Distribution temps-fréquence à noyau radialement Gaussien : optimisation pour la classification par le critère d'alignement noyau-cible

Paul Honeine, Cédric Richard

Institut Charles Delaunay (FRE CNRS 2848) - LM2S - Université de Technologie de Troyes
12 rue Marie Curie, BP 2060, 10010 Troyes cedex - fax. 03.25.71.56.99
paul.honeine@utt.fr, cedric.richard@utt.fr

Résumé – Cet article traite de l'ajustement des paramètres des distributions temps-fréquence pour la résolution d'un problème de classification de signaux. On s'intéresse en particulier à la distribution à noyau radialement Gaussien. On exploite le critère d'alignement noyau-cible, développé pour la sélection du noyau reproduisant dans le cadre des méthodes à noyau. Celui-ci présente l'intérêt de ne nécessiter aucun apprentissage de la statistique de décision. On adapte le critère d'alignement noyau-cible au noyau radialement Gaussien, en détournant une technique classique de réduction de termes interférentiels dans les représentations temps-fréquence. On illustre cette approche par des expérimentations de classification de signaux non-stationnaires.

Abstract – In this article, we design optimal time-frequency distributions for classification. Our approach is based on the kernel-target alignment criterion, which has been investigated in the framework of kernel-based machines for selecting optimal reproducing kernels. One of its main interests is that it does not need any computationally intensive training stage and cross-validation process. We take advantage of this criterion to tune radially Gaussian kernel, and consider a classical optimization technique usually used for reducing interference terms of time-frequency distributions. We illustrate our approach with some experimental results.

1 Introduction

Les distributions temps-fréquence, en particulier de la classe de Cohen, fournissent des outils puissants adaptés à l'analyse des signaux non-stationnaires. Parmi celles-ci, les distributions à noyau radialement Gaussien offrent une diversité de représentations pouvant satisfaire à un vaste choix d'applications. On cherche par exemple à optimiser la représentation temps-fréquence d'un signal donné, par une réduction des termes interférentiels [1, 2]. On peut encore estimer ses paramètres afin de faciliter la résolution d'un problème de détection ou de classification de signaux non-stationnaires [3, 4].

A l'exception de quelques travaux récents [5, 6], ces différentes approches n'ont pas exploité les récentes avancées des méthodes de reconnaissance de formes à noyau reproduisant. Celles-ci sont particulièrement attractives en raison de leur complexité algorithmique réduite, et parce qu'elles profitent des nouvelles avancées en théorie statistique de l'apprentissage. On a récemment proposé dans [7, 8] un cadre général pour la mise en œuvre des méthodes à noyau dans le domaine temps-fréquence grâce à un choix approprié du noyau reproduisant. Dans [9], on a soulevé le problème de sélection d'une distribution temps-fréquence adaptée à la résolution d'un problème de classification de signaux. La stratégie proposée ici repose sur un critère développé dans le cadre des méthodes à noyau : l'alignement noyau-cible [10].

Dans le présent article, on exploite ce critère pour ajuster les paramètres de la distribution temps-fréquence à noyau radialement Gaussien. Pour ce type de distributions, on montre que la maximisation de ce critère se réduit à un problème d'optimisation équivalent à celui présenté dans [1] pour l'analyse de signaux. On peut alors adapter l'algorithme de descente de gradient, proposé dans ce même article, à notre problème de classification.

La suite de cet article est organisée ainsi. A la section suivante, on présente les distributions temps-fréquence de la classe de Cohen, et leurs usages dans le cadre des méthodes de reconnaissance de formes à noyau. En particulier, on s'intéresse à la distribution à noyau radialement Gaussien. A la section 3, on adapte le critère d'alignement noyau-cible à cette dernière. On illustre notre approche par des simulations à la section 4, et on compare les résultats à ceux obtenus avec la distribution de Wigner.

