

# Support Vector Machines for Hyperspectral Image Classification with Spectral-based kernels

Grégoire MERCIER\* and Marc LENNON\*†

\*GET - ENST Bretagne, dpt ITI, CNRS FRE 2658 TAMCIC, team TIME.

Technopole Brest-Iroise, CS 83818; 29238 Brest cedex - France

e-mail: gregoire.mercier@enst-bretagne.fr

†AvelMor, Technopole Brest-Iroise, Place Nicolas Copernic, 29280 Plouzane - France

**Abstract**—Support Vector Machines (SVM) have been recently used with success for the classification of hyperspectral images. This method appears to be a robust alternative for pattern recognition with hyperspectral data: since the method is based on a geometric point of view, no statistical estimation has to be achieved. Then, SVM outperforms classical supervised classification algorithms such as the maximum likelihood when the number of spectral bands increases or when the number of training samples remains limited.

Nevertheless, those kernel-based methods do not take into consideration the bio-physical meaning of the spectral signature. In fact, kernels are functions that are based on the quadratic distance between support vectors. Then, some modified kernels are presented to take into consideration the spectral similarity between support vectors to outperform SVM-based classification of hyperspectral data cube. Those kernels (that still suit Mercer's conditions) are based on the use of spectral angle to evaluate the distance between support vectors.

Classifiers to compare have been applied to an image from the CASI sensor including 17 bands from 450 to 950nm representing an intensive agricultural region (Brittany, France). It appears that those kernels reduces false alarms that were induced by illumination effects with classical kernels.

## I. INTRODUCTION

Support Vector Machines (SVM) [1] have been recently introduced in the statistical learning theory domain [2] for regression and classification problems, and applied to the classification of multispectral [3] and hyperspectral [4], [5], [6] images.

The technique consists in finding the optimal separation surface between classes thanks to the identification of the most representative training samples of the side of the class. These samples are called *support vectors*. If the training data set is not linearly separable, a kernel method is used to simulate a non-linear projection of the data in a higher dimension space, where the classes are linearly separable. Besides, unless statistical estimations, a small number of training samples (understood that those are representative) is enough to find the support vectors. Then, this kind of classifier reveals very interesting properties for hyperspectral image processing: it does not suffer from the Hughes phenomenon (which is: for a limited number of training samples, the classification rate decreases as the dimension increases) and it may perform class separation even with means very closed to each other with a small number of training samples. This separability remains

quite difficult even with techniques dedicated to hyperspectral data such as Spectral Angle Mapping or Spectral Unmixing.

However, separability measures are based on dot product or geometric distance between vectors. Those approaches do not take into consideration spectral meaning and behavior. Even if an object is observed with several illumination conditions, its spectral signature remains of the same shape and has to be classified the same way. Then, it is proposed to integrate spectral knowledge into SVM classifiers for processing hyperspectral data cubes.

It enables the classification results to be improved for thematic classification of hyperspectral data cube. The process has been applied on hyperspectral images from the CASI sensor.

## II. THE SVM APPROACH

The complete mathematical formulation of SVM can be found in [1], [2]. We just give a brief description of the classification process.

### A. SVM basis

A two-class classification problem can be stated the following way:  $N$  training sample are available and can be represented by the set pairs  $\{(y_i, \mathbf{x}_i), i = 1, 2, \dots, N\}$  with  $y_i$  a class label of value  $\pm 1$  and  $\mathbf{x}_i \in \mathbb{R}^n$  feature vector with  $n$  components. The classifier is represented by the function  $f(\mathbf{x}; \boldsymbol{\alpha}) \rightarrow y$  with  $\boldsymbol{\alpha}$  the parameters of the classifier.

The SVM method consist in finding the optimum separating hyperplan so that:

- 1) Samples with labels  $y = \pm 1$  are located on each side of the hyperplane;
- 2) The distance of the closest vectors to the hyperplane in each side of maximum. These are called support vectors and the distance is the optimal margin (see Fig. 1-a-).

The hyperplane is defined by  $\mathbf{w} \cdot \mathbf{x} + b = 0$  where  $(\mathbf{w}, b)$  are the parameters of the hyperplane. The vectors that are not on this hyperplane lead to:  $\mathbf{w} \cdot \mathbf{x} + b \geq 0$  and allow the classifier to be defined as:  $f(\mathbf{x}; \boldsymbol{\alpha}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$ . The support vectors lie on two hyperplanes, which are parallel to the optimal hyperplane, of equation:  $\mathbf{w} \cdot \mathbf{x} + b = \pm 1$ .

