OPTIMAL SELECTION OF TIME-FREQUENCY REPRESENTATIONS FOR SIGNAL CLASSIFICATION: A KERNEL-TARGET ALIGNMENT APPROACH

Paul Honeiné(1)*, Cédric Richard*(2)*, Patrick Flandrin*(3)*, Jean-Baptiste Pothin*(2)

(1)Sonalyse, Pist Oasis, 131 impasse des palmiers, 30319 Alès, France
(2)ISTIT (FRE CNRS 2732), Troyes University of Technology, BP 2060, 10010 Troyes cedex, France
(3)Laboratoire de Physique (UMR CNRS 5672), École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon, France

ABSTRACT

In this paper, we propose a method for selecting time-frequency distributions appropriate for given learning tasks. It is based on a criterion that has recently emerged from the machine learning literature: the kernel-target alignment. This criterion makes possible to find the optimal representation for a given classification problem without designing the classifier itself. Some possible applications of our framework are discussed. The first one provides a computationally attractive way of adjusting the free parameters of a distribution to improve classification performance. The second one is related to the selection, from a set of candidates, of the distribution that best facilitates a classification task. The last one addresses the problem of optimally combining several distributions.

1. INTRODUCTION

Time-frequency and time-scale distributions provide a powerful tool for analyzing nonstationary signals. They can be set up to support a wide range of tasks depending on the user's information need. As an example, there exist classes of distributions that are relatively immune to interference and noise for analysis purpose [1, 2, 3]. There are also distributions that maximize a contrast criterion between classes to improve classification accuracy [4, 5, 6]. Over the last decade, a number of new pattern recognition methods based on reproducing kernels have been introduced. The most popular ones are Support Vector Machines (SVM), kernel Ficher Discriminant Analysis (kernel-FDA) and kernel Principal Component Analysis (kernel-PCA) [7]. They have gained wide popularity due to their conceptual simplicity and their outstanding performance [8]. Despite these advances, there are few papers other than [9, 10] associating time-frequency analysis with kernel machines. Clearly, time-frequency analysis still has not taken advantage of these new information extraction methods, although many efforts have been focused to develop task-oriented signal representations.

We begin this paper with a brief review of the related work [10]. We show how the most effective and innovative kernel machines can be configured, with a proper choice of reproducing kernel, to operate in the time-frequency domain. In the above cited paper, however it was posed as an open question how to objectively pick time-frequency distributions that best facilitate the classification task at hand. An interesting solution has recently been developed within the area of machine learning through the concept of kernel-target alignment [11]. This criterion makes possible to find the optimal reproducing kernel for a given classification problem without designing the classifier itself. In this paper, we discuss three applications of the alignment criterion to select time-frequency distributions that best suit a classification task. The first one provides a computationally attractive way of adjusting the free parameters of a distribution. The second one is related to the selection of the best distribution from a set of candidate distributions. The last one addresses the problem of optimally combining several distributions to achieve improvements in classification performance.

2. BACKGROUND ON KERNEL MACHINES

In this section, we concisely review the fundamental building blocks of kernel machines, mainly the definition of reproducing kernel Hilbert spaces, the kernel trick and the representer theorem. Let $\mathcal{X}$ be a subspace of $\mathcal{L}_2(\mathbb{C})$, the space of finite-energy complex signals. A kernel is a function κ from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{C}$, with hermitian symmetry. The following two definitions provide the basic concept of reproducing kernels [12].

Definition 1. *A kernel $\kappa(x_i, x_j)$ is said to be positive definite on $\mathcal{X}$ if the following is true:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \, \overline{a}_j \, \kappa(x_i, x_j) \geq 0, \tag{1}$$

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $a_1, \ldots, a_n \in \mathbb{C}$.

Definition 2. *Let $(\mathcal{H}, \langle \cdot\,, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{C}$. The function $\kappa(x_i, x_j)$ from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{C}$ is the reproducing kernel of $\mathcal{H}$ if, and only if,*

- *the function $\kappa_{x_i} : x_j \mapsto \kappa_{x_i}(x_j) = \kappa(x_i, x_j)$ is in $\mathcal{H}$, for all $x_i \in \mathcal{X}$;*
- *$\psi(x_i) = \langle \psi(\cdot), \kappa_{x_i}(\cdot) \rangle_{\mathcal{H}}$, for all $x_i \in \mathcal{X}$ and $\psi \in \mathcal{H}$.*