2 Distribution temps-fréquence et méthodes à noyau

Une distribution temps-fréquence de la classe de Cohen, caractérisée par une fonction de paramétrisation Φ_σ, est définie pour un signal x par

$$C_x^\sigma(t,f) = \iint \Phi_\sigma(\nu,\tau)\ A_x(\nu,\tau)\ \mathrm{e}^{-j2\pi(f\tau-t\nu)}\mathrm{d}\nu\ \mathrm{d}\tau,$$

où $A_x = \int x(t+\tau/2)\overline{x(t-\tau/2)}\ \mathrm{e}^{j2\pi\nu t}\mathrm{d}t$ est la fonction d'ambiguïté du signal x. Une distribution particulière est la distribution de Wigner, obtenue en considérant une fonction de paramétrisation unité, soit $\Phi_\sigma(\nu,\tau) = 1$ sur tout le plan Doppler-retard. Les propriétés de la distribution sont déterminées par la fonction de paramétrisation considérée, que l'on cherche à optimiser pour une application donnée. Toutefois, la résolution de ce problème

se heurte au nombre élevé de paramètres libres. Pour remédier à cela, une approche souvent privilégiée consiste à imposer un caractère passe-bas au filtre Φ_σ et le paramétrer selon une fonction radialement Gaussienne [1], définie par $\Phi_\sigma(r,\theta) = \mathrm{e}^{-r^2/2\sigma^2(\theta)}$ en coordonnées polaires dans le plan Doppler-retard. Le problème d'optimisation est alors réduit à la détermination de la largeur de bande $\sigma(\cdot)$, unidimensionnelle, qui détermine la forme de la fonction de paramétrisation, et par conséquent les propriétés de la distribution temps-fréquence correspondante.

Dans le cadre général des méthodes à noyau, les performances d'une règle de décision sont largement influencées par le noyau reproduisant considéré. Celui-ci correspond à un produit scalaire des données dans un espace transformé, obtenu par une transformation non-linéaire de l'espace des observations. Pour l'analyse de signaux non-stationnaires, il est naturel de considérer les distributions de la classe de Cohen pour représenter les signaux. Pour tout couple de signaux (x_i, x_j), le noyau reproduisant associé à la distribution temps-fréquence $C_{x,\sigma}$ donnée est définie par le produit scalaire des représentations de deux signaux, $C_{x_i,\sigma}$ et $C_{x_j,\sigma}$, à savoir

$$\kappa_\sigma(x_i,x_j) = \iint |\Phi_\sigma(\nu,\tau)|^2 \; A_{x_i}(\nu,\tau) \; \overline{A_{x_j}(\nu,\tau)} \; \mathrm{d}\nu \; \mathrm{d}\tau.$$

Il est souvent plus commode d'exprimer celui-ci en coordonnées polaires, en particulier pour les distributions temps-fréquence à noyau radialement Gaussien. Dans ce dernier cas, on écrit le noyau reproduisant selon

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

où les fonctions d'ambiguïté sont exprimées en coordonnées polaires. L'usage de ce noyau reproduisant permet aux diverses méthodes de reconnaissance de formes à noyau d'opérer sur les représentations temps-fréquence à noyau radialement Gaussien, comme étudié dans le cas général dans [7]. Nous allons à présent considérer la mise en œuvre du critère d'alignement noyau-cible dans le domaine temps-fréquence grâce à ce noyau reproduisant.

3 Alignement et distributions radialement Gaussiennes

Le critère d'alignement noyau-cible est une mesure de similarité entre les données transformées selon l'application non linéaire associée au noyau reproduisant, et les étiquettes donnant la classe des signaux. Pour un problème à 2 classes, et étant donné un ensemble d'apprentissage $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ de n signaux x_k accompagnés chacun de leur étiquette $y_k = \pm 1$, le critère d'alignement noyau-cible est défini par

$$\mathcal{A}(K_\sigma, K_c) = \frac{\langle K_\sigma, K_c\rangle_F}{n\|K_\sigma\|_F}, \tag{1}$$

