SIGNAL-DEPENDENT TIME-FREQUENCY REPRESENTATIONS FOR CLASSIFICATION USING A RADIALLY GAUSSIAN KERNEL AND THE ALIGNMENT CRITERION

Paul Honeine, Cédric Richard

Institut Charles Delaunay (ICD-LM2S), FRE CNRS 2848, Université de technologie de Troyes, BP 2060, 10010 Troyes cedex, France

Abstract

We propose a method for tuning time-frequency distributions with radially Gaussian kernel within a classification framework. It is based on a criterion that has recently emerged from the machine learning literature: the kernel-target alignement. Our optimization scheme is very similar to that proposed by Baraniuk and Jones for signal dependent time-frequency analysis. The relevance of this approach of improving time-frequency classification accuracy is illustrated through examples.

Time-Frequency Kernel Machines

Time-frequency distributions : Cohen's class is defined by the distributions of the form

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

where $A_x(\nu,\tau)$ denotes the narrow-band ambiguity function of x, and $\Phi(\nu,\tau)$ is a parameter function, both expressed in rectangular coordinates.

Tunable distributions : While the two-dimensional function Φ determines the properties of the distribution, one often seeks to parameterize it with a one-dimensional function, say σ, and denote it Φ_σ. This is the essence of the RGK time-frequency distribution.

RGK time-frequency distribution : The parameter function is defined by

$$\Phi_\sigma(r,\theta) = \mathrm{e}^{-r^2/2\sigma^2(\theta)},$$

in polar coordinates, with $\theta = \arctan(\tau/\nu)$ and $r = \sqrt{\nu^2+\tau^2}$. The function $\sigma(\cdot)$ is called the spread function. It determines the shape of Φ_σ in the ambiguity plane, and thus the properties of the time-frequency distribution.

Kernel machines : Kernel machines are non-linear pattern recognition techniques obtained from classical linear ones by using the kernel trick and a reproducing kernel. The latter corresponds to an inner product in a transformed space. Taking advantage of new theoretical advances, kernel machines are attractive by their reduced algorithmic complexity, mainly due to the *kernel trick*. This key idea exploits the fact that a great number of pattern recognition techniques does not depend explicitly of the data itself, but rather of their inner products. A generalization of these are the reproducing kernels, corresponding to an inner product of implicitly transformed data, while every reproducing kernel determines the transformation, up to a unitary transformation.

Time-frequency kernel machines : Cohen's class distributions are specific signal transformations. Given any pair of signals (x_i, x_j), the reproducing kernel associated to such spaces of transformations can be expressed by

$$\kappa(x_i,x_j) = \iint |\Phi(\nu,\tau)|^2\, A_{x_i}(\nu,\tau)\, \overline{A_{x_j}(\nu,\tau)}\, \mathrm{d}\nu\, \mathrm{d}\tau,$$

in rectangular coordinates, or equivalently in polar coordinates

$$\kappa(x_i,x_j) = \iint r\, |\Phi(r,\theta)|^2\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)}\, \mathrm{d}r\, \mathrm{d}\theta.$$

RKG reproducing kernel :

$$\kappa_\sigma(x_i,x_j) = \iint r\, A_{x_i}(r,\theta)\, A_{x_j}(r,\theta)\, \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\, \mathrm{d}r\, \mathrm{d}\theta.$$

The use of this reproducing kernel allows a wide class of pattern recognition methods to operate on the RGK distribution, as previously studied for other time-frequency distributions.

In what follows, we consider a criterion initially proposed within the framework of kernel machines, in order to optimize the parameters of the RGK reproducing kernel, and therefore the corresponding RGK distribution.

Kernel-target alignment

Classification of signals : Consider a 2-class classification problem of signals, from a training set $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ of n signals x_k with their labels $y_k = \pm 1$. Let K_σ be the Gram matrix of the training set, whose (i,j)-th entry is $\kappa_\sigma(x_i,x_j)$, and K_t the target matrix whose (i,j)-th entry is $y_i y_j$ (product of the outputs of the ideal classifier, given the input x_i and x_j).

Kernel-target alignment score : To measure the similarity between the reproducing kernel and the class labels, we consider the kernel-target alignment, defined by

$$\mathcal{A}(K_\sigma,K_t) = \frac{\langle K_\sigma, K_t\rangle_F}{\|K_\sigma\|_F \|K_t\|_F}, \qquad (1)$$

where $\langle\cdot,\cdot\rangle_F$ designates Frobenius scalar product, and $\|\cdot\|_F$ its norm.

Kernel-target alignment criterion : Cristianini *et al.* proposed to select appropriate reproducing kernels by maximizing this score. Theoretical and experimental results show that good generalization performance may be expected by using kernels with large alignment score. Note that this criterion does not require any computational intensive stage for designing and testing classifiers.

Classification-Dependent Time-Frequency Distribution

Optimal RGK time-frequency distribution : The optimal spread function $\sigma^*(\cdot)$ is determined by maximizing the alignment score, with $\sigma^* = \arg\max_\sigma \mathcal{A}(K_\sigma,K_t)$.

Constrained optimization problem : The optimization problem above is equivalent to maximizing the numerator of (1) subject to a constant denominator, namely,

$$\max_\sigma \sum_{i,j=1}^n y_i\, y_j\, \kappa_\sigma(x_i,x_j) \qquad \text{subject to} \qquad \sum_{i,j=1}^n \kappa_\sigma(x_i,x_j)^2 = V_0, \qquad (2)$$

where V_0 is a preset normalization parameter.

