SEGMENTATION NON CANONIQUE POUR LA CLASSIFICATION SUPERVISEE PAR ARBRES

L. HAYOUN[1] P. COMON[2] F. NIVELLE[1]

[1] Thomson Missile Electronics, 23 rue Pