classification algorithms, but the classification map was not completely satisfying when investigated visually. In the experimental analysis conducted in [25], it is observed that SVM leaded to slightly higher classification accuracies than (MLP) NN. For both classifiers, the accuracy depends on factors such as the number of hidden nodes in the case of NN, and kernel parameters in the case of SVM. Thus, the model selection phase is fundamental for obtaining good results, but the training time required by the SVM is less than the one taken by NN.

SVM can be particularly effective also in the analysis of very high resolution (VHR) images. The typical poor spectral resolution of VHR images requires the extraction of additional features (e.g., texture and geometric measures) to characterize the objects present in the scene under investigation and to discriminate different land-cover classes. Different features modeling objects at different scales are generally necessary for an adequate characterization of the information classes [27], thus resulting in classification problems characterized by large dimensional feature spaces (with some analogies with the problems related to the classification of hyperspectral image). The study proposed in [27] points out that SVM can be effectively applied to the classification of VHR images using a feature extraction block that aims at adaptively modeling the spatial context of each pixel according to a hierarchical multilevel segmentation of the scene. A similar approach can also be adopted for the joint classification of SAR and optical data with SVM, as presented in [29]. In [30], an analysis is proposed on the joint use of hyperspectral and LIDAR data for the classification of complex forest areas. The experimental results obtained in [29-30] show that SVMs are effective for combining multisensor data in complex classification problems and outperforms other more traditional classifiers.

Table 1 – Selected papers related to the application of SVM to classification of different kinds of RS data

Authors	Description	RS data
J. A. Gualtieri and S. Chettri [13]	In this paper, the authors introduce SVM for the classification of RS data. In particular they applied SVM to hyperspectral data acquired by NASA's AVIRIS sensor and the commercially available AISA sensor. The authors discuss the robustness of SVM to the course of dimensionality (Hughes phenomenon).	AVIRIS (224 spectral bands) and AISA (20-40 bands)
F. Melgani, L. Bruzzone [16]	This paper addresses the problem of the classification of hyperspectral remote sensing images by SVMs. The authors propose a theoretical discussion and experimental analysis aimed at understanding and assessing the potentialities of SVM classifiers in hyperdimensional feature spaces. Then, they assess the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hypersubspaces of various dimensionalities. To sustain such an analysis, the performances of SVMs are compared with those of two other nonparametric classifiers (i.e., radial basis function neural networks and the K-nearest neighbor classifier). Four different multiclass strategies are analyzed and compared: the one-against-all, the one-against-one, and two hierarchical tree-based strategies.	AVIRIS (224 spectral bands)
G. Camps-Valls, L. Bruzzone [17]	This paper presents the framework of kernel-based methods in the context of hyperspectral image classification, illustrating from a general viewpoint the main characteristics of different kernel-based approaches and analyzing their properties in the hyperspectral domain. In particular, we assess performance of regularized radial basis function neural networks (Reg-RBFNN), standard support vector machines (SVMs), kernel	AVIRIS (224 spectral bands)

