Classification basée sur l'extraction conjointe d'attributs dans le plan temps-fréquence selon un critère d'information mutuelle

Edith GRALL- MAËS, Pierre BEAUSEROY

Laboratoire de Modélisation et Sûreté des Systèmes, Université de Technologie de Troyes
12 rue Marie Curie, BP 2060, 10010 Troyes Cedex, France

Edith.Grall-Maes@univ-troyes.fr, Pierre.Beauseroy@ univ-troyes.fr

Résumé – La méthode proposée concerne la classification de signaux basée sur une extraction automatique et conjointe d'attributs, dans le cas de processus non stationnaires uniquement décrits à l'aide d'une base d'exemples étiquetés. Les attributs sont définis par le résultat de transformations appliquées à la distribution du Wigner-Ville du signal à classer. Chaque transformation est sélectionnée au sein d'une famille de transformations paramétrées. Les valeurs des paramètres sont optimisées afin de maximiser l'information discriminante portée conjointement par les attributs, compte tenu d'une base d'apprentissage. L'information discriminante est mesurée à l'aide d'un critère d'information mutuelle, basé sur l'estimation des lois de distribution conjointes des attributs. Afin de prendre en compte toute l'information portée par les attributs et d'assurer une cohérence avec la phase d'extraction, le classifieur utilise également l'estimation des lois de probabilités. Le principe obtenu présente l'intérêt de ne faire aucune hypothèse sur les lois suivies par les attributs conditionnellement à chacune des classes. La méthode a été appliquée à un problème de classification de signaux de l'électroencéphalogramme du sommeil. De bonnes performances ont été obtenues à partir de l'extraction conjointe de deux attributs.

Abstract – The presented method deals with the classification of signals using an automatic and joint features extraction, for non stationary processes only described through training data. The features are defined by the result of transformations applied to the signal Wigner-Ville distribution. Each transformation is chosen from a family of parametrized transformations. The parameters values are optimized using training data for maximising the discriminant information contained jointly in the features. The measure of the discriminant information is based on a mutual information criterion, which uses estimation of the joint probability densities. To take into account all the information contained in the features and fit with the extraction rule, the classifier is also based on the estimated densities. An advantage of this method is that no hypothesis is done about the conditional distributions over the class. The method has been applied to a classification problem of sleep electroencephalogram signals. Good results have been obtained using a joint extraction of two features.

1. Introduction

Les représentations temps-fréquence constituent un outil important pour l'analyse et la classification des signaux non stationnaires [1]. Une classe importante de détecteurs est basée sur la corrélation temps-fréquence qui consiste à comparer la représentation du signal à classer à une représentation de référence [2]. Le détecteur ainsi obtenu possède une structure imposée qui est linéaire, ce qui peut constituer une limitation pour certains problèmes. En outre, lorsque le processus considéré est simplement décrit à l'aide d'une base d'apprentissage, la référence doit être estimée à partir de cette base, en tenant compte de la difficulté inhérente à la dimension importante des données fournies par les représentations temps-fréquence.

Une autre approche consiste à construire un classifieur basé sur des attributs extraits de la représentation temps-fréquence. Dans ce cadre il existe différentes méthodes qui utilisent des informations connues a priori, par exemple pour la détection de signaux à modulation de fréquence linéaire ou hyperbolique [3-4]. D'autres méthodes sont basées sur une extraction arbitraire de signatures ou d'attributs [5-6].

Une méthode antécédante à celle qui est proposée a pour objectif l'extraction d'attributs permettant la caractérisation de processus uniquement décrits à l'aide d'une base d'exemples étiquetés et en l'absence de connaissance statistique a priori [7]. Tout attribut est défini comme le résultat généré par une transformation appliquée à la distribution de Wigner-Ville du signal à classer. Chaque transformation est sélectionnée au sein d'une famille de transformations paramétrées, afin de maximiser l'information discriminante portée par l'attribut. Le caractère discriminant est estimé à partir d'une base d'apprentissage à l'aide d'un critère d'information mutuelle.