The maximization of the margin with the equations of the two support vector hyperplanes leads to the following

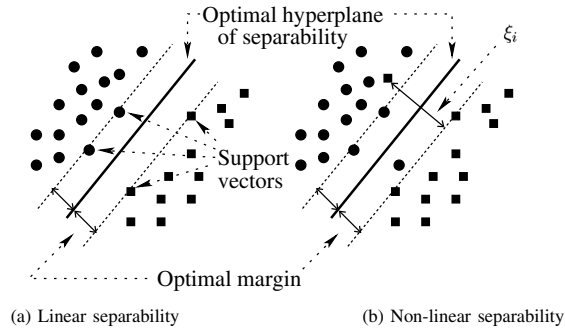


Fig. 1. SVM classifier.

constrained optimization problem:

$$\min \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\} \text{ with } y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, i = 1, \dots, N. \quad (1)$$

### B. Non-linear separability

If the training samples are not linearly separable (see Fig.1-b-), a regularization parameter  $C$  and error variables  $\varepsilon_i$  are introduced in (1) in order to reduce the weighting of misclassified vectors. This optimization problem can be solved using Lagrange multipliers and then becomes:

$$\left\{ \begin{array}{l} \min \left\{ \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i,j=1}^N \lambda_i \lambda_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right\}, \\ 0 \leq \lambda_i \leq C, \quad \forall i = 1, 2, \dots, N, \\ \sum_{i=1}^N \lambda_i y_i = 0, \quad \forall i = 1, 2, \dots, N, \end{array} \right. \quad (2)$$

where the  $\lambda_i$  are the Lagrangian multipliers and are non-zero only for the support vectors. Thus, hyperplane parameters  $(\mathbf{w}, b)$  and the classifier function  $f(\mathbf{x}; \mathbf{w}, b)$  can be computed by optimization process. The free source code `svmlight` [7] has been adapted to the classification of hyperspectral imagery and used in this study.

### C. Multiple class separation

SVM are designed to solve two-class problems. Two approaches can be used for a  $M$ -class problem:

- 1) so called *one against all*:  $M$  classifiers are iteratively applied on each class against all the others.
- 2) so called *one against one*:  $\frac{M(M-1)}{2}$  classifiers are applied on each pair of classes, the most often computed label is kept for each vector.

We have used the second approach, although needing more SVM to be applied, that allows the computing time to be decreased because the complexity of the algorithm depends strongly on the number of training samples.

## III. KERNEL-BASED SVM

### A. Non-linear classifier

SVM can be generalized to compute nonlinear decision surfaces in  $n$ . The method consists in projecting the data in a higher dimension space where they are considered to

become linearly separable. SVM applied in this space lead to the determination of nonlinear surfaces in the original space. Actually, the projection can be simulated using a kernel method.

It can be noticed that only dot products  $(\mathbf{x}_i \cdot \mathbf{x}_j)$  are involved in (2). If  $\mathbf{x} \in n$  is projected in a higher-dimension space  $\mathcal{H}$  with a non-linear function  $\Phi : n \rightarrow \mathcal{H}$ , then  $(\mathbf{x}_i \cdot \mathbf{x}_j)$  is replaced by  $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ . The kernel function  $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$  is introduced in (2) and do not require explicit knowledge of  $\Phi(\cdot)$ .

Then, the non-linear classifier can be expressed as:

$$f(\mathbf{x}; \alpha) = \text{sgn} \left( \sum_{i=1}^{N_S} \lambda_i y_i K(\mathbf{s}_i, \mathbf{x}) + b \right), \quad (3)$$

where the  $\mathbf{s}_i$  are the  $N_S$  support vectors.

