Optimisation d'ondelettes pour la classification

Marie-Françoise Lucas[1], Eric Hitti[2], Christian Doncarli[1]

[1]Institut de Recherche en Cybernétique de Nantes (IRCCyN), U.M.R. 6597
Ecole Centrale de Nantes, Université de Nantes, Ecole des Mines de Nantes
1, rue de la Noë, BP 92101, 44321 Nantes Cedex 03, France

[2]Faculté de Pharmacie - LMPP, 2 avenue Pr. Léon Bernard 35043 Rennes cedex

Marie.lucas@irccyn.ec-nantes.fr, Eric.Hitti@univ-rennes1.fr
Christian.Doncarli@irccyn.ec-nantes.fr

Résumé –
Dans un contexte de classification supervisée où l'on dispose d'une population d'apprentissage expertisée, on recherche, parmi un ensemble de représentations temps-échelle discrètes, la représentation la plus adaptée aux signaux à classer, en optimisant la séquence génératrice à l'origine de la décomposition (et donc la forme de l'ondelette mère). Pour cela, on définit un critère de contraste portant sur la population d'apprentissage, et on montre que sa maximisation conduit à la minimisation du taux de mal classés sur une population test disjointe.

Abstract –
In a supervised classification context, be given an expertised learning set of signals, we search among a set of discret time-scale representations the most adapted to the signals to classify, by optimizing the filter used for the decomposition (and then the mother wavelet). In this aim, we define a contrast criteria calculated on the learning set, and we show that his maximisation corresponds to minimise the bad-classified rate on an independent testing set.

1 Introduction

Le cadre de ce travail est celui de la classification supervisée de signaux non stationnaires dans le cas où on ne dispose pas de modèle mais où l'on utilise une transformée du signal, et notre objectif est d'optimiser cette transformée vis à vis des classes d'apprentissage. Parmi de telles approches, nous citerons en particulier une approche temps-fréquence [1], réalisant une optimisation conjointe du noyau et de la distance par minimisation de la probabilité estimée de l'erreur de classification, et une approche temps-échelle [2] recherchant au sens du critère de Fisher une meilleure base d'ondelettes à partir d'une décomposition en paquets d'ondelettes. Dans cette dernière méthode cependant, le choix de l'ondelette mère n'entre pas dans le cadre de la procédure d'optimisation, mais est fixé a priori.

Nous nous intéressons ici aux transformées sur des bases discrètes d'ondelettes (ou sur des frames) correspondant à une meilleure base de paquets d'ondelettes. Notre objectif est d'optimiser la séquence h génératrice de la décomposition. Nous proposons une procédure générale de classification dans laquelle (i) la séquence h (et donc la meilleure base qu'elle induit) est optimisée et (ii) le critère de contraste utilisé correspond à la probabilité estimée de l'erreur de classification, en reprenant les résultats généraux de [1].

Ce papier est organisé de la façon suivante. Après avoir précisé la position du problème (section 2) nous présentons en section 3 la paramétrisation utilisée pour la représentation en ondelettes à optimiser. Dans la section 4, nous définissons le critère de contraste, dont l'optimisation conduit à la minimisation de la probabilité d'erreur de classification. En section 5, nous vérifions, sur des signaux synthétiques, l'hypothèse de gaussianité, et nous montrons la pertinence de la représentation et du critère de décision choisis en observant l'adéquation du critère de contraste calculé sur la population d'apprentissage avec les résultats de la classification obtenus sur une large population de signaux tests. Enfin, nous donnons les résultats de la procédure d'optimisation.

2 Position du problème

Typiquement, une procédure de classification avec optimisation est constituée d'un espace de représentation $\mathcal{R}$, d'une règle de décision dans cet espace et d'un critère dont on attend qu'il donne une estimation fidèle de la capacité de la représentation optimisée à bien classer. On dispose d'un ensemble d'apprentissage $\aleph$ comportant des données labellisées, et les classes ont des coûts d'erreur de classification égaux. Nous utiliserons les notations suivantes : c désigne le nombre de classes, $\aleph_i$ l'ensemble des signaux x d'apprentissage pour la classe ω_i. Dans l'espace de représentation $\mathcal{R}$, les données sont notées z, $\mathcal{Z}_i$ désignant l'ensemble d'apprentissage de la classe ω_i.