Néanmoins l'optimisation individuelle des attributs ne permet pas de limiter la redondance de l'information portée par les différents attributs et l'accroissement du nombre d'attributs peut s'avérer sans intérêt pour la classification ultérieure. En outre, selon le classifieur choisi, l'information discriminante portée par les attributs peut n'être que partiellement prise en compte. Pour maximiser l'information discriminante portée par un ensemble d'attributs et la prendre en compte lors de la classification, il est nécessaire d'extraire

conjointement les attributs et d'associer un principe de classification adapté. Le principe d'une telle méthode est développé dans le second paragraphe. La méthode est ensuite appliquée à un problème de classification de complexes K et d'ondes delta dans l'électroencéphalogramme du sommeil ; les résultats de cette application sont présentés dans le paragraphe 3. Les conclusions sont ensuite exposées.

2. Principe de la méthode

2.1 Principe général de la méthode d'extraction

Le principe général de la méthode consiste à considérer *d* familles $\mathcal{F}_{X_i}$(*i=1..d*) d'attributs X_i, et à déterminer les attributs $\tilde{X}_i$ qui maximisent l'information discriminante qu'ils portent conjointement.

Chaque famille d'attributs est définie par un ensemble de transformations paramétrées. Les valeurs des paramètres des transformations associées aux attributs sont déterminées pour maximiser la mesure du caractère discriminant des *d* attributs conjoints à partir d'une base d'exemples étiquetés.

Les transformations doivent être définies par un nombre limité de paramètres, afin de réduire le nombre de degrés de liberté du problème.

Le caractère discriminant est mesuré à l'aide d'un critère d'information mutuelle. Il présente l'intérêt de ne faire aucune hypothèse sur la nature des lois de probabilité des attributs conditionnellement à la classe d'appartenance.

2.2 Familles d'attributs considérées

Afin de caractériser des processus non stationnaires, les données sur lesquelles sont appliquées les transformations sont les échantillons de la distribution de Wigner-Ville, car elles fournissent une description conjointe en temps et en fréquence. Les transformations choisies permettent d'identifier des régions (domaines restreints) du plan temps-fréquence pour lesquelles il existe des propriétés très dépendantes de la classe. Elles peuvent s'exprimer par la composition de deux opérations. La première consiste à sélectionner une région temps-fréquence en pondérant la distribution de Wigner-Ville par une fenêtre bidimensionnelle. La seconde consiste à passer de la « distribution de Wigner-Ville restreinte » à la valeur de l'attribut à l'aide d'une application, qui permet d'exprimer une caractéristique précise de la représentation restreinte.

Les applications considérées sont définies par l'énergie (E), l'espérance temporelle ($\hat{t}$) et l'espérance fréquentielle ($\hat{f}$) de la représentation restreinte. Elles ont été fixées a priori et choisies de manière à fournir une description de l'énergie du signal et de sa répartition le long de l'axe temporel et de l'axe fréquentiel.

Les fenêtres bidimensionnelles de pondération permettent de mettre en valeur le contenu du signal spécifiquement dans les régions définies par les fenêtres. Les fenêtres choisies sont des fonctions gaussiennes définies par leur position centrale en temps et en fréquence (*t* et *f*), leurs dispersions (*A* et *B*) et leur orientation (ω). L'expression analytique est donnée par la relation (1) :

$$\pi(t,f) = exp\left(-\left(t^2\left(\frac{\cos^2\omega}{A^2}+\frac{\sin^2\omega}{B^2}\right)+f^2\left(\frac{\sin^2\omega}{A^2}+\frac{\cos^2\omega}{B^2}\right)+ft\sin 2\omega\left(\frac{1}{A^2}-\frac{1}{B^2}\right)\right)\right) \quad (1)$$

En outre, cette fonction a la propriété d'être dérivable par rapport à chacune des variables, ce qui constitue un avantage important pour la résolution d'un problème d'optimisation.

