Extraction d'attributs discriminants par optimisation de fonctions paramétrées

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@utt.fr, Pierre.Beauseroy@utt.fr

Résumé – Une méthode est proposée pour extraire automatiquement des attributs discriminants dans le cas d'un processus décrit à l'aide d'une base d'exemples étiquetés. Les attributs sont sélectionnés, à l'aide de familles de fonctions paramétrées, en déterminant les paramètres optimaux par rapport à un critère de séparabilité des classes. Les fonctions paramétrées choisies mesurent des caractéristiques correspondant aux moments d'ordre 0 ou 1 d'une représentation uni- ou bi-dimensionnelle pondérée. L'aspect continu des fonctions paramétrées permet d'explorer un ensemble infini d'attributs et d'éviter de traiter un problème de complexité combinatoire. Le critère mesurant la séparabilité des classes est basé sur les matrices de dispersion, et permet la sélection conjointe d'attributs. L'élaboration d'un classifieur linéaire, adapté aux attributs extraits est proposé. La méthode est appliquée à des signaux simulés décrits par leur représentation temporelle.

Abstract – A method is proposed for automatic extraction of discriminant features for processes described by sample sets. The features are selected from sets of parameterized mappings, choosing optimal parameters according to a criterion measuring the class separability. Parameterized mappings measure characteristics that are the zero- and first-order moments of a one- or two-dimensional weighted representation. The continuous parameterized mappings provide infinite feature sets and avoid a combinatorial problem. The criterion measuring the class separability uses scatter matrices and is adapted to the joint feature extraction. A linear classifier adapted to the extrated features is proposed. The method is applied to simulated signals described by their temporal representation.

1 Introduction

Pour procéder à la classification de signaux aléatoires, il est fréquent de faire précéder la phase de classification par une phase d'extraction d'attributs discriminants. Celle-ci permet de réduire la dimension de l'espace de représentation original, et fournit un nouvel espace de représentation compact. Les attributs, extraits à une fin de classification, présentent par ailleurs un intérêt pour la description du processus et la compréhension de phénomènes sous-jacents au problème étudié.

L'espace de représentation original à partir duquel les attributs sont extraits peut tout simplement correspondre aux données mesurées ou bien être un espace transformé facilitant la mise en évidence d'attributs discriminants. L'extraction est parfois facilitée par l'utilisation de connaissances *a priori* du processus. Généralement le nombre d'attributs proposés est grand, puis il est réduit à l'aide d'une phase de sélection d'attributs réellement discriminants. Compte tenu du problème d'explosion combinatoire qui apparaît lors de la sélection de m attributs parmi n, différentes approches permettant d'éviter la recherche exhaustive, telles que des méthodes sous-optimales [6], ou des méthodes d'optimisation de type "Branch and Bound" [7] ont été proposées.

L'approche proposée ici consiste à sélectionner des attributs dans des familles de fonctions paramétrées, en maximisant un critère mesurant la séparabilité des classes estimé à partir d'une base d'exemples étiquetés. Les attributs étant sélectionnés dans un ensemble infini de fonctions continues, leur extraction n'implique pas la résolution d'un problème combinatoire.

Les attributs mesurent des caractéristiques paramétrables dans des représentations uni- ou bi-dimensionnelles pondérées. Chaque attribut est défini par la caractéristique qu'il mesure et les valeurs des paramètres qui lui sont associées. L'extraction des attributs au sein d'un ensemble de familles consiste donc à déterminer les paramètres optimaux de façon à ce que les attributs maximisent un critère de mesure de séparabilité des classes. Lorsque les attributs extraits sont destinés à réaliser la classification de réalisations, l'élaboration d'un classifieur exploitant l'information discriminante est nécessaire.

La section 2 présente le principe de la méthode, les fonctions paramétrées, et le critère de mesure de séparabilité des classes. La section 3 décrit le classifieur. Des résultats obtenus sur un problème simulé sont ensuite exposés dans la section 4.

2 Extraction d'attributs

2.1 Principe

La méthode consiste à considérer d familles F_k $(k = 1..d)$ de fonctions, chacune paramétrée par un vecteur $\theta_k \in U_k \subset \mathbf{R}^n$, et à déterminer le vecteur $\tilde{\theta} = [\tilde{\theta_k}]_{(k=1..d)}$ optimal par rapport à un critère mesurant la séparabilité des classes.

Chaque famille F_k est définie par

$$F_k = \{f_{\theta_k} | \theta_k \in U_k \subset \mathbf{R}^n\} \tag{1}$$

où $f_{\theta_k} : L^2(\mathbf{R}^m) \rightarrow \mathbf{R}$ est une fonction

- à appliquer à la représentation R_x de la donnée observée x,

– mesurant une caractéristique fixée *a priori* de la représentation R_x pondérée,
– paramétrée par le vecteur $\theta_k = [p_{k,1}, p_{k,2} \dots p_{k,n}]$ où chaque $p_{k,i}$ est un paramètre de la fonction f_{θ_k}.

Ainsi pour une représentation R_x (de dimension m) l'attribut (de dimension 1) de la famille F_k et associé au vecteur θ_k est donné par $X_k = f_{\theta_k}(R_x)$.

Dans le cas où la donnée observée est un signal $x(t)$, la représentation R_x peut être tout simplement la représentation temporelle $x(t)$ ou être une autre représentation unidimensionnelle, telle que la représentation fréquentielle. La représentation peut aussi être bidimensionnelle en prenant une distribution temps-fréquence, qui permet de décrire conjointement le contenu temporel et fréquentiel du signal [5]. En pratique les représentations sont des données discrètes, de dimension m ; néanmoins la méthode peut aussi s'appliquer à des données continues.

Le vecteur optimal $\tilde{\theta}$ est celui qui vérifie :

$$\Phi_{S,F}(\tilde{\theta}) > \Phi_{S,F}(\theta) \quad \forall\, \theta = [\theta_k]_{(k=1..d)} \tag{2}$$

où Φ est un critère mesurant la séparabilité des classes, S est un ensemble d'apprentissage étiqueté donné, et F est un ensemble des familles F_k fixées *a priori*. Les familles engendées par des fonctions paramétrées par un vecteur θ_k de variables continues sont de dimension infinie. La recherche du vecteur optimal $\tilde{\theta}$ se ramène alors à la résolution d'un problème d'optimisation, et non à la résolution d'un problème combinatoire comme dans le cas d'une sélection d'attributs dans un ensemble donné de dimension finie. Le vecteur optimal étant déterminé en considérant les k familles conjointement, la dépendance statistique des données est prise en compte.

Dans une démarche où le nombre d'attributs k n'est pas déterminé *a priori* il est possible d'augmenter graduellement le nombre de familles en adoptant par exemple une sélection séquentielle des familles, ou des algorithmes plus complexes. La mesure du critère à chaque étape permet de savoir si l'ajout d'attributs apporte réellement de l'information discriminante.

2.2 Fonctions paramétrées

Les fonctions paramétrées f_{θ_k} doivent posséder un nombre limité de paramètres afin de restreindre la complexité du problème et obtenir un bon pouvoir de généralisation. Des liens entre pouvoir de généralisation, complexité et taille de l'ensemble d'apprentissage ont en effet été établis en théorie de l'apprentissage [8].

Les fonctions f_{θ_k} proposées sont similaires à celles utilisées dans [4]. Elles correspondent aux moments d'ordre 0 ou 1 de la représentation originale, uni- ou bidimensionnelle, pondérée par une fenêtre, qui est une fonction gaussienne. Ces fonctions ont été choisies car elles correspondent à des attributs usuels dans de nombreuses applications et leur interprétation est aisée, ce qui a son importance quand les attributs sont utilisés pour aider à la compréhension de phénomènes sous-jacents au processus étudié. En outre les familles fournissent une grande diversité d'attributs tout en possédant un nombre restreint de paramètres.

Dans le cas de données monodimensionnelles où la représentation est $x(t)$, la famille F_1 regroupant les moments d'ordre 1 est définie à l'aide des fonctions f_{θ_1} :

$$f_{\theta_1}(x) = \frac{\sum_j\ j\Delta t\ \psi_{\theta_1}(j\Delta t) x(j\Delta t)}{\sum_j\ \psi_{\theta_1}(j\Delta t) x(j\Delta t)} \tag{3}$$

où Δt est la période d'échantillonnage du signal, et ψ_{θ_1} est une fenêtre gaussienne paramétrée par son centre m_1 et sa dispersion A_1 :

$$\psi_{\theta_1}(j\Delta t) = \psi_{[A_1,m_1]}(j\Delta t) = \exp\left(-\frac{(j\Delta t - m_1)^2}{A_1^2}\right). \tag{4}$$