où $\langle\cdot,\cdot\rangle_F$ et $\|\cdot\|_F$ sont le produit scalaire de Frobénius et la norme de Frobénius, K_σ la matrice de Gram de terme général $\kappa_\sigma(x_i,x_j)$ pour $i,j = 1,\ldots,n$, et $K_c = \boldsymbol{y}\,\boldsymbol{y}^t$ la matrice cible avec $\boldsymbol{y} = [y_1 \ldots y_n]^t$. Cristianini *et al.* suggèrent dans [10] d'utiliser le critère d'alignement afin de rechercher le noyau reproduisant le mieux adapté à la résolution d'un problème de classification donné. Cette démarche présente l'intérêt de ne nécessiter aucun coûteux apprentissage de la règle de décision, la sélection du noyau étant pratiquée *a priori*. Un lien direct avec l'erreur de généralisation assure la pertinence du critère.

La paramétrisation optimale est obtenue par maximisation de l'alignement, à savoir $\sigma^* = \arg\max_\sigma \mathcal{A}(K_\sigma, K_c)$, ce qui correspond à la maximisation du numérateur de l'expression (1), sous contrainte que son dénominateur soit constant. On peut alors écrire le problème d'optimisation sous contrainte suivant :

$$\max_\sigma \sum_{i,j=1}^{n} y_i\, y_j\, \kappa_\sigma(x_i,x_j), \tag{2}$$

sous la contrainte

$$\sum_{i,j=1}^{n} (\kappa_\sigma(x_i,x_j))^2 = V_0, \tag{3}$$

où V_0 est un paramètre de normalisation. En développant la fonction objectif à maximiser, on aboutit à

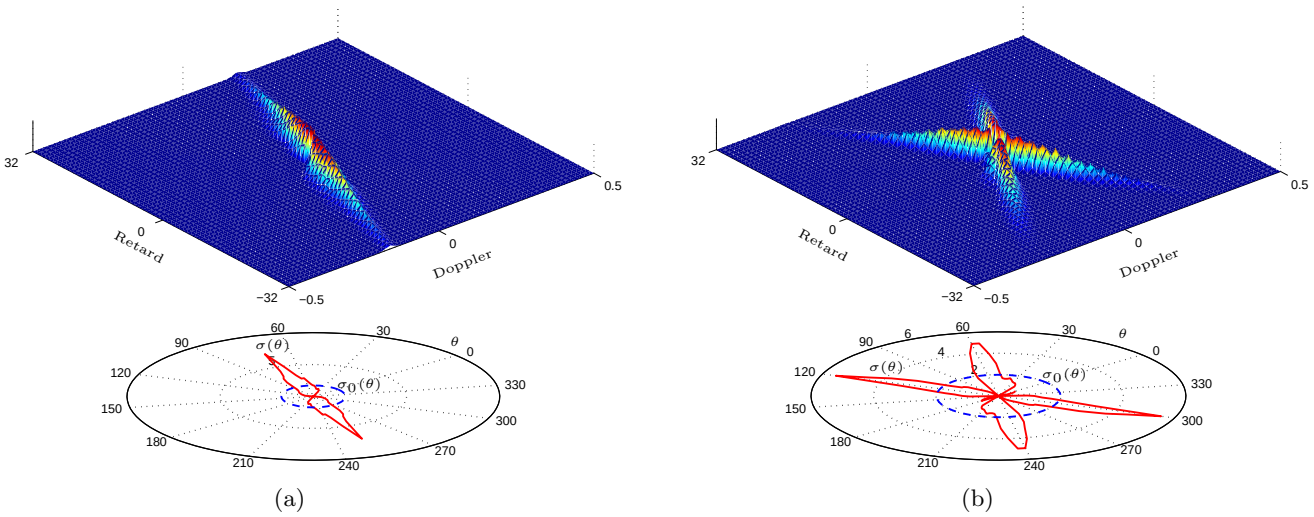
$$\begin{aligned}
&\sum_{i,j=1}^{n} y_i y_j\, \kappa_\sigma(x_i,x_j) \\
&\quad= \sum_{i,j=1}^{n} y_i y_j \iint r \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}} \; \mathrm{d}r \; \mathrm{d}\theta \\
&\quad= \iint r \Big[\sum_{i,j=1}^{n} y_i y_j \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \Big] \; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}} \; \mathrm{d}r \; \mathrm{d}\theta
\end{aligned} \tag{4}$$