It can be shown that every positive definite kernel is the reproducing kernel of a unique Hilbert space of functions from $\mathcal{X}$

to $\mathbb{C}$, called *reproducing kernel Hilbert space*. Reciprocally, every reproducing kernel is a positive definite kernel. A proof of this may be found in [12]. From the second point of definition 2 results a fundamental property of reproducing kernel Hilbert space. Replacing $\psi(\cdot)$ by $\kappa_{x_j}(\cdot)$, we obtain

$$\kappa(x_j, x_i) = \langle \kappa_{x_j}(\cdot), \kappa_{x_i}(\cdot) \rangle_{\mathcal{H}} \tag{2}$$

for all x_i, $x_j \in \mathcal{X}$, which is the origin of the now generic term *reproducing kernel* to refer to κ. Denoting by $\varphi(\cdot)$ the map that assigns to each x the kernel function $\kappa(x, \cdot)$, equation (2) implies that $\kappa(x_j, x_i) = \langle \varphi(x_j), \varphi(x_i) \rangle_{\mathcal{H}}$. The kernel then evaluates the inner product of any pair of elements of $\mathcal{X}$ mapped to $\mathcal{H}$ without any explicit knowledge of $\varphi(\cdot)$. This key idea is known as the *kernel trick* because it can be used to transform linear algorithms expressed only in terms of inner products into nonlinear ones.

The representer theorem [13], like the kernel trick, is a quintessential building block for kernel machines. Consider a training set $\mathcal{A}_n$ consisting of n input-output pairs (x_i, y_i). This theorem states that any function $\psi^*(\cdot)$ of $\mathcal{H}$ minimizing a regularized cost function of the form

$$J((x_1, y_1, \psi(x_1)), \ldots, (x_n, y_n, \psi(x_n))) + g(\|\psi\|^2_{\mathcal{H}}), \tag{3}$$

with $g(\cdot)$ a monotone increasing function on $\mathbb{R}_+$, can be expressed as a kernel expansion in terms of available data

$$\psi^*(x) = \sum_{i=1}^{n} a_i^* \, \kappa(x, x_i). \tag{4}$$

Applications of this theorem include SVM, kernel-PCA and kernel-FDA [7]. In the next section, we show how kernel machines can be configured, with a proper choice of reproducing kernel, to operate in the time-frequency domain.

3. TIME-FREQUENCY REPRODUCING KERNELS

For reasons of conciseness, we restrict ourselves to the Cohen class of time-frequency distributions. They can be defined as

$$C_x^{\Phi}(t, f) = \iint \Phi(\nu, \tau) \, A_x(\nu, \tau) \, e^{-2j\pi(f\tau + \nu t)} \, d\nu \, d\tau, \tag{5}$$

where $A_x(\nu, \tau)$ denotes the narrow-band ambiguity function of x, and $\Phi(\nu, \tau)$ is a parameter function. Conventional pattern recognition algorithms applied directly to time-frequency representations consist of estimating $\Psi^*(t, f)$ in the statistics

$$\psi^*(x) = \langle \Psi^*, C_x^{\Phi} \rangle = \iint \Psi^*(t, f) \, C_x^{\Phi}(t, f) \, dt \, df \tag{6}$$

to optimize a criterion of the general form (3). Examples of cost functions include the maximum output variance for PCA, the maximum margin for SVM, and the maximum Fisher criterion for FDA. It is apparent that this direct approach is computationally demanding because the size of C_x^{Φ} grows quadratically in the length of the input signal x. Faced with such prohibitive computational costs, an attractive alternative is to make use of the kernel trick and the representer theorem, if possible, with the following kernel

$$\kappa_{\Phi}(x_i, x_j) = \langle C_{x_i}^{\Phi}, C_{x_j}^{\Phi} \rangle. \tag{7}$$

Writing condition (1) as $\| \sum_i a_i \, C_{x_i}^{\Phi} \|^2 \geq 0$, which is indeed satisfied, we verify that κ_{Φ} is a positive definite kernel. We denote by $\mathcal{H}_{\Phi}$ the unique reproducing kernel Hilbert space associated with κ_{Φ}. This argument shows that (7) can be associated with any kernel machine reported in the literature to perform pattern recognition in the time-frequency domain. Thanks to the representer theorem, the solution $\psi^*(x)$ admits a time-frequency interpretation, $\psi^*(x) = \langle \Psi^*, C_x^{\Phi} \rangle$, with

$$\Psi^* = \sum_{i=1}^{n} a_i^* \, C_{x_i}^{\Phi}. \tag{8}$$

This equation is obtained by combining (4) and (6). The question of how to select C_x^{Φ} is still open. The next section brings some elements of answer in a binary classification framework.

