

Robust Support Vector Method for Hyperspectral Data Classification and Knowledge Discovery

Gustavo Camps-Valls, *Member, IEEE*, Luis Gómez-Chova, Javier Calpe-Maravilla, *Member, IEEE*, José D. Martín-Guerrero, Emilio Soria-Olivas, Luis Alonso-Chordá, José Moreno, *Member, IEEE*

Abstract—In this paper, we propose the use of Support Vector Machines (SVM) for automatic hyperspectral data classification and knowledge discovery. In the first stage of the study, we use SVMs for crop classification and analyze their performance in terms of efficiency and robustness, as compared to extensively used neural and fuzzy methods. Efficiency is assessed by evaluating accuracy and statistical differences in several scenes. Robustness is analyzed in terms of (a) suitability to working conditions when a feature selection stage is not possible, and (b) performance when different levels of Gaussian noise are introduced at their inputs. In the second stage of this work, we analyze the distribution of the support vectors (SV) and perform sensitivity analysis on the best classifier in order to analyze the significance of the input spectral bands.

For classification purposes, six hyperspectral images acquired with the 128-band HyMAP spectrometer during the DAISEX-1999 campaign are used. Six crop classes were labelled for each image. A reduced set of labelled samples is used to train the models and the entire images are used to assess their performance.

Several conclusions are drawn: (1) SVMs yield better outcomes than neural networks regarding accuracy, simplicity and robustness; (2) training neural and neurofuzzy models is unfeasible when working with high dimensional input spaces and great amounts of training data; (3) SVMs perform similarly for different training subsets with varying input dimension, which indicates that noisy bands are successfully detected; and (4) a valuable ranking of bands through sensitivity analysis is achieved.

Index Terms—Hyperspectral imagery, crop classification, knowledge discovery, Support Vector Machines, neural networks.

I. INTRODUCTION

THE information contained in hyperspectral data about the chemical properties of the surface allows the characterization, identification, and classification of the surface features by means of recognition of unique spectral signatures, with improved accuracy and robustness. Pattern recognition methods have proven to be effective techniques in applications of this

kind [1]. In fact, classification of surface features in satellite imagery is one of the most important applications of remote sensing. Nevertheless, it is often difficult and time-consuming to develop classifiers by hand, so many researchers have turned to techniques from the fields of statistics and machine learning in order to automatically generate both supervised and unsupervised classifiers. Unsupervised methods are not sensitive to the number of labelled samples since they work on the whole image, but the correspondence between clusters and classes is not ensured. Consequently, supervised methods are preferable when the desired input-output mapping is well-defined and a data set of true labels is available, as occurs in our case study. However, the main problem with supervised methods is that the learning process depends heavily on the quality of the training data set and the input space dimensionality [2]. In both cases, data in the input space is represented in the form of an N -dimensional vector for each pixel, where N is the number of spectral bands. Certainly, the quality of data and the high dimension of the input space are main issues to be addressed, given the high cost of true sample labelling, the high number of spectral bands, and the high variability of the earth's surface. Therefore, *robust* methods for hyperspectral data classification are needed, as they are insensitive both to noise and to the high input dimension.

In order to circumvent problems when dealing with a high dimensional input space, in practice, a preprocessing (feature selection/extraction) stage is often introduced. Numerous feature selection methods have been proposed in the literature: Principal Component Analysis (PCA) [3], wavelet transforms [4], modular neural networks [5], linear filtering [6], etc. However, the design and application of this stage is time-consuming, scenario-dependent, and sometimes needs *a priori* knowledge. Consequently, many advanced supervised methods have been developed to tackle the problem of automatic hyperspectral data classification with a simple feature selection or without one: statistical approaches [7], fuzzy models [8], projection pursuit classifiers [9], radial basis function (RBF) neural networks [10], [11], [12], multilayer perceptrons [13], [14], [15], [16], genetic algorithms [17], self-organizing maps [18], etc. Few works, however, have benchmarked state-of-the-art nonlinear classifiers for hyperspectral imagery.

In this context, Support Vector Machines (SVM) have recently been proposed as efficient (non-linear) supervised classification and regression tools [19], [20]. SVMs are not drastically affected by the *curse of dimensionality*, or its Hughes attendant [21], and offer solutions with an explicit

Manuscript received April 23, 2003; revised November 29, 2003; accepted January 20, 2004.

G. Camps-Valls, L. Gómez-Chova, J. Calpe-Maravilla, J. D. Martín-Guerrero and E. Soria-Olivas are with Digital Signal Processing Group, GPDS (<http://gpds.uv.es>). Electronics Department, Facultat de Física, Universitat de València. C/ Dr. Moliner, 50. 46100 Burjassot (València) Spain. E-mail: gcamps@uv.es.

L. Alonso-Chordá and J. Moreno are with Laboratory for Earth Observation (LEO), Department of Thermodynamics, Facultat de Física, Universitat de València. C/ Dr. Moliner, 50. 46100 Burjassot (València) Spain.

dependence on the most informative patterns in the data. These characteristics make them suitable for hyperspectral data classification and knowledge discovery¹, respectively. Previous works using SVMs have shown successful classification performance of multispectral [6], [23], [24], [25], [26], [27] and hyperspectral [28], [29], [30] data. Nevertheless, further work must be carried out in order to study robustness in noisy situations (presence of redundant bands), high dimensional input spaces, and changing environments (several images). In addition, few efforts have been made to compare SVMs with other widely used pattern recognition methods, such as RBF networks or soft computing approaches, and little attention has been done to analyzing the structure of the final model. The latter is a very interesting option because, once the model is built, its parameters (weights in the case of neural networks, or support vectors in the case of SVMs) contain valuable information about the problem. Analysis of the model can allow us to perform a ranking of the available features and, consequently, to gain knowledge in the problem by identifying relevant or meaningless features.

In this paper, we extend the work presented in [31], [32] and propose the use of SVMs to develop robust crop cover classifiers and to obtain an interpretable thematic map of the crops on the scenes using hyperspectral imagery. The work can be divided in two stages, as follows:

- 1) *Classification*. We first compare SVMs to other well-known, machine learning methods such as multilayer perceptrons (MLP) [33], Radial Basis Function (RBF) neural networks [34], and Co-Active Neuro-Fuzzy Inference Systems (CANFIS) [35]. Comparison is carried out in terms of accuracy and robustness regarding the input space dimension and presence of noisy bands.
- 2) *Knowledge discovery*. We analyze the distribution of the support vectors in the input spaces and perform sensitivity analysis on the best classifier in order to attain a ranking of the input bands significance. Some physical conclusions are drawn.

The paper is outlined as follows. In Section II and III, data collection and the experimental setup are presented, respectively. The classification methods used are described in Section IV, with special emphasis on SVMs. The classification results are presented in Section V. In Section VI, we analyze the model structure. In Section VII, we end this paper with some conclusions and a proposal for future work.

II. DATA COLLECTION

This work is a contribution to the Digital Airborne Imaging Spectrometer Experiment (DAISEX) project, funded by the European Space Agency (ESA) within the framework of its Earth Observation Preparatory Program during 1998, 1999, and 2000 (more details at <http://io.uv.es/projects/daisex/>). Three data acquisition campaigns were carried out in the area

of Barrax (Spain). We have used six hyperspectral images acquired with the HyMAP spectrometer during the DAISEX-1999 campaign. This instrument provides 128 bands across the reflective solar wavelength region of $0.4\mu\text{m} - 2.5\mu\text{m}$ with contiguous spectral coverage (except in the atmospheric water vapour absorptions bands), bandwidths around 16 nm, very high signal to noise ratio, and a spatial resolution of 5m. Fig. 1 shows the distribution of the HyMAP channels over the spectral domain.

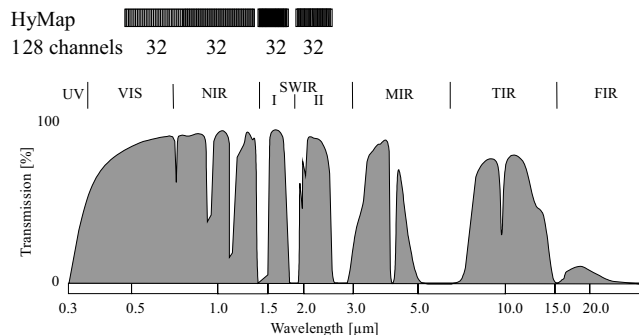


Fig. 1. HyMAP spectral channels and the atmospheric transmission over the different spectral ranges.

The HyMAP images acquired in the DAISEX-1999 campaign correspond to two consecutive days (one flight on the first day and two flights on the second day). A flight consisted of two overpasses, one in the North-South direction (BAR1) and the other in the East-West direction (BAR2), which yielded six images of the same area (Fig. 2). The acquisition of images from the same area in two days under different illumination conditions ensures the robustness of the results. One acquisition of the two flight directions took place at solar noon (at 12:00 UTC or 14:00 local time) assuring a recording of the hot-spot conditions. Two further acquisitions were planned at minus three hours (9:00 UTC) and plus three hours (15:00 UTC) from solar noon. This makes it possible to see angular reflectance changes not only with a view angle but also with an illumination angle (three solar elevation/azimuth angles). Table I shows the data acquisition program.