Objective functional : By expanding (2), we can write

$$\sum_{i,j=1}^n y_i\, y_j\, \kappa_\sigma(x_i,x_j) = \iint r\Big[\sum_{i,j=1}^n y_i\, y_j\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)}\Big]\, \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\, \mathrm{d}r\, \mathrm{d}\theta.$$

We obtain the same form of objective functional to be maximized as the one obtained by Baraniuk and Jones for signal dependent time-frequency analysis, which was $\iint r|A_x(r,\theta)|^2\, \mathrm{e}^{-r^2/\sigma^2(\theta)}\, \mathrm{d}r\, \mathrm{d}\theta$, where the signal dependent term $|A_x(r,\theta)|^2$ is substituted by the equivalent representation $\sum_{i,j} y_i y_j\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)}$. Since the latter depends only on the training signals and their labels, we can evaluate it prior to any optimization scheme. Exactly the same algorithm previously proposed for signal analysis can then be used to solve this problem, for signal classification purpose. In particular, we relax the computationally expensive constraint in (2) by substituting it with a constraint on the volume of the parameter function, i.e., $\int \sigma^2(\theta)\mathrm{d}\theta = V_0'$ as recommended by Baraniuk and Jones.

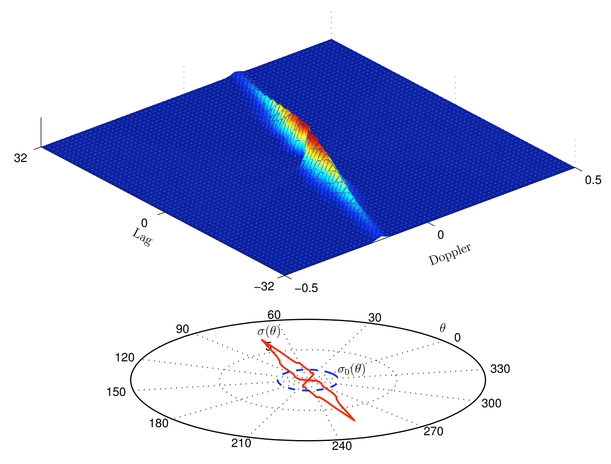
The algorithm :

```
Initialisation
      Compute the equivalent representation of the training set :
            Ψ(r,θ) = Σ_{i,j=1}^n y_i y_j A_{x_i}(r,θ) conj(A_{x_j}(r,θ))
At each iteration k+1, repeat
    1.Evaluate the gradient of the functional at the spread vector
            ∇f_σ = [∂f_σ/∂σ(0), ··· , ∂f_σ/∂σ(l−1)], where ∂f_σ/∂σ(θ) = (2Δ_r^2/σ^3(θ)) Σ_r r^3 Ψ(r,θ) e^{−(r Δ_r)^2/σ^2(θ)}
    2.Update the spread vector with a gradient ascent scheme : σ_{k+1} = σ_k + μ_k ∇f_{σ_k}
    3.Project into feasible set by rescaling :  σ_{k+1} = σ_{k+1} · V_0'/‖σ_{k+1}‖^2.
```

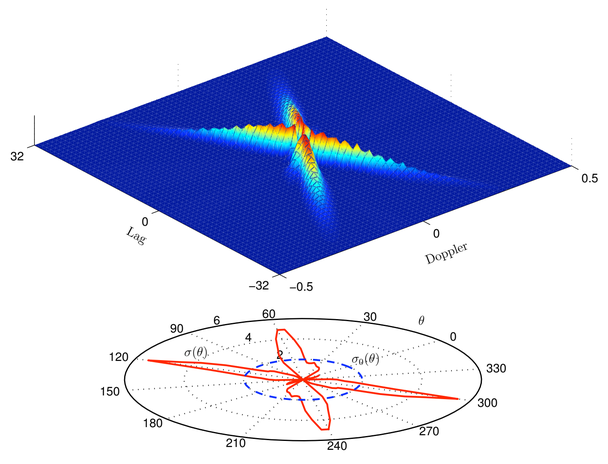
Simulations

We illustrate the proposed approach with two classification problems. They consist of two sets of 200 signals of 64 samples with a linear frequency modulation (chirp), in an additive white Gaussian noise of variance 4. In the first case, signals have an increasing modulation (in normalized frequency), from 0.1 to 0.25 for the first class, and from 0.25 to 0.4 for the second one. In the second case, the signals have an increasing modulation for the first class, from 0.1 to 0.4, and a decreasing modulation, from 0.4 to 0.1 for the second class. By applying the proposed optimization algorithm to each case, we got optimal RGK functions that correspond to relevant regions, in terms of classification, of the ambiguity domain. This is represented in Figure (a) and in Figure (b).

The relevance of using the kernel-target alignment criterion to improve classification is illustrated with an SVM classifier associated with the Wigner distribution or the optimal RGK distribution, for each studied case. In the tables, we represent the error rate, estimated on a test set of 2000 signals, and the number of support vectors, both averaged over 20 realizations. Note that the optimal RGK distribution minimizes the classification error, and also results in almost half the number of support vectors as compared to the Wigner distribution. This is mainly due to the optimality of the resulting distribution on the one hand, and on the other hand to its regularity, i.e., robustness caused by reduced interference terms.



(a)



(b)

	Error rate (%)	Number of SV
Wigner distribution	23.9 ± 1.3	172.6 ± 5.5
Optimal distribution	21.4 ± 2.0	87.5 ± 4.1

	Error rate (%)	Number of SV
Wigner distribution	19.4 ± 1.5	163.7 ± 4.6
Optimal distribution	17.8 ± 1.5	84.5 ± 6.1

www.tfd.fr