4. KERNEL-TARGET ALIGNMENT

The alignment criterion is a measure of similarity between two reproducing kernels, or between a reproducing kernel and a target function [11]. Given a training set $\mathcal{A}_n$, the alignment of kernels κ_1 and κ_2 is defined as follows

$$A(\kappa_1, \kappa_2; \mathcal{A}_n) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}, \tag{9}$$

where $\langle \cdot\,, \cdot \rangle_F$ is the Frobenius inner product between two matrices, and K_1 and K_2 are the Gram matrices with respective entries $\kappa_1(x_i, x_j)$ and $\kappa_2(x_i, x_j)$, for all i, $j \in \{1, \ldots, n\}$. The alignment then is simply the correlation coefficient between the bidimensional vectors K_1 and K_2.

For binary classification purpose, the decision statistic should satisfy $\psi(x_i) = y_i$, where y_i is the class label of x_i. By setting $y_i = \pm 1$, the ideal Gram matrix would be given by

$$K^*(i, j) = \langle \psi(x_i), \psi(x_j) \rangle = \begin{cases} \;\;\,1 & \text{if} \;\; y_i = y_j \\ -1 & \text{if} \;\; y_i \neq y_j, \end{cases} \tag{10}$$

in which case $\sqrt{\langle K^*, K^* \rangle_F} = n$. In [11], Cristianini *et al.* propose maximizing the alignment with the target K^* in order to determine the most relevant reproducing kernel for a given classification task. The ease with which this criterion can be estimated using only training data, prior to any computationally intensive training, makes it an interesting tool for kernel selection. Its relevance is supported by the existing connection between the alignment score and the generalization performance of the resulting classifier. This has motivated various computational methods of optimizing kernel alignment, including metric learning [14], eigendecomposition of the Gram matrix [11, 15] and linear combination of kernels [16, 17]. We will focus on the latter of these issues, which consider the kernel expansion

$$\kappa_\alpha(x_i, x_j) = \sum_{k=1}^{m} \alpha_k \kappa_k(x_i, x_j) \tag{11}$$

and study the problem of choosing the α_k's to maximize the kernel-target alignment. A positivity constraint on these coefficients is imposed to ensure the positive definiteness of κ_α. Some more or less efficient algorithms have been proposed in the literature. In [16], it has been shown that a concise analytical solution exists in the $m = 2$ case:

$$(\alpha_1^*, \alpha_2^*) = \begin{cases} (\alpha_1, \alpha_2) & \text{if} \quad \alpha_1, \alpha_2 > 0 \\ (1, 0) & \text{if} \quad \alpha_2 \leq 0 \\ (0, 1) & \text{if} \quad \alpha_1 \leq 0, \end{cases} \tag{12}$$

with

$$\alpha_1 = \frac{1}{2} \frac{\langle K_1, K^* \rangle_F - 2\langle K_1, K_2 \rangle_F \alpha_2}{\|K_1\|_F^2 + \lambda}$$

$$\alpha_2 = \frac{1}{2} \frac{(\|K_1\|_F^2 + \lambda)\langle K_2, K^* \rangle_F - \langle K_1, K_2 \rangle_F \langle K_1, K^* \rangle_F}{(\|K_1\|_F^2 + \lambda)(\|K_2\|_F^2 + \lambda) - \langle K_1, K_2 \rangle_F^2},$$

where $\lambda \geq 0$ arises from a regularization constraint penalizing $\|\alpha\|^2$. To combine more than 2 kernels, we opted for a *branch and bound* approach. It starts from the best available kernel, and selects from the remaining kernels the one which best increases the alignment criterion. This procedure is iterated until no improving candidates can be found.

5. TIME-FREQUENCY FORMULATION

By placing time-frequency based classification within the larger framework of kernel machines, we can take advantage of concepts and tools that have been developed above. In this section, we focus on selecting time-frequency distributions appropriate for binary classification tasks. That is, we consider the maximization problem

$$\Phi^* = \arg\max_{\Phi} \frac{\langle K_\Phi, K^* \rangle_F}{n\sqrt{\langle K_\Phi, K_\Phi \rangle_F}}, \tag{13}$$

where K_Φ is the Gram matrix associated with C_x^Φ. We also discuss how to improve performance by optimally combining several time-frequency distributions.