TABLE I
SOLAR POSITION DURING IMAGE ACQUISITIONS.

Image	Date	Hour	Solar Position	
			Elevation	Azimuth
BAR1_12	06/03/1999	11:52	73.03	168.49
BAR2_12	06/03/1999	12:08	73.25	181.31
BAR1_09	06/04/1999	8:01	37.51	88.75
BAR2_09	06/04/1999	8:16	38.64	91.17
BAR1_15	06/04/1999	14:58	50.07	258.35
BAR2_15	06/04/1999	15:11	47.56	260.86

The calibration in the reflecting region of the HyMAP spectrometer was made during the flight and using the vicarious calibration [36]. The atmospheric correction was based on the MODTRAN atmospheric model with the software package ATCOR-A (atmospheric correction, airborne version) by the

¹A *knowledge discovery* learning scheme is constituted by a preprocessing stage, a *data mining* step (in which a classifier is developed), and a model analysis phase. In such an approach, the last objective is "to process the data in order to extract valid, novel, potentially useful, and ultimately understandable structure in data" [22].

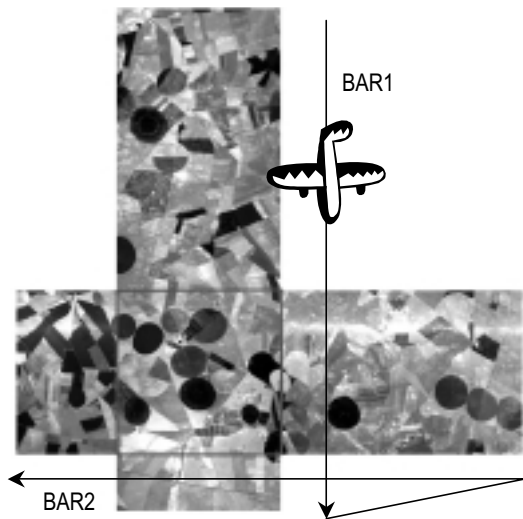


Fig. 2. Flight pattern followed for DAISEX-1999 at Barrax (Spain). The red-box indicates the area selected for this work.

Deutsches Zentrum für Luft-und Raumfahrt (DLR) [37]. At the same time as one aerial campaign took place, another one took place on the ground level [38] with the acquisition of atmospheric measurements, spectral measurements at the surface, measures of temperature, and samples of vegetation. The use of a Global Positioning System (GPS) plus location of Ground Control Points (GCP) enabled an accurate geocoding of the HyMAP images.

After data acquisition, a preliminary test was carried out to measure the quality of data. No significant signs of coherent noise were found. In order to analyze the incoherent noise, we represented boxplots of the reflectance mean values (\hat{x}) with standard deviation (σ_x) and the factor σ_x/\hat{x} for pixels from the same crop. Despite the fact that noise was in general very low, it was far too high in some bands. A high level of noise was found at bands 1 ($0.40 \mu\text{m}$), 65 ($1.49 \mu\text{m}$) and 128 ($2.48 \mu\text{m}$) for DAISEX-1999 (Fig. 3). Bands 2, 66, 67, and 97 were also considered noisy bands due to their high variability. In fact, The HyMap-1999 bands 1 and 128 were no longer available in DAISEX-2000 due to the high level of noise they suffered. This issue constitutes an *a priori* difficulty for models that take into account all available bands.

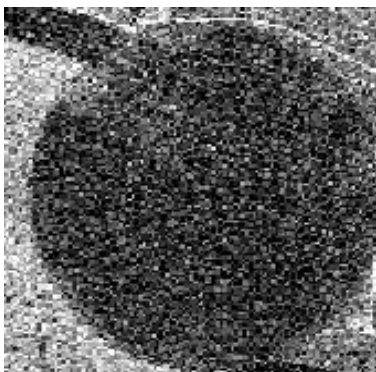


Fig. 3. Noise analysis in a HyMAP image. Incoherent noise observed in Band 65 ($1.487 \mu\text{m}$) for alfalfa crop.

For classification purposes, six different classes were identified in the designated area (corn, sugar beets, barley, wheat, alfalfa, and soil), which were labelled from #1 to #6, respectively. In this sense, the task is referred to as a multiclassification pattern recognition problem. The samples were chosen to have good spatial coverage so the natural variability of the vegetation could be ensured. Three types of units were chosen for sampling, which were based on the type of variability that they represent: full covered fields (alfalfa, wheat, and barley), sparsely vegetated fields (corn and small sugar beets), and bare soil fields. Corn was in an early stage of maturity with fields from two-leaf to five-leaf corn. Bare soils ranged from compacted marly soil to wide surfaces of red clay soil (smooth as well as rough). Alfalfa was representative of a homogeneous canopy. Sugar beets were in an early stage of phenology and showed small coverage and the soil was rather heterogeneous (Fig. 4).

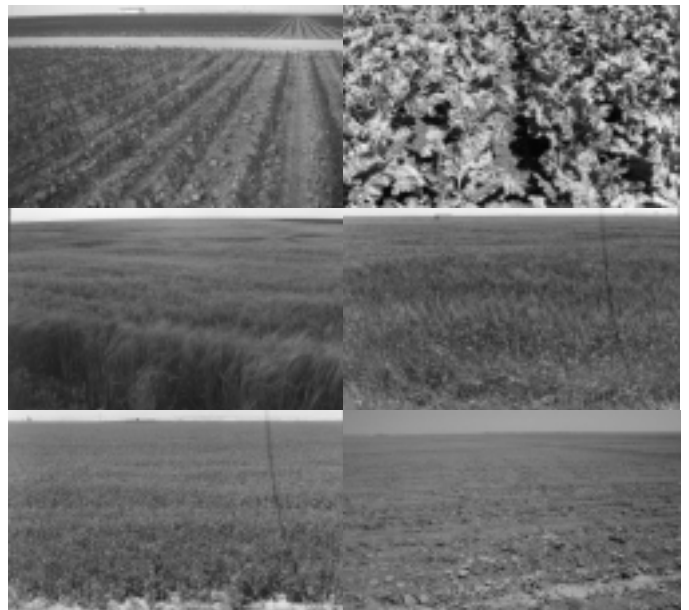


Fig. 4. Pictures of different representative fields at Barrax during the DAISEX-1999 campaign considered for classification purposes. From top to bottom and from left to right: corn, sugar beets, barley, wheat, alfalfa, and soil.

III. EXPERIMENTAL SETUP

Once the desired input-output mapping for training and validation are defined, a feature selection stage is usually used to reduce the dimension of the input space. This can make the training process feasible and improve results by removing noisy bands. However, the design and application of dimension-reduction techniques is time-consuming and scenario-dependent, which are obvious problems to circumvent. In fact, we are not only interested in the classification accuracy provided by each method but also in their suitability to working conditions when a feature selection stage is not possible. The high amount of data potentially available generates problems for data processing. In that sense, providing automatic classification procedures to a ground browsing system could aid in this task. We have simulated these possible

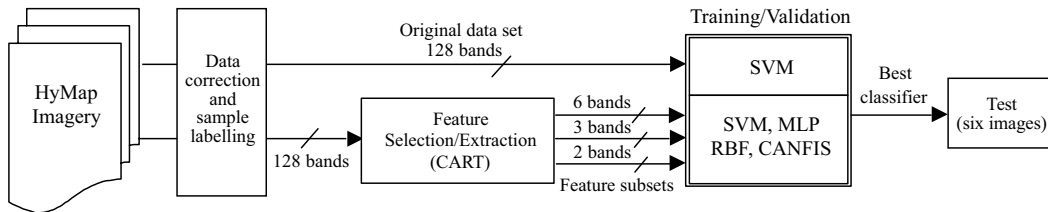


Fig. 5. Diagram of the hyperspectral data classification process. A training data set is extracted from the six collected images and then a CART-based feature selection stage yields three representative subsets (consisting of 6, 3 and 2 bands, respectively). An additional scenario considering the whole training data set (128 bands) is incorporated into the training process as an additional scenario. Four classifiers are thus implemented and tested in the six whole images.

situations by considering models with and without a feature selection stage. The proposed learning scheme is shown in Fig. 5. In this approach, a preprocessing stage selects different subsets of representative bands, which are used to train the selected classifiers. A review of the methods used is provided in the next section. Additionally, SVMs are used to train models with all available bands. Finally, models are compared in terms of robustness and accuracy.

In a previous work [39], we presented a dimensionality reduction strategy to eliminate redundant information and a subsequent selection of the most discriminative features based on Classification And Regression Trees (CART). CARTs allow non-linear feature selection by inspecting main and surrogate splits. In CART, we can control the complexity of the final tree and select the smaller tree with the lowest error. Each variable in the tree has a score of importance that is based on how often and with what significance it served as primary or surrogate splitter throughout the tree. Confidence on this analysis can be ensured since the classification rates of the best CART achieved average recognition rates higher than 91% in the validation set, suggesting that the underlying differences between classes were captured. The introduction of linear rules in each node of the tree improved results but made it more difficult to “illuminate” the model. This work yielded three subsets of representative features (six, three and two bands) that constitute three different pattern recognition problems, respectively. The subset consisting of six bands is shown in Table II. The subset with three bands used 6, 22 and 99 reflectance bands and the subset with only two bands was formed by bands 17 and 22. We can conclude that analysis of the CART model is time-consuming and needs an expert user with in-depth knowledge of the problem at hand. These drawbacks are usually shared by a great number of supervised feature selection models. Therefore, despite the reliable feature selection extracted from CART, a method that is less sensitive to input space dimension would be beneficial.

For classification purposes, two data sets (training and validation sets) were built consisting of 150 labelled samples *per* class. Finally, a test set consisting of the true map of the scene over complete images was used to select the best model. In each of the six images (700×670 pixels), the total number of test samples is 327,336 (corn 31,269; sugar beets 11,322; barley 124,768; wheat 53,400; alfalfa 24,726; and soil 81,851) and the rest is considered to be unknown.

TABLE II
CHARACTERISTICS OF THE REPRESENTATIVE BANDS EXTRACTED THROUGH ANALYZING CART SURROGATE AND MAIN SPLITS.

Bands	Wavelength [μm]	Band width [μm]	Characteristics
6	0.5030	0.0160	Leaf pigments (carotenes and chlorophyll s).
17	0.6710	0.0156	Chlorophyll-a maximum absorption.
22	0.7470	0.0155	Red edge (change Visible-Near Infrared). Leaf Area Index.
24	0.7770	0.0164	Beginning of Near InfraRed (NIR) with high reflectance and low absorbance. Leaf biomass and structure.
99	1.9860	0.0215	Water absorption. Soil moisture and leaf water content.
118	2.3210	0.0200	Water absorption. Dry matter and soil minerals.

IV. MACHINE LEARNING METHODS

In this work, four classification approaches (both neural and kernel methods) have been used. Since neural networks have been extensively employed in hyperspectral data classification and SVMs are relatively new in this field, only a brief background is provided regarding the former.

A. Neural Networks

The traditional model of a feedforward multilayer neural network, commonly known as multilayer perceptron (MLP), is composed of a fully-connected layered arrangement of artificial neurons in which each neuron of a given layer feeds all the neurons of the next layer [33] (Fig. 6(a)). An MLP for multiclassification requires an output node for each class if no output coding is performed. Training of the network can be accomplished using the *backpropagation* learning algorithm [40].

In a Radial Basis Functions (RBF) neural network, notationally, the sigmoid-shape activation function of an MLP is substituted by a Gaussian function (Fig. 6(b)). The learning rule to update weight and variance vectors can be derived

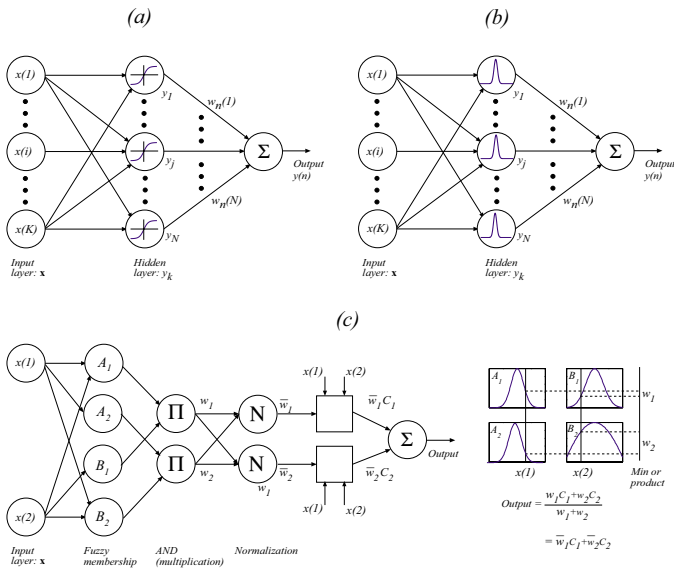


Fig. 6. Schematic of the neural networks used in this work. (a) In an MLP, each neuron passes the weighted sum of its inputs through a sigmoid-shape function (e.g. hyperbolic tangent). The output of a neuron in a given layer acts as an input to neurons in the next layer. In the network illustration, each line represents a synaptic connection. (b) In an RBF neural network, the sigmoidal activation function of an MLP is replaced by a Gaussian function with adjustable widths and centers. (c) A two-input, one-output CANFIS network and an illustration of output calculation.

by using the *delta rule*. Gaussian-like RBFs are local, i.e. give a significant response only in a neighbourhood near the centre. These features induce good mappings but, in turn, may produce overfitting and yield poor results with uncertain inputs (noisy environments).

A very promising paradigm in machine learning is constituted by the neurofuzzy approach in which, fuzzy logic and neural networks are combined. The Co-Active Neuro-Fuzzy Inference Systems (CANFIS) model integrates adaptable fuzzy inputs with a modular neural network to rapidly and accurately approximate complex functions (Fig. 6(c)). Fuzzy inference systems are also valuable as they combine the explanatory nature of rules (membership functions, MF) with the power of neural networks. These kinds of networks solve problems more efficiently than neural networks when the underlying function to model is highly variable or locally extreme since, in those cases, MLP or RBF networks attempt to discover a global optimization. The fundamental component of CANFIS is a fuzzy axon which applies membership functions to the inputs. Basically, two membership function types can be used (Gaussian or generalized bell). Fuzzy axons are valuable because their MF can be modified through backpropagation during network training to expedite the convergence. A second advantage is that fuzzy synapses aid in characterizing inputs that are not easily discretized. The second major component of CANFIS is a modular network that applies functional rules to the inputs. Two fuzzy structures are mainly used; the Tsukamoto model and the Sugeno (TSK) model. Finally, a combiner is used to apply the MF outputs to the modular network outputs. The combined outputs are then channeled through a final output layer and the error is backpropagated

to both the MF and the modular network. Full details of this network can be found in [35].

B. Support Vector Machines

Neural networks and other gradient-descent based methods are trained in order to minimize the so-called *empirical risk*, i.e. the error in the training data set and, therefore, follow the Empirical Risk Minimization (ERM) principle. However, to attain significant results in the validation set (“out-of-sample” dataset), stopping-criteria or pruning techniques must be used. On the other hand, SVMs have been recently proposed as an efficient method for pattern classification and nonlinear regression. Their appeal lies in their strong connection to the underlying statistical learning theory where an SVM is an approximate implementation of the method of structural risk minimization (SRM) [19]. This principle states that a better solution (in terms of generalization capabilities) can be found by minimizing an upper bound of the generalization error.

SVMs have many attractive features. For instance, the solution of the quadratic programming (QP) problem is globally optimized while, with neural networks, the gradient based training algorithms only guarantee finding a local minima. In addition, SVM, can handle large input spaces, which is especially convenient when working with hyperspectral data, can effectively avoid overfitting by controlling the margin, and can automatically identify a small subset made up of informative points, namely *support vectors* (SV). Consequently, they have been used for particle identification, face recognition, text categorization, time series prediction, engine knock detection, bioinformatics, database marketing, etc. The reader can visit <http://www.kernel-machines.org> for introductory tutorials, publications and software resources.

Support Vector methods report four basic characteristics:

- *High generalization capabilities.* The classification methodology attempts to separate samples belonging to different classes by tracing maximum margin hyperplanes, known as Optimal Decision Hyperplanes (ODH) (Fig. 7(a)). Therefore, the global optimization ensures good *a priori* generalization (performance on previously unseen data) capabilities. Maximizing the distance of samples to the ODH is equivalent to minimizing the norm of \mathbf{w} and this becomes the first term in the minimizing functional. For better manipulation of this functional, the squared norm $\|\mathbf{w}\|^2$ is preferred.
- *Slack variables.* When data are not linearly separable, SVMs relax the constraints by introducing positive slack variables ξ_i (allowed errors) for each sample i [41]. The cost associated to each sample is included in the functional to be minimized. Thus, for an error to occur, the corresponding ξ_i must exceed unity, so $\sum_i \xi_i$ is an upper bound on the number of training errors. Hence, a natural way to assign an extra cost for errors is to change the objective function to be minimized from $\|\mathbf{w}\|^2/2$ to $\|\mathbf{w}\|^2/2 + C \sum_i \xi_i$, where C is a parameter to be chosen by the user. A larger C corresponds to assigning a higher penalty to errors.

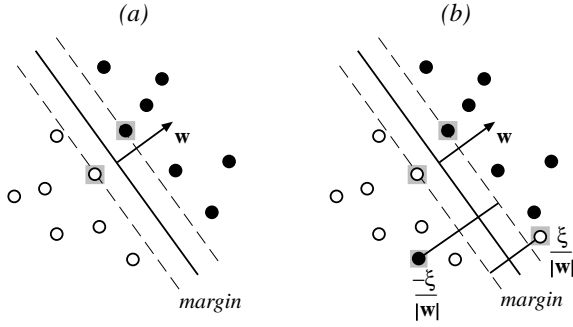


Fig. 7. (a) The Optimal Decision Hyperplane (ODH) in a linearly separable problem. Optimal margin hyperplane is equivalent to minimizing $\|\mathbf{w}\|$. Only support vectors (gray-squared samples) are necessary to define the ODH. (b) Linear decision hyperplanes in nonlinearly separable data can be handled by including slack variables ξ_i to allow classification errors.

- *Feature spaces.* SVMs can also build non-linear decision functions by transforming input data \mathbb{R}^N to a high (possibly infinite) dimensional feature space (\mathbb{R}^H , $H > N$) where data are linearly separable. This rather old trick [42] was used in [43] to accomplish nonlinear SVMs in a straightforward way (Fig. 8). The basic idea of this method is that data appear in the training algorithm in the form of dot products, $\mathbf{x}_i \cdot \mathbf{x}_j$. Therefore, if data are previously mapped ϕ to some other Euclidean space \mathcal{H} , they appear again in the form $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. One does not need to know explicitly the mapping ϕ consequently, but only the kernel function $K(\cdot, \cdot)$. The pair $\{\mathcal{H}, \phi\}$ will exist with the properties described above if Mercer's conditions are satisfied [44].

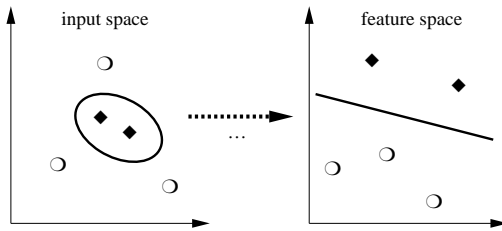


Fig. 8. The so-called "kernel trick" consists of mapping the training samples into a higher-dimensional feature space via a non-linear function ϕ and constructing a separating hyperplane with maximum margin there. This yields a non-linear decision boundary in input space. The figure has been adapted from [20].

- *Support Vectors.* The decision hyperplane is constituted by a linear combination of (few) non-linearly transformed input space samples called support vectors. In order to solve the minimizing functional, restrictions are introduced in it through Lagrange multipliers, α_i . After solving this quadratic problem with linear restrictions, only examples with non-zero α_i count in the solution (the support vectors). This feature reports some advantages in order to analyze the usually complex model, as will be shown in Section VI.

Once the basic ideas underlying SVMs have been presented, in the following sections we provide the standard formulations of the binary classification and multiclassification approaches.

1) *Two-class SVM formulation:* Given a labeled training data set $((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n))$, where $\mathbf{x}_i \in \mathbb{R}^N$ and $y_i \in \{+1, -1\}$ and a nonlinear mapping, $\phi(\cdot)$, usually to a higher dimensional space, $\mathbb{R}^N \xrightarrow{\phi(\cdot)} \mathbb{R}^H$ ($H > N$), the SVM method solves:

$$\min_{\mathbf{w}, \xi_i, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \right\} \quad (1)$$

subject to the following constraints:

$$y_i(\phi^T(\mathbf{x}_i)\mathbf{w} + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n \quad (2)$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n \quad (3)$$

where \mathbf{w} and b define a linear regressor in the feature space, which is nonlinear in the input space. In addition, ξ_i and C respectively are a positive slack variable and the penalization applied to errors (Fig. 7(b)). The parameter C can be regarded as a regularization parameter that affects the generalization capabilities of the classifier and is selected by the user.

An SVM is trained to construct a hyperplane $\phi^T(\mathbf{x}_i)\mathbf{w} + b = 0$ for which the margin of separation is maximized. Using the method of Lagrange multipliers, this hyperplane can be represented as:

$$\sum_i \alpha_i y_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) = 0 \quad (4)$$

where the auxiliary variables α_i are Lagrange multipliers. Its solution reduces to:

Maximize:

$$L_d \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad (5)$$

subject to the constraints:

$$0 \leq \alpha_i \leq C, \quad (6)$$

$$\sum_i \alpha_i y_i = 0 \quad (7)$$

Using the Karush-Kuhn-Tucker (KKT) Theorem, the solution is a linear combination of the training examples that lie closest to the decision boundary. Only these examples, affect the construction of the separating hyperplane.

The mapping ϕ is performed in accordance with Cover's theorem, which guarantees that patterns, that are non-linearly transformed to a high-dimensionality space, are linearly separable there. Working with high dimension converted patterns would, in principle, constitute an intractable problem but all the ϕ mappings used in the SVM learning occur in the form of an inner product. Accordingly, the solution is to replace all the occurrences of an inner product resulting from two mappings with the kernel function K defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \quad (8)$$

Then, without considering the mapping ϕ explicitly, a non-linear SVM can be constructed by selecting the proper kernel.

2) *Multiclass SVM formulation*: One approach to solving K -class problems is by considering the problem as a collection of binary classification problems and then to construct K classifiers (one for each class). The k th classifier constructs a hyperplane between class n and the $k - 1$ other classes. A majority vote across the classifiers or some other measure can then be applied to classify a new sample. Alternatively, $\binom{k(k-1)}{2}$ hyperplanes can be constructed, separating each class from all the others and then applying a similar voting scheme. In conclusion, among the binary settings, we can describe several approaches: the comparison of each class against all the others [45], known as one-vs-all; the comparison of each class against all the other classes individually [46], known as all-pairs, or the comparison of a subset of classes against the rest of them using error correcting codes [47]. The last one represents specific cases of the previous two.

In this paper, we have adopted the multiclassification approach, which is formulated as follows. Given a classifier (\mathbf{w}^j, b^j) , $j \in \{0, \dots, k - 1\}$ for each class, in order to assign a sample \mathbf{x} to a certain k th class, we must calculate the output of the k classifiers and select the one with the highest output. We must thus solve the following convex problem:

$$\min_{\mathbf{w}^j, b^j, \xi_i^{j,m}} \frac{1}{2} \sum_{k=0}^{k-1} \|\mathbf{w}^j\|^2 + C \sum_{j=0}^{k-1} \sum_{m=0, m \neq j}^{k-1} \sum_{i=1}^{n_j} \xi_i^{j,m} \quad (9)$$

subject to the following restrictions:

$$(\mathbf{w}^j)^T \phi(\mathbf{x}_i^j) + b^j - (\mathbf{w}^m)^T \phi(\mathbf{x}_i^j) + b^m \geq 1 - \xi_i^{j,m} \quad (10)$$

$$\xi_i^{j,m} \geq 0, \quad (11)$$

where \mathbf{x}_i^j represents the sample i from class j , $\forall j = 0, \dots, k - 1$, $\forall m = 0, \dots, k - 1$ ($m \neq j$), and $\forall i = 1, \dots, n_j$. If the problem is separable, all $\phi(\mathbf{x}_i^j) = 0$ and, in addition, restriction (10) indicates that output provided by the classifier j (\mathbf{w}^j, b^j) to the n_j samples \mathbf{x}_i^j must be greater than the one provided by the rest $k - 1$ classifiers, assuring a minimum margin between samples belonging to different classes. The minimizing functional guarantees that the margin is maximum. Auxiliar variables $\xi_i^{j,m}$ have been introduced, as in $k=2$, in order to solve non-linearly separable problems.

We then proceed as in the two-class case. First, we obtain the minimizing functional introducing the linear restrictions through Lagrange multipliers. We then use the KKT conditions to obtain Wolfe's dual problem as the maximizing functional which only depends on the Lagrange multipliers $\alpha_i^{j,m}$. The non-linear transformation $\phi(\cdot)$ appears again in such a way that it is not necessary to know its explicit form and thus, we can work with the Reproducing Kernels in Hilbert Spaces (RKHS). See [48] for full details.

V. CLASSIFICATION RESULTS

A. Model development

As regards the MLP and RBF models, we varied the number of hidden neurons (< 100 to avoid overfitting), the weight initialization range and the learning rate (between 0.01 and 3) in order to determine the best topology. A great amount of CANFIS models were developed by varying the number

(2-8) and structure (Bell and Gaussian) of the MF and the fuzzy model (TSK and Tsukamoto), along with the number of hidden layers (2-5) and step size (0.001-0.1). The momentum term remained constant and equal to zero.

In the case of SVMs, nonlinear classifiers were obtained by taking the dot product in kernel-generated spaces. The following RKHS have been used in this work:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
- Gaussian (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

Note that one or more free parameters must be previously settled in the nonlinear kernels (polynomial degree d , Gaussian width γ) together with the *penalization* parameter C . In all cases, we considered equiprobable classes for training and validation and thus no individual penalization parameter was used [49]. However, the test set contains highly unbalanced classes and thus, the latter practice could improve results if the training process were intentionally driven by priors. However, this would not be a fair assumption for our purposes, i.e. achieving an automatic scenario-independent classifier.

The selection of the best subset of free parameters is usually done by cross-validation methods but this can lead to poor generalization capabilities and lack of representation. We alleviated this problem by using the 8-fold cross-validation method² with the training data set.

Many discriminative methods, including neural networks and SVMs, are often more accurate and efficient when dealing with only two classes. For large numbers of classes, higher-level multiclass methods utilize these two-class classification methods as the basic building blocks, namely "one-against-the-rest" procedures. However, such approaches lead to suboptimal solutions when dealing with multiclass problems and the well-known problem of the "false positives". Therefore, we have used a multiclassification scheme for all the methods.