Chaque attribut est alors défini par cinq paramètres décrivant la fenêtre de pondération, et un type d'application choisi parmi trois possibles (E, $\hat{t}$ ou $\hat{f}$). Par exemple, l'attribut donné par l'énergie et en utilisant une fenêtre de pondération π s'exprime par la relation 2 :

$$E_{\pi(A,B,\omega)}(t,f) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\pi(t'-t,f'-f)W_x(t',f')dt'df' \quad (2)$$

Une famille est engendrée par tous les attributs pour un type (E, $\hat{t}$ ou $\hat{f}$) et un domaine $[t_0\pm\Delta t]x[f_0\pm\Delta f]$ fixés a priori. Ce domaine de variation associé au couple (*t,f*) permet d'associer un caractère local en temps et en fréquence à la famille. Un exemple de famille d'attributs est donné par :

$$\mathcal{F}_X = \left\{ X = E_\pi(t,f) \middle| (A,B,\omega) \in \Re^3, (t,f) \in [t_0 \pm \Delta t]x[f_0 \pm \Delta t] \right\} \quad (3)$$

2.3 Recherche des attributs optimaux

La recherche des attributs optimaux $\tilde{X}_i$(*i=1..d*) consiste à déterminer les paramètres $\tilde{t}_i, \tilde{f}_i, \tilde{A}_i, \tilde{B}_i$ et $\tilde{\omega}_i$ qui leur sont associés.

La sélection des attributs est basée sur un critère d'information mutuelle. En effet l'information mutuelle $I(C; X_1, X_2, ..X_d)$, où *C* est la variable aléatoire correspondant à la classe d'appartenance et X_i (*i=1..d*) sont les *d* variables aléatoires corrrespondant aux attributs, constitue une mesure de l'information portée conjointement par les *d* attributs sur la classe d'appartenance [8]. Cette mesure est basée sur la loi de distribution conjointe des variables aléatoires. Ce critère ne fait aucune hypothèse sur la nature des lois, au contraire de nombreux autres qui font l'hypothèse de distributions normales.

Le problème d'optimisation à résoudre consiste alors à trouver les attributs $\tilde{X}_i$ tels que

$$I(C; \tilde{X}_1, \tilde{X}_2, ..\tilde{X}_d) = \max_{X_1, X_2, ..X_d} I(C; X_1, X_2, ..X_d) \quad (4)$$

En supposant la probabilité de chaque classe connue ou estimée, la maximisation de $I(C; X_1, X_2, ..X_d)$ se ramène à la minimisation de l'entropie conditionnelle $H(C \mid X_1, X_2, ..X_d)$. La résolution de ce problème nécessite l'estimation de cette entropie. Elle s'obtient en discrétisant

les densités de probabilités conjointes calculées à l'aide de l'estimateur de Parzen :

$$H(C \mid X_1, X_2, ..X_d) = -\sum_i P_i \sum_{x_1} \sum_{x_2} ... \sum_{x_d} p_i(\mathbf{x}) \log \frac{P_i p_i(\mathbf{x})}{p(\mathbf{x})} \Delta x \tag{5}$$

où

- i représente l'indice de la classe
- P_i la probabilité de la classe i
- $\Delta x = \Delta x_1 \Delta x_2 .. \Delta x_d$
- x_i et Δx_i correspondent aux valeurs et à la largeur des intervalles obtenus par discrétisation de la variable aléatoire X_i
- $p(\mathbf{x}) = p(x_1, x_2, ..x_d)$ est l'estimation de la densité de probabilité conjointe en $(x_1, x_2, ..x_d)$

et $p_i(\mathbf{x}) = p_i(x_1, x_2, ..x_d)$ celle de la densité de probabilité conditionnelle à la classe i.

Les densités de probabilité sont estimées à partir de la base d'exemples étiquetés servant à décrire le processus. Pour que cette estimation soit représentative du processus, le nombre d'exemples doit être assez important par rapport à la dimension d.

Compte tenu de la forme de l'estimateur de Parzen et de la forme des fenêtres, les dérivées partielles des densités de probabilité par rapport aux variables associées aux fenêtres gaussiennes peuvent s'exprimer analytiquement. Le gradient de l'entropie conditionnelle admet donc une expression analytique. Par conséquent il est possible de déterminer les valeurs des paramètres optimaux en effectuant une optimisation numérique à l'aide d'un algorithme de type quasi-Newton.

2.4 Construction du classifieur

Le principe d'extraction d'attributs est basé sur l'optimisation de l'information discriminante, mesurée à partir des lois de distribution conjointes conditionnellement à chacune des classes. Le classifieur qui permet de passer de l'espace formé par les d attributs à la décision doit prendre en compte le mieux possible l'ensemble de cette information. Ceci assure en effet que le caractère optimal obtenu dans l'étape d'extraction n'est pas modifié dans l'étape de décision.

