15

Using Kernel Methods in a Learning Machine Approach for Multispectral Data Classification. An Application in Agriculture

Adrián González*, José Moreno†, Graham Russell‡and Astrid Márquez*
*Romulo Gallegos University, Venezuela
†Central University of Venezuela
‡University of Edinburgh
Scotland, UK

Abstract

Most pattern recognition applications within the Geoscience field involve the clustering and classification of remote sensed multispectral data, which basically aims to allocate the right class of ground category to a reflectance or radiance signal. Generally, the complexity of this problem is related to the incorporation of spatial characteristics that are complementary to the nonlinearities of land surface heterogeneity, remote sensing effects and multispectral features. The present chapter describes recent developments in the performance of a kernel method applied to the representation and classification of agricultural land use systems described by multispectral responses. In particular, we focus on the practical applicability of learning machine methods to the task of inducting a relationship between the spectral response of farms land cover to their informational typology from a representative set of instances. Such methodologies are not traditionally used in agricultural studies. Nevertheless, the list of references reviewed here show that its applications have emerged very fast and are leading to simple and theoretically robust classification models. This chapter will cover the following phases: a)learning from instances in agriculture; b)feature extraction of both multispectral and attributive data and; c) kernel supervised classification. The first provides the conceptual foundations and a historical perspective of the field. The second belongs to the unsupervised learning field, which mainly involves the appropriate description of input data in a lower dimensional space. The last is a method based on statistical learning theory, which has been successfully applied to supervised classification problems and to generate models described by implicit functions.

1. Introduction

A farming type or modality is a representation of a population of farms that share the same n dimensional traits. Typically, farming system studies seek to define separate groups of farms by looking for a natural structure among the observations. The objective is to maximize homogeneity within clusters and heterogeneity between them (Dixon et al., 2001; Hair et al., 1998). Information about properties of farming systems such as censuses and surveys have long been the most widely used instruments to gather data on agrarian activities; indeed, historically they have proved to be a useful means of gaining knowledge of such diverse agrarian features as: dominant patterns of farm activities and household livelihoods, including field crops, livestock, trees, aquaculture, grazing and forest areas, crop-livestock integration, technology, farm size and land tenure, to mention but a few. Nevertheless, the high requirements in terms of human and monetary resources of censuses and surveys prevent their application with the frequency and extent required to tackle the complexity of many agricultural issues.

The rapid development shown by land observation satellites over the last three decades has made a great deal of information about land surfaces available. This has widely been used to study land cover changes by the general model of pattern recognition process; which can be divided into a sequence of three main elements: a) generation of input random vectors with the information to be classified (sensor); b) translation of data into a statistically independent representation code that preservs their most relevant characteristics (feature extraction); and c) a system that, based on extracted features, can develop a function space where an operator might be built to serve as an answer predictor to any input generated by the sensor (classification). In this sense, within the field of pattern recognition, one of the most studied subjects is the idea of approximating relationships from the within-farm land surface processes and their emerging spectral response; using methods that can fit the complexity of these processes. This is vitally important for the study of crop-livestock production systems, given that these are critical to the livelihood of an important portion of the rural population at a worldwide level (Bouwman et al., 2005; Seré & Steinfeld, 1996). In addition, projections indicate that the demand for livestock food products is increasing globally (Delgado et al., 1999; Wint et al., 2000), and concern about the potential response of these systems is generally justified.

On this issue, a problem that remains open is the spatial monitoring of crop-livestock systems especially for those involving open range feeding, from which sometimes only time- and site-specific data can be approximated through field methods. These are usually not cost effective and suffer from poor spatial resolution. It is also true, in a broader context, that public availability of space-based remote sensing has helped with the monitoring of land surface biophysical properties. Some approaches have been concerned with the correction of observational data to create valued-added time series (Gleason et al., 2002; Green & Hay, 2002). Others in turn stress the use of optical, thermal and microwave data to model atmospheric and soil moisture (Dubayah, 1992; McVicar & Jupp, 2002); exploiting radiative transfer theory to estimate biophysical properties of vegetation (Goel, 1987; Myneni et al., 1992; Wylie et al., 2002); and macroscale modelling (Asrar & Dozier, 1994; Kimes et al., 2002). In summary, most methodologies monitor and map land surface processes by classification or detecting change (Song et al., 2001). Nevertheless, there is no evidence of using the gathered spectral data in recognising patterns associated with agricultural land management where an optimal discrimination of pixel mixture might be inferred beyond a training set. It has been in this context that the general aim of this chapter was defined to provide a unified framework and examples of using learning machines to accomplish the task of pattern recognition for complex mosaics of within-farm land use in crop-livestock systems from multi-spectral data.

These methodologies are based on feature induction from a training set by establishing a separating hyperplane between any two classes whose margin is maximum. Additionally, they include the inherent advantage of kernel functions, through which solutions are not built in the input space but into one with a higher dimensionality. In this feature space, it is possible that linear functions are enough to separate classes; given that input data are taken to this space by a nonlinear transformation whose diversity adds richness to the process of finding - if it exists - a solution. This flexibility is considered critical within the field of learning machines, to deal with complex task of using multispectral data for pattern recognition in crop-livestock systems.

2. Historical elements of statistical learning from instances in agriculture

2.1 General problem

The process of estimating an unknown input-output dependence and generalising it beyond a limited training set of observations is acknowledged as learning from instances, which had its origin in the pioneering work of Rosenblatt (1958). During the 1960's the application of this paradigm was seriously hampered as a result of the work of Minsky & Papert (1969). By this time it was thought that complex applications in the real world would require representational hypotheses much more expressive than linear functions, given that the target concept could not normally be represented as a simple linear combination of data attributes. As a result, some fields of study such as learning machine and pattern recognition were negatively affected, preventing their use on applied research including farming systems. It was subsequently demonstrated that the theories of Minsky & Papert were wrong.

Creating typologies of farming systems has been one of the major approaches within the field of agricultural systems in which research has been conducted. This paradigm mainly refers to those methods characterised by inductive non-supervised clustering of farms within a taxonomy; where farm likeness is represented according to a finite set of m-dimensional variables (Berdegue & Escobar, 1990; Köbrich et al., 2003; Kostrowicki, 1977). During the 70's most of the learning techniques used in the agricultural system field were influenced by the wave of learning linear decision surfaces (Capillon, 1985; Hart, 1990; Kostrowicki, 1977). That kind of representation was preferred given that its theoretical properties were well understood. After the 80s, researchers trying to move away from the limitations of linear models started using non-linear models in decision trees decision trees and artificial neural networks. These techniques were rapidly employed within the agriculture domain. However, the main problems of these approaches were their theoretical weakness and that their solution space had many local minima.

The consolidation and application of statistical learning theory during the mid-90's allowed the development of efficient algorithms to learn non-linear functions. These ideas completely recast the pioneering work of Rosenblatt (1958); and were theoretically supported in statistical learning theory (Vapnik, 1995, 1998; Vapnik & Chernovenkis, 1974). Vapnik and Chervonenkis formalised the learning problem as a function estimation; where given an empirical data set generated by a regular stochastic distribution, the algorithm pursues the extraction of regularities in the data by a general model of learning that might be summarised in a sequence of components: a) an input vector generator; b) a system that produces an output value and c) a linear machine.

Contrasting with the statistical learning theory, which appeared on the scene quite recently, another current solution implementation is based on kernel functions (Aronszajn, 1950; Mercer, 1909), which were first studied about a century ago, and which have been playing an important role in increasing the representational capacity of the solutions especially in agricultural applications involving remote sensing. Their use within the learning task relates closely to data pre-processing; and along with the learning machine, constitutes a compact body.

2.2 Particular cases

Supervised and unsupervised learning are among the most investigated applications in agriculture. The former approach pursues building relations between input vectors and target outputs. The outputs may be expressed at different scales: categorically or numerically, corresponding to classification and regression problems respectively. The unsupervised approach, rather than approximating input data to a target label, seeks to approximate data by similarity expressions, generally distance functions, from which groups of data that resemble each other can be built. This paradigm is usually referred to as clustering (Bishop, 2006).

The remote sensing works of Hermes et al. (1999) and Huang et al. (2002) are precursors of the classification approach in agriculture, where, given a spatially dispersed set of pixels, different forms of land cover (closed forest, open forest and woodland) are classified according to their spectral response. Other research of this kind includes the work of Keuchel et al. (2003) which progressively compares land cover classification using three methods (support vector machines, maximum likelihood and iterated conditional models); and the work of Su et al. (2007) which uses the multi-angle approach and its corresponding spectro-radiometer image to accurately map grassland types by support vector machines. A good application of learning machines on the regression problem is the work of Yang et al. (2007) within the forestry field. In that research, the target vector used was eddy covariance-based gross primary production (GPP) and three remotely sensed variables (land surface temperature, enhanced vegetation index and land cover) in order to predict flux-based GPP at a continental scale.

Regarding the clustering problem in the unsupervised ground, Diez et al. (2006) combined a kernel-based similarity function and a support vector machine to permit the identification of public beef product preferences stratified by market segment. In addition, within the unsupervised family can be found density estimators, which mainly project data from a high onto a lower dimensional space to determine its distribution in the input space in order to add visual richness to the solutions represented (Bishop, 2006).