L'espace de représentation $\mathcal{R}$ correspond ici à la projection des signaux sur une meilleure base de paquets d'ondelettes (le terme d'ondelette étant pris dans un sens large). Cette base est engendrée par une séquence génératrice h, dépendant d'un vecteur paramètre θ (cf. § 3). L'espace $\mathcal{R}$ est donc paramétré par θ.

La règle de décision que nous avons retenue ici est la règle du plus proche représentant, (le représentant de la classe ω_i étant donné par la moyenne des éléments de $\mathcal{Z}_i$ et noté $\overline{z}_i$). Une donnée inconnue z sera affectée à la classe ω_{i_0} telle que $i_0 = \arg\min_i d(z, \overline{z}_{i=1,\ldots,c})$ où d est la distance retenue.

Le choix laissé à l'utilisateur concerne le filtre $h(\theta)$. Ce choix sera optimisé afin de minimiser la probabilité d'erreur de classification.

3 Transformée en ondelettes à optimiser

3.1 Meilleure base de paquets d'ondelettes

Un arbre de paquets d'ondelettes [3] est défini à partir d'une séquence h génératrice d'une multirésolution de $L^2(\mathbb{R})$. C'est un ensemble d'atomes temps-fréquence, organisés selon un arbre binaire dont les noeuds (paquets d'ondelettes) seront notés Ψ_j^p, où j définit le niveau de résolution et p la bande de fréquence analysée. Les coefficients de la décomposition sur les paquets sont obtenus par :

$$\begin{array}{l} c_0^0[k] = x[k] \\ c_{j+1}^{2p}[k] =\downarrow 2[c_j^p * \overline{h}][k] \\ c_{j+1}^{2p+1}[k] =\downarrow 2[c_j^p * \overline{g}][k] \end{array} \qquad (1)$$

$$\text{en notant} \quad \overline{f}[k] = f[-k]$$

où $g[k] = (-1)^{1-k} h[1-k]$ dans le cas d'ondelettes orthogonales. Dans le cas où h n'induit pas une multirésolution mais un *tight frame*, l'équation (1) réalise (avec la même définition de g) une décomposition sur des fonctions qui ne sont plus des ondelettes au sens strict du terme, mais que nous appellerons également ondelettes.

La décomposition étant fortement redondante, on extrait une base discriminante selon l'algorithme de Saito et Coifman [2] dont le principe est le suivant : étant donnée une population d'apprentissage $\aleph$, on calcule pour chaque noeud Ψ_j^p son pouvoir discriminant vis à vis de $\aleph$ et on recherche l'ensemble des noeuds les plus discriminants et constituant une base, selon la stratégie classique de Wickerhauser et Coifman [5]. La décomposition de x de longueur N sur la meilleure base (ou frame) est définie par :

$$BP_x^h = \{c_j^p[k]\}_{k=0,N/2^j-1}, (j,p)/\psi_j^p \in \text{meilleure base}\} \qquad (2)$$

l'exposant h indiquant que la meilleure base dépend de la séquence génératrice. C'est cette décomposition BP_x^h, éventuellement réduite à ses composantes les plus discriminantes, que nous appellerons l'individu z dans $\mathcal{R}$.

3.2 Paramétrisation de la séquence h

Les fonctions de la meilleure base sont complètement déterminées par la donnée de h et par la procédure de sélection réalisée à partir de la population d'apprentissage selon le schéma décrit ci-dessus. C'est donc le paramètre θ dont dépend ce filtre qu'il s'agit d'optimiser. Pour engendrer un *tight frame* ou une base orthogonale d'ondelettes [4], les coefficients de h doivent respecter $M/2+1$ contraintes (pour un filtre de longueur M), que nous qualifierons de contraintes structurelles :

$$\begin{array}{l} \sum_n h[n] = \sqrt{2} \\ \sum_n h^2[n] = 1 \\ \sum_n h[n] \cdot h[n-2k] = 0 \text{ si } k \neq 0 \end{array} \qquad (3)$$