	Fisher discriminant (KFD) analysis, and regularized AdaBoost (Reg-AB).	
G..M. Foody, A. Mathur [19]	In this paper, an approach for multiclass classification of airborne sensor data by a single SVM analysis is evaluated against a series of classifiers that are widely used in remote sensing, with particular regard to the effect of training set size on classification accuracy. In addition to the SVM, the same datasets were classified using a discriminant analysis, decision tree, and multilayer perceptron neural network. For each classification technique, accuracy was positively related with the size of the training set. In general, the most accurate classifications were derived from the SVM approach.	Airborne Thematic Mapper (ATM) (11 spectral bands, spatial resolution of 5m)
C. Huang, L.S. Davis, J.R.G. Townshend [20]	This paper introduces the theory of SVM and provides an experimental evaluation of its accuracy, stability, and training speed in deriving land cover classifications from satellite images. SVM algorithm is compared with other supervised algorithms: maximimum likelihood (ML) classifier, neural network classifier, and decision tree classifier.	(Spatially degraded) Landsat Thematic Mapper (TM)
G. Zhu, D. G. Blumberg [21]	This paper presents a study on the mapping of urban environments using ASTER data and SVM-based classification algorithms. A case study of the classification of the area of Beer Sheva, Israel is presented. Field validation shows that the classification is reliable and precise.	Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER)
L. Su, M. J. Chopping, A. Rango, J. V. Martonchik, D. P. C. Peters [22]	This paper present a study on mapping and monitoring the desert environment using SVM for the analysis of Multi-angle Imaging Spectro-Radiometer (MISR) RS data. Many classification experiments have been implemented to find the optimal combination of MISR multi-angle data for maximizing the classification accuracy.	Multi-angle Imaging Spectro-Radiometer (MISR)
J. Keuchel, S. Naumann, M. Heiler, A. Siegmund [23]	This paper presents three different approaches to the classification of satellites images: maximum likelihood classifier, SVM, and iterated conditional model (ICM) to perform contextual classification using Markov random field model. The classification algorithms were applied to a Landsat 5 TM image of Tenerife, the largest of the canary Island.	Landsat 5 TM
B. Dixon, N. Candade [25]	This paper presents a study on the comparison between SVM and NN for the classification of RS data. An experimental analysis is carried on Landsat 5 TM data, acquired in the South West of Florida. The obtained results confirmed that SVM and NN outperform the traditional ML classifier. SVM classification resulted slightly more accurate than NN, but SVM required much less computational effort in the training phase.	Landsat 5 TM
L. Bruzzone, L. Carlin [27]	This paper proposes a system for the classification of VHR images. The proposed system is made up of two main blocks: 1) a feature-extraction block that aims at adaptively model the spatial context of each pixel according to a hierarchical multilevel segmentation of the scene and 2) a classification block based on SVM. Experimental results on VHR images confirm the effectiveness of the proposed system.	Quickbird

B. Waske, S. Van der Linden [29]	This paper presents a strategy for the joint classification of multiple segmentation levels from multisensor imagery, using SAR and optical data. The two datasets are separately segmented at different scale levels and independently classified by two SVM-based classifiers. The fusion strategy is based on the application of an additional classifier, which takes in input the soft output of the pre-classified results of the two datasets. The obtained experimental results show that the useful combination of multilevel-multisensor data is feasible with machine learning techniques like SVM and Random forest.	Multitemporal SAR data and Landsat 5 TM
M. Dalponte, L. Bruzzone, and D. Gianelle [30]	In this paper, the authors propose an analysis on the joint use of hyperspectral and light detection and ranging (LIDAR) data for the classification of complex forest areas. In greater detail, they present: 1) an advanced system for the joint use of hyperspectral and LIDAR data in complex classification problems; 2) an investigation on the effectiveness of the very promising SVM and Gaussian ML with leave-one-out covariance algorithm for the analysis of forest areas characterized from a high number of species; and 3) an analysis of the effectiveness of different LIDAR returns and channels for increasing the classification accuracy obtained with hyperspectral images.	Hyperspectral (126 spectral bands) and LIDAR (mean density of 5.6 points per square meter)

The RS literature related to SVM is not limited to the use of this approach on different data and different application domains. Recently, more advanced SVM-based classifiers have been developed for facing complex problems related to the properties of remote sensing images. A list of relevant papers that introduced advanced techniques based on SVM for the classification of RS data is reported in Table 2. These papers represent the most recent (and in some cases on-going) research activities in this field and give insight about the research direction for the next years.

In this context, it is worth mentioning the semi-supervised SVM classifiers [37-43], which are devised for addressing ill-posed problems characterized by a very small ratio between the number of available training samples and the number of features by reinforcing the learning procedure with the use of unlabeled samples. It is worth noting, that even if SVMs have very good generalization capability, they cannot model the classification problem when very few training samples are available (“strongly” ill-posed problems). In these cases, the exploitation of the unlabeled samples to enrich the information of the training samples can result in a significant improvement in the model estimation. The first work on Semisupervised SVM in RS was presented in [37, 38]. The presented semisupervised SVM (S^3VM) is based on transductive inference that exploits a specific iterative algorithm which gradually searches a reliable separating hyperplane in the kernel space with a process that incorporates both labeled and unlabeled samples in the training phase. In [39], an S^3VM classification technique is proposed, where the learning phase is performed by optimizing the objective function directly in the primal formulation (without exploiting the dual representation that can be obtained with Lagrange multipliers). In [40], the Laplacian SVM technique [41] is introduced in the RS community. This technique adopts an additional regularization term on the geometry of both labeled and unlabeled samples by using the graph Laplacian. This method follows a non-iterative optimization procedure in contrast to most transductive learning methods and provides out-of-sample predictions in contrast to graph-based approaches. Experimental results confirm the effectiveness of S^3VM techniques for solving ill-posed remote-sensing classification problems. In general S^3VM provides higher accuracy and better generalization ability than standard supervised SVM. In this respect, a more detailed picture of the status on the research on the application of S^3VM to hypedimensional problems can be found in [43].

Other studies address the inclusion of the spatial-context information of the single pixel in the SVM classification process. To this end, [44] proposes a framework for applying the maximum a posteriori (MAP) estimation principle in remote sensing image segmentation, which incorporates contextual and geometrical information in the SVM classification process by means of Markov random field (MRF). In [45], the use of composite kernels is introduced in remote sensing to adopt different kernel functions for different subsets of features to combine spatial and spectral information in an effective way. In [47], a context-sensitive semisupervised SVM is proposed, which exploits the contextual information of the pixels during the learning phase, in order to improve the robustness to possible mislabeled training patterns (which are not unlikely to be present in the reference data due to different kinds of errors that may occur in the collection of labeled samples). This is achieved according to both the design of a semisupervised procedure and the definition of a contextual term in the cost function associated with the learning of the classifier. In the experimental analysis, the authors also studied the robustness to mislabeled training patterns of some widely used supervised and semisupervised classification algorithms (i.e., conventional SVM, progressive semisupervised SVM, Maximum Likelihood, and *k*-Nearest Neighbor). Thanks to their high generalization capability, SVM-based approaches resulted more robust than other classification approaches (e.g., statistical approaches) to the presence of mislabeled training patterns.

The study in [48] addresses the problem of automatic updating the land-cover maps by using RS images periodically acquired over the same investigated area under the hypothesis that a reliable ground truth is not available for all the considered acquisitions. The problem is modeled under the domain-adaptation framework by introducing a novel method designed for land-cover map updating, which is based on a domain-adaptation SVM (DASVM) technique. Given two RS images I_1 and I_2 acquired over the same area at different times (t_1 and t_2, respectively), the goal of the DASVM is to obtain an accurate classification of I_2 by exploiting the labeled training samples from reference image I_1 and the unlabeled samples from the new image I_2. The DASVM algorithm is based on an iterative process, which starts by training an SVM classifier with the original training samples of I_1 and gradually introduces semilabeled samples of I_2 and erases the original training samples. At convergence a final classification function ruled only by semilabeled samples at time t_2 is obtained. In addition, the authors propose a circular accuracy assessment strategy for the validation of the results obtained by domain-adaptation classifiers when no reference data for the considered image I_2 are available.

Another recent and promising approach to the analysis RS data is associated with active learning [49-50], which allows an interactive classification of RS images. The active learning approach is based on the iteration on three different conceptual steps. In the first step the learning process queries unlabeled samples to select the most informative ones; in the second step the supervisor (e.g., the user) provides a label to the selected samples interacting with the system; and in the third step the learner updates the classification rule by retraining with the updated training set. In [49], it is noted that SVMs are particularly suited to active learning since they are characterized by a small set of support vectors (SVs) which can be easily updated over successive learning iterations. Moreover, one of the most efficient query functions is based on the selection of the sample closest to the separating hyperplane defined at the considered iteration. For additional information about recent developments in kernel methods for the analysis of RS images, we refer the reader to [51].

Table 2 – Relevant papers about advanced techniques based on SVM for the classification of RS data

Authors	Description
L. Bruzzone, M. Chi, M. Marconcini [38]	This paper introduces a semisupervised classification method that exploits both labeled and unlabeled samples for addressing ill-posed problems with SVMs. The proposed method exploit specific iterative algorithms which gradually search a reliable separating hyperplane in the kernel space with a process that incorporates both labeled and unlabeled samples in the training phase. The authors propose a novel modified transductive SVM classifier designed for addressing ill-posed remote-sensing problems, which has the following properties: 1) it is based on a novel transductive procedure that exploits a weighting strategy for unlabeled patterns, based on a time-dependent criterion; 2) is able to mitigate the effects of suboptimal model

	selection (which is unavoidable in the presence of small-size training sets); and 3) can address multiclass cases.
M. Chi, L. Bruzzone [39]	This paper addresses classification of hyperspectral remote-sensing images with kernel-based methods defined in the framework of semisupervised SVM (S^3VMs). In particular, the authors analyzed the critical problem of the nonconvexity of the cost function associated with the learning phase of S^3VMs by considering different (S^3VMs) techniques that solve optimization directly in the primal formulation of the objective function. As the nonconvex cost function can be characterized by many local minima, different optimization techniques may lead to different classification results. The presented techniques are compared with S^3VMs implemented in the dual formulation in the context of classification of real hyperspectral remote sensing images.
L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, J. Calpe [40]	This letter presents a semisupervised method based on kernel machines and graph theory for remote sensing image classification. The support vector machine (SVM) is regularized with the unnormalized graph Laplacian, thus leading to the Laplacian SVM (LapSVM). The method is tested in the challenging problems of urban monitoring and cloud screening, in which an adequate exploitation of the wealth of unlabeled samples is critical.
A. A. Farag, R. M. Mohamed, A. El-Baz [44]	This paper proposes a complete framework for applying the maximum a posteriori (MAP) estimation principle in remote sensing image segmentation. The MAP principle provides an estimate for the segmented image by maximizing the posterior probabilities of the classes defined in the image. The posterior probability can be represented as the product of the class conditional probability (CCP) and the class prior probability (CPP). For the CCP, a supervised algorithm which uses the SVM density estimation approach is proposed. For the CPP estimation, Markov random field (MRF) is a common choice which incorporates contextual and geometrical information in the estimation process.
G. Camp-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J Calpe-Maravilla [45]	This letter presents a framework of composite kernel machines for enhanced classification of hyperspectral images. This novel method exploits the properties of Mercer's kernels to construct a family of composite kernels that easily combine spatial and spectral information. This framework of composite kernels demonstrates: 1) enhanced classification accuracy as compared to traditional approaches that take into account the spectral information only: 2) flexibility to balance between the spatial and spectral information in the classifier; and 3) computational efficiency.
M. Marconcini, G. Camps-Valls, L. Bruzzone [46]	This letter presents a novel composite semisupervised SVM for the spectral–spatial classification of hyperspectral images. In particular, the proposed technique exploits the following: 1) unlabeled data for increasing the reliability of the training phase when few training samples are available and 2) composite kernel functions for simultaneously taking into account spectral and spatial information

	included in the considered image. Experiments carried out on a hyperspectral image pointed out the effectiveness of the presented technique, which resulted in a significant increase of the classification accuracy with respect to both supervised SVMs and progressive semisupervised SVMs with single kernels, as well as supervised SVMs with composite kernels.
L. Bruzzone, C. Persello [47]	This paper presents a novel context-sensitive semisupervised SVM (CS^4VM) classifier, which is aimed at addressing classification problems where the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns). Unlike standard context-sensitive methods, the proposed CS^4VM classifier exploits the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase to improve the robustness to possible mislabeled training patterns. This is achieved according to both the design of a semisupervised procedure and the definition of a novel contextual term in the cost function associated with the learning of the classifier. In order to assess the effectiveness of the proposed CS^4VM and to understand the impact of the addressed problem in real applications, the authors also present an extensive experimental analysis carried out on training sets that include different percentages of mislabeled patterns having different distributions on the classes. In the analysis they also study the robustness to mislabeled training patterns of some widely used supervised and semisupervised classification algorithms (i.e., conventional SVM, progressive semisupervised SVM, Maximum Likelihood, and *k-N*N)
L. Bruzzone, M. Marconcini [48]	In this paper, the authors address automatic updating of land-cover maps by using remote-sensing images periodically acquired over the same investigated area under the hypothesis that a reliable ground truth is not available for all the considered acquisitions. The problem is modeled in the domain-adaptation framework by introducing a novel method designed for land-cover map updating, which is based on a domain-adaptation SVM technique. In addition, a novel circular accuracy assessment strategy is proposed for the validation of the results obtained by domain-adaptation classifiers when no ground-truth labels for the considered image are available.
D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, W. J. Emery [50]	In this paper, an active learning method is proposed for the semi-automatic selection of training sets in remote sensing image classification. The method adds iteratively to the current training set the unlabeled pixels for which the prediction of an ensemble of classifiers based on bagged training sets show maximum entropy. This way, the algorithm selects the pixels that are the most uncertain and that will improve the model if added in the training set. The user is asked to label such pixels at each iteration. Experiments were carried out using SVM.