In summary, these methodologies are based on feature induction from a representative set of instances, where it may be possible to produce a model able to generalise beyond the training instances. In this way a description of relationships present in the original data is possible, and their representation is simplified at the same time that their main features are preserved. Today, there is still a wide usage of linear paradigms in farming systems studies (Dobremez & Bousset, 1995; Köbrich et al., 2003; Milá et al., 2006)while extensive applications of linear machine techniques in agriculture are still scarce. The forerunners have shown that models generated are flexible, theoretically robust and provide expressive solutions. Some of the preliminary results of the present topic may be found in González et al. (2007). For those seeking a deep understanding in the machine learning field the following publications are suggested: Bishop (2006); Cristianini & Shawe-Taylor (2000); Shawe-Taylor & Cristianini (2006) and Vapnik (1995, 1998).

3. Feature extraction of both multispectral and attributive data

Feature extraction constitutes an important task within multidimensional crop-livestock pattern classification. The idea behind it is, among others, to isolate those statistical characteristics of the data that portray essential elements of them, and to provide a better understanding about the underlying processes that generate the data (Guyon & Elisseeff, 2003). Feature extraction is also very effective for avoiding the redundancy that characterises crop-livestock systems (crop production, land use, livestock production, management, etc) by finding meaningful projections, of even low dimensional input data, into a feature space. Principal components analysis (PCA) is one of the standard techniques to obtain features from input data (Jolliffe, 2002). This is achieved by maximising the projected variance onto mutually orthogonal eigenvectors along the directions of higher eigenvalues through iterative algorithms that minimise information losses. PCA basically performs a linear decomposition of input vectors, into a space whose coordinate system is hierarchically organised by data variability (Bishop, 2006).

Feature extraction through principal component analysis (also referred to as the *Karhunen-Loève* transform) can be traced back to the pioneering work of Pearson (1901) and Hotelling (1933a,b). Today PCA is one of the feature extraction methods most used in farming systems (Berdegue & Escobar, 1990; Köbrich et al., 2003), and there has been considerable research surrounding the application of this technique in different topics of pattern recognition (Bishop, 2006; Duda et al., 2001; Jolliffe, 2002). Basically, the method involves the finding of a lower dimensional space by the orthogonal transformation of the coordinate system where a given data set is described, with the aim of identifying directions of maximum variability. Let us consider a set of observations such that:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \tag{1}$$

where X is the original data set $m \times n$ matrix, n is the number of samples, which conform m-dimensional vectors ($\alpha = x_1 \dots x_m \in \mathbb{R}^N$) of random variables in an arbitrary space. These vectors are linearly decomposed into another coordinate system whose first axis is a projection of each observation and respond to the linear function $\alpha_1^T x$. This new $m = 1$-dimensional subspace is oriented to the direction where the elements of X show their highest variability.

The subsequent axes are orthogonally aligned in X to the next highest direction through recursive linear decompositions until m vectors have been aligned $\alpha_m^T x$. The axes of this new coordinate system are organized hierarchically according to data variability, and are normally referred to as principal components. It might happen that those components in directions of very low variability are practically constant for all vectors (Jolliffe, 2002), and can be eliminated since they do not contribute new information. Therefore, a substantial dimensionality reduction ($<< m$) of the problem is usually achieved, given that typically a few axes are enough to retain most of the data structure, if this exists.

Generally the feature extraction and dimensionality reduction proceeds as described above. However, it is worth pointing out the following observations: to obtain the new coordinate system data must be projected to the direction aligned with the maximum variance; this best fit axis passes through the mean of the data cloud which is given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x \tag{2}$$

In order to establish this direction, data is projected onto the $d = 1$-dimensional vector whose scalar value projection is defined by $\alpha_1^T x$ with a projected data variability such that:

$$\frac{1}{n} \sum_{i=1}^{n} \{\alpha_1^T x - \alpha_1^T \bar{x}\}^2 = \alpha_1^T S \alpha_1 \tag{3}$$

Variability maximisation is pursued in such a way that the sum of squares of element on α_1 equals 1 ($\alpha_1^T S\alpha_1 = 1$), where S is defined by:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_n - \bar{x})(x_n - \bar{x})^T \tag{4}$$

At this stage, the main task is the minimisation of redundancy present in the covariance and maximisation of useful information provided by the variance. Diagonal elements of the covariance matrix summarise the data dynamic of interest as long as they are high. Otherwise, they are associated with noise. Maximisation of $\alpha_1^T S\alpha_1$ is performed incorporating a Lagrange multiplier λ:

$$\alpha_1^T S\alpha_1 + \lambda_1(1 - \alpha_1^T\alpha_1) \tag{5}$$

whose derivative with respect to α_1 yields:

$$S\alpha_1 = \lambda_1\alpha_1 \tag{6}$$

Considering that the eigenvalues are ordered in a decreasing sequence ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$) being $\lambda' = \lambda_{max}$ and proceeding by mathematical induction, it is assumed that principal components from 1 to $m-1$ can be found along the first $m-1$ directions of eigenvectors. The principal component m_{th} is constrained to be orthogonal to such directions. In in the variance expression in this direction $\alpha_1 \cdots \alpha_{m-1} = 0$. So maximising S subject to this condition and being a unitary vector $|\alpha| = 1$, or $S\alpha = 1$

$$\alpha_1^T S\alpha_1 = \lambda_1 \tag{7}$$

Hence, the principal component m^{th} can be found along with the eigenvalue m^{th} and it can be established that the variance equals the eigenvalue m^{th} when α_1 is aligned to the direction of the m^{th} principal component (Bishop, 2006; Jolliffe, 2002).

In the literature correlation and covariance matrices can be presented as alternatives. To be completely accurate, the covariance matrix is the mean scalar product of patterns minus the mean, while the correlation matrix is a standardized version of the covariance matrix, given that the correlation originates from the mean scalar products of the patterns divided by the product of the standard deviation of patterns (Field, 2005). When this kind of analysis is performed from centred data ($\sum_{i=1}^{m} x_i = 0$) both matrices are equivalent.

Principal component analysis has been shown to be a very powerful technique for finding orthogonal derived variables that in succession maximise the variance of a given data set (Jolliffe, 2002; Mardia et al., 1979). However, sources of nonlinearities and complexities in real-world problems might require to be hypothetised in sub-spaces much richer than a linear combination of features (Cristianini & Shawe-Taylor, 2000). Therefore, nonlinear generalisations of principal components analysis play an important role in pattern analysis through the inclusion of kernel functions.

PCA has performed well in previous studies related to farming systems, especially for dimensionality reduction and for interpreting multiple crop-livestock signals (Köbrich et al., 2003). However, crop-livestock system variables interact in a non-linear dynamic, which in turn usually produces complex outcomes of landscape heterogeneity, livestock activity, and vegetation interactions. In consequence, most of these crop-livestock systems traits are subject to limited description within the second order correlation approach of linear PCA. One solution to this problem is the generalisation of linear PCA setting to an application of kernel principal component analysis (KPCA) (Schölkopf et al., 1998). This algorithm combines the simplicity of linear PCA with the capability of integral operators, known as kernel functions; to express data from input space as dot products in the feature space. This method enables the construction of nonlinear versions of the original variables in a high dimensional context (Shawe-Taylor & Cristianini, 2006).

3.1 Coping with non linearities

The kernel "trick" permits the generalising of any algorithm that uniquely depends on inner products (Aizerman et al., 1964). This approach has proven to be particularly helpful for those statistical problems that involve feature extraction (Schölkopf et al., 1998); classification (Boser et al., 1992); regression (Williams, 1998) and clustering (Crammer & Singer, 2002; Graepel & Obermayer, 1998). Generally it can be said that kernel methods serve to induct non-linear functions in feature spaces usually of high dimensionality, and also may be incorporated into the dual form of most algorithms in such a way that it is not necessary to calculate explicitly the transformation to the feature space (Shawe-Taylor & Cristianini, 2006).

A result of the inclusion of the kernel idea within the dual representation, is that the computation task is not affected by the feature space dimensionality (Cristianini & Shawe-Taylor, 2000), and given that the gram-matrix is the unique information used in the feature space, the amount of work required to calculate the inner product is not necessarily proportional to the feature number. Thus the use of kernels can be seen as a means to establish an implicit correspondence between the original data and the feature space, without the limitations associated with the computation of such correspondence.

Within a broad context, the study of statistical aspects of pattern analysis has been approached from two main paradigms: the Bayesian approach (Duda et al., 2001) and empirical processes (Vapnik, 1995). Boser et al. (1992) pioneered the merging of kernel methods and statistical learning theory (empirical processes approach) through large margin classifiers. However, most of the theoretical development on kernel methods has its origin in the research of Mercer (1909) and Aronszajn (1950) where fundamental issues of Mercer's theory and Hilbert's spaces were treated respectively. After the crisis of the main linear approaches commonly used in the learning machine field (Fisher, 1936; Rosenblatt, 1958) as a result of the publication of Minsky & Papert (1969), one of the alternatives proposed was the threshold multilayer structures, which led to the development of neural networks (generalised perceptron) with associated algorithm as back propagation (Hertz et al., 1991) .