Il reste donc $M/2-1$ degrés de liberté (contraintes utilisateur) pour choisir les coefficients du filtre h. Dans le contexte de la classification, la seule contrainte utilisateur va être la minimisation de la probabilité d'erreur de classification. A titre d'exemple [4] , pour $M=6$, il y a 4 contraintes structurelles, et il reste 2 degrés de liberté ($\theta = [a \quad b]$) pour définir h :

$$\begin{array}{l} i = 0,1 : \\ \quad h[i] = [(1 + (-1)^i \cos(a) + \sin(a)) \\ \qquad (1 - (-1)^i \cos(b) - \sin(b)) \\ \qquad +(-1)^i 2 \sin(b)\cos(a)](4/\sqrt{2}) \\ \\ i = 2,3 : \\ \quad h[i] = [(1 + \cos(a-b) + (-1)^i \sin(a-b))](2/\sqrt{2}) \\ \\ i = 4,5 : \\ \quad h[i] = (1/\sqrt{2}) - h(i-4) - h(i-2) \end{array} \qquad (4)$$

C'est cette structure que nous avons utilisée pour réaliser les simulations présentées en section 5.

4 Critère de probabilité d'erreur

4.1 Définition du critère

Une erreur de classification se produit chaque fois qu'un individu z est affecté à ω_j alors qu'en réalité il appartient à $\omega_{i \neq j}$. La probabilité d'erreur correspondante est :

$$P_e(j|i) = P(z \text{ affecté à } \omega_j | z \in \omega_{i \neq j})$$

et la probabilité d'erreur totale s'exprime alors :

$$P_e = \frac{1}{c} \sum_i \sum_{j \neq i} P_e(j|i)$$

Si on appelle $d_{j|i}^\theta$ la variable aléatoire "*distance d'un individu de la classe i au représentant de la classe j*", et $e_{j|i}^\theta = d_{j|i}^\theta - d_{i|i}^\theta$, alors les signaux mal classés selon la règle du plus proche représentant correspondent aux réalisations négatives de $e_{j|i}^\theta$. Sous l'hypothèse que $e_{j|i}^\theta$ est à distribution gaussienne (cf. § 5), $P_e^\theta(j|i) = Prob(e_{j|i}^\theta < 0)$ s'écrit :

$$P_e^\theta(j|i) = \frac{1}{\sqrt{(2\pi)}\sigma_{j|i}^\theta} \int_{-\infty}^0 \exp -\frac{1}{2}\left(\frac{u - m_{j|i}^\theta}{\sigma_{j|i}^\theta}\right)^2 du = Q\left(\frac{m_{j|i}^\theta}{\sigma_{j|i}^\theta}\right) \qquad (5)$$

avec :

$$\begin{array}{l} Q(u) = \frac{1}{\sqrt{(2\pi)}} \int_u^{+\infty} \exp(\frac{-u^2}{2}) du \\ m_{j|i}^\theta = E[e_{j|i}^\theta] \qquad (\sigma_{j|i}^\theta)^2 = var[e_{j|i}^\theta] \end{array}$$

En utilisant les valeurs $\widehat{m}^{\theta}_{j|i}$, $\widehat{\sigma}^{\theta}_{j|i}$ estimées sur la population d'apprentissage, la probabilité d'erreur estimée, fonction de θ et de la distance retenue, s'écrit :

$$\widehat{P}_e(\theta, d) = \frac{1}{c} \sum_i \sum_{j \neq i} Q(\frac{\widehat{m}_{j|i}}{\widehat{\sigma}_{j|i}}) \quad (6)$$

Dans la pratique, nous retiendrons le critère suivant :

$$CPE(\theta, d) = -\log(\widehat{P}_e(\theta, d)) = -\log(\frac{1}{c} \sum_i \sum_{j \neq i} Q(\frac{\widehat{m}^{\theta}_{j|i}}{\widehat{\sigma}^{\theta}_{j|i}})) \quad (7)$$

4.2 Mise en oeuvre

Le schéma général de la procédure d'optimisation se présente de la manière suivante :

- pour différentes distances d, pour différentes longueurs M de séquence h, maximiser le critère vis-à-vis de θ
- retenir la meilleure combinaison (d, M, θ).