4. Discussion and Conclusion

In this chapter we presented a review on SVMs in the classification of remote-sensing data, recalling their theoretical formulation, and discussing the motivations at the basis of their use in remote sensing. We presented a literature survey about the adoption of SVMs for the analysis of different kinds of RS images. We observed a large variety of studies published on the use of SVMs for the analysis of different kinds of RS data, which confirm that SVMs represent a valuable and effective tool for the analysis of RS data and can be used in many different applications in the context of RS. We observed that one of the most appealing properties of SVM for the classification of RS data is its high generalization capability and robustness to the Hughes effect, which allow SVMs to operate in large dimensional feature spaces with few training samples. For this reason, SVMs represent an effective choice for the classification of hyperspectral data. Nevertheless, the SVM approach turned out to be particularly effective also in the classification of very high resolution (VHR) images, which typically require the extraction of several additional features to characterize and discriminate the different land-cover classes. Thus, both the classification of VHR and hyperspectral images typically result in classification problems characterized by large dimensional feature spaces. Moreover, thanks to its distribution-free approach and the capability to cope with strongly non-linear problems by means of the kernel function, SVMs are a valuable tool also for the classification of data acquired by different information sources.

In addition, we pointed out the most recent works about the development of advanced SVM-based techniques for the analysis of RS data. Among these developments, we recall semisupervised and domain-adaptation SVM, techniques based on SVM that exploit the spatial-context information, and active learning methods. Semisupervised SVMs have shown to be effective in exploiting both labeled and unlabeled samples for the learning of the classification algorithm, further augmenting the generalization capability and the robustness to the Hughes phenomenon with respect to standard supervised SVM. Domain-adaptation SVM resulted effective for addressing the problem of automatic updating land-cover maps by using RS images periodically acquired over the same investigated area. Context-sensitive techniques based on SVM have been proposed for both regularizing the classification map (exploiting the context information in the classification phase) or for improving the robustness to mislabeled training samples (using the context information in the learning phase of the algorithm). Another promising approach is active learning, which allows one an interactive analysis of RS data, by driving the user to label unlabeled samples that are selected by a query function as most informative.

We can conclude that the SVM approach showed to be very promising for the classification of RS data and recent works demonstrate that SVM can be used as basis for the development of advanced techniques for solving specific RS problems or for exploiting particular properties of the RS data. However, still effort should be devoted to the development of advanced techniques that can effectively extract useful information from the rich and complex data acquired by the last generation of RS sensors. Moreover, effort is required also for applying the SVM-based approaches developed in the research activities in real-world RS problems. Indeed, at the present, the most of the real problems related to RS image classification are still solved with standard classifiers (like maximum likelihood or *k*-NN) that, even if simple, cannot guarantee the accuracy and generalization capabilities of SVMs in complex problems.