The other approach was data pre-processing: in other words, the projection of data into a higher dimensional space to increase the computational power by including redundancies in their representation and assuring an effective feature extraction process from very complex data. An interesting alternative method to accomplish the above task, was the use of kernel methods, whose functions and corresponding feature spaces theory derive from integral operators studies (Aronszajn, 1950; Berg et al., 1984; Sahitoh, 1988). The inclusion of these constructs into a nonlinear generalisation of principal components analysis was led by Schölkopf et al. (1998). One of the main achievements of the study was to express the feature extraction based on eigen-decomposition, as a process that pursues the finding of orthonormalized directions in a kernel-defined feature space by dual representation, along which data variability is maximised.

Nonlinear PCA might be expressed as an eigenvalue problem. Consider a feature space $\mathcal{H}$ associated to the input space $\mathbb{R}^m$ by a non-linear transformation:

$$\Phi : X \Rightarrow \mathcal{H}, \qquad x \Rightarrow \Phi(x) \tag{8}$$

The feature space $\mathcal{H}$ can show an arbitrarily large dimensionality ($m \times m$), that is potentially infinite. Assuming that in this space data are centred according to $\sum_{i=1}^{m} x_i = 0$, the covariance matrix can be written in $\mathcal{H}$ as following:

$$Cov = \frac{1}{p} \sum_{i=1}^{p} \Phi(x^j)\Phi(x^j)^T \tag{9}$$

Having a feature space that possesses infinite dimensions, $\Phi(x^j)\Phi(x^j)^T$ can be considered the linear operator in $\mathcal{H}$ that performs the transformation $x \Rightarrow \left\langle \Phi(x^j)\Phi(x^j)^T \cdot x \right\rangle$. The main objective then consists of finding the solution to an eigenvalue problem that satisfies $\lambda v = Cov\, v$, without working explicitly in the feature space. By analogy to the input space analysis, all solutions v with $\lambda \neq 0$ are encountered in the sub-space generated by $\Phi(x^1), \ldots, \Phi(x^p)$. This includes two helpful implications:

1. The following equation can be used:

$$\lambda \langle \Phi(x^n) \cdot v \rangle = \langle \Phi(x^n) \cdot Cov\, v \rangle \qquad \forall n = 1, \ldots, p \tag{10}$$

2. Provided that $\lambda \geq 0$ are found subject to the existence of non null eigenvectors $v \in \mathcal{H} \setminus \{0\}$; and given that coefficients belonging to $\alpha_i (i = 1, \cdots, p)$ are determined by linear combinations of $\Phi(x^n)$, v can be written as:

$$v = \sum_{i=1}^{p} \alpha_i \Phi(x^i) \tag{11}$$

These expressions can be merged by substituting both into $\lambda v = Cov\, v$ and multiplying both sides by $\Phi(x)^T$ in order to express them as kernel terms $K(x^i, x^j) = \Phi(x^i)^T \Phi(x^j)$:

$$\lambda \sum_{i=1}^{p} \alpha_i \left\langle \Phi(x^n) \cdot \Phi(x^i) \right\rangle = \frac{1}{p} \sum_{i=1}^{p} \alpha_i \left\langle \Phi(x^n) \cdot \sum_{i=1}^{p} \Phi(x^j) \left\langle \Phi(x^j) \cdot \Phi(x^i) \right\rangle \right\rangle \qquad \forall\, n = 1, \ldots, p \tag{12}$$

which in terms of the matrix (Gram $p \times p$) notation, integrated by the elements $K_{ij} = \left\langle \Phi(x^i) \cdot \Phi(x^j) \right\rangle$, the equation for all n are consolidated in:

$$p\lambda K\alpha = K^2\alpha \tag{13}$$

where α represents the column vector integrated by elements $\alpha_1, \cdots, \alpha_p$. Finding solutions to the previous equation requires an eigenvalue problem to be solved:

$$p\lambda\alpha = K\alpha \qquad \forall \lambda \neq 0 \tag{14}$$

It can be demonstrated that this simplification (removing K from both sides) leads to (14) without those K that showed zero eigenvalues, not affecting the projection of principal components and bringing all useful solutions from (13). So if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the eigenvalues of K ($p\lambda$ solutions) and $\alpha^1, \cdots, \alpha^p$ the whole corresponding eigenvectors set, being λ_q the last non-zero eigenvalue (assuming that Φ is not identically 0). The condition of unitary norm ($\langle v^n \cdot v^n \rangle = 1$) for corresponding vectors in the feature space leads to the following solution of normalisation over $\alpha^1, \cdots, \alpha^q$ when (11) and (14) are used:

$$\begin{aligned} 1 &= \sum_{i,j=1}^{p} \alpha_i^n \alpha_j^n \left\langle \Phi(x^i) \cdot \Phi(x^j) \right\rangle = \sum_{i,j=1}^{p} \alpha_i^n \alpha_j^n K_{ij} \\ 1 &= \langle \alpha^n \cdot K\alpha^n \rangle = \lambda_n \langle \alpha^n \cdot \alpha^n \rangle \end{aligned} \tag{15}$$

The principal components projections can be calculated by projecting an x test point with an image $\Phi(x)$ onto eigenvectors v in the feature space with $n = 1, \cdots, q$; and expressing them in kernel notation using (11); that way principal components can be extracted:

$$\langle v^n \cdot \Phi(x) \rangle = \sum_{i=1}^{p} \alpha_i^n \left\langle \Phi(x^i) \cdot \Phi(x) \right\rangle \tag{16}$$

These are the non-linear principal components or features corresponding to Φ (Bishop, 2006; Schölkopf et al., 1998).

To illustrate the above descriptions, differences in performance between linear (LPCA) (Hotelling, 1933a,b; Pearson, 1901) and kernel principal component analysis (KPCA) (Schölkopf et al., 1998) will be depicted in the following lines, based on the effectiveness of extracted features to yield meaningful and compact farm groups (dependent variable) within unsupervised classification by hierarchical clustering procedures (Johnson, 1967; Ward, 1963), using as few principal components as possible. For the purpose of this illustration, meaningful groups were defined as those clusters whose means were significantly different from each other, showing strong similarities within groups and possessing high variability between groups. Such estimations were based on a discriminant analysis approach (Fisher, 1936) using the statistics of Wilks' lambda ($W\lambda$), Hotelling's test (T^2), Pillai's trace test (P); Roy's maximum root (RM); and average squared canonical correlation (r^2) using data from farming systems located in the central plains of Venezuela.

An example of a comparison between the best performing configuration of kernel methods and the linear approach whose feature extraction required six principal directions is presented in Table 1. The profiles of clustering performance after discriminant analysis for the Gaussian kernel show that means of farm classes of the selected variables were different in the population given the closeness of Wilks' lambda statistic to zero and comparatively higher values of the Pillai, Hotelling and Roy tests with respect to the linear and polynomic approaches. Also, classification based on Gaussian feature extraction, showed higher average squared canonical correlations (r^2) supporting the idea of well separated groups accounting for a high percentage (69%) of the total variance explained.

The percentage of farms classified correctly was slightly higher when feature extraction was performed using polynomic kernels compared to inserting linear and Gaussian kernels. However, this feature extraction method did not provide enough information to find directions in the feature space along which farm groups were as well separated as with the Gaussian kernel. Even so, its performance was much better than classification based on linearly extracted feature vectors.

Kernel	%*C*	$W\lambda$	*PT*	T^2	*RM*	r^2
Linear	88.3	0.15	1.21	3.36	2.30	0.60
Gaussian	90.3	0.09	1.38	3.50	2.43	0.69
Polynomic	91.5	0.11	1.31	3.83	1.94	0.65

%C: percentage classified correct; $W\lambda$: Wilks' lambda; *PT*: Pillai's trace; T^2: Hotelling's test; *RM*: Roy's minimum root; r^2: squared average canonical correlation

Table 1. Impact of kernel function on clustering performance using linear, Gaussian and polynomic approaches of feature extraction, after stepwise discriminant analysis for a group of farms in Venezuela.