La statistique de décision la plus appropriée est celle basée sur l'estimation de la loi de distribution des attributs, obtenue pour l'estimation de l'entropie conditionnelle :

$$\lambda(\mathbf{x}) = \frac{P_0 p_0(\mathbf{x})}{P_1 p_1(\mathbf{x})} \underset{D_1}{\overset{D_0}{\gtrless}} \mu \tag{6}$$

où $\mathbf{x} = (x_1, x_2, ..., x_d)$

et μ est un seuil conditionné par le taux de fausse alarme

3. Application à l'EEG du sommeil

3.1 Introduction

La méthode décrite a été appliquée pour la classification de complexes K et d'ondes delta dans l'électroencéphalogramme du sommeil. Le complexe K est un signal transitoire important dans l'étude du sommeil. Néanmoins sa détection automatique est difficile car il peut prendre des formes très diverses, et parfois très similaires à celle des ondes delta.

La base de données utilisée est composée de 292 complexes K (classe C_0) et de 314 ondes delta (classe C_1).

L'expérience de classification effectuée était basée sur l'extraction conjointe de deux attributs. Tout d'abord, un couple de familles de transformations a été choisi. Pour cela, différents couples ont été considérés et les attributs conjoints associés à l'information mutuelle maximale ont été recherchés. Le couple sélectionné est celui qui est associé à la plus grande information mutuelle maximale. Parmi les deux familles sélectionnées, l'une correspond à des transformations de type énergie et l'autre à des transformations de type espérance temporelle.

3.2 Résultats

L'objectif de l'étude était tout d'abord d'évaluer l'intérêt d'utiliser une extraction conjointe. Une comparaison a donc été effectuée avec une extraction non conjointe, qui consiste à déterminer les attributs indépendamment en résolvant un problème d'optimisation par attribut. D'autre part l'utilisation d'un classifieur linéaire a été comparée à celle d'un classifieur basé sur l'estimation de la densité de probabilité, afin d'évaluer l'intérêt de choisir un classifieur adapté au principe d'extraction. Quatre configurations (extraction conjointe / indépendante, classifieur linéaire / basé sur les densités de probabilité) ont donc été considérées.

La figure 1 représente les lois de probabilités conditionnelles des attributs obtenus par une extraction conjointe, en utilisant l'ensemble de la base de signaux.

Les risques d'erreur de classification pour les quatre configurations ont été estimés selon une procédure de leave-one-out. Elle consiste à classer chaque échantillon à partir de l'ensemble d'extraction et de classification élaboré en utilisant uniquement les autres exemples de la base. Les résultats de classification sont donnés dans le tableau 1 et les courbes COR représentées sur la figure 2.

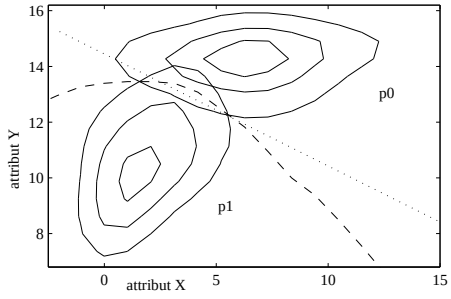

FIG. 1 : densités de probabilités conditionnelles estimées et frontière des deux classifieurs

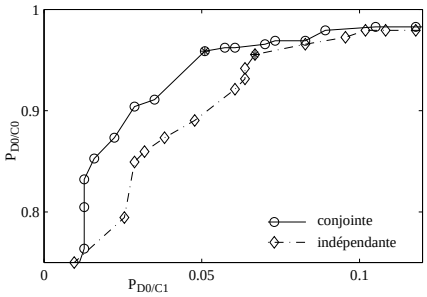

FIG. 2 : courbes COR (cas du classifieur basé sur la ddp) comparaison des extractions conjointe et indépendante

TAB. 1 : Résultats de classification associés au risque d'erreur minimum dans chacun des cas

extraction classifieur	indépend. linéaire	indépend. ddp	conjointe linéaire	conjointe ddp
erreur (%)	5,8	5,6	5,6	4,6
P D1/C0 (%)	4,1	4,4	4,1	4,1
P D0/C1 (%)	7,3	6,7	7,0	5,1

3.3 Analyse

Le tableau 1 montre que le classifieur basé sur la loi de probabilité permet de mieux exploiter l'information discriminante que le classifieur linéaire. En effet dans le cas du classifieur basé sur la loi de probabilité la forme de la frontière est adaptée à la forme des distributions, comme dans l'exemple présenté à la figure 1.

L'extraction conjointe permet d'améliorer les résultats par rapport à une extraction non conjointe (tableau 1 et figure 2). Lors d'une étude antérieure, des résultats avaient été obtenus en utilisant l'extraction indépendante de trois attributs et un classifieur linéaire. Le risque d'erreur minimum estimé était de 4,9% [7]. De meilleures performances (4,6%) sont obtenues avec seulement deux attributs à l'aide d'une extraction conjointe et d'un classifieur basé sur la loi de probabilité.

L'extraction conjointe, lorsqu'elle est associée à un classifieur approprié, présente donc un réel intérêt. Les performances présentées sont également supérieures à celles figurant dans différentes publications et obtenues avec d'autres méthodes [9-11]. La méthode proposée, bien que limitée en pratique à un faible nombre d'attributs à cause des problèmes d'estimation des densités de probabilité, permet donc d'obtenir d'excellents résultats de classification.

4. Conclusion

La méthode proposée permet d'élaborer un classifieur dans le cas de processus non stationnaires décrits par une base d'exemples. Son principe consiste à extraire de manière automatique et conjointe des attributs discriminants et à construire un classifieur adapté. Elle présente l'avantage de ne pas imposer une structure au classifieur, et d'être indépendante de la dimension des données fournies par la distribution de Wigner-Ville.

Chacun des attributs extrait est défini par une transformation paramétrée, sélectionnée au sein d'une famille afin de maximiser l'information discriminante portée conjointement par les attributs. Les principes d'extraction et de classification sont basés sur l'estimation des lois de probabilités. La méthode obtenue présente donc l'intérêt de ne faire aucune hypothèse sur les lois suivies par les attributs conditionnellement à chacune des classes.

Cette méthode a été validée dans le cadre du problème de la classification de complexes K et d'ondes delta dans l'électroencéphalogramme du sommeil. La classification basée sur l'extraction conjointe de deux attributs a fourni de très bonnes performances.

Références

[1] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Proc. Magazine*, pp. 21-67, 1992.

[2] P. Flandrin, "A time-frequency formulation of optimum detection," *IEEE Trans. on Acous.,Speech, Signal Processing*, vol. 36, pp.1377-1384, 1988.

[3] S. Barbarossa, "Analysis of multicomponent LFM signals by a combined Wigner-Hough transform," *IEEE Transactions on Signal Processing*, vol. 43, pp.1511-1515, 1995.

[4] A. Papandreou, S. Kay and G. F. Boudreaux-Bartels, "The use of hyperbolic time-frequency representations for optimum detection and parameter estimation of hyperbolic chirps," *IEEE-SP TFTS*, pp. 369-372, 1994.

[5] M. Shamsollahi, L. Senhadji et R. Lebouquin-Jeannes, "Détection de signatures temps-fréquence sur des crises d'épilepsie," *GRETSI*, pp. 1367-1370, 1997.

[6] R. Abeysekera and B. Boashash, "Methods of signal classification using the images produced by the Wigner-Ville distribution," *Pattern recognition letters* 12, pp.717-729, 1991.

[7] E. Grall-Maës and P. Beauseroy, "Features extraction for signal classification based on Wigner-Ville distribution and mutual information criterion", IEEE-SP TFTS, pp.589-592, Pittsburgh, 1998.

[8] T.M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.

[9] A. Da Rosa, B. Kemp, T. Paiva, F. Lopes da Silva, and H. Kamphuisen. "A model-based detector of vertex sharp waves and K-complexes in sleep electroencephalogram," *Electroencephalography and clinical neurophysiology*, vol. 78, pp.71-79, 1991.

[10] I.N. Bankman, V.G. Sigillito, R.A.Wise, and P.L. Smith, "Feature-based detection of K-complex wave in the human electroencephalogram using neural networks," *IEEE Trans. on Biomedical Engineering*, vol. 39, pp.1605-1610, 1992.

[11] B. Jansen, and P. Desai. "K-complex detection using multi-layer perceptron and recurrent network," International Journal of Bio-medical computing, vol. 37, pp.249-257, 1994.