All neural models were developed in MATLAB[®] environment (Mathworks, Inc). Since the computational burden was very high, *m-files* were translated to MEX-files and the programs were run on fast workstations. In the case of SVM, we used the OSU implementation, which is available from http://www.eleceng.ohio-state.edu/~maj/osu_svm/.

B. Model comparison

Table III shows the average recognition rate (ARR[%]) of the six images in training, validation, and test sets. The ARR% is calculated as the rate of correctly classified samples over the total number of samples averaged over the six available images. Section III contains details on the training, validation and test sets.

Some conclusions can be drawn from Table III. SVMs perform better than neural networks in all scenarios. Moreover, when a feature selection stage is not possible, and thus 128 bands should be used, the computational burden involved in the training process of neural networks make these methods unfeasible. In contrast, SVMs are not drastically affected

²The 8-fold cross-validation uses 7/8 of the data for training and 1/8 for validation purposes. This procedure is repeated eight times with different validation sets.

TABLE III

AVERAGE RECOGNITION RATES (ARR [%]) OF THE SIX IMAGES IN TRAINING, VALIDATION, AND TEST SETS FOR DIFFERENT MODELS. THE FOUR SUBSETS (128, 6, 3, 2 BANDS) ARE EVALUATED, ALL OF THEM CONTAINING 150 SAMPLES PER CLASS. THE COLUMN "FEATURES" GIVES SOME INFORMATION ABOUT THE FINAL MODELS. FOR THE CASE OF SVMs, WE INDICATE IN BRACKETS THE PENALIZATION PARAMETER, THE KERNEL USED AND ITS OPTIMAL PARAMETERS (POLYNOMIAL ORDER d OR GAUSSIAN WIDTH γ), AND THE RATE OF SUPPORT VECTORS, RESPECTIVELY. BOLD FACE FONT IS USED TO INDICATE THE BEST KERNEL IN EACH SUBSET. FOR THE CASE OF NEURAL NETWORKS, WE INDICATE THE NUMBER OF INPUT \times HIDDEN \times OUTPUT NODES.

METHOD	FEATS.	TRAIN.	VALID.	TEST
SVM128	Linear	99.89	98.78	95.45
SVM128	Polynomial (5.59, 4, 12.11%)	100	98.78	95.53
SVM128	RBF	100	97.78	94.13
SVM6	Linear	99.89	99.33	94.44
SVM6	Polynomial (20.57, 4, 8.67%)	99.79	99.44	96.44
SVM6	RBF	100	98.78	94.87
SVM3	Linear	89.00	87.22	81.31
SVM3	Polynomial	88.89	87.44	82.03
SVM3	RBF (35.94, 10^{-5} , 12.88%)	91.22	91.00	85.16
SVM2	Linear	89.11	88.33	81.42
SVM2	Polynomial	89.11	88.33	82.55
SVM2	RBF (43.29, 10^{-2} , 16.88%)	89.11	89.11	82.68
MLP128	-	-	-	-
MLP6	6 \times 5 \times 6	99.33	99.44	94.53
MLP3	3 \times 25 \times 6	90.22	87.67	82.97
MLP2	2 \times 27 \times 6	88.00	85.67	81.95
RBF128	-	-	-	-
RBF6	6 \times 16 \times 6	98.88	98.80	94.10
RBF3	3 \times 31 \times 6	88.20	87.00	81.44
RBF2	2 \times 18 \times 6	87.33	85.25	81.62
CANFIS128	-	-	-	-
CANFIS6	6 \times 2 \times 7 \times 6	98.68	96.66	94.22
CANFIS3	3 \times 3 \times 12 \times 6	89.20	88.77	81.64
CANFIS2	2 \times 8 \times 15 \times 6	86.33	86.00	81.82

by input dimension and presence of noisy bands. This has sometimes led to the idea that a feature selection is not necessary when working with SVMs, which is not completely true, as shown in [20], [50]. In noisy applications, a feature selection is not only recommendable but mandatory, since it could remove undesired features and better results could thus be obtained. In our case study, no numerical (ARR<3%) or statistical (κ scores in the range [0.6,0.8]) differences are found between SVMs with and without a step for dimensionality reduction prior to classification. This indicates that noisy bands have been successfully identified and their contribution to the final decision attenuated without decreasing the recognition rate. Therefore, two preliminary conclusions can be extracted:

- 1) SVMs have proven to be efficient models that inherently detect noisy features.
- 2) A feature selection step slightly improves results.

This induces a clear trade-off: we could obtain good results by using an SVM without a preliminary feature selection stage or, we could (slightly) improve results by including a dedicated

feature selection step, which is time-consuming and requires more effort. Depending on the application requirements, the user could choose between these two options.

In the same table, we also observe that, as the dimension of the input space is lower, neural networks degrade more rapidly than SVMs do. In that sense, the complexity³ of all models increases as the input dimension decreases. In fact, RBF kernels and more than 15% of SVs are strictly necessary to attain significant results with less than six bands. Despite the fact that the polynomial kernel has been claimed to be specially well-suited for hyperspectral data classification [28], it has yielded results similar to the ones for the linear kernel in our case (see the next section for details).

Table IV shows the confusion matrix of an image provided by the best classifier (SVM with polynomial kernel, 6 bands). We also include the two methods of calculation classification accuracy: users accuracy and producers accuracy for each class. Users accuracy (UA[%]) calculates correctly classified samples in a desired class over the total samples in that desired class, and provides an indication of errors of case omission. Producers accuracy (PA[%]) is the calculation of correctly classified samples in a predicted class over the total samples in that predicted class. High rates of users and producers accuracies (UA>90%, PA>84%) are achieved for all classes but SVMs misclassify almost 6% of bare soils (class #6) as corn (class #1), which is due to the fact that corn is in an early stage of maturity.

TABLE IV

CONFUSION MATRIX ALONG WITH THE USERS ACCURACY (UA%) AND PRODUCERS ACCURACY (PA%) YIELDED BY THE BEST SVM CLASSIFIER IN THE TEST SET (WHOLE SCENE).

Desired class	Predicted class						UA[%]
	#1 Corn	#2 Sugar beets	#3 Barley	#4 Wheat	#5 Alfalfa	#6 Soil	
#1	31188	67	7	1	0	6	99.74
#2	23	11256	43	0	0	0	99.42
#3	732	702	120874	1993	18	449	96.88
#4	12	108	320	52956	4	0	99.17
#5	28	106	140	36	24413	3	98.73
#6	4914	1003	1539	190	15	74190	90.64
PA%	84.53	85.00	98.33	95.98	99.85	99.39	

Figure 9 shows the original and the classified samples for one of the collected images. Corn classification seems to be the most troublesome. The reason for that is the presence of a whole field of two-leaf corn in the early stage of maturity, where soil was predominant and was not accounted for the reference labelled image. The confusion matrix supports this conclusion as most of the errors are committed with the bare soil class.

³We evaluate the model's complexity in terms of the kernel used and the number of SVs in the SVM approach, and in terms of the number of hidden neurons in the neural networks. We have based this decision on the works [45], [51], [20], where an intuitive relation between neural networks and Support Vector Machines is sketched.

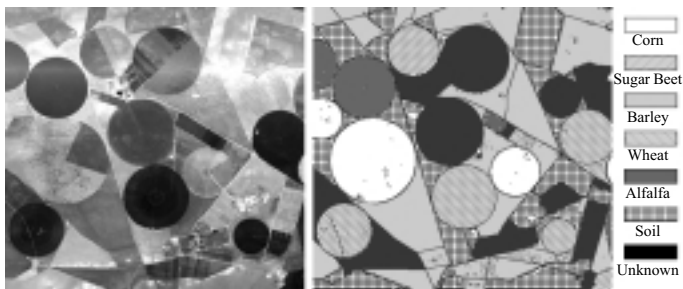


Fig. 9. (a) RGB composite of the red, green and blue channels from 128-bands HyMAP image taken in June, 1999 of Barrax (Spain). (b) Map of the whole image classified with the labels of the classes of interest.

C. Effect of free parameters

In order to develop a support vector classifier, the penalization parameter C and the kernel parameters must be tuned. It is a common practice to trying exponentially increase sequences of C in order to identify good parameters ($C = 10^{-2}, 10^{-1}, \dots, 10^6$). In our case study, good results were achieved in the range of $C \in [1, 100]$. Nevertheless, this parameter showed relatively high variability for each scenario. In general, as the input dimension increased, the necessary penalization parameter decreased. This could be related to the fact that more information was added and thus, a lower penalization of errors was necessary. However, this must be assessed in other applications and scenarios.

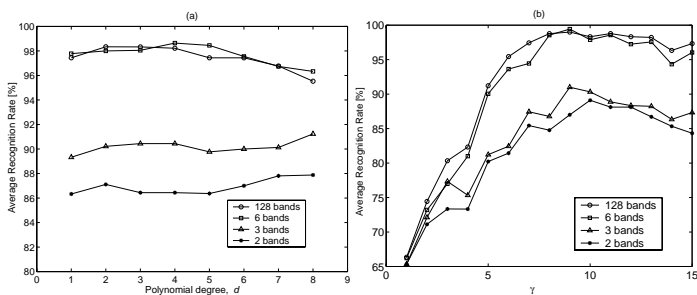


Fig. 10. Impact of the (a) polynomial and (b) RBF kernel parameters on the overall performance.