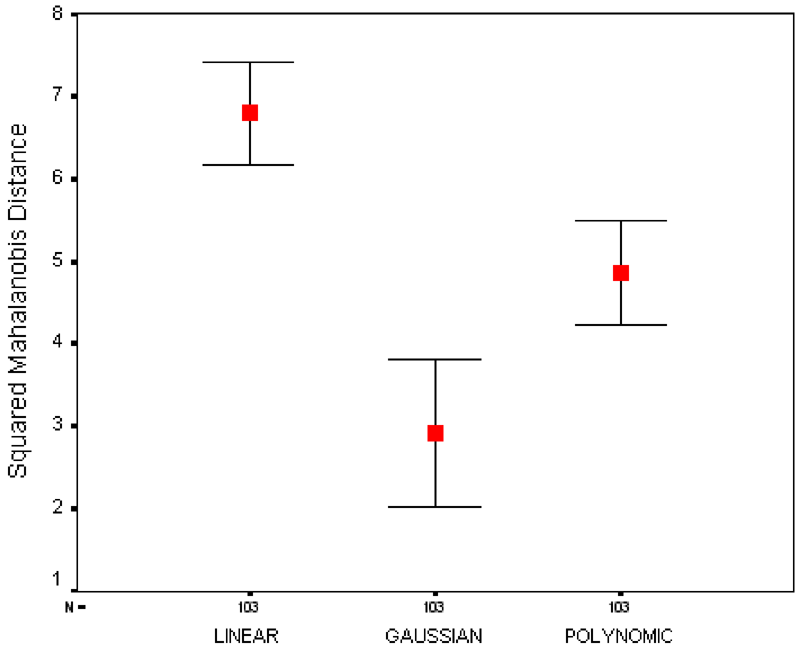
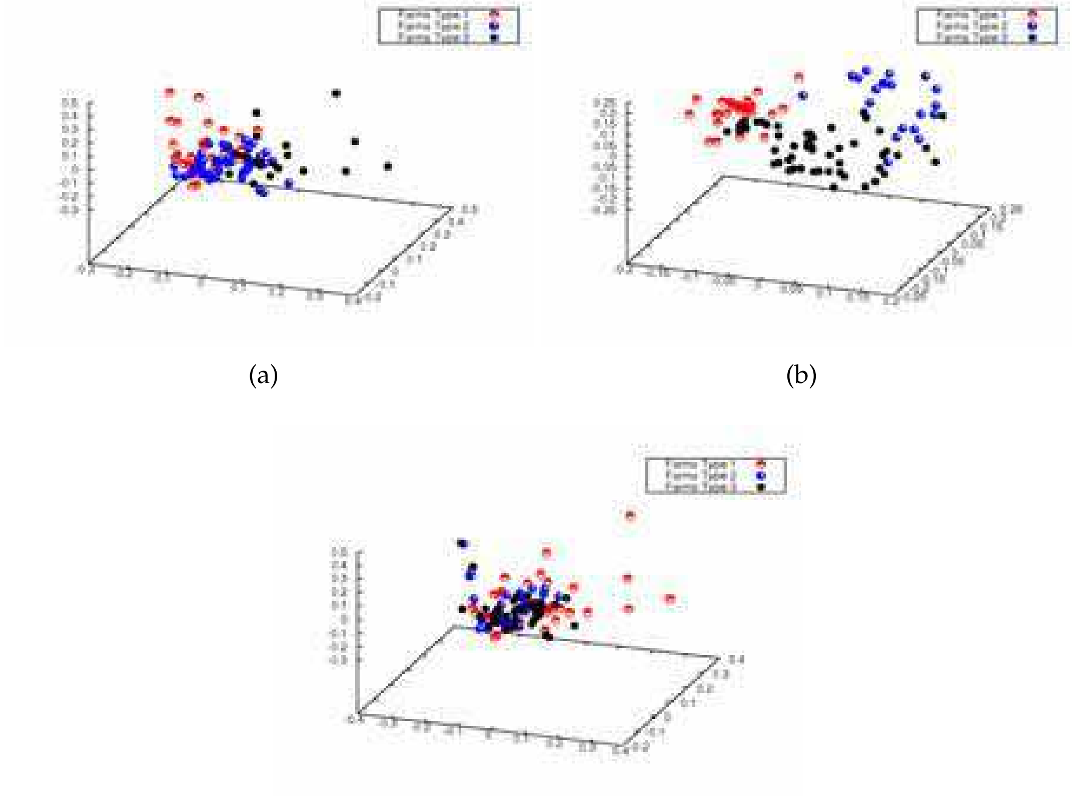


Fig. 1. Adjusted means and confidence intervals (95%) of squared Mahalanobis distance by selected feature extraction methods (linear, gaussian and polynomic) for a group of farms in Venezuela, after stepwise discriminant analysis.

Within canonical discriminant analysis, if a farm belongs to a particular class, it must fulfill some distance constraints with respect to the centroid of its class and projections of these groups onto some discriminant direction are expected to be compact and to show minimum overlaps. Hence, an easy way to assess the compactness of a given class is to look at the proximity of an observation set to its class-centroid. A visual approximation of these differences can be seen in Fig.1, where squared adjusted means of the Mahalanobis distance and their respective confidence intervals (95%) are shown for each feature extraction method. As can be observed, clusters segmented from feature vectors extracted by the linear approach and the Gaussian kernel were shown to be comparatively more scattered with respect to the clusters achieved from the polynomic feature extraction method, which showed a higher proximity (minimium distance) between a within-class object and its cluster centroid.

This effect is illustrated in Fig. 2, where farm objects are projected onto their first three principal directions with different levels of class overlap for the three feature extraction methods used. Only one of the three algorithms (Gaussian kernel) leads to a classification model that describes in a suitable way (without overlapping) the groups suggested by the instances cloud. The linear and polynomic-kernel methods were completely ineffective for cluster separation. This is mainly due to the topology of the sample covariance matrix as a result of the effect that the feature extraction method had on class-object component coordinates.



(a)

(b)

(c)

Fig. 2. Projection onto the three first principal components by farm class for linear (a), Gaussian (b), and polynomic kernel (c).),feature extraction approaches.

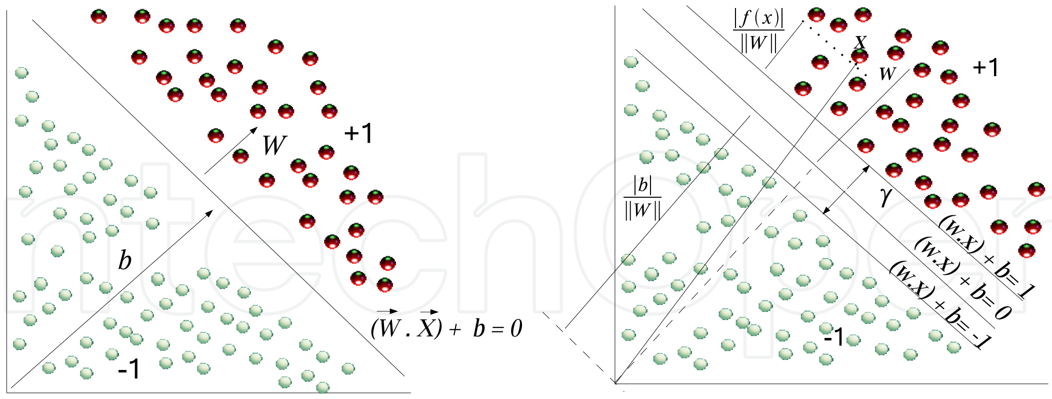
4. Kernel supervised classification of multispectral data

Once farm labels (informational classes) have been generated as illustrated in previous section, multispectral data can be used to perform supervised classification of farms' spectral responses. Traditionally, multispectral data, such as those from the Landsat series of satellites, have been used for mapping geology, geobotany, forestry, agriculture, soil and land cover. They have rarely been used to identify continuous pixel groups integrated in a class such as a farm, which is a mosaic of land covers. However, kernel methods coupled to a maxim-margin classifier can achieve the difficult task of discriminating farm types using their land cover spectral response as recorded in a satellite image as indicators. The resultant representation is flexible, uniform over the pattern presented, and preserves the topology of the input space.

4.1 Classification

Spatial land cover classification has been mainly approached through the following paradigms: maximum likelihood classifier (MLC) (Strahler, 1980); fuzzy clustering (Kosko & Isaka, 1993); and artificial neural networks (ANN) (Miller et al., 1995). However, farm classes are abstractions which are sometimes difficult to observe directly, and this leads to a number of limitations of these methods. For instance, MLC methods are not free from distribution assumptions, given their parametric premises. Fuzzy clustering represents the solutions in terms of probabilities, where both fuzzy rules and membership functions are subjected to the bias of the interpreter. The ANN method has theoretical weaknesses because of its black box character, preventing the proper repeatability of the results. The presence of local minima and of the time-consuming training process (referred to as lack of convergence) are also significant limitations.