On retrouve la fonction objectif à maximiser présentée dans [1], à savoir $\iint r|A_x(r,\theta)|^2 \; \mathrm{e}^{-r^2/\sigma^2(\theta)} \; \mathrm{d}r \; \mathrm{d}\theta$, où la partie dépendante du signal, $|A_x(r,\theta)|^2$, est substituée par la représentation équivalente $\sum_{i,j} y_i y_j \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)}$ qui ne dépend que de l'ensemble d'apprentissage, signaux et étiquettes. Il est à noter que cette dernière peut être évaluée préalablement à toute optimisation. On peut alors avoir recourt à l'algorithme d'optimisation de descente de gradient proposé dans [1], avec la même complexité calculatoire une fois la représentation équivalente évaluée. Pour cela, on relâche la contrainte (3), coûteuse en temps de calcul, en la remplaçant par la contrainte sur le volume de la fonction de paramétrisation, selon $\int \sigma^2(\theta)\mathrm{d}\theta = V_0'$, comme préconisé dans [1]. Dans ce qui suit, on présente la mise en œuvre de l'algorithme proposé.

L'algorithme nécessite une discrétisation du plan Doppler-retard, que l'on opère comme proposée dans [1]. Le noyau reproduisant est alors donné par

$$\kappa_\sigma(x_i,x_j) = \sum_{r,\theta} r \, A_{x_i}(r,\theta) \, \overline{A_{x_j}(r,\theta)} \; \mathrm{e}^{-(r\Delta_r)^2/\sigma^2(\theta)}.$$

pour la distribution temps-fréquence à noyau radialement Gaussien, avec $\Delta_r = 2\sqrt{\pi/l}$, l étant la longueur des signaux échantillonnés. En reprenant la fonction objective dans (4), le problème d'optimisation s'écrit en coor-



(a) (b)

FIG. 1 – Résultats obtenus pour la 1ère application (a) et la 2ème application (b). En haut : fonction de paramétrisation optimale résultante. En bas : son contour en rouge, et le contour initial en bleu, en coordonnées polaires.

données polaires selon

$$\max_{\sigma} \sum_{r,\theta} r \Big[\sum_{i,j=1}^{n} y_i y_j \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \Big] \; \mathrm{e}^{-(r\Delta_r)^2/\sigma^2(\theta)}, \tag{5}$$

sous la contrainte

$$\sum_{\theta} \sigma^2(\theta) = V_0'.$$

Pour résoudre ce problème d'optimisation avec contrainte, on considère l'algorithme de type descente de gradient alternée suivant. A l'itération $k+1$, on opère dans un premier temps une mise à jour de la solution selon

$$\sigma_{k+1}(\theta) = \sigma_k(\theta) + \mu_k \, \frac{\partial f}{\partial \sigma_k(\theta)},$$

où μ_k est un paramètre contrôlant la vitesse de convergence, et f la fonction objective à maximiser dans (5), et dont le gradient évalué en $\sigma_k(\theta)$ est défini par le vecteur $\Big[\frac{\partial f}{\partial \sigma_k(0)}, \cdots, \frac{\partial f}{\partial \sigma_k(l-1)}\Big]$, avec

$$\frac{\partial f}{\partial \sigma_k(\theta)} = \frac{2\Delta_r^2}{\sigma_k^3(\theta)} \sum_{r} r^3 \; \Psi(r,\theta) \; \mathrm{e}^{-(r\,\Delta_r)^2/\sigma^2(\theta)}.$$

Dans cette expression, la représentation équivalente, donnée par l'expression

$$\Psi(r,\theta) = \sum_{i,j} y_i y_j A_{x_i}(r,\theta) \overline{A_{x_j}(r,\theta)}, \tag{6}$$

est évaluée préalablement à la phase d'optimisation. Dans une seconde étape, on prend en compte la contrainte en projetant la solution sur l'ensemble des fonctions admissibles, ce qui revient à normaliser $\sigma_{k+1}(\theta)$ à chaque itération selon $\|\sigma_{k+1}(\theta)\|/V_0'$.