Figure 10 shows the impact of kernel parameters in the average recognition rate. Several conclusions can be drawn from this analysis:

- *Polynomial kernel.* As shown in Section V-B, the parameter to be defined for this kernel is the polynomial order, d , which was varied in the range 1 to 8, as suggested in the literature [41]. Fig. 10(a) shows the influence of this parameter on the overall performance in the four scenarios considered. A non-linear ($d > 1$) approach results in better overall performance in the four scenarios. This result was expected since boundaries between classes are presumed to be non-linear. In general, better results are achieved as d increases, especially significant for the cases of three and two bands (optimal $d = 8$). This fact agrees with the results provided in [52], [24]. On the other hand, when using more than six bands, a local maximum of the recognition rate is observed around

$d = 4$, which also agrees with the work of Cortes and Vapnik [41]. In fact, the obtained results in this paper confirm the hypothesis presented in [24] by which high dimensional input spaces can be mapped into linear ones by using relatively low-order polynomial orders.

- *RBF kernel.* We varied the γ parameter between 1 and 15 according to preliminary studies [52], [24]. Fig. 10(b) shows the influence of this parameter on the overall performance. In all scenarios, we obtain similar behaviour; the ARR[%] increases as γ rises in the range 1-8. A global maximum is observed at $\gamma = 9$ (three bands) and $\gamma = 10$ (two bands). Therefore, as the number of input variables is increased, smoother solutions (lower values of γ) become necessary.

D. Robustness in noisy conditions

Since neither numerical nor statistical differences have been observed between neural networks and SVMs, we decided to test the robustness of the classifiers. We tested the robustness capabilities over the best classifiers (6 bands) by introducing Gaussian noise with zero mean and standard deviation σ , $\mathcal{N}(0, \sigma)$, together with the inputs. This simulates situations such as labelling errors, sparse classification boundaries or sensitivity of the classifier to exact input values. The results are shown in Fig. 11.

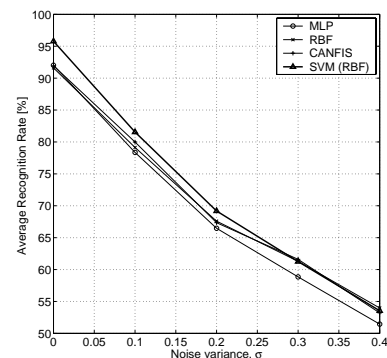


Fig. 11. Evaluation of the recognition rate in the validation set when additive Gaussian noise with zero mean and standard deviation σ is introduced in the best classifier. Test was repeated 100 times, which represents a reasonable confidence margin for the measured recognition rate.

An exponential decay of the overall performance is observed as the noise variance is increased. SVMs perform slightly better than neural networks, and this difference is constant as a function of σ . All kernels worked similarly but the performance of RBF was slightly better in the range $\sigma \in [0, 0.4]$ (results not shown). When $\sigma > 0.2$, RBF kernel deteriorates its performance more rapidly than linear or polynomial kernels do. This is explained by the direct influence of C and γ in the solution. Smooth solutions allow good results without noise ($\sigma = 0$) but deteriorates the overall performance in the presence of high noise levels due to the local mapping provided by RBF kernels. However, high levels of additive noise ($\sigma > 0.3$) are non-standard situations and thus we can claim that RBF kernels yield accurate (noise free, $\sigma = 0$) and robust (moderate noise levels, $\sigma < 0.2$) solutions.

VI. KNOWLEDGE DISCOVERY

SVMs have demonstrated to be well-suited techniques in classification and regression tasks with an additional advantage: their solution is a linear combination of some (non-linearly transformed) training vectors and thus its analysis can provide added knowledge about the problem. In this communication, we perform (1) a geometrical analysis of the input space and its relationship to the critical samples and (2) a sensitivity analysis of the best overall classifier.

A. Geometry analysis in input spaces

The classifier using 6 bands was formed by 78 support vectors (SVs), namely 8.67% of the whole training data set, which indicates that a very reduced subset of examples is necessary to attain significant results. However, in order to analyze the geometry of the entire input space, it is more convenient to use the SVM trained with all available bands (SVM128, 12.11% SVs). By visual inspection of the distribution of support vectors in each class, we observed that classes #1 (unmatured corn) and #6 (soil) could be discriminated mainly with high reflectance values in bands 1-20. This can be explained since plants present chlorophyll and other pigment absorptions in these visible spectrum bands. This result matches perfectly with the ones obtained from the CART selection (see Table II) in which bands 6, 17 and 22 provide information on cellular pigments and chlorophyll-a maximum absorption, respectively. With a similar analysis, class #5 (alfalfa crops with very homogeneous green cover) can be successfully identified with high values in bands 20-60. This could be due to the fact that plants present high reflectance in these near infrared spectrum bands. CART also selected representative bands in this spectrum bandwidth; band 22 is related to canopy maturity, and band 24 provides more reflectance and less absorbance due to leaf structure.

B. Sensitivity analysis

Sensitivity analysis is used to study the influence of input variables on the dependent variable and consists of evaluating the changes in training error that would result if an input were removed from the model. This measure, commonly known as *delta error* in the literature, produces a valuable ranking of the relevance of input variables. An additional sensitivity measure, called *Average Absolute Gradient (AAG)* has been computed. This measurement is based on perturbing an input and monitoring model outputs and is extensively described in [53].

In Figure 12, we show the ranking of variables according to these two common sensitivity measures for the best SVM with six input features⁴. An additional measure of the feature relevance is to consider all partitions whose order comprises more than 90% of the relative relevance. In our case study, an order of $m = 5$ was found. This indicates that partitions with higher order contribute with few information, i.e. five variables

⁴Models were re-trained after each feature selection run, as proposed in [54], [55]. This methodology ensures effectiveness in the feature selection process.

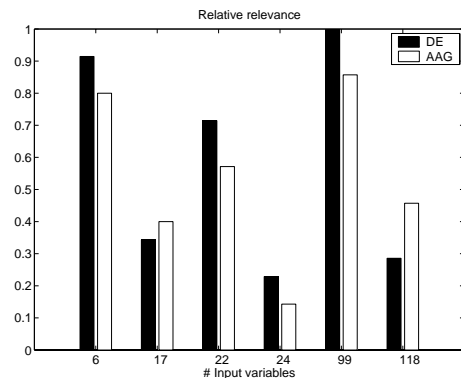


Fig. 12. Ranking provided by sensitivity measures of the best SVM (polynomial kernel) using six input bands.

are enough to describe the problem accurately. In addition, a significant difference is found between $m = 6$ and $m = 3$ regarding the 0.9-quantile of the feature selection problem (1 and 0.75, respectively), which could explain the lower results obtained when using less than six variables. These results become more evident when one compares the information contained in the subset of two input variables where the 0.9-quantile falls to 0.31.

From a physical viewpoint, the six selected spectral bands are related to: cellular pigments (carotenoids) absorption; chlorophyll absorption; red edge (change visible/near-infrared) related with canopy maturity; leaf structure at the beginning of near-infrared with more reflectance and less absorbance; and water absorption bands due to soil moisture and leaf water content. The selection of these bands is consequent with the characteristics of the crop fields to be classified in the spectral domain. See Fig. 13 for proper analysis.

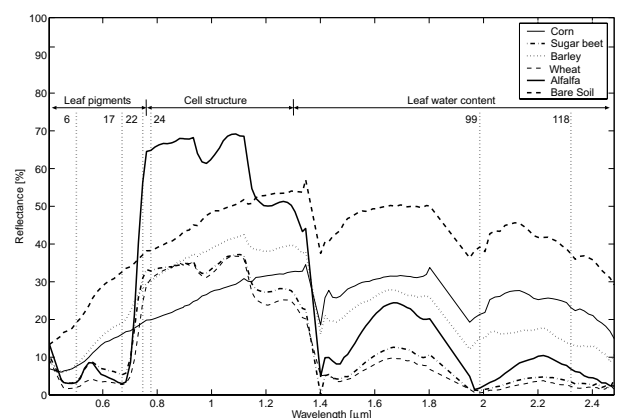


Fig. 13. Reflectance [%] curves of all the crops in the scene and the selected subset of six spectral bands extracted from atmospherically corrected HyMAP data.

VII. DISCUSSION AND CONCLUSIONS

In this communication, we have proposed the use of kernel methods for both hyperspectral data classification and knowledge discovery. In the first stage of the study, we used SVMs for crop classification and analyzed their performance

in terms of efficiency and robustness, as compared to other well-known neural approaches. Several tests have provided useful information about possible limitations of classifiers working with and without a feature selection stage. In the second stage of this work, we analyzed the distribution of SVs in the input space, and then performed a sensitivity analysis of the best method.