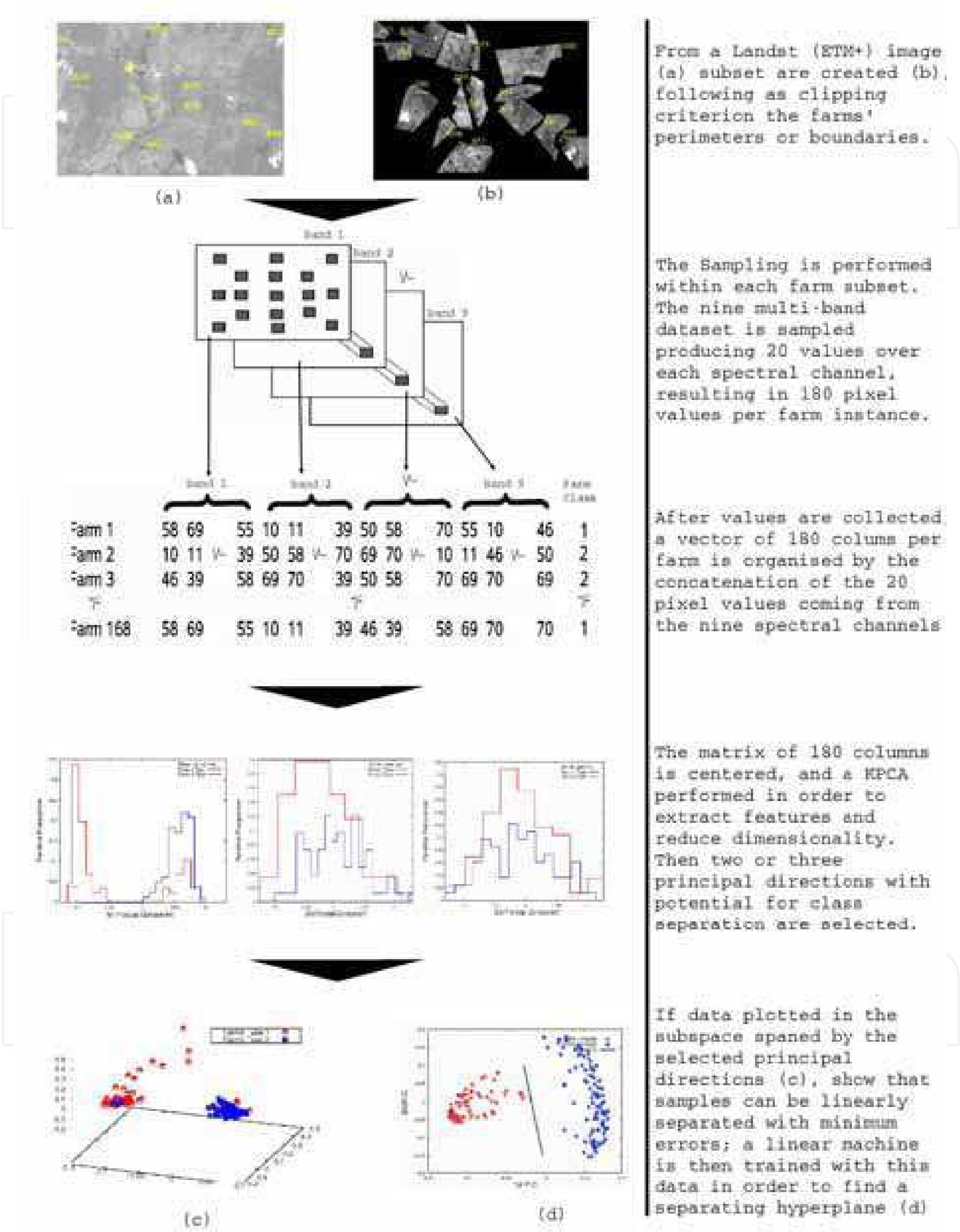
There are two main practical approaches to induce linear classifier parameters; on the one hand are those methods based on modelling conditional density functions (generative models) such as: linear discriminant analysis (Fisher, 1936; Lachenbruch, 1975) and Naive Bayes Classifier (Domingos & Pazzani, 1997); on the other hand, there are those that pursue the maximization of the outputs quality over a training set (discriminative models). These devices include: logistic regression (Hosmer & Lemeshow, 2000), the perceptron (Rosenblatt, 1958) and support vector machine (Vapnik, 1995; Vapnik & Chernovenkis, 1974). The main characteristic of support vector machines is that they seek to find a maximal margin hyperplane (Fig. 3). This is achieved using optimization procedures that can place severe computational demands. These problems were central to developing the kernel-adatron method which takes advantage of the adatron simplicity (Anlauf & Biehl, 1989), generalizing it to operate in a high dimensional feature space by the introduction of kernel functions. It solves the optimization problem of the Lagrangian formalism performing the margin-maximization through the application of a gradient ascent algorithm, resulting in an enhanced capability to learn nonlinear boundaries with a rate of convergence that is exponentially fast.



(a) (b)

Fig. 3. Toy example of decision boundaries between classes; modified from Cristianini & Shawe-Taylor (2000).

Practical applications of these approaches can be addressed through the "hybrid" algorithm known as the kernel-adatron, first proposed by Friest et al. (1998). This uses a classifier that is based on a linear decision function whose estimated output is given by $y = f(\vec{x} \cdot \vec{w}) =$



From a Landst (ETM+) image (a) subset are created (b), following as clipping criterion the farms' perimeters or boundaries.

The Sampling is performed within each farm subset. The nine multi-band dataset is sampled producing 20 values over each spectral channel, resulting in 180 pixel values per farm instance.

	Band 1	Band 2	…	Band 9	Farm Class
Farm 1	58 69 55	10 11 39	50 58 70	55 10 46	1
Farm 2	10 11 … 39	50 58 … 70	69 70 … 10	11 46 … 50	2
Farm 3	46 39 58	69 70 39	50 58 70	69 70 69	2
…		…			…
Farm 168	58 69 55	10 11 39	46 39 58	69 70 70	1

After values are collected a vector of 180 colums per farm is organised by the concatenation of the 20 pixel values coming from the nine spectral channels

The matrix of 180 columns is centered, and a KPCA performed in order to extract features and reduce dimensionality. Then two or three principal directions with potential for class separation are selected.

If data plotted in the subspace spaned by the selected principal directions (c), show that samples can be linearly separated with minimum errors; a linear machine is then trained with this data in order to find a separating hyperplane (d)

Fig. 4. Landsat image segmentation procedure.

$f(\sum_i x_i w_i)$, where $\vec{x}$ represents the input feature vector to the classifier; $\vec{w}$ is the vector of weights defining the separating boundary and f is a function that projects input values x on w. In this way input patterns are linearly separated by dividing the input space with a hyperplane (Fig. 3).

Fig. 4 depicts a Landsat image segmentation procedure used in a learning machine classification context. As can be seen, the nine multi-band raster dataset is sampled producing a collection of pixel values over each band, following an amplified von Neumann vicinity in a pre-selected area of interest within the farm's perimeter. This training data set was used as input to a dimension reduction procedure, using principal component analysis with kernel (KPCA).

Using kernels to learn potential nonlinear representation hypotheses based on the function of the form $f(x) = \sum_i^n \alpha_i y_i K(x_i, x) + b$, essentially involves the simulation of the nonlinear projection of the input data in a higher dimensional space (Schaback & Wendland, 2006):

$$\begin{aligned} \Phi : S \in \mathbb{R}^d &\rightarrow \mathcal{F} \in \mathcal{H} \\ x &\mapsto \Phi(x) \end{aligned} \tag{17}$$

where $\mathcal{F}$ denotes a feature space; and, $\mathcal{H}$ represents a dot product space, within which, a learning relationship could be induced between a pattern $\Phi(x)$ and a label y. In this way, having as theoretical context Mercer's theorem (Aizerman et al., 1964; Mercer, 1909); (18) represents the kernel matrix, where each entry is a measure of similarity between two objects. Thus, a symmetric function $K(x_i, x)$ was a kernel if it fulfilled Mercer's condition, i.e. the function K is (semi) positive definite. When this is the case there exists a mapping ϕ such that it is possible to write $K(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(x_i) \cdot \phi(x) \rangle$.

$$K(x_i, x) \triangleq \langle \phi(x_i) \cdot \phi(x) \rangle \Rightarrow \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \dots \\ K(x_2, x_1) & \ddots & \\ \vdots & & \end{bmatrix} \tag{18}$$

The kernel represents a dot product on a feature space $\mathcal{F}$ into which the original vectors were mapped (Fig. 5). In this way a kernel function defines an embedding of memory patterns into (high or infinite dimensional) feature vectors and allows the algorithm to be carried out in this space without the need to represent it explicitly (Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002). Further details on the way this procedure was implemented are outside the scope of this paper. Nevertheless, for those seeking deeper understanding of the ideas behind kernel-based learning theory there are fuller descriptions in Aizerman et al. (1964); Aronszajn (1950); Mercer (1909) and Schölkopf & Smola (2002). Applications of kernel methods and learning machines may also be reviewed in García & Moreno (2004a,b,c)