Dans le cas de données bidimensionnelles où la représentation est la distribution de Wigner-Ville W_x [1, 2], les fonctions f_{θ_1}, correspondant aux moments d'ordre 1 par rapport au temps, sont données par :

$$f_{\theta_1}(W_x) = \frac{\sum_k \sum_l t_k\ \psi_{\theta_1}(t_k,f_l) W_x(t_k,f_l)}{\sum_k \sum_l \psi_{\theta_1}(t_k,f_l) W_x(t_k,f_l)} \tag{5}$$

où $t_k = k\Delta t$ et $f_l = l\Delta f$ représentent les points d'échantillonnage dans le plan temps-fréquence, et $\psi_{\theta_1} = \psi_{[A_1,B_1,\omega_1,t_1,f_1]}$ est la fenêtre gaussienne de centre de gravité (t_1,f_1), de dispersions A_1 et B_1, et d'orientation ω_1 :

$$\psi_{(A_1,B_1,\omega_1,t_1,f_1)}(t,f) = \psi_{(A_1,B_1,\omega_1,0,0)}(t-t_1, f-f_1) \tag{6}$$

avec

$$\begin{aligned}\psi_{(A,B,\omega,0,0)}(t,f) = \exp\Big(-\Big(&t^2\left(\tfrac{\cos^2\omega}{A^2} + \tfrac{\sin^2\omega}{B^2}\right) + \\ &f^2\left(\tfrac{\sin^2\omega}{A^2} + \tfrac{\cos^2\omega}{B^2}\right) + ft\sin 2\omega\left(\tfrac{1}{A^2} - \tfrac{1}{B^2}\right)\Big)\Big).\end{aligned} \tag{7}$$

2.3 Critère de séparabilité des classes

Le critère utilisé pour mesurer la séparabilité des classes est basé sur les matrices de dispersion [3]. Il combine la matrice de variance-covariance intra-classes S_w et la matrice de variance-covariance inter-classes S_b, et est donné par :

$$J = tr(S_w^{-1} S_b) \tag{8}$$

où

$$S_w = \sum_{i=1}^{L} P_i E\left\{(\mathbf{X} - M_i)(\mathbf{X} - M_i)^T | \omega_i\right\} = \sum_{i=1}^{L} P_i \Sigma_i \tag{9}$$

$$S_b = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T \tag{10}$$

avec

$$M_0 = E\{\mathbf{X}\} \tag{11}$$

et L le nombre de classes. Les M_i et Σ_i sont les vecteurs moyens et les matrices de variance-covariance des attributs relatifs à la classe i, et dépendent du vecteur de paramètres θ.

Le problème donné par la relation (2) est résolu à l'aide d'un estimateur $\widehat{J}$ du critère J, lui-même basé sur des estimateurs des vecteurs moyens et des matrices de variance-covariance.

Ce critère fournit une bonne mesure de la séparabilité des classes pour autant que les distributions soient unimodales et non enchevêtrées [3], ce qui peut présenter une limitation. En revanche, contrairement à d'autres critères nécessitant l'estimation des densités de probabilité et fortement dépendant du nombre d'attributs, J peut être estimé correctement même lorsque le nombre d d'attributs considéré est grand. En effet la variance des estimateurs des vecteurs moyens et des matrices de variance-covariance dépend surtout du nombre d'échantillons.

3 Classification

Les attributs étant généralement extraits pour la classification, l'élaboration d'un classifieur exploitant le mieux possible l'information contenue dans les attributs est nécessaire.

Il est montré dans [3] que l'information de discrimination mesurée à l'aide du critère défini par (8) est parfaitement contenue dans un espace de dimension $L-1$. En effet, la matrice $S_w^{-1}S_b$ est de rang $L-1$ et les attributs peuvent être projetés dans l'espace engendré par les $L-1$ vecteurs propres de $S_w^{-1}S_b$ correspondant aux valeurs propres non nulles, sans que la valeur du critère ne soit modifiée.

Dans le cas particulier de deux classes, l'information de discrimination mesurée est parfaitement conservée lorsque les attributs sont projetés sur le vecteur $V = S_w^{-1}(M_2 - M_1)$. Le classifieur obtenu correspond au classifieur linéaire optimal lorsque le critère de contraste maximisé est une fonction de la forme $f(\eta_0,\eta_1,P_0\sigma_0 + P_1\sigma_1)$ où $\eta_{i=0,1}$ et $\sigma_{i=0,1}$ sont les moments d'ordre 1 et 2 de la statistique de décision conditionnellement aux classes.