Several conclusions can be drawn from this work:

- *Accuracy.* SVMs yield better outcomes than neural networks in terms of recognition and misrecognition rates. Despite the fact that the differences between methods are neither numerical nor statistical in the training and validation sets, better results are obtained in the test set, which indicates that a stable model has been obtained. We have shown both the best overall models for each scenario and the average recognition rates *per class*. The latter gives some insight into the class complexity, i.e. the complexity of the mapping necessary to attain significant results. An additional advantage of the use of SVMs is that good results are obtained with less influence by the input dimension.
- *Simplicity and computational cost.* We used three-layer (input, hidden and output) neural networks architecture, which is enough for classifying hyperspectral imageries [56] (except for the case of the CANFIS model, which uses five layers). Training neural models became unfeasible with 128 input bands. This is an important benefit of using SVMs. On one hand, the input dimensionality does not have a dramatic influence on the computational cost. On the other hand, the computational cost involved in training an SVM increases in polynomial time with the number of training samples since a restriction for each sample is included in the minimizing functional. This is not specially dramatic in our application where we have used only 900 samples for training, which did not take more than a few seconds using the MATLAB OSU implementation. These characteristics make the SVM approach well-suited to this problem.
In addition, training a neural network requires tuning several parameters such as the transfer function, the cost function, the training algorithm, the network architecture, learning parameters such as the learning rate and the momentum term, the number of epochs, and defining a stopping criterion. For training the SVM, one only has to select a kernel function, its free parameters and the regularized constraint C . There are many reasons to select the RBF kernel *a priori*: it has less numerical difficulties, and only the Gaussian width has to be tuned. The use of the RBF kernel, implicitly converts the SVM into a regularized RBF neural network but with the additional advantage that the centers of the Gaussians are tuned automatically. In addition, sigmoid kernels behave like RBF for certain parameters [57], [58] but unfortunately, they are non-positive definite kernels in all situations, which precludes their practical application [20].
- *Robustness to input space dimension.* In our thematic application, SVMs have performed similarly in the four

classification scenarios, which indicates that noisy bands have been successfully detected. This, in turn, leads to the conclusion that SVMs are well-suited techniques in applications where the number of potentially useful input features is high and a feature selection stage is not possible or is unadvisable given the application technical specifications. The issue of feature selection in the SVM framework has received attention in the recent years [50], [59]. The fact that SVMs are not *drastically* affected by the input space dimensionality has sometimes led to the wrong idea that a feature selection is not necessary at all. The SRM principle ensures certain robustness to outliers or abnormal samples in the distribution inherently, but the selection of the optimal subset of training features is still an unsolved problem in the literature. We can state that in most applications, the success of machine learning is strongly affected by data quality (redundant, noisy or unreliable information) and thus a feature selection is not only recommendable but mandatory. Nevertheless, in our specific application, where the sensor provides high resolution imageries and few features (only 7 out of 128) can be treated as effective disturbing or noisy samples, a feature selection step is not strictly necessary and good results can be obtained by using an SVM without a feature selection stage. We have also shown how the inclusion of a feature selection stage (CART) has removed undesired features, thereby obtaining (slightly) better results. However, this improvement does not compensate the effort (Section V-D).

- *Robustness to outliers.* In general, for any real-world application, observations are always subject to noise or outliers. Outliers may occur for various reasons, such as erroneous measurements or noisy phenomenon appearing in the tail portion of some noise distribution functions. When the obtained observations contain noise or outliers, the learning process, being unaware of those situations, may try to fit that unwanted data and this behavior may lead to a corrupted approximation function. This phenomenon is often called overfitting, which can usually lead to the loss of generalization performance in the test phase. The issue of robustness to outliers has been dealt with the SVM literature, which has also been assessed in this paper. Basically, a more stable solution is obtained using SVMs over neural networks.
- *Interpretability.* An additional advantage found in the SVM framework is that the solution is expressed as a (nonlinear) function of the most representative input space samples in the distribution and thus, the analysis of these samples adds some knowledge gain about the problem. From the sensitivity analysis of the classifier, specific bands have been identified as especially relevant and a physical interpretation has been provided for our specific application.

In conclusion, SVMs have proven to be very efficient in different situations when a feature selection phase is not possible. This method has tolerated the presence of ambiguous patterns and features in our data set. The fact that we have

obtained simple solutions (low rate of SVs) can induce a good method for compression of hyperspectral images with minimal loss of critical information. In [60], a support vector regressor has been presented for 2D image coding. Presently, we are extending the use of SVMs as image compression tools to the 3D hypercube case in order to provide an efficient method for Level Two product users. Additionally, a very interesting possibility in the DAISEX project consists of validating results with HyMAP imagery from other campaigns (2000 and 2003) in order to achieve robust and automatic, multi-temporal classification methods. Finally, inclusion of spatial information to the automatic classifier could improve the results, as suggested in [29].

ACKNOWLEDGEMENTS

This research has been partially supported by the Information Society Technologies (IST) programme of the European Community. The results of this work will be applied in the "Smart Multispectral System for Commercial Applications" project (SmartSpectra, www.smartspectra.com).

All the data used were acquired in the Scientific Analysis of the European Space Agency (ESA) Airborne Multi-Annual Imaging Spectrometer Campaign DAISEX (Contract ESA/ESTEC 15343/01/NL/MM).

REFERENCES

- [1] P. Swain, *Remote Sensing: The Quantitative Approach*. New York, NY: McGraw-Hill, 1978, ch. Fundamentals of pattern recognition in remote sensing, pp. 136–188.
- [2] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873–885, 1989.
- [3] C. Bachmann, T. Donato, G. M. Lamela, W. J. Rhea, M. H. Bettenhausen, R. A. Fusina, D. K. R., J. H. Porter, and B. R. Truitt, "Automatic classification of land cover on Smith Island, VA, using HyMAP imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2313–2330, 2002.
- [4] T. Moon and E. Merenyi, "Classification of hyperspectral images using wavelet transforms and neural networks," in *Proceedings of the Annual SPIE Conference*, 1995, p. 2569.
- [5] P. Blonda, V. laForgia, G. Pasquariello, and G. Satalino, "Feature extraction and pattern classification of remote sensing data by a modular neural system," *Optical Engineering*, vol. 35, pp. 536–542, 1996.
- [6] M. Lennon, G. Mercier, and L. Hubert-Moy, "Classification of hyperspectral images with nonlinear filtering and support vector machines," in *IEEE International Geoscience and Remote Sensing Symposium. IGARSS'02*, vol. 3, Toronto, Canada, Jun 2002, pp. 1670–1672.
- [7] F. Melgani and S. R. Serpico, "A statistical approach to the fusion of spectral and spatio-temporal contextual information for the classification of remote-sensing images," *Pattern Recognition Letters*, vol. 23, pp. 1053–1061, 2002.
- [8] A. Bardossy and L. Samaniego, "Fuzzy rule-based classification of remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 2, pp. 362–374, Feb. 2002.
- [9] D. Trizna, C. Bachmann, M. Sletten, N. Allan, J. Toporkov, and R. Harris, "Projection pursuit classification of multiband polarimetric SAR land images," in *International Geoscience and Remote Sensing Symposium. IGARSS*, Nov. 2001, pp. 2380–2386.
- [10] L. Bruzzone and D. Fernandez-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, pp. 1179–1184, 1999.
- [11] G. Giacinto, F. R., and L. Bruzzone, "Combination of neural and statistical algorithms for supervised classification of remote-sensing images," *Pattern Recognition Letters*, vol. 21, no. 5, pp. 399–405, 2000.
- [12] L. Bruzzone and R. Cossu, "A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 9, pp. 1984–1996, 2002.
- [13] D. L. Civco, "Artificial neural networks for land-cover classification and mapping," *International Journal of Geophysical Information Systems*, vol. 7, no. 2, pp. 173–186, 1993.
- [14] P. Dreyer, "Classification of land cover using optimized neural nets on SPOT data," *Photogrammetric Engineering and Remote Sensing*, vol. 59, no. 5, pp. 617–621, 1993.
- [15] H. Bischof and A. Leona, "Finding optimal neural networks for land use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 1, pp. 337–341, 1998.
- [16] H. Yang, F. van der Meer, W. Bakker, and Z. J. Tan, "A back-propagation neural network for mineralogical mapping from AVIRIS data," *International Journal of Remote Sensing*, vol. 20, no. 1, pp. 97–110, 1999.
- [17] N. Harvey, S. Brumby, S. Perkins, J. Szymanski, J. Theiler, J. Bloch, R. Porter, M. Galassi, and A. C. Young, "Image feature extraction: GENIE vs conventional supervised classification techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 4, pp. 393–404, 2002.
- [18] E. Merenyi, R. B. Singer, and W. H. Farrand, "Classification of the LCVF AVIRIS test site with a Kohonen artificial neural network," in *Proceedings of the Fourth Annual JPL airborne earth science workshop*, 1993, pp. 117–120.
- [19] V. N. Vapnik, *Statistical Learning Theory*. New York: John Wiley & Sons, 1998.
- [20] B. Schölkopf and A. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press Series, 2001.
- [21] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.
- [22] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, "Mathematical programming for data mining: formulations and challenges," *INFORMS Journal on Computing*, vol. 11, no. 3, pp. 217–238, 1999. [Online]. Available: citeseer.nj.nec.com/bradley98mathematical.html
- [23] J. Pierce, M. Diaz-Barrios, J. Pinzon, S. L. Ustin, P. Shih, S. Tournois, P. J. Zarco-Tejada, V. C. Vanderbilt, and G. L. Perry, "Using support vector machines to automatically extract open water signatures from POLDER multi-angle data over boreal regions," in *International Geoscience and Remote Sensing Symposium. IGARSS'02*, 2002, pp. 2349–2350.
- [24] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.
- [25] M. Azimi and S. A. Zekavat, "Cloud classification using support vector machines," in *IEEE International Geoscience And Remote Sensing Symposium. IGARSS'00*, vol. 2, Hawaii, USA, July 2000, pp. 669–671.
- [26] S. Perkins, N. Harvey, S. Brumby, and K. Lacker, "Support vector machines for broad area feature extraction in remotely sensed images," in *Proceedings of SPIE 4381*, April 2001.
- [27] L. Bruzzone and F. Melgani, "Support vector machines for classification of hyperspectral remote-sensing images," in *IEEE International Geoscience and Remote Sensing Symposium. IGARSS'02*, vol. 3, Toronto, Canada, Jun 2002, pp. 1670–1672.
- [28] J. A. Gualtieri and R. F. Crompt, "Support vector machines for hyperspectral remote sensing classification," in *Proceedings of the SPIE, 27th AIPR Workshop*, Feb. 1998, pp. 221–232.
- [29] J. A. Gualtieri, S. R. Chettri, R. F. Crompt, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *Proceedings of The 1999 Airborne Geoscience Workshop*, Feb. 1999.
- [30] J. Zhang, Y. Zhang, and T. Zhou, "Classification of hyperspectral data using support vector machine," in *IEEE International Conference on Image Processing*, 2001, pp. 882–885.
- [31] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, and J. Moreno, "Support vector machines for crop classification using hyperspectral data," in *1st Iberian Conference on Pattern Recognition and Image Analysis*. Mallorca, Spain: Lecture Notes in Computer Science. Springer-Verlag, Jun 2003, pp. 134–141.
- [32] —, "Kernel methods for HyMap imagery knowledge discovery," in *SPIE International Symposium Remote Sensing*, Barcelona, Spain, Set 2003.
- [33] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice Hall, 1999.