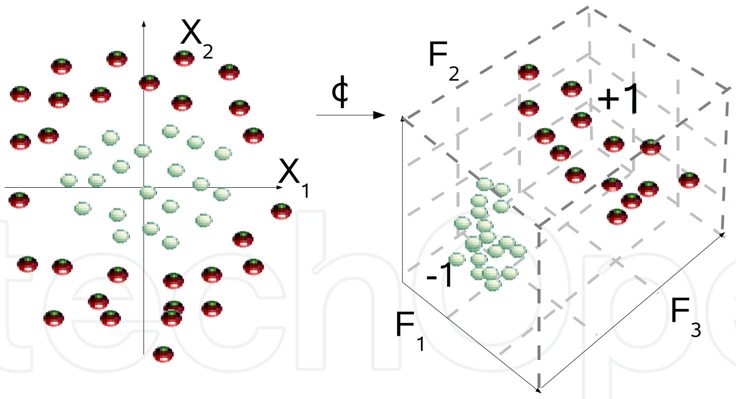
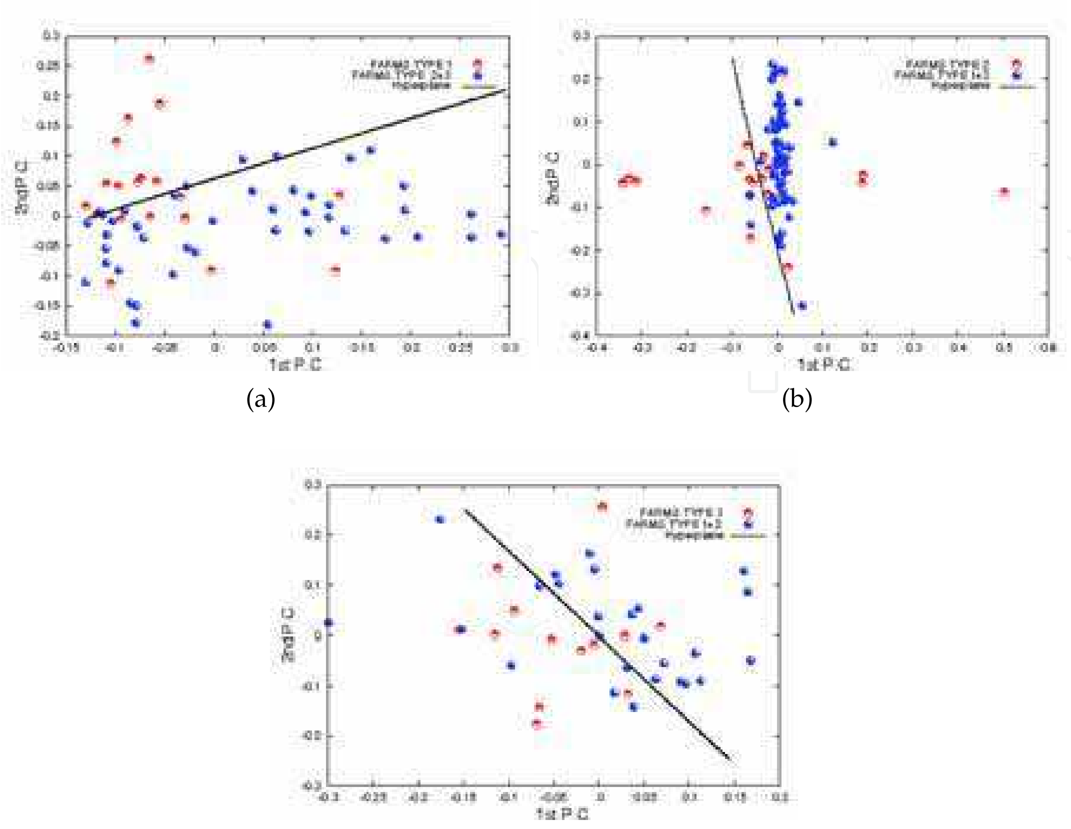


Fig. 5. Toy example illustration of the effect of mapping a simple binary problem to a higher dimensional feature space on the ability to separate complex relations.

The basic KA algorithm is a binary classifier that makes use of an optimization procedure based on the descent gradient to find the maxim-margin hyperplane that separates two groups. For the classification of farms from a multi-class problem (existence of three or more informational categories), a one against the rest strategy might be adopted. Basically, three machines (one per each class) can be trained organized in such an assembly that the class of interest is compared against the other two (Fig. 6).
Table 2 presents the performance accuracy of the three KA machines trained for an experimental group. As can be seen the KA appears to be more sensitive for class 1, given the highest accuracy reached, and its degree of overlap seems to be with class 3. This may be explained by the levels of farming intensification observed in farm class 1, with an important degree of fragmentation of the land cover mosaic, which probably facilitated its differentiation from those instances that resemble the more natural scenes typical of the less intensive farm classes 2 and 3 (Drury, 2001). The tendency to wrongly allocate farm type 3 as class 1 might be because these groups of farms share similar attributes in their proportions of pasture, forage and forest cover. Misclassification between classes 2 and 3 can be explained by the lack of anthropogenic changes leading to the occupation of less discrete areas of the feature space as a function of the natural environment context (Richards and Jia, 2006; Landgrebe, 2007).
In this kind of study, farms are seen as bags of pixels representing different land covers in a space where each dimension is associated with a spectral channel. Because this vector space was sensibly transformed by non linear feature extraction to improve representation, and with this to ensure equivalent land covers mapped to similar feature vectors, it is possible to reach an acceptable level of accuracy with this approach.



(a)

(b)

(c)

Fig. 6. Separating hyperplanes for farms class 1 (a) and 3 (c) using a Gaussian kernel ($\sigma = 200$), and class 2 (b) using a polynomial kernel (order= 3; $\sigma = 4$).

	KA	*Predicted*				
		Class 1	*Class* 2	*Class* 3	Σ	Accuracy (%)
Actual	*Class* 1	35	0	4	39	89.8
	Class 2	2	34	6	42	80.95
	Class 3	3	2	17	22	77.27
	Σ	40	36	27	103	
	Accuracy (%)	87.5	94.44	62.96		*Overall* *Accuracy*(%) 83.49

KA: Kernel Adatron

Table 2. Confusion matrix for the segmentation of three farm categories trained on 14, 20, and 16 cases for class 1, 2, and 3 respectively using the KA machine.

Finding these separating decision functions on the segmentation of farm classes is particularly significant given the non-stationary spatial behaviour of the spectral response of this kind of object; and because of the small training set size with respect to the dimensionality of the input space. Another important consideration is that this approach only focuses on extreme samples for its training, making possible the derivation of comparable levels of performance at a lower cost. This fact would confirm the argued advantages of previous applications of kernel methods in the land use domain, in which decision functions have been induced without any other *a priori* knowledge about the land cover than labels (Huang et al., 2002; Zhu & Blumberg, 2002). This implies a considerable resource saving in practical application to livestock systems monitoring.

4.2 The multispectral data

The use of multispectral data to distinguish one type of land cover from another, has been an effective way of linking anthropomorphic intervention to a physical environment, particularly within the agricultural sector (Campbell, 2002). For instance, Wylie et al. (2002) combined optical and thermal data to estimate biophysical properties of vegetation. Other approaches use the land cover mosaic, to induct farm typologies based on their relative spectral similarities, as in the case of Duvernoy (2000). The popularity of using visible and near infrared (VIR) imagery on the classification of areas covered by agricultural activities, is because plant cell structures, morphology, chlorophyll and other pigments have a marked effect on this wavelength range (Drury, 2001), and on the temperature brightness of thermal infrared (TIR) radiation incident on living plants (Rees, 2007).

The configuration of multispectral sensors, such as Landsat 7 Enhanced Thematic Mapper Plus (ETM+), is particularly well suited to perceive the energy field, in the form of VIR and TIR radiation emanating from vegetation covers (Richards & Jia, 2006). This feature makes many multispectral data sensible to spatial patterns tied to crop calendars, and vegetative growth-lessening as a result of phenophases (Campbell, 2002; Richards & Jia, 2006). For instance, the spectral bands per pixel in Landsat 7 are delineated by six VIR bands, where band 6 is split into two channels defined by filters that control the radiance that reach the sensor; and a panchromatic band (Barsi et al., 2003; Heckenlaible et al., 2007). These radiometric features make Landsat 7 a good choice within the context of farming system research at household resolution level. The precision to which this sensor registers the radiation power, for a particular pixel in a given wavelength is 8 bits (256 levels) (Richards & Jia, 2006). This feature enhances the ability of the sensor to distinguish the spectral responses from different materials, when human-scale factors such as agriculture need to be addressed (Campbell, 2002; Landgrebe, 2007).

As with radiometric resolution, the spatial resolution of Landsat 7, which ranges from 15 to 60 meters per pixel across all the spectral bands, is rich (small or fine) compared to farms, which are the usual objects of study in farming systems research and where a pixel smaller than the agricultural field to be studied is usually preferred (Landgrebe, 2007). To these spatial characteristics of Landsat, should be added its scanning features, whose cover swath is 185 km^2, which means that each scene sample observes an area of 34.225 km^2. Such an overlay represents an advantage for farming system research given the scale of the typical study area (10.000 km^2), and because the whole can be extracted from one image. However depending on the size of the farms under study misclassification risk might occur, from the impact that spatial resolution has on the separability of informational classes (Landgrebe, 2007). Spatial resolution has been shown to have a significant influence on spectral class separability because

of the hierarchy that generally characterizes informational categories (Campbell, 2002; Rees, 2007); and there is reason to believe that similar effects occur with collections of land cover such as farms (Landgrebe, 2007). For studies of farming systems, the spatial resolution of Landsat 7 (ETM+) data, might be too fine for the purposes of classification, in the sense that sometimes it is desirable to have pixel sizes smaller, but not excessively smaller, than the field under study, because too fine a resolution may lead to pixels that spectrally do not represent the field of interest but only part of it. In farm classification, most of the time interest is focussed on pixels that integrate across what is desired to be called a field, which in this study would be a farm, rather than a small part of a particular cover of crop, grassland or forest.
From that viewpoint, an alternative possibility is to use a source of data with a coarse spatial resolution, such as the Moderate Resolution Image Spectrometer (MODIS) (NASA, 2008). This sensor is one of the principal instruments aboard EOS[1] AM-1 (TERRA); and its spatial resolution ranges from 500m to 1 km, with a viewing swath width of 2.330 km. The possibility that the use of this sensor would lead to an improvement in farm classification accuracy, is in line with the reviews of Landgrebe (2007) and Drury (2001) in the sense that compared with Landsat 7, each pixel in MODIS would be made up of a mixture of "Landsat-size" pixels on ccategories such as grass and crops that may lead to an improved representation of a farm as a field of interest. The advantages of MODIS are not restricted to its spatial resolution; its spectral resolution, 36 channels covering from visible to thermal infrared spectral regions, also presents some benefit compared to the 7 bands of Landsat. This spectral richness wshould increase the accuracy of discrimination of complex classes, because of the high volume space. Evidence for the significance of spectral resolution on discrimination accuracy comes from Bazi & Melgani (2006); Foody & Mathur (2004); Melgani & Bruzzone (2004) and Muñoz-Marí et al. (2007). They exploit high spectral resolution sources, spreading the data out as much as possible in the feature space to make the most of the spectral richness, that generally results in small classification errors.

5. References

Aizerman, M., Braverman, E. & Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning, *Automations and Remote Control* **25**: 821–837.

Anlauf, J. & Biehl, M. (1989). The adatron: an adaptative perceptron algorithm, *Europhysics Letters* **10**: 687–692.

Aronszajn, N. (1950). Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68**: 337–404.

Asrar, G. & Dozier, J. (1994). *EOS-science strategy for the Earth observing system*, AIP Press, Woodbury, NY.

Barsi, J., Schott, J., Palluconi, F., Helder, D., Hook, S., Markham, B., Chander, G. & O'Donnell, E. (2003). Landsat tm and etm+ thermalband calibration, *Canadian Journal of Remote Sensing* **29**(2): 141–153.

Bazi, Y. & Melgani, F. (2006). Toward an optimal SVM classification system for hyperspectral remote sensing images, *IEEE Transaction on Geoscience and Remote Sensing* **44**(11): 3374–3385.

Berdegue, J. & Escobar, G. (1990). Conceptos y metodologias para la tipificación de sistemas de finca, *Tipificación de Sistemas de Producción Agrrícola*, RIMISP, Santiago de Chile, pp. 13–43.

[1] Earth Observation System

Berg, C., Christensen, J. & Ressel, P. (1984). *Harmonic analysis on semigroups*, Springer-Verlag, New York, USA.

Bishop, C. (2006). *Pattern recognition and machine learning*, Springer, Singapure.

Boser, B., Guyon, M. & Vapnik, V. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* pp. 144–152.

Bouwman, A., der Hoek, K. V., Eickhout, B. & Soenario, I. (2005). Exploring changes in world ruminant production sytems, *Agricultural Systems* **84**(2): 121–153.

Campbell, J. (2002). *Introduction to remote sensing*, Taylor & Francis, London, UK.

Capillon, A. (1985). Connaitre la diversité des exploitations: un préalable á la recherche de réferences techniques regionales, *Agriscope* **6**: 31–40.

Crammer, K. & Singer, Y. (2002). Pranking with ranking, *in* T. Dietterich, S. Becker & Z. Ghahramani (eds), *Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, pp. 641–647.

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, UK.

Delgado, C., Rosengrant, M., Steinfeld, H., Ehui, S. & Courbois, V. (1999). Livestock to 2020. the next food revolution, *Food, Agriculture and the Environment Discussion Paper 28*, International Food Policy Research Institute, Food and Agriculture Organization of the United Nations, International Livestock Research Institute, Washington, DC.

Diez, J., Coz, J., Bahamonde, A., nudo, C. S., Olleta, J., Macie, S., Campo, M., Panea, B. & Alberti, P. (2006). Identifying market segments in beef: breed, slaughter weight and ageing time implications, *Meat Science* **74**(4): 667–675.

Dixon, J., Gulliver, A. & Gibbon, D. (2001). *Farming systems and poverty. Improving farmer's livelihoods in a changing world*, FAO and World Bank, Malcolm Hall, chapter 1st.

Dobremez, L. & Bousset, J. (1995). *Rendre compte de la diversité des exploitations agricoles. Une démarche d'analyse par exploration conjointe de sources statistiques, comptables et technico-économiques*, Cemagref, France.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple bayesan classifier under zero-one loss, *Machine Learning* **29**: 103–137.

Drury, S. (2001). *Image Interpretation in Geology*, 3rd edn, Nelson Thornes Ltd, Cheltenham, UK.

Dubayah, R. (1992). Estimating net solar radiation using landsat thematic mapper and digital elevation data, *Water Resourses Research* **28**: 2469–2484.

Duda, R., Hart, P. & Stork, D. (2001). *Pattern classification*, 2nd. edn, John Wiley & Sons, INC, New York, USA.

Duvernoy, I. (2000). Use of land cover model to identify farm types in the Misiones agrarian frontier (Argentina), *Agricultural Systems* (64): 137–149.

Field, A. (2005). *Discovering statistics using SPSS*, 2nd edn, SAGE Publications, London, UK.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**: 179–188.

Foody, M. & Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, *Remote Sensing of Environment* **93**: 107–117.

Friest, T., Campbell, C. & Cristianini, N. (1998). *The kernel-adatron: A fast and simple learning procedure for support vector machines*, In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan-Kaufmann, San Francisco, USA.

García, C. & Moreno, J. (2004a). *The hopfield associative memory network: Improving performance with the kernel "trick"*, In: C. Lemaître; C.A. Reyes and J.A. Gonzalez (Eds). IBERAMIA 2004, LNAI 3315, Springer-Verlag, Berlin, Germany, pp. 871–880.

García, C. & Moreno, J. (2004b). *Kernel based method for segmentation and modeling of magnetic resonance images*, In: C. Lemaître; C.A. Reyes and J.A. Gonzalez (Eds). IBERAMIA 2004, LNAI 3315, Springer-Verlag, Berlin, Germany, pp. 636–645.

García, C. & Moreno, J. (2004c). *The kernel hopfield memory network*, In: P.M.A. Sloot; B. Chopard and A.G. Hoekstra (Eds). ACRI 2004, LNCS 3305, Springer-Verlag, Berlin, Germany, pp. 755–764.

Gleason, A., Prince, S., Goetz, S. & Small, J. (2002). Effects of orbital drift on land surface temperature measured by avhrr thermal sensors, *Remote Sensing of Environment* **79**: 147–165.

Goel, N. (1987). Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data, *Remote Sensing Review* **3**: 1–212.

González, A., Russell, G., Márquez, A., Moreno, J., García, C., Domínguez, C., Colmenares, O. & Machado, J. (2007). Supervised farm classification from remote sensing images based on the kernel adatron algorithm, *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS07)*, Barcelona, Spain.

Graepel, T. & Obermayer, K. (1998). Fuzzy topographic kernek clustering, *in* W. Bauer (ed.), *Proceeding of the 5th GI Workshop Fuzzy Neuro Systems '98*, pp. 90–97.

Green, R. & Hay, S. (2002). The potential of pathfinder avhrr data for providing surrogate climatic variables across africa and europe for epidemiological applications, *Remote Sensing of Environment* **79**: 166–175.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**: 1157–1182.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1998). *Multivariate data analysis*, 5th edn, Prentice Hall, Upper Saddle River, New Jersey.

Hart, R. (1990). Componentes, subsistemas y propiedades del sistema finca como base para un método de clasificación, *Tipificación de Sistemas de Producción Agrrícola*, RIMISP, Santiago de Chile, pp. 45–61.

Heckenlaible, D., Meyerink, A., Torbert, C. & Lacasse, J. (2007). Landsat 7 (L7) enhanced thematic mapper plus(ETM+ level zero-r distribution product (LORP) data format control book (DFCB), *Technical report*, Department of the Interior U.S. Geological Survey, Sioux Falls, South Dakota.

Hermes, L., Frieauff, J., Puzicha, J. & Buhmann, J. (1999). Support vector machines for land usage classification in landsat tm imagery, *Geoscience and Remote Sensing Symposium*, Vol. 1 of *IGARSS 99*, IEEE international, pp. 348–350.

Hertz, J., Krogh, A. & Palmer, R. (1991). *Introduction to the theory of neural computation*, Addison-Wesley.

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn, Wiley, New York.

Hotelling, H. (1933a). Analysis of a complex of statistical variables into principal components, *The Journal of Educational Psychology* **24**(6): 417–441.

Hotelling, H. (1933b). Analysis of a complex of statistical variables into principal components, *The Journal of Educational Psychology* **24**(7): 498–520.

Huang, C., Davis, L. & Townshend, J. (2002). An assesment of support vector machines for land cover classification, *International Journal of Remote Sensing* **23**: 725–749.

Johnson, S. (1967). Hierarchical clustering schemes, *Psychometrika* **32**(3): 241–254.