### B. Usual kernels

Every function  $K(\cdot, \cdot)$  that satisfies Mercer's conditions may be considered as an eligible kernel. The Mercer's conditions state as:

$$\forall g(\cdot) \in \mathcal{L}^2(n) \text{ so that } \int g(\mathbf{x})^2 d\mathbf{x} \text{ is finite,}$$

$$\text{then } \int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0. \quad (4)$$

A great number of kernels exist and it is difficult to explain their individual characteristics. As shown in [8], two kinds of kernels may be defined:

- 1) *Local kernels*. Only the data that are close or in the proximity of each others have an influence on the kernel values. Basically, all kernels that are based on a distance function are local kernels. Examples of typical local kernels are:

$$\text{Radial basis: } K(\mathbf{x}, \mathbf{x}_i) = \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2),$$

$$\text{KMOD: } K(\mathbf{x}, \mathbf{x}_i) = \exp\left(\frac{1}{1+\|\mathbf{x}-\mathbf{x}_i\|^2}\right) - 1,$$

$$\text{Inverse multiquadric: } K(\mathbf{x}, \mathbf{x}_i) = \frac{1}{\sqrt{(\|\mathbf{x}-\mathbf{x}_i\|^2+1)}}.$$

- 2) *Global kernels*. Samples that are far away from each others still have an influence on the kernel value. All kernels based on the dot-product are global:

$$\text{Linear: } K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x} \cdot \mathbf{x}_i,$$

$$\text{Polynomial: } K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i + 1)^p,$$

$$\text{Sigmoid: } K(\mathbf{x}, \mathbf{x}_i) = \tanh(\mathbf{x} \cdot \mathbf{x}_i + 1).$$

### C. Spectral Kernels

The local kernels are based on a quadratic distance evaluation between two samples. In order to fit hyperspectral point of view, it is of interest to consider new criteria that take into consideration spectral signature concept. Spectral angle (SA)  $\alpha(\mathbf{x}, \mathbf{x}_i)$  is defined in order to measure the spectral difference between  $\mathbf{x}$  and  $\mathbf{x}_i$  while being robust to differences of the overall energy (e.g. illumination, shadows...).

$$\alpha(\mathbf{x}, \mathbf{x}_i) = \arccos \left( \frac{\mathbf{x} \cdot \mathbf{x}_i}{\|\mathbf{x}\| \|\mathbf{x}_i\|} \right). \quad (5)$$

The local kernels listed earlier that are based on quadratic distance (Radial basis function, KMOD, Inverse multiquadric) may then be defined with SA distance. Since  $|\alpha(\mathbf{x}, \mathbf{x}_i)|$  is bounded and positive, spectral angle based kernels satisfy the Mercer's condition (4).

An other measure of interest is the *Spectral Information Divergence* (SID) proposed in [9], [10] which may be thought of as a Kullback-Leibler divergence of spectral signatures.  $SID(\mathbf{x}, \mathbf{x}_i) = \mathcal{D}(\mathbf{x}||\mathbf{x}_i) + \mathcal{D}(\mathbf{x}_i||\mathbf{x})$  with  $\mathcal{D}(\mathbf{x}||\mathbf{x}_i)$  defined as:

$$\mathcal{D}(\mathbf{x}||\mathbf{x}_i) = \sum_{\ell=1}^n p_{\mathbf{x}}(\ell) \log \left( \frac{p_{\mathbf{x}}(\ell)}{p_{\mathbf{x}_i}(\ell)} \right),$$

with  $p_{\mathbf{x}}(\ell) \stackrel{\text{def}}{=} x(\ell) / \sum_{k=1}^n x(k)$ . Once again, some spectral divergence-based local kernels may be defined. Since SID is similar to mutual information, it can be bounded by the log of the dynamic of  $\mathbf{x}$ ; then SID-based kernels satisfy Mercer's conditions.

#### D. Mixture of kernels

As in [11], it is of interest to consider the reflectance scale of spectral signature but also their spectral shape. Then, linear mixture of kernels can fit the dual point of view: similarity according to the dot product or euclidian distance and also, similarity according to the spectral shape (SA or SID). Mixture of kernels may be defined as [8]:

$$K(\mathbf{x}, \mathbf{x}_i) = \mu K_a(\mathbf{x}, \mathbf{x}_i) + (1 - \mu) K_b(\mathbf{x}, \mathbf{x}_i), \quad (6)$$

where  $K_a(\cdot, \cdot)$  and  $K_b(\cdot, \cdot)$  are two kernels (*e.g.* local, global, SA or SID-based). Since  $K_a(\cdot, \cdot)$  and  $K_b(\cdot, \cdot)$  satisfy Mercer's conditions, all linear combinations are eligible for kernels.

## IV. EXPERIMENTS AND CONCLUSION

SVM classifications have been applied to a hyperspectral image from the airborne CASI sensor with 17 spectral bands from 450 to 950nm. The ground resolution is two meters and the image has been calibrated to reflectance by the means of the empirical line method, as shown on Fig. 2.