Pour un θ fixé, les différentes étapes du calcul du critère sont résumées ci-dessous :

- calcul de $h(\theta)$ selon la définition 4
- décomposition des individus de la population d'apprentissage $\aleph$ sur la base de paquets d'ondelettes induite par $h(\theta)$ selon l'équation 1
- recherche de la meilleure base $BP^h_x(\theta)$; $\mathcal{Z}$ est la projection de $\aleph$ sur cette base
- calcul des $d^{\theta}_{j|i}$
- estimation des $\widehat{m}^{\theta}_{j|i}$, $\widehat{\sigma}^{\theta}_{j|i}$
- calcul du critère selon l'équation 7.

5 Résultats et conclusion

La pertinence de la représentation et du critère ainsi que les résultats de l'optimisation sont évalués sur des simulations : on dispose de 2 classes de chirps (signaux utilisés et décrits dans [1]) bruités avec un RSB de 0dB, 50 individus par classe dans la population d'apprentissage et 1000 individus par classe dans la population test (disjointe). L'équation des signaux est donnée par :

$$\begin{aligned} &\forall N \in [0; N-1] \\ &x[k] = A \sin 2\pi[f_0 k + \psi_0] \\ &\quad + B \sin 2\pi[\tfrac{f_2 - f_1}{2N} k^2 + f_1 k + \psi_1] + b[k] \end{aligned}$$

Les deux classes diffèrent par leur distribution de la fréquence f_2, les valeurs et distributions des différents paramètres $(A, f_0, \psi_0, B, f_1, f_2, \psi_1, N)$ étant :
$\{1; 0.25; \mathcal{U}[0; 1]; 1; 0.4; \mathcal{U}[0.10; 0.20]; \mathcal{U}[0; 1]; 128\}$
pour la classe ω_1 et
$\{1; 0.25; \mathcal{U}[0; 1]; 1; 0.4; \mathcal{U}[0.25; 0.35]; \mathcal{U}[0; 1]; 128\}$
pour la classe ω_2.

Le premier point de l'analyse des résultats concerne l'hypothèse sous-jacente de gaussianité de l'erreur $e^{\theta}_{j|i}$ permettant d'exprimer le critère de probabilité d'erreur selon l'équation 7. Pour un espace de représentation donné déterminé par un paramétrage θ fixé, la figure 1 montre la distribution de la variable aléatoire $e^{\theta}_{1|2} = d^{\theta}_{1|2} - d^{\theta}_{2|2}$ sur 1000 réalisations de signaux : on constate que l'hypothèse de gaussianité est réaliste.

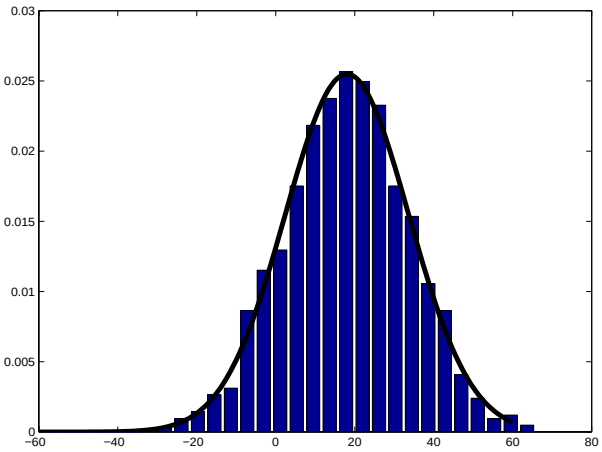

FIG. 1 – Distribution de $e^2_{1|2}$ estimée sur une population de 1000 signaux : l'hypothèse de gaussianité est réaliste.