4 Application à un problème simulé

4.1 Problème

La méthode est appliquée à un problème simulé, en prenant la représentation temporelle du signal, dans le cas d'un problème à deux classes équiprobables. Le processus aléatoire X considéré est tel que chaque réalisation est un vecteur $x^{(n)} = [x_1^{(n)} \dots x_j^{(n)} \dots x_J^{(n)}]$ où $x_j^{(n)}$, l'échantillon correspondant à la réalisation n et à l'instant $j\Delta t$, est donné par :

$$x_j^{(n)} = x^{(n)}(j\Delta t) = \sum_{i=1}^{4} 10 \exp\left(-\frac{\left(j\Delta t - t_i^{(n)}\right)^2}{\left(\sigma_i^{(n)}\right)^2} \right) \quad (12)$$

où

- $\sigma_i^{(n)}$ suit une loi uniforme $U(1{,}5\,,4{,}5)$,
- $t_i^{(n)}$ suit une loi normale $N(m_{i,c},16)$,
- $i = 1..4$ correspond au numéro du noyau gaussien,
- $c = 0..1$ correspond à la classe.

Les valeurs $m_{i,c}$ sont $m_{1,0} = 8$, $m_{1,1} = 11$, $m_{2,0} = 25$, $m_{2,1} = 28$, $m_{3,0} = 45$, $m_{3,1} = 48$, $m_{4,0} = 59$, $m_{4,1} = 62$. Le signal moyen de chaque classe ainsi que des exemples sont représentés sur la figure 1.

4.2 Extraction des attributs

L'extraction de d attributs, pour d compris entre 1 et 6, a été effectuée en considérant 3 cas de familles :

- uniquement des attributs correspondant à des moments d'ordre 0,
- uniquement des attributs correspondant à des moments d'ordre 1,
- $d-1$ attributs correspondant à des moments d'ordre 1 et un attribut correspondant à un moment d'ordre 0.

L'optimisation permettant de déterminer les attributs a été réalisée en considérant une base d'apprentissage de 2000 signaux.

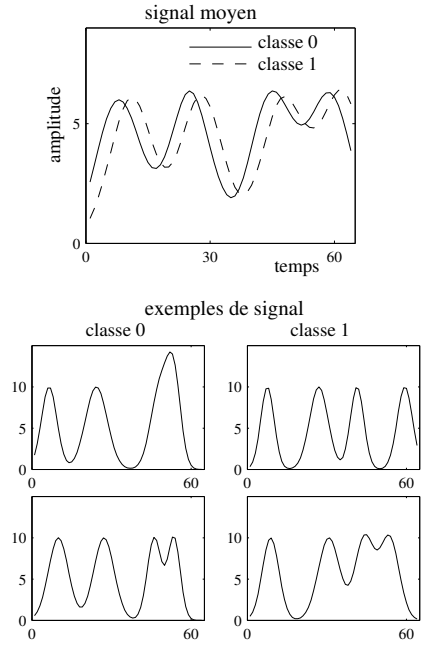

FIG. 1: *Signaux moyens et exemples de signaux*

Les valeurs du critère $\widehat{J}$ obtenues sont reportées sur la figure 2. La valeur maximum est atteinte dans le cas d'attributs correspondant à des moments d'ordre 1, ce qui est cohérent avec le problème posé. La valeur est croissante pour un nombre d'attributs compris entre 1 et 4, et stagne ensuite.

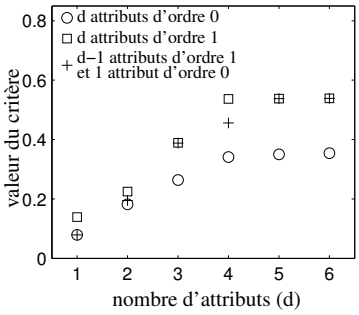

FIG. 2: *Valeur du critère de mesure de séparabilité des classes*

4.3 Classification

Les performances du classifieur utilisant les attributs extraits, et présenté dans le paragraphe 3 ont été estimées. Un tel classifieur a été construit pour les 3 cas de familles décrits dans le paragraphe précédent. Deux essais ont été réalisés. L'un avec une base d'apprentissage de $N = 200$ signaux (répartis équitablement dans les deux classes) et l'autre avec $N = 2000$.
Une base de test comprenant 2000 signaux a été utilisée pour les deux essais afin d'estimer les erreurs de classification définies par $P_E = [P(D_0/C_1) + P(D_1/C_0)]/2$.

Afin de pouvoir évaluer les performances du classifieur obtenu, le résultat théorique de classification a été déterminé en utilisant les valeurs de $t_i^{(n)}$ générées. Le classifieur de Bayes donne un taux d'erreur de 23,5%. Un autre élément de comparaison a été déterminé en construisant un classifieur linéaire utilisant les 64 échantillons du signal. L'erreur de classification

estimée à l'aide de la base de test, dans le cas où l'apprentissage a été réalisé avec la base de 200 signaux (respectivement 2000), est de 27,9% (respectivement de 26%).

Les résultats de classification obtenus avec les classifieurs utilisant les attributs extraits sont reportés sur les figures 3 et 4.

