The angular kernel in machine learning for hyperspectral data classification

Paul Honeine, Cédric Richard

THE ANGULAR KERNEL IN MACHINE LEARNING FOR HYPERSPECTRAL DATA CLASSIFICATION

Paul Honeine

Institut Charles Delaunay (UMR CNRS 6279)
Université de technologie de Troyes
10010 Troyes, France

Cédric Richard

Laboratoire H. Fizeau (UMR CNRS 6525, OCA)
Université de Nice Sophia-Antipolis
06108 Nice, France

ABSTRACT

Support vector machines have been investigated with success for hyperspectral data classification. In this paper, we propose a new kernel to measure spectral similarity, called the angular kernel. We provide some of its properties, such as its invariance to illumination energy, as well as connection to previous work. Furthermore, we show that the performance of a classifier associated to the angular kernel is comparable to the Gaussian kernel, in the sense of *universality*. We derive a class of kernels based on the angular kernel, and study the performance on an urban classification task.

Index Terms— Hyperspectral data, spectral angle, SVM, reproducing kernel, machine learning

1. INTRODUCTION

Hyperspectral images are now widely available, owing to the development of remote sensing sensors with an improvement in both spectral and spatial resolutions. For instance, airborne sensors provide hyperspectral images with more than a hundred spectral bands and a spatial resolution up to one meter per pixel. Such resolution allows classification of urban structures by virtue of, on the one hand the spatial visual-perception, and on the other the spectral physical-features. Finer resolution provides an increase in the dimensionality of the processed data, allowing for a better discrimination between different classes of data, e.g. between trees, roads, bricks, etc. Furthermore, only a limited set of observations with labels is available. However, constructing a classification rule based on a small training set in a high dimensional space is an ill-posed problem.

Kernel-based methods provide the opportunity to overcome these problems, with the Support Vector Machines (SVM) which take advantage of the combination of the regularized structure of the decision rule and the elegant use of a reproducing kernel to measure the similarity between data, independent of their dimension. Recently, SVM were investigated for hyperspectral data classification, and have proven to provide high performance of detection and discrimination. Initially, conventional kernels were used such as the Gaussian kernel [1, 2], or adjusted to select optimal spectral bands [3], or even combined using spatial and spectral information [4]. Taking into account the spectral signature concept with an invariance to overall energy (e.g. illuminations), the spectral angle [5] as a measure of distance has been adapted to operate on a Gaussian kernel in [6, 7, 8].

In this paper, we propose the *angular kernel*, a new measure of similarity between two spectra which is insensitive to their energies. We provide some of its properties and connections to previous work. Performance associated to this kernel are studied in the light of the *universality* property, comparing it to the consistency well-known kernels such as the classical Gaussian kernel. We derive a class of kernels based on the angular kernel, and illustrate their performance on a real hyperspectral image for classification of urban data. But before, we review the concept of kernel-based machines for hyperspectral data classification.

1.1. Kernel-based machines for hyperspectral data

Pioneered by Vapnik's SVM [9], kernel-based machines have proven to be successful in many pattern recognition problems. The key issue behind the high generalization ability of SVM is maintained by a complexity control of the solution, tunable by a regularization parameter (often denoted C) which controls the tradeoff between the model simplicity and the fitness to the training data. Furthermore, a central characteristic of these machines is that they can be expressed in terms of inner products of input data. Replacing these inner products with a *reproducing kernel* provides an efficient way to implicitly map the data into a high, even infinite, dimensional space and apply the original algorithm in this space. Hence, performance depends crucially on the chosen reproducing kernel.

By Mercer's theorem, reproducing kernels are positive semi-definite functions, hence can be expressed as an inner product in a high-dimensional feature space. An easy way to construct valid (reproducing) kernels is to apply rules for

The authors wish to thank the University of Pavia and the HySenS project, for providing the data which made this work possible, and Prof. Paolo Gamba for sharing such data.

Table 1. Some simple rules for engineering a valid kernel from available ones, with $\beta_k, c \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}$.

Rule	Expression
R1. Linear combination	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_k \beta_k \kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
R2. Positive Offset	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + c$
R3. Product	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_k \kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
R4. Exponential	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{1}{\sigma^2}\kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)$
R5. Normalization	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sqrt{\kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_i)\,\kappa_k(\boldsymbol{x}_j, \boldsymbol{x}_j)}}$

engineering more complicated kernels from simple ones. Some basis rules are enumerated in Table 1. The first three rules can be combined into the rule: a positive-coefficient polynomial of a reproducing kernel is a valid one. These rules allow to generate most well-known kernels from the linear kernel $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. For instance applying rule R3, or combining rules R2 and R3, with $\kappa_k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ gives respectively the homogeneous and inhomogeneous polynomial kernels, while rule R5 provides the normalized linear kernel $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / \|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|$. The exponential kernel $\exp(\frac{1}{\sigma^2}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$ results from rule R4, while the Gaussian kernel $\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ is obtained from normalizing the exponential kernel with rule R5.

In order to take into account the nature of spectral characteristics in the hyperspectral data, the spectral angle [10] as a measure of distance is extensively used in the literature, thanks to its invariance to the spectral energy, e.g. illumination. It is defined between two spectra $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as

$$\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \arccos\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}\right), \tag{1}$$

where $\|\cdot\|$ is the Euclidean distance and $\langle \cdot, \cdot \rangle$ its inner product. In order to provide a kernel based on this measure, most work consider it[1] as a distance, and adapt any distance-based kernel for this purpose [6]. The most investigated kernel is the Gaussian kernel [8], of the form

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\tfrac{1}{2\sigma^2}\,\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \tag{2}$$

or substituting the angle with its square value in [7].

2. THE ANGULAR KERNEL

Since each spectrum is positive by nature, as well as the ratio in (1), all spectra lie in the positive orthant[2]. We define in this orthant, denoted hereafter by $\mathcal{X}$, the angular kernel as a similarity measure between two spectra, with

$$\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) = \arccos\left(-\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}\right), \tag{3}$$

[1] In fact, they use the absolute value of the spectral angle. However, this quantity is nonnegative for (positive-value) spectral data.

[2] An orthant is the analogue in high dimensional spaces of a quadrant in the plane or an octant in three dimensions.

for any pair $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X}^2$. This kernel corresponds to a monotonic increasing transformation of the normalized linear kernel, with values ranging between $\pi/2$ and π.

Proposition 1. *The angular kernel defined in* (3) *is a valid reproducing kernel.*

Proof. To prove this, recall from trigonometric identities the expansion of the $\arccos$ function into an infinite series:

$$\begin{aligned} \arccos z &= \frac{\pi}{2} - \arcsin z \\ &= \frac{\pi}{2} - \sum_{k=0}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2(2k+1)} z^{2k+1}, \end{aligned}$$

for any $|z| \leq 1$. By substituting z with the normalized linear kernel, we obtain the expansion of the angular kernel:

$$\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\pi}{2} + \sum_{k=0}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2(2k+1)} \left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}\right)^{2k+1} \tag{4}$$

This expansion is a positive-coefficient polynomial of the normalized linear kernel, resulting from rules R1, R2, and R3. Thus, the angular kernel is a valid reproducing kernel. □

Therefore, one can use the angular kernel with any kernel-based learning machine, in order to adapt them for (hyper-) spectral data. Next, we give some insights on the geometric structure of the feature space associated to the angular kernel.

2.1. Properties of the angular kernel and its feature space

Before proceeding, we establish the connection between the angular kernel and the spectral angle defined in (1), the latter being a distance. For this purpose, recall the trigonometric identity $\arccos(-u) = \pi - \arccos(u)$. Thus, we have

$$\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) = \pi - \theta(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{5}$$

This equivalence will be useful throughout this paper.

The space associated to the angular kernel has a very rich structure. From the expansion (4), it is obvious that the dimension of the feature space is infinite. Let $\phi(\cdot)$ denotes the map induced by this reproducing kernel, mapping the input space to the feature space, i.e. $\phi : \mathcal{X} \to \mathcal{H}$. The norm of the image of any mapped data is

$$\|\phi(\boldsymbol{x}_i)\|^2_{\mathcal{H}} = \alpha(\boldsymbol{x}_i, \boldsymbol{x}_i) = \arccos(-1) = \pi.$$

Therefore, all data of the input space are mapped onto the sphere of radius $\sqrt{\pi}$. Moreover, since $\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ on $\mathcal{X}^2$, the images lie in the positive orthant. The distance is defined as the norm of the difference, with its square value

$$\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|^2_{\mathcal{H}} = 2\,(\pi - \alpha(\boldsymbol{x}_i, \boldsymbol{x}_i)) = 2\,\theta(\boldsymbol{x}_i, \boldsymbol{x}_i),$$

where the last equality follows from (5). In other words, the square distance is equal to twice the spectral angle. Since $\alpha(\boldsymbol{x}_i, \boldsymbol{x}_i) \in [\pi/2 \; \pi]$, such a distance is upper bounded by $\sqrt{\pi}$. We refer the interested reader to [11] for a deeper understanding of the geometry of the feature space.

2.2. Performance associated to the angular kernel

All these properties provided so far do not give any information about the performance associated to the use of the angular kernel in machine learning, e.g. SVM for classification. Kernels based on the exponential function are the most used ones, such as the Gaussian and exponential kernels. The performance associated to these kernels is often assigned to their expansion in terms of an infinite series of monomials, with fast falling weightings. The angular kernel shares with these kernels such a property, as given in the expansion (4), and thus is likely to give comparable performance.

The generalization abilities of machine learning classifiers of SVM type, independent of the learning scheme, are studied in [12], using the concept of *universal* kernels. The authors show that there exists a certain class of kernels that are consistent for a large class of classification problems, provided a suitably chosen regularization. This class of so-called universal kernels, includes the Gaussian and the exponential kernels. This is formalized here for the angular kernel.

Proposition 2. *The angular kernel is a universal kernel on every compact subset of* $\mathcal{X}$.

Sketch of proof. From (4), the angular kernel takes the form

$$\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=0}^{\infty} a_k \left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|} \right)^k ,$$

with $a_k > 0$ for all $k \geq 0$. Due to [12, Corollary 10], we get the universality of the kernel. □

The universality of the angular kernel means that the functions of its associated feature space are capable of approximating all continuous functions on compact subsets in $\mathcal{X}$.

2.3. A class of angular kernels

In previous section, we compared the proposed kernel to classical Gaussian and exponential kernels, despite the fact that the angular kernel has no tunable parameter. This property may be advantageous, since we do not need a tuning step to adapt the kernel to the problem under consideration. Next, we provide a class of kernels (with tunable parameters) based on the angular kernel, following the same scheme provided in Section 1.1.

The (homogeneous) polynomial kernel associated to the angular kernel takes the form

$$\alpha_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j))^p ,$$

for $p \in \mathbb{N}_+$. An inhomogeneous counterpart of this kernel can be given by $(\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) + c)^p$, for any positive c. The exponential kernel is defined as the exponential of the angular kernel, up to a multiplicative bandwidth parameter, namely

$$\alpha_e(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\tfrac{1}{\sigma^2} \alpha(\boldsymbol{x}_i, \boldsymbol{x}_j)\right) .$$

It is worth noting that it is more convenient to use *angular* values for the bandwidth parameter σ^2, for instance $\sigma^2 = \pi$ which results into a kernel with values within $[\sqrt{e} \;\; e]$.

In order to construct the equivalent of the Gaussian kernel for the class of angular kernels, we apply the normalization rule R5 to the above exponential kernel. We obtain

$$\begin{aligned} \alpha_G(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \frac{\alpha_e(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sqrt{\alpha_e(\boldsymbol{x}_i, \boldsymbol{x}_i)\, \alpha_e(\boldsymbol{x}_j, \boldsymbol{x}_j)}} \\ &= \exp\left(-\tfrac{1}{\sigma^2}(\pi - \alpha(\boldsymbol{x}_i, \boldsymbol{x}_j))\right) \\ &= \exp\left(-\tfrac{1}{\sigma^2}\theta(\boldsymbol{x}_i, \boldsymbol{x}_j))\right) , \end{aligned}$$

where the identity (5) is applied. The resulting kernel is equivalent to the Gaussian kernel based on the spectral angle (2), and extensively used in the literature on hyperspectral data. In other words, the angular kernel (3) can be considered as a *linear* counterpart of the Gaussian kernel (2).

3. EXPERIMENTAL RESULTS

Data sets are taken with the ROSIS-03 (Reflective Optical System Imaging Spectrometer) provided by the HySenS project. The original hyperspectral image is of the University of Pavia, Italy, with 610-by-340 pixels and 103 frequency bands. For experiments, we took a sub-image of 250-by-250 pixels representing the south-east of the original image, illustrated in Figure 1 (left). Ground truth information about 6 classes are included to train and test the classifiers, as given in Table 2 and illustrated in Figure 1 (middle and right).

For experimentations, an off-the-shelf SVM classifier is used, and applied here in a one-against-all scheme: binary classifiers are trained on each class against the others, while each test observation is assigned to the class with the maximum output. Preliminary experiments were conducted in order to adjust the regularization term in SVM, C, by a search over a logarithmic grid $[10^{-3} \; 10^{3}]$ with increment 10^{-k}. For the angular Gaussian kernel, we need also to adjust the bandwidth parameter, σ^2, which is determined over a grid search over $[\pi/2^6 \;\; \pi/2]$, with increment of the form $\pi/2^k$.

Experiments were conducted on different kernels from the class of angular kernels. Table 2 summarizes the results associated to three kernels, and give the misclassification error

Table 2. The 6 classes with the ratio of train/test samples, and the misclassification error rates associated to angular kernels.