Le deuxième aspect concerne la pertinence de la représentation associée au critère que nous avons défini : à chaque valeur du critère (fonction de θ) correspond un espace de représentation des signaux. La démarche proposée est pertinente si la maximisation du critère sur la population d'apprentissage correspond à la sélection d'un espace de représentation d'autant plus apte à bien classer les individus d'une population test, c'est-à-dire ayant une bonne capacité de généralisation. L'illustration de ces performances est donnée figure 2, où chaque point a pour ordonnée la valeur du critère calculé sur la population d'apprentissage et pour abscisse le taux d'individus mal classés obtenu sur un ensemble de test disjoint. L'ensemble des points est généré pour une part (points symbolisés par des '.') en modifiant un filtre h de longueur 6 (défini par l'équation 3) par une sélection aléatoire du paramètre $\theta = [a, b]$ (cf § 3.2), pour une autre part ('+' sur la figure) en utilisant des filtres prédéterminés de différentes longueurs et générateurs d'ondelettes classiques (Daubechies, splines, coiflets). On peut constater que la pertinence du critère et de la représentation associée est vérifiée dans la mesure où globalement sa maximisation sur l'ensemble d'apprentissage correspond à la minimisation du taux de mal classés sur l'ensemble test. Le dernier point abordé ici concerne le résultat de l'optimisation. Le cercle de la figure 2 indique la position correspondant à l'optimum $h(\theta_{opt})$ obtenu lors de la procédure d'optimisation décrite au § 4.2 et réalisée dans le cas où h est de longueur 6. Le point de départ choisi pour l'optimisation $\theta_{init} = [0, 0]$ correspond à l'ondelette de Haar visualisée figure 3. La valeur du critère en ce point initial est 0.4, et le taux de mal classés correspondant sur la population test est de 15%. Pour l'optimum obtenu ($\theta_{opt} = [0.2908, -0.3477]$), la valeur du critère est de 6.68 et le taux de mal classes sur la population test est de 2.3%. L'ondelette optimale correspondante est représentée figure 4.

On constate donc que l'optimisation de la forme de l'on-

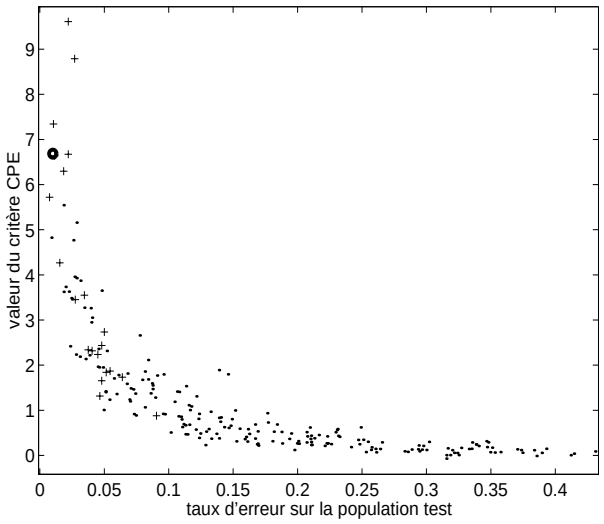

FIG. 2 – Capacité de généralisation de la représentation associée au critère : globalement, maximiser le critère CPE sur la population d'apprentissage équivaut à minimiser le taux de mal classés sur une population test. Chaque point correspond à une séquence h de longueur 6 fonction de θ tiré aléatoirement (.), ou à une séquence h de longueur ≥ 6 génératrice d'une ondelette classique (+). Résultat de l'optimisation : le cercle (o) correspond à l'optimum $h(\theta_{opt})$ obtenu par optimisation dans le cas où h est de longueur 6.

delette mère à travers le paramètre θ permet d'améliorer de façon conséquente les performances de classification par rapport aux résultats fournis par une ondelette classique comme l'ondelette de Haar (et que l'on peut utiliser sans connaissance *a priori* sur les signaux), sachant que dans les deux cas l'espace de représentation des signaux est obtenu après la recherche d'une meilleure base (au sens de la classification) de paquets d'ondelettes. Ainsi, l'optimisation de la forme de l'ondelette mère constitue un apport complémentaire substantiel à une simple recherche de meilleure base.