Jolliffe, I. (2002). *Principal components analysis*, Springer, New York, USA.

Keuchel, J., Naumann, S., Heiler, M. & Siegmund, A. (2003). automatic land cover analysis for tenerife by supervised classification using remotely sensed data, *Remote Sensing of Environment* **86**(4): 530–541.

Kimes, D., Gastellu-Etchegorry, J. & Esteve, P. (2002). Recovery of forest canopy characteristics through inversion of a complex 3d model, *Remote Sensing of Environment* **79**: 320–328.

Köbrich, C., Rehman, T. & Khan, M. (2003). Typification of farming systems for constructing representative farm models: two illustrations of the application of multi-variate analyses in chile and pakistan, *Agricultural Systems* **76**(1): 141–157.

Kosko, B. & Isaka, S. (1993). Fuzzy logic, *Scientific American* **271**: 76–81.

Kostrowicki, J. (1977). Agricultural typology concept and method, *Agricultural System* **2**: 33–45.

Lachenbruch, P. (1975). *Discriminant analysis*, Hafner Press, New York, USA.

Landgrebe, D. (2007). Multispectral thematic mapping of land areas, some fundamentals, *IEEE Geosciences and Remote Sensing Society Newsletter* (145): 11–15.

Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate analysis*, Academic Press, London, UK.

McVicar, T. & Jupp, D. (2002). Using covariates to spatially interpolate moisture availability in the murray-darling basing. a novel use of remotely sensed data, *Remote Sensing of Environment* **79**: 199–212.

Melgani, F. & Bruzzone, L. (2004). Classification of hyperpectral remote sensing images with support vector machines, *IEEE Transactions on Geosciences and Remote Sensing* **42**(8): 1778–1790.

Mercer, J. (1909). *Functions of positive and negative type and their connection with the theory of integral equations*, Philosophical Transactions of the Royal Socciety of London, London.

Milá, M., Bartolomé, J., Quintanilla, R., García-Cachán, M., Espejo, M., Herráiz, P., Sánchez-Recio, J. & Piedrafita, J. (2006). Structural characterisation and typology of beef cattle farms of spanish wooded rangelands (dehesas), *Livestock Science* **99**: 197–209.

Miller, D., Kaminsky, E. & Rana, S. (1995). Neural network classification of remote-sensing data, *Computers and Geosciences* **21**: 377–386.

Minsky, M. & Papert, S. (1969). *Perceptrons: an introduction to computational geometry*, MIT Press.

Muñoz-Marí, J., Bruzzone, L. & Camps-Valls, G. (2007). A support vector domain description approach to supervised classification of remote sensing images, *IEEE Transaction on Geosciences and Remote Sensing* **45**(8): 2683–2692.

Myneni, R., Asrar, G. & Hall, F. (1992). A three-dimensional radiative transfer method for optical remote sensing of vegetated land surface, *Remote Sensing of Environment* **41**: 105–121.

NASA (2008). Moderate Resolution Image Spectrometer (modis), *Online. Retrieved April 25, 2008 from http://modis.gsfc.nasa.gov/*.

Pearson, K. (1901). On lines and planes of closets fit to points in space, *Phylosophical Magazine* **2**: 559–572.

Rees, W. (2007). *Physical principles of remote sensing*, Cambridge University Press, Cambridge, UK.

Richards, J. & Jia, X. (2006). *Remote sensing digital image analysis*, Springer-Verlag, Berlin, Germany.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review* **65**(6): 386–408.

Sahitoh, S. (1988). *Theory of reproducing kernels and its applications*, Logman Scientific & Technical, Harlow, UK.

Schaback, R. & Wendland, H. (2006). *Kernel techniques: from machine learning to meshless methods*, Acta numerica (2006), Cambridge University Press.

Schölkopf, B. & Smola, A. (2002). *Learning with kernels. Support vector machines, regularization, optimization, and beyond*, The MIT Press, Cambridge, Massachusetts, USA.

Schölkopf, B., Smola, A. & Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation.* **10**: 1299–1319.

Seré, C. & Steinfeld, H. (1996). *World livestock production systems: Current status, issues and trends*, FAO Animal Production And Health Paper, Rome, Italy.

Shawe-Taylor, J. & Cristianini, N. (2006). *Kernel methods for pattern analysis*, Cambridge Univ. Press, Cambridge, UK.

Song, C., Woodcock, C., Seto, K., Lenney, M. & Macomber, S. (2001). Classification and change detection using landsat tm data: When and how to correct atmospheric effects?, *Remote Sensing of Environment* **75**: 230–244.

Strahler, A. (1980). The use of prior probabilities in maximun likelihood classification of remotely sensed data, *Remote Sensing of Environment* **10**: 135–163.

Su, L., Chopping, M., Rango, A., Martonchik, J. & Peters, D. (2007). Support vector machines for recognition of semi-arid vegetation types using misr multi-angle imagery, *Remote Sensing of Environment* **107**(1-2): 299–311.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer-Verlag, New York, USA.

Vapnik, V. (1998). *Statistical learning theory*, John Wiley & Sons.

Vapnik, V. & Chernovenkis, A. (1974). *Theory of pattern recognition*, Nauka, Moscow.

Ward, J. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**(301): 236–244.

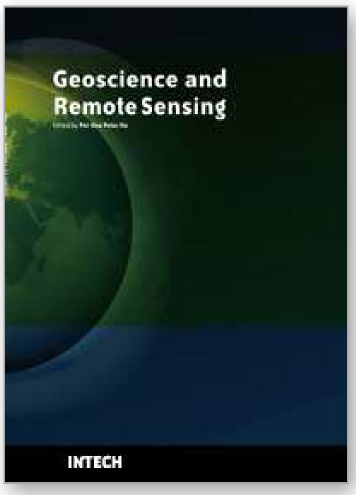
Williams, C. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond, *in* M. Jordan (ed.), *Learning and inference in graphical models*, Kluwer.

Wint, W., Slingenbergh, J. & Rogers, D. (2000). Livestock distribution, production and diseases, *Towards a global livetock atlas*, Food and Agriculture Organisation of the United Nations, Rome. www.fao.org/ag/againfo/resources/en/glipha/default.html.

Wylie, B., Meyer, D., Tieszen, L. & Mannel, S. (2002). satellite mapping of surface biophysical parameters at the biome scale over the north american grasslands, *Remote Sensing of Environment* **79**: 266–278.

Yang, F., Ichii, K., White, M., Hashimoto, H., Michaelis, A., Votava, P., Zhu, A., Huete, A., Running, S. & Nemani, R. (2007). Developing a continental-scale measure of gross primary production by combining modis and americaflux data through support vector machine approach, *Remote Sensing of Environment* **110**(1): 109–122.

Zhu, G. & Blumberg, D. (2002). Classification using ASTER data and SVM algorithms; the case study of Beer Sheva, Israel, *Remote Sensing of Environment* **80**: 233–240.



Geoscience and Remote Sensing
Edited by Pei-Gee Peter Ho

ISBN 978-953-307-003-2
Hard cover, 598 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

Remote Sensing is collecting and interpreting information on targets without being in physical contact with the objects. Aircraft, satellites ...etc are the major platforms for remote sensing observations. Unlike electrical, magnetic and gravity surveys that measure force fields, remote sensing technology is commonly referred to methods that employ electromagnetic energy as radio waves, light and heat as the means of detecting and measuring target characteristics. Geoscience is a study of nature world from the core of the earth, to the depths of oceans and to the outer space. This branch of study can help mitigate volcanic eruptions, floods, landslides ... etc terrible human life disaster and help develop ground water, mineral ores, fossil fuels and construction materials. Also, it studies physical, chemical reactions to understand the distribution of the nature resources. Therefore, the geoscience encompass earth, atmospheric, oceanography, pedology, petrology, mineralogy, hydrology and geology. This book covers latest and futuristic developments in remote sensing novel theory and applications by numerous scholars, researchers and experts. It is organized into 26 excellent chapters which include optical and infrared modeling, microwave scattering propagation, forests and vegetation, soils, ocean temperature, geographic information , object classification, data mining, image processing, passive optical sensor, multispectral and hyperspectral sensing, lidar, radiometer instruments, calibration, active microwave and SAR processing. Last but not the least, this book presented chapters that highlight frontier works in remote sensing information processing. I am very pleased to have leaders in the field to prepare and contribute their most current research and development work. Although no attempt is made to cover every topic in remote sensing and geoscience, these entire 26 remote sensing technology chapters shall give readers a good insight. All topics listed are equal important and significant.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Adrian Gonzalez, Jose Moreno, Graham Russell and Astrid Marquez (2009). Using Kernel Methods in a Learning Machine Approach for Multispectral Data Classification. An Application in Agriculture, Geoscience and Remote Sensing, Pei-Gee Peter Ho (Ed.), ISBN: 978-953-307-003-2, InTech, Available from: http://www.intechopen.com/books/geoscience-and-remote-sensing/using-kernel-methods-in-a-learning-machine-approach-for-multispectral-data-classification-an-applica

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai

www.intechopen.com

Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821