- [34] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press—Oxford, 1996.
- [35] J. Jyh-Shing Roger, S. Chuen-Tsai, and M. Eiji, *Neuro-Fuzzy and Soft-Computing*. Englewood Cliffs, NJ: Prentice Hall, 1997.
- [36] A. Müller, R. Richter, and U. Heiden, "Vicarious calibration of imaging spectrometers in the reflective region," in *Final Results Workshop on DAISEX, ESA/ESTEC*. Noordwijk, The Netherlands: ESA Publications Division., 2001.
- [37] R. Richter, "Atmospheric correction methodology for imaging spectrometer data," in *Final Results Workshop on DAISEX, ESA/ESTEC*. Noordwijk, The Netherlands: ESA Publications Division., 2001.
- [38] J. Moreno, V. Caselles, J. Martínez-Lozano, J. Melia, J. Sobrino, A. Calera, F. Montero, and J. Cisneros, "The measurement programme at Barrax," in *Final Results Workshop on DAISEX, ESA/ESTEC*. Noordwijk, The Netherlands: ESA Publications Division., 2001.
- [39] L. Gómez-Chova, J. Calpe, E. Soria, G. Camps-Valls, J. D. Martín, and J. Moreno, "CART-based feature selection of hyperspectral images for crop cover classification," in *IEEE International Conference on Image Processing*, Barcelona, Spain, Set 2003.
- [40] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986, vol. 1.
- [41] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273 – 297, 1995.
- [42] A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.
- [43] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *5th Annual ACM Workshop on COLT*, D. Haussler, Ed. Pittsburgh, PA: ACM Press, 1992, pp. 144–152. [Online]. Available: <http://www.clopinet.com/isabelle/Papers/colt92.ps>
- [44] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. New York: Interscience Publications. John Wiley, 1953.
- [45] B. Schölkopf, K.-K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. N. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, Nov. 1997.
- [46] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The annals of statistics*, vol. 26, no. 2, pp. 475–471, 1998.
- [47] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, no. 2, pp. 263–286, 1995.
- [48] J. Weston and C. Watkins, "Multi-class Support Vector Machines," Royal Holloway - University of London. Dpt. of Computer Science. Egham, Surrey TW20 0EX, England, Tech. Rep. CSD-TR-98-04, May 1998, <http://citeseer.nj.nec.com/8884.html>.
- [49] Y. Lin, Y. Lee, and G. Wahba, "Support Vector Machines for classification in nonstandard situations," University of Wisconsin-Madison, Department of Statistics TR 1016, 2000.
- [50] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *NIPS*, 2000, pp. 668–674. [Online]. Available: citeseer.nj.nec.com/article/weston01feature.html
- [51] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp. 121–167, 1998. [Online]. Available: [/papers/Burges98.ps.gz](http://papers/Burges98.ps.gz)
- [52] J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *NASA Goddard Space Flight Center Applied Information Sciences Branch - Code 935*, 2000.
- [53] G. B. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade*. Springer-Verlag, Berlin, Heidelberg, 1998.
- [54] A. N. Refenes, A. Zapranis, and G. Francis, "Stock performance modeling using neural networks: A comparative study with regression models," *Neural Networks*, vol. 7, no. 2, pp. 375–388, 1994.
- [55] W. S. Sarle, "How to measure importance of inputs?" Available at <ftp://ftp.sas.com/pub/neural/importance.html>, SAS Institute Inc., Cary, NC, USA, 2000.
- [56] J. D. Paola and R. A. Schowengerdt, "A review and analysis of back-propagation neural networks for classification of remotely sensed multi-spectral imagery," *International Journal of Remote Sensing*, no. 16, pp. 3033–3058, 1995.
- [57] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.
- [58] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," National Taiwan University, Department of Computer Science and Information Engineering, Tech. Rep., 2003, available at <http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf>.
- [59] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *Journal of Machine Learning Research*, no. 3, pp. 1357–1350, 2003.
- [60] J. Robinson and V. Kecman, "Combining Support Vector Machine learning with the discrete cosine transform in image compression," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 950–958, 2003.

Gustavo Camps-Valls received a PhD degree in Physics (2002) from the Universitat de València, Spain. He is currently an associate professor in the Department of Electronics Engineering at the Universitat de València. His research interests are neural networks and kernel methods for hyperspectral data classification, health sciences, and safety-related areas. Visit <http://www.uv.es/~gcamps> for more information.

Luis Gómez-Chova received his B.Sc. with first class honours in 2000 in Electronics from the Universitat de València, Spain. He was awarded by the Spanish Ministry of Education (MECD) with the National Award for Electronics Engineering. Since 2000 he has been with the Department of Electronic Engineering. Last year, he read his Master Thesis. He is pursuing his PhD degree with a research scholarship from the MECED.

Javier Calpe-Maravilla received his B.Sc. in 1989 and his Ph.D. degree in 1993 in Physics from the Universitat de València, Spain. Since 1991 he has been with the Department of Electronics Engineering at the Universitat de València, where he belongs to the Digital Signal Processing Group, GPDS. He is an Assistant Professor and Head of Department. He holds an industrial patent, has co-authored more than 47 papers in scientific magazines and 114 communications to conferences, worked on 21 projects with private companies, and 8 with public funds, including co-leading one funded by the European Union. His research activities include DSP, its industrial applications, and smart sensors.

José David Martín-Guerrero received the MS Degree in Electronics Engineering in 2001 from the Universitat de València, Spain. He is currently working towards his PhD degree in Artificial Intelligence methods. His research interests include neural networks, fuzzy logic and statistical methods. He is a Member of the European Neural Network Society.

Emilio Soria-Olivas is an Assistant Professor at the University of València. He obtained his PhD degree in Electronics Engineering in 1997. His research is centered mainly in the analysis and applications of adaptive and neural systems.

Luis Alonso-Chordá received his B.Sc. in 1999 in Physics from the Universitat de València, Spain. Since 1999 he has been with the Department of Thermodynamics. Currently, he is pursuing his PhD degree in Physics working on the geometric correction problem of airborne and spaceborne remote sensing imagery. Other research interests include remote sensing of canopy chlorophyll fluorescence, and scaling effects in RS imagery.

José Moreno is a Professor of Earth Physics at the Universitat de València, Spain. His main work is related to modelling and monitoring land surface processes. He has been involved in many international projects and research networks, being a PI for ENVISAT and CHRIS/PROBA projects, and was responsible for the DAISEX and SPARC ESA campaigns. During 1995-1996 he was a visiting scientist at NASA/JPL. Author of many publications in the field, including several book chapters, Dr. Moreno has served as Associate Editor for IEEE Transactions of Geosciences and Remote Sensing (1994-2000), and has been a member of the European Space Agency Earth Sciences Advisory Committee (1998-2002), the Space Station Users Panel, and other international advisory committees.