D'autre part, en revenant sur la figure 2, on voit que la solution optimale (notée 'o' sur cette figure) recherchée dans un espace restreint correspondant à des séquences génératrices de faible longueur, et bien que peu régulière localement, se comporte très souvent mieux (en terme de capacité de généralisation) que celles (notées '+') correspondant à des ondelettes classiques définies à partir de séquences génératrices de longueur ≥ 6, ayant traditionnellement de bonnes propriétés de régularité, et pour lesquelles on a également recherché la meilleure base de paquets.

L'aspect localement irrégulier de l'ondelette obtenue s'explique par le fait qu'il s'agit d'une réalisation d'une séquence aléatoire, liée a la réalisation de l'ensemble d'apprentissage retenu. Toutefois, il est probable qu'une version plus régulière conduirait à de meilleurs résultats sur une population test. C'est pourquoi dans la suite de ce travail on étudiera les propriétés statistiques de la séquence aléatoire obtenue (ondelette optimale) : d'une part moyenne et dispersion pour des apprentissages de taille donnée et d'autre part, propriétés asymptotiques lorsque la taille de l'apprentissage devient infinie. Enfin on s'intéressera à la capacité de généralisation (concentration et convexité moyennes de la ligne obtenue dans le plan critère/performance) en fonction de la taille de l'apprentissage.

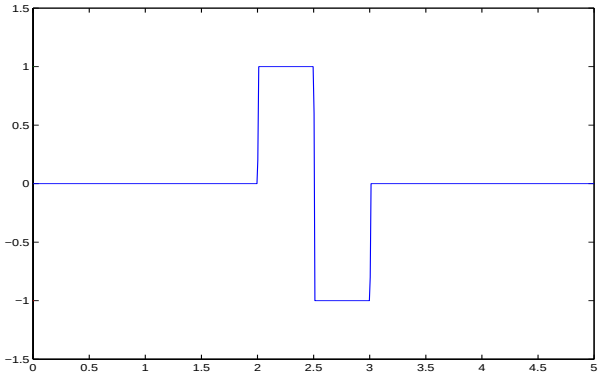

FIG. 3 – Fonction ondelette mère (ondelette de Haar) obtenue avec la séquence génératrice initiale $h(\theta_{init})$ où $\theta_{init} = [0, 0]$, point de départ de la procédure d'optimisation.

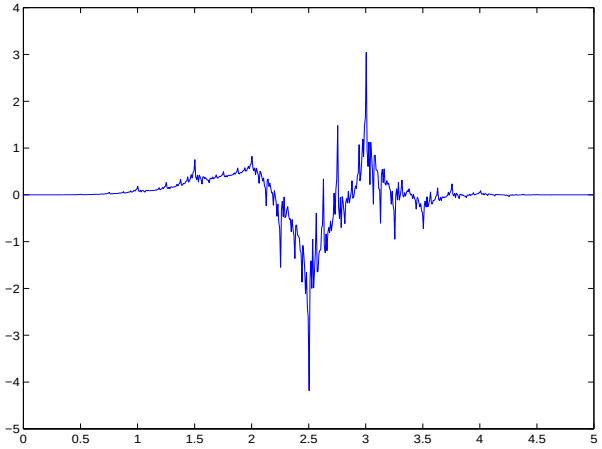

FIG. 4 – Fonction ondelette mère obtenue avec la séquence génératrice optimale $h(\theta_{opt})$ donnée par la procédure d'optimisation.

Références

[1] M. Davy, C. Doncarli, F. Boudreaux. *Improved Optimisation of Time-Frequency Based Classifiers.* IEEE Signal Processing letters, Vol 8 n°2, pp. 52-57, 2001.

[2] N. Saito, R.R. Coifman. *Local discriminant bases.* Wavelet applications in Signal and Image Processing II,A.F. Laine and M.A. Unser, Eds. Proc. SPIE vol 2303, 1994.

[3] S. Mallat. *A wavelet tour of signal processing.* Academic Press, 1998.

[4] C.S. Burrus, R.A. Gopinath, H. Guo. *Introduction to wavelets and wavelet transforms.* Prentice Hall, 1998.

[5] R.R. Coifman, M.V. Wickerhauser. *Entropy based algorithm for best basis selection.* IEEE Transaction on information theory, 38 :713-718, 1992.