Before proceeding, note that the experiments were run on 64-sample data generated according to the hypothesis test

$$\begin{cases} \omega_0 : x(t) = w_0(t) \\ \omega_1 : x(t) = w_1(t) + e^{2j\pi[\phi(t) + \phi_0]}, \end{cases} \tag{14}$$

where $\phi(t)$ is a quadratic phase modulation and ϕ_0 the initial phase. The noises $w_0(t)$ and $w_1(t)$ are zero-mean, Gaussian and white with variances σ_0^2 and σ_1^2, respectively. They were fixed to 2.25 for the first two experiments, and ϕ_0 was considered a random variable uniformly distributed over $[0, 2\pi[$. In the third experiment, σ_0^2 and σ_1^2 were set to 9 and 4, respectively, and ϕ_0 was fixed to 0. For each experiment, a training set $\mathcal{A}_{200}$ of size 200 was generated with equal priors. A test set $\mathcal{T}_{1000}$ of 1000 examples was also created to estimate the generalization performance of kernel-optimal SVM classifiers trained on $\mathcal{A}_{200}$.

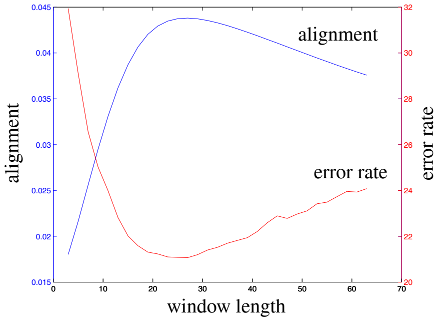

Fig. 1. Adjustment of the window size of a spectrogram using the kernel-target alignment. Comparison with the error rate of a SVM classifier.

5.1. Parameter setting

The first illustration deals with parameter setting of time-frequency distributions. Without any loss of generality, we address the problem of adjusting the window size of a spectrogram S_x with a view to maximize classification accuracy. The reproducing kernel is then defined as

$$\kappa_{\text{sp}}(x_i, x_j) = \langle S_{x_i}, S_{x_j} \rangle. \tag{15}$$

Figure 1 shows, as a function of the window size, the kernel-target alignment of κ_{sp} over the training set $\mathcal{A}_{200}$. It also includes the error rate of a SVM classifier trained and tested on $\mathcal{A}_{200}$ and $\mathcal{T}_{1000}$, respectively. We note that the maximum alignment is obtained with a window size of 27, and coincides with the lowest error rate. This shows that with a high alignment on the training set, we can expect a good generalization performance of a kernel-based classifier.

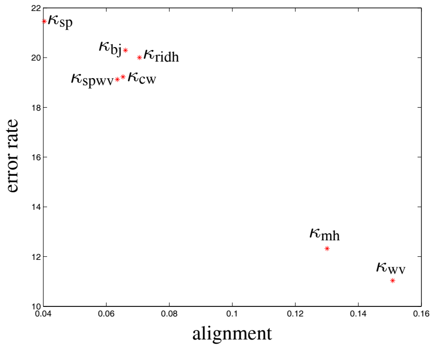

Fig. 2. Alignment and error rate for different kernels.

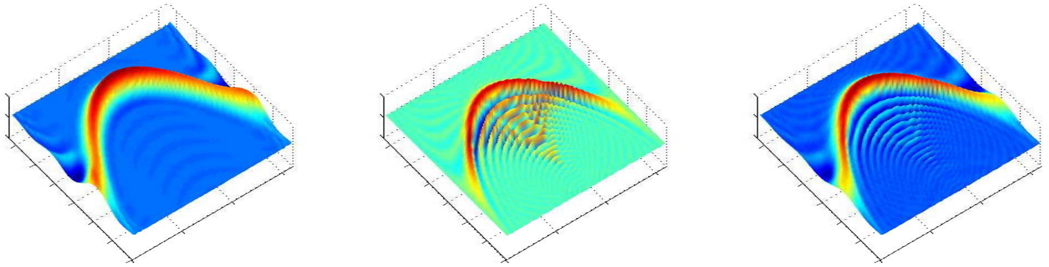

Fig. 3. Smoothed pseudo-Wigner (left), Wigner (middle), and composite associated with the kernel $\kappa_{\text{spwv}} + 0.208\,\kappa_{\text{wv}}$ (right). Here these distributions are applied to the signal to be detected.