Class-name	#train	#test	α	α_3	α_G
Asphalt	210	2238	7.9 %	6.3 %	11.3 %
Meadow	188	4259	3.9 %	3.9 %	3.9 %
Tree	144	889	4.0 %	4.2 %	3.8 %
Metal sheet	129	647	0.1 %	0.1 %	0.2 %
Brick	93	741	4.0 %	15 %	8.8 %
Shadow	12	84	1.0 %	0.5 %	4.7 %
		Overall error:	**4.4 %**	**5.2 %**	**5.7 %**

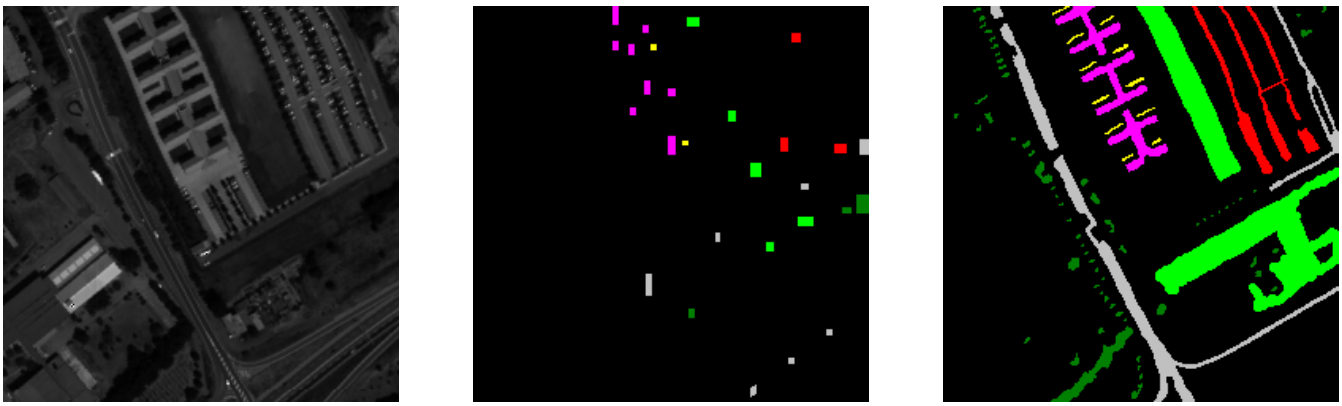

Fig. 1. The hyperspectral image (slice at mid spectral-band) considered in this paper (left), with the spatial distribution of the training (middle) and test (right) datasets. The legend is indicated in Table 2.

rates. The polynomial kernel with degrees ranging from 2 to 10 was used, with the cubic one α_3 giving the best performance. Both the exponential and the Gaussian kernels give comparable results, given in the table for the optimal pair of parameters (C, σ^2). The angular kernel gives slightly better classification performance for almost all the 6 classes, and an overall better classification rate, even though the other kernels have been tune to their best parameter values.

4. CONCLUSION

This paper has addressed the problem of classification of hyperspectral data, providing a new class of kernels for machine learning. The analysis has been carried out on the angular kernel, enumerating some of its properties and giving connections to other kernels. Moreover, we showed that this is a *universal kernel*, resulting into the consistency of the obtained classifier. Preliminary experimental results indicate the adequacy of such reproducing kernels for the classification of hyperspectral data.

5. REFERENCES

[1] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[2] K. Rajpoot and N. Rajpoot, "SVM optimization for hyperspectral colon tissue cell classification," in *Proc. 7th International Conference on Medical Image Computing and Computer Assisted Intervention*. 2004, vol. 3217, pp. 829–837, Springer.

[3] Baofeng Guo, Steve R. Gunn, Robert I. Damper, and James D. B. Nelson, "Customizing kernel functions for svm-based hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 622–629, 2008.

[4] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *Geoscience and Remote Sensing Letters, IEEE*, vol. 3, no. 1, pp. 93–97, 2006.

[5] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, J. P. Barloon, and A. F. H. Goetz, "The spectral image processing system (sips): Interactive visualization and analysis of imaging spectrometer data," *Remote Sensing Environ*, vol. 2-3, no. 44, pp. 145–163, 1993.

[6] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Proc. International Geoscience And Remote Sensing Symposium (IGARSS '03)*, July 2003, vol. 1, pp. 288–290 vol.1.

[7] M. Fauvel, J. Chanussot, and J.A. Benediktsson, "Evaluation of kernels for multiclass classification of hyperspectral remote sensing data," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, May 2006, vol. 2.

[8] M. N. M. Sap and M. Kohram, "Spectral angle based kernels for the classification of hyperspectral images using support vector machines," in *Proc. Second Asia International Conference on Modelling & Simulation (AMS)*, Washington, DC, USA, 2008, pp. 559–563, IEEE Computer Society.

[9] Vladimir N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[10] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 7, pp. 1552–1565, July 2004.

[11] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernal functions," *Neural Netw.*, vol. 12, no. 6, pp. 783–789, 1999.

[12] Ingo Steinwart, "On the influence of the kernel on the consistency of support vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 67–93, 2002.