It appears that using spectral knowledge into SVM classification reduces false alarms for thematic classification. For instance, artificial forest area, which is difficult to classify since trees are small and there is a lot of shadows, has been correctly classified with spectral-angle based kernel. Also, fields are classified with homogeneous area which suit thematic mapping for land use.



Fig. 2. Casi image to process.

Fig. 3 shows the results of several classification achieved in order to identify artificial forest, several kinds of fields, water, roads and wasteland. With classical kernels, classes are not detected correctly: see the hedges, the forest area and over classifications in the wastelands with complex texture.

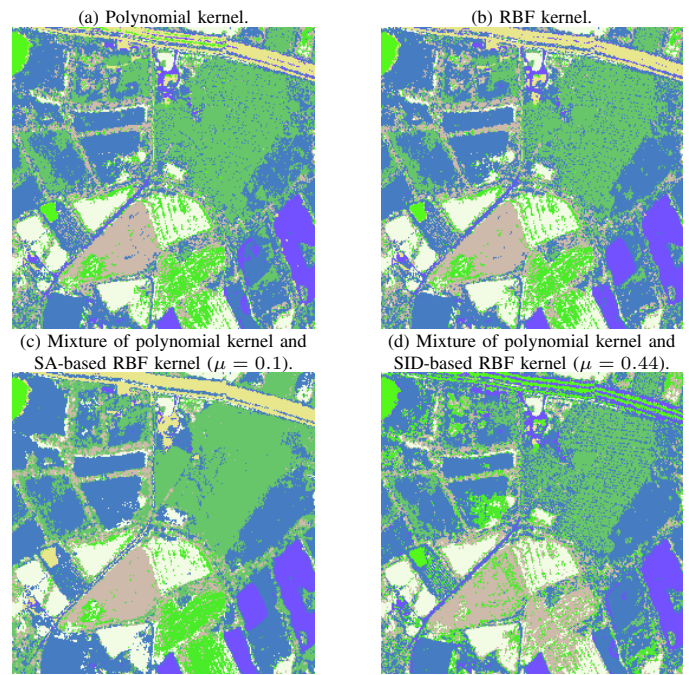


Fig. 3. Resulting classifications with several kernels.

In conclusion, it appears that spectral-based kernel yields better results than SID-based kernel from a thematic point of view. Actually, mixture between spectral-based and quadratic-based kernels achieves the best balance between a geometric point of view and a spectral characterization for integrating bio-physical similarity for the classification of hyperspectral data cube.

## REFERENCES

- [1] C. J. Burges, "A tutorial on support vector machines for pattern recognition," in *Data mining and knowledge discovery*, U. Fayyad, Ed. Kluwer Academic, 1998, pp. 1–43.
- [2] V. N. Vapnick, *Statistical Learning Theory*. John Wiley and Sons Inc., 1998.
- [3] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assesment of support vector machines for land cover classification," *Int. J. Remote sensing*, vol. 23, no. 4, pp. 725–749, 2002.
- [4] J. A. Gualtieri and R. F. Crompt, "Support vector machines for hyperspectral remote sensing classification," in *Proceedings of the SPIE*, vol. 3584, 1999, pp. 221–232.
- [5] F. Melgani and L. Bruzzone, "Support vector machines for classification of hyperspectral remote sensing images," in *IGARSS*, 2002.
- [6] M. Lennon, G. Mercier, and L. Hubert-Moy, "Classification of hyperspectral images with nonlinear filtering and support vector machines," in *IGARSS*, 2002.
- [7] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT-press, 1999, ch. 9, pp. 41–56, see also: <http://kernel-machines.org>.
- [8] G. Smits and E. Jordaen, "Improved svm regression using mixtures of kernels," in *IJCNN*, 2002.
- [9] C. Chang, "Spectral information divergence for hyperspectral image analysis," in *IGARSS*, 1999, pp. 509–511.
- [10] —, "An information theoretic-based measure for spectral similarity and discriminability," *IEEE transactions on geoscience and remote sensing*, vol. 46, no. 5, pp. 1927–1932, August 2000.
- [11] M. Lennon, G. Mercier, and L. Hubert-Moy, "Nonlinear filtering of hyperspectral images with anisotropic diffusion," in *IGARSS*, 2002.