On insiste sur le fait que la représentation $\Psi(r,\theta)$ peut être calculée dans une étape d'initialisation. De plus, l'expression (6) se prête à un calcul itératif ne nécessitant pas de conserver en mémoire l'ensemble des fonctions d'ambiguïté des signaux de la base d'apprentissage. Une fois la représentation $\Psi(r,\theta)$ obtenue, la technique d'optimisation devient indépendante de la taille de l'ensemble d'apprentissage. Ceci n'est pas le cas pour les approches telles que [3], qui nécessitent l'évaluation des représentations temps-fréquence de chaque élément de l'ensemble d'apprentissage, à chaque itération de l'algorithme d'optimisation. Pour cette raison certainement, les auteurs de [3] se restreignent à un ensemble d'apprentissage de 15 signaux pour chaque classe.

4 Expérimentations

On considère successivement deux problèmes de classification de deux familles de 200 signaux de taille 64, à modulation fréquentielle linéaire, noyés dans un bruit blanc Gaussien de variance 4. Ceci correspond à un rapport signal-bruit, défini par le rapport des puissances du signal utile et du bruit, de l'ordre de -8 dB. La première application concerne des signaux à modulation fréquentielle croissante, de 0.1 à 0.25 pour la première classe, et de 0.25 à 0.4 pour la seconde, en échelle fréquentielle normalisée. La figure 1(a) en haut présente la fonction de paramétrisation à profil Gaussien ainsi obtenue, ce qui montre la pertinence de cette région dans le plan Doppler-retard pour la classification. La figure 1(a) en bas illustre son contour en rouge $\sigma(\theta)$, ainsi que le contour initial $\sigma_0(\theta)$ en bleu avant optimisation. Ce dernier est déterminé par la contrainte de volume, que l'on a fixé à $V_0' = 2$. Dans une seconde application, on propose d'étudier le cas où les régions des signaux des deux classes sont distinctes dans le plan Doppler-retard. On considère pour cela des signaux comportant une modulation fréquentielle linéairement croissante pour la première classe, de 0.1 à 0.4, alors qu'elle décroît de 0.4 à 0.1 pour la seconde classe. Dans la figure 1(b), on représente la fonction de paramétrisation ainsi obtenue. Elle correspond à un filtrage de l'information pertinente des deux régions d'intérêt pour la classification.

Pour illustrer la maximisation de l'alignement, on représente sur la figure 2 l'évolution moyenne sur 20

	1ère application		2ème application	
	Taux d'erreur (%)	Nombre de SV	Taux d'erreur (%)	Nombre de SV
Distribution de Wigner	19.41 ± 1.23	161.9 ± 4.97	19.41 ± 1.34	164.3 ± 5.62
Distribution optimale	16.89 ± 2.01	65.2 ± 4.46	17.81 ± 1.72	83.85 ± 5.65

TAB. 1 – Comparaison du taux d'erreur (%) et du nombre de vecteurs support (SV) pour un classifieur SVM associé d'une part à la distribution de Wigner, et d'autre part à la distribution optimale, pour chacune des deux applications.