References

1. J.A. Benediktsson, P.H. Swain, and O.K. Ersoy, "Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-28, no. 4, pp. 540-552, July 1990.
2. P.D. Heermann, N. Khazenie, "Classification of multispectral remote sensing data using a back-propagation neural network", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 1, pp. 81 - 88, Jan. 1992.
3. H. Bischof, W. Schneider, A.J. Pinz, "Multispectral classification of Landsat-images using neural networks", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 3, pp. 482 - 490, May 1992.
4. L. Bruzzone, D.F. Prieto, "A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, Part 2, pp. 1179 - 1184, March 1999.
5. L. Bruzzone, M. Marconcini, U. Wegmüller, and A. Wiesmann, "An Advanced System for the Automatic Classification of Multitemporal SAR Images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1351 - 1362, June 2004.
6. S.B. Serpico, and F. Roli, "Classification of multisensory remote-sensing images by structured neural networks", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 33, No. 3, pp. 562-578, May 1995.
7. L. Bruzzone, D.F. Prieto, S.B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, No. 3, Part 1, pp. 1350 – 1359, May 1999.
8. J.A. Benediktsson, I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion", *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 37, No. 3, Part 1, pp. 1367 - 1377, May 1999.
9. L.O. Jimenez, A. Morales-Morell, A. Creus, "Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks", *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 37, No. 3, Part 1, pp. 1360 - 1366, May 1999.
10. V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
11. J.A. Gualtieri and R.F. Cromp, "Support Vector Machines for Hyperspectral Remote Sensing Classification", 27'th AIPR Workshop, *Proc. of the SPIE*, vol. 3584, pp. 221-232, 1998.
12. J.A. Gualtieri, S.R. Chettri, R.F. Cromp, and L.F. Johnson, "Support Vector Machine Classifiers as Applied to AVIRIS Data", Summaries of the Eighth JPL Airborne Earth Science Workshop, JPL Publication 99-17, pp. 217-227, 1999, ftp://popo.jpl.nasa.gov/pub/docs/workshops/99_docs/toc.html.
13. J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data", *Proc. of IEEE-IGARSS 2000*, Hawaii, pp. 813-815, 2000.
14. F. Melgani, L. Bruzzone, "Support vector machines for classification of hyperspectral remote-sensing images", *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2002)*, Toronto, Canada, pp. 506-508, June 2002.
15. G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J.D. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, No. 7, pp. 1530 - 1542 , July 2004.
16. F. Melgani, L. Bruzzone, "Classification of hyperspectral remote-sensing images with support vector machines", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, No. 8, pp. 1778-1790, August 2004.
17. G. Camps-Valls, L. Bruzzone, "Kernel-based methods for hyperspectral images classification", *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 43, No. 6, pp. 1351-1362, June 2005.

18. M. Fauvel, J.A. Benediktsson, J. Chanussot, J.R. Sveinsson, “Spectral and Spatial Classification of Hyperspectral Data Using SVMs and Morphological Profiles”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 46, No. 11, pp. 3804-3814, November 2008.
19. G.M. Foody, A. Mathur, “A relative evaluation of multiclass image classification by support vector machines”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, No. 6, pp. 1335 - 1343, June 2004.
20. C. Huang, L.S. Davis, J.R.G. Townshend, “An assessment of support vector machines for land cover classification”, *Int. journal of Remote Sensing*, Vol. 23, No. 4, pp. 725-749, 2002.
21. G. Zhu, D. G. Blumberg, “Classification using ASTER data and SVM algorithms; The case study of Beer Sheva, Israel”, *Remote Sensing of Environment*, Vol. 80, No. 2, pp. 233-240, May 2002.
22. L. Su, M. J. Chopping, A. Rango, J. V. Martonchik, D. P. C. Peters, “Support vector machines for recognition of semi-arid vegetation types using MISR multi-angle imagery”, *Remote Sensing of Environment*, Vol. 107, No. 1-2, pp. 299-311, March 2007.
23. J. Keuchel, S. Naumann, M. Heiler, A. Siegmund, “Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data”, *Remote Sensing of Environment*, Vol. 86, No. 4, pp. 530-541, 2003.
24. I. Lizarazo, “SVM-based segmentation and classification of remotely sensed data”, *int. journal of Remote Sensing*, Vol. 29, No. 24, pp. 7277–7283, December 2008.
25. B. Dixon, N. Candade, “Multispectral landuse classification using neural networks and support vector machines: one or the other, or both?”, *Int. journal of Remote Sensing*, Vol. 29, No. 4, pp. 1185–1206, February 2008.
26. H. Carrao, P. Goncalves, M. Caetano, “Contribution of multispectral and multitemporal information from MODIS images to land cover classification”, *Remote Sensing of Environment*, Vol. 112, No. 3, pp. 986–997, March 2008.
27. L. Bruzzone, L. Carlin, “A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 44, No. 9, 2006, 2587-2600.
28. B. Waske, J.A. Benediktsson, “Fusion of Support Vector Machines for Classification of Multisensor Data”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 45, No. 12, December 2007, pp. 3858-3866.
29. B. Waske, S. Van der Linden, “Classifying Multilevel Imagery From SAR and Optical Sensors by Decision Fusion ”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 46, No. 5, pp. 1457-1466, May 2008.
30. M. Dalponte, L. Bruzzone, and D. Gianelle, “Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 46, No. 5, pp. 1416-1427, May 2008.
31. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
32. N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge, U.K.: University press, 1995.
33. B. Schölkopf and A. Smola, *Learning With Kernels*, Cambridge, MA: MIT Press, online: http://www.learning-with-kernels.org, 2002.
34. Chih-Wei Hsu; Chih-Jen Lin, “A comparison of methods for multiclass support vector machines”, *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 415 - 425, March 2002.
35. Lili Cheng, Jianpei Zhang, Jing Yang, Jun Ma, “An Improved Hierarchical Multi-Class Support Vector Machine with Binary Tree Architecture”, *International Conference on Internet Computing in Science and Engineering*, 2008.
36. G. F. Hughes, “On the mean accuracy of statistical pattern recognition,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, January 1968.