5.2. Selection of a distribution

The second illustration is concerned with the selection of a distribution from a set of candidates. The latter consists of the following distributions: Wigner (κ_{wv}), smoothed pseudo-Wigner (κ_{spwv}), Margenau-Hill (κ_{mh}), Choï-Williams (κ_{cw}), Born-Jordan (κ_{bj}), reduced interference with Hanning window (κ_{ridh}), and spectrogram (κ_{sp}). Figure 2 shows the performance averaged over 50 independent realizations of the training and test sets. It provides the alignment of the above-mentioned kernels over $\mathcal{A}_{200}$, versus the error rate of a SVM classifier trained and tested on $\mathcal{A}_{200}$ and $\mathcal{T}_{1000}$, respectively. The apparent relationship between these two criteria emphasizes once more the relevance of the kernel-target alignment.

5.3. Combination of distributions

The last illustration focuses on the combination of time-frequency distributions to achieve improvements in classification performance. This problem was addressed with the kernel-based process (11)-(12), which was applied to the above-described set of candidate distributions. Kernels κ_{spwv} and κ_{wv} were successively selected. The kernel-target alignment increased from 0.1039 to 0.1076, while the error rate of the SVM classifier reduced from 4.7% to 3.2%. Figure 3 presents the composite time-frequency distribution, applied here to the signal to be detected.

Another experimentation was carried out by adding the short-time Fourier transform to the above-mentioned set of quadratic distributions. Note that $\kappa_{\text{stft}}(x_i, x_j) = \langle x_i, x_j \rangle$ for a normalized window. Kernels κ_{stft} and κ_{spwv} were successively chosen, for a final alignment of 0.1698 and an error rate of 2.7%. This result is consistent with statistical decision theories since the log-likelihood ratio for the detection problem under consideration involves both linear and quadratic components of the observation.

6. CONCLUSION

In this paper, we showed that specific reproducing kernels allow any kernel machine to operate on time-frequency representations. We also proposed a method, based on the kernel-target alignment, for selecting or combining time-frequency distributions to achieve improvements in classification performance. All these links offer new perspectives in the field of non-stationary signal analysis since they provide an access to the most recent methodological and theoretical developments of pattern recognition and statistical learning theory.

7. REFERENCES

[1] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by reassignment methods," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[2] D. Jones and R. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Transactions on Signal Processing*, vol. 43, no. 10, pp. 2361–2371, 1995.

[3] J. Gosme, C. Richard, and P. Gonçalvès, "Adaptive diffusion of time-frequency and time-scale representations: a review." *IEEE Transactions on Signal Processing*, vol. 53, no. 11, 2005.

[4] L. Atlas, J. Droppo, and J. McLaughlin, "Optimizing time-frequency distributions for automatic classification," in *Proc. SPIE*, vol. 3162, 1997, pp. 161–171.

[5] C. Heitz, "Optimum time-frequency representations for the classification and detection of signals," *Applied Signal Proceedings*, vol. 3, pp. 124–143, 1995.

[6] M. Davy, C. Doncarli, and G. Boudreaux-Bartels, "Improved optimization of time-frequency based signal classifiers," *IEEE Signal Processing Letters*, vol. 8, no. 2, pp. 52–57, 2001.

[7] K. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–202, 2000.

[8] V. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

[9] M. Davy, A. Gretton, A. Doucet, and P. Rayner, "Optimised support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 442–445, 2002.

[10] P. Honeiné, C. Richard, and P. Flandrin, "Reconnaissance des formes par méthodes à noyau dans le plan temps-fréquence," in *Proc. Colloque GRETSI*, Louvain-la-Neuve, Belgium, 2005, pp. 969–972.

[11] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[12] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[13] B. Schölkopf, R. Herbrich, and R. Williamson, "A generalized representer theorem," NeuroCOLT, Royal Holloway College, University of London, UK, Tech. Rep. NC2-TR-2000-81, 2000.

[14] G. Wu, E. Y. Chang, and N. Panda, "Formulating distance functions via the kernel trick," in *Proc. 11th ACM International conference on knowledge discovery in Data mining*, 2005, pp. 703–709.

[15] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "On the extensions of kernel alignment," Dept. Comput. Sci., University of London, Tech. Rep. 120, 2002.

[16] J.-B. Pothin and C. Richard, "Kernel machines : une nouvelle méthode pour l'optimisation de l'alignement des noyaux et l'amélioration des performances," in *Proc. Colloque GRETSI*, Louvain-la-Neuve, Belgium, 2005, pp. 1133–1136.

[17] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing kernel alignment over combinations of kernels," Department of Computer Science, University of London, Tech. Rep. 121, 2002.