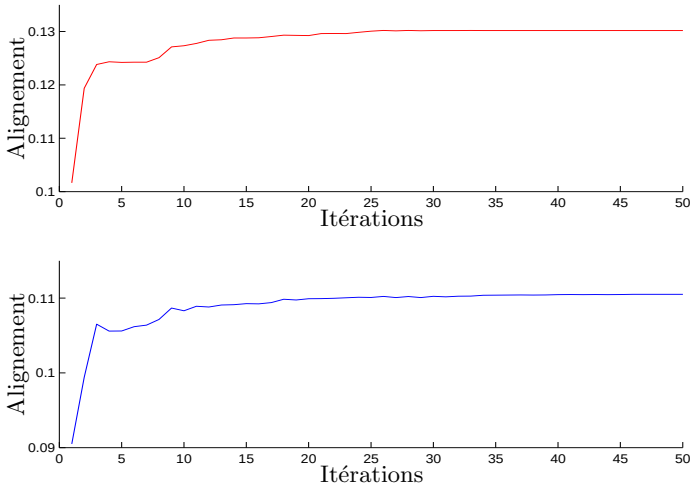


FIG. 2 – Evolution de l'alignement, moyenné sur 20 réalisations, pour la 1ère (en haut) et la 2ème (en bas) application.

réalisations de ce paramètres au cours des itérations, pour chacune des deux applications. Ceci met en évidence la convergence de l'algorithme vers une valeur maximale de l'alignement, malgré la substitution de la contrainte $\|K_\sigma\| = V_0$ par $\int \sigma^2(\theta)\,\mathrm{d}\theta = V_0'$.

Afin d'illustrer la pertinence de cette stratégie, on propose d'estimer l'erreur de classification obtenue à partir d'un classifieur de type Support Vector Machines (SVM), associée à chacune des distributions temps-fréquence : la distribution de Wigner et la distribution optimale. Le tableau 1 présente, en les moyennant sur 20 réalisations, le taux d'erreur obtenu sur un ensemble de test de 2000 signaux, et le nombre de support vecteurs correspondant. Non seulement la distribution optimale minimise l'erreur de classification, mais aussi conduit à une division par deux environ du nombre de vecteurs support. Ceci est principalement dû, d'une part au caractère optimal de la distribution ainsi obtenue, et d'autre part au caractère régularisant de ce traitement.

5 Conclusion

Le critère d'alignement noyau-cible s'avère très pertinent pour paramétrer des distributions temps-fréquence dans le cadre de problèmes de classification. Dans le cas particulier d'une distribution à noyau radialement Gaussien, nous avons montré que la maximisation de ce critère se réduit à un problème d'optimisation classique, permettant ainsi de reprendre des techniques de résolution qui ont fait leurs preuves dans la communauté temps-fréquence. La pertinence de notre approche est soutenue par des expérimentations, qui montrent une amélioration sensible des performances de classification de SVM, ainsi qu'une diminution du nombre de vecteurs support.

Références

[1] R. Baraniuk and D. Jones, "Signal-dependent time-frequency analysis using a radially gaussian kernel," *Signal Processing*, vol. 32, no. 3, pp. 263–284, 1993.

[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] M. Davy and C. Doncarli, "Optimal kernels of time-frequency representations for signal classification," in *Proc. of the IEEE International Symposium on Time-Frequency and Time-Scale analysis*, (Pittsburgh, USA), pp. 581–584, IEEE Signal Processing Society, Oct. 1998.

[4] C. Doncarli and N. Martin, *Décision dans le plan temps-fréquence*. Paris : Hermès Sciences, Traité IC2, 2004.

[5] 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.

[6] A. Rakotomamonjy, X. Mary, and S. Canu, "Non-parametric regression with wavelet kernels," *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, pp. 153–163, 2005.

[7] P. Honeine, C. Richard, and P. Flandrin, "Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence," in *Actes du XX*ème *Colloque GRETSI sur le Traitement du Signal et des Images*, (Louvain-la-Neuve, Belgium), 2005.

[8] P. Honeine, C. Richard, and P. Flandrin, "Time-frequency learning machines," *IEEE Transactions on Signal Processing*, 2006. (in press).

[9] P. Honeine, C. Richard, P. Flandrin, and J.-B. Pothin, "Optimal selection of time-frequency representations for signal classification : A kernel-target alignment approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toulouse, France), May 2006.

[10] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," in *Innovations in Machine Learning : Theory and Application* (D. Holmes and L. Jain, eds.), pp. 205–255, Springer Verlag, 2006.