37. L. Bruzzone, M. Chi, M. Marconcini, "Transductive SVM for Semisupervised Classification of Hyperspectral Data," *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005)*, Seoul, Korea, July 2005.
38. L. Bruzzone, M. Chi, M. Marconcini, "A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 44, No. 11, pp. 3363-3373, 2006.
39. M. Chi, L. Bruzzone, "Semi-supervised Classification of Hyperspectral Images by SVMs Optimized in the Primal", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 45, No. 6, Part 2, pp. 1870-1880, 2007.
40. L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, J. Calpe, "Semisupervised Image Classification With Laplacian Support Vector Machines", *IEEE Geoscience and Remote Sensing Letters*, Vol. 5, No. 3, pp. 336-340, July 2008.
41. M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," J. Mach. Learn. Res., vol. 7, pp. 2399–2434, Nov. 2006.
42. G. Camps-Valls, T. Bandos Marsheva, D. Zhou, "Semi-Supervised Graph-Based Hyperspectral Image Classification", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 45, No. 10, pp. 3044-3054, October 2007.
43. L. Bruzzone, M. Chi, M. Marconcini, "Semisupervised Support Vector Machines for Classification of Hyperspectral Remote Sensing Images", in *Hyperspectral Data Exploitation: Theory and Applications*, Ed: C-I. Chang**,** *John Wiley & Sons, Inc.*, 2007, chapter 11, pp. 275-311.
44. A.A. Farag, R. M. Mohamed, A. El-Baz, "A unified framework for MAP estimation in remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 7, pp. 1617-1634, July 2005.
45. G. Camp-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J Calpe-Maravilla, "Composite Kernels for Hyperspectral Image Classification" *IEEE Geoscience and Remote Sensing Letters*, Vol. 3, No. 1, January 2006.
46. M. Marconcini, G. Camps-Valls, L. Bruzzone, "A Composite Semisupervised SVM for Classification of Hyperspectral Images", *IEEE Geoscience and Remote Sensing Letters*, Vol. 6, No. 2, pp. 234-238, 2009.
47. L. Bruzzone, C. Persello, "A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples", *IEEE Transactions on Geoscience and Remote Sensing*, 2009, in press.
48. L. Bruzzone, M. Marconcini, "Toward an Automatic Updating of Land-Cover Maps by a Domain Adapatation SVM Classifier and a Circular Validation Strategy", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, No. 4, pp. 1108-1122, 2009.
49. P. Mitra, B. U. Shankar, S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines", *Pattern Recognition Letters*, vol.25, pp. 1067-1074, 2004.
50. D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, W. J. Emery, "Active Learning of Very-High Resolution Optical Imagery with SVM: Entropy vs Margin Sampling", in Proc. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2008)*, Boston, USA, pp. IV-73 - IV-76, July 2008.
51. G. Camps-Valls and L. Bruzzone, "Kernel Methods for Remote Sensing Data Analysis", John Wiley & Sons, Inc., 2009, in press.