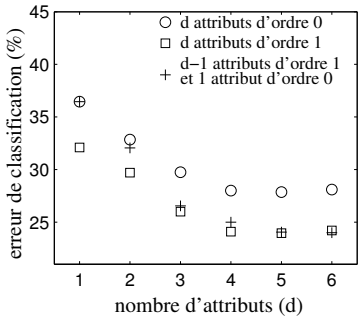

FIG. 3: *Résultats de classification dans le cas d'une base d'apprentissage de taille 2000.*

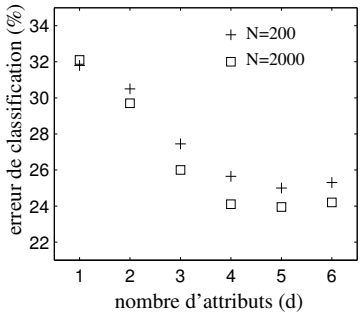

FIG. 4: *Résultats de classification dans le cas d'attributs qui sont des moments d'ordre 1.*

Les meilleurs résultats sont obtenus pour 4 ou 5 attributs dans le cas où tous les attributs sont des moments d'ordre 1. En ayant effectué l'apprentissage avec la base de taille 200 et celle de taille 2000, l'erreur de classification est respectivement de 25% et de 23,9%, ce qui est proche du résultat théorique.

Les fenêtres gaussiennes obtenues pour 5 attributs qui sont des moments d'ordre 1 sont reportées sur la figure 5. Ces fenêtres permettent de mettre en évidence les différents noyaux gaussiens présents dans le signal.

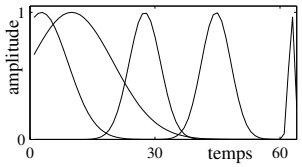

FIG. 5: *Fenêtres gaussiennes obtenues pour 5 attributs qui sont des moments d'ordre 1*

5 Conclusion

La méthode proposée permet d'extraire de façon automatique des attributs discriminants à partir d'un ensemble de fonctions paramétrées mesurant des moments d'ordre 0 ou 1 d'une représentation uni- ou bidimensionnelle pondérée. Les attributs optimaux sont extraits selon un critère mesurant la séparabilité des classes basé sur les matrices de variance-covariance.

Les attributs sont sélectionnés dans un ensemble infini engendré par des fonctions paramétrées continues, à l'aide d'un processus d'optimisation. Le processus de sélection d'attibuts est donc très différent de celui utilisé pour la sélection de m attributs parmi n, qui implique un problème de complexité combinatoire. Les attributs sont extraits de façon conjointe au sein d'un ensemble de familles, et donc la dépendance statistique des données est prise en considération. Compte tenu du problème de généralisation, le nombre d'attributs pouvant être extraits est uniquement limité par le rapport entre le nombre de paramètres à optimiser et la taille de la base d'apprentissage. Les attributs proposés sont aisés à interpréter et adaptés à de nombreux problèmes pratiques.

Dans le cas de deux classes, l'information discriminante mesurée par le critère peut parfaitement être prise en compte par un classifieur linéaire. Le classifieur obtenu ne constitue toutefois pas un classifieur linéaire des données originales en raison de la nature des attributs. Les attributs proposés fournissent en effet des fonctions non linéaires et non quadratiques des données initiales, ne pouvant pas être intégrées dans des classifieurs simples appliqués directement aux données originales.

L'application de la méthode proposée à un exemple simulé a montré qu'il est possible d'extraire des attributs discriminants, et représentatifs du problème étudié.

La méthode peut sans difficulté être généralisée à d'autres types de familles d'attributs.

Références

[1] T.A.C.M. Claasen and W.F.G. Mecklenbräuker. The Wigner distribution - A tool for time-frequency signal analysis - part I : Continuous-time signals. *Philips J. Res.*, 35:217–250, 1980.

[2] T.A.C.M. Claasen and W.F.G. Mecklenbräuker. The Wigner distribution - A tool for time-frequency signal analysis - part II : Discrete-time signals. *Philips J. Res.*, 35:276–300, 1980.

[3] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 1990.

[4] E. Grall-Maës and P. Beauseroy. Mutual Information-Based Feature Extraction on the Time-Frequency Plane. *IEEE Transactions on Signal Processing*, 50(4):779–790, 2002.

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

[6] A.N. Mucciardi, and E.E. Gose. A comparison of seven techniques for choosing subsets of pattern recognition properties. *IEEE Transactions Computers*, pages 1023–1031, 1971.

[7] P.M. Narendra, and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions Computers*, pages 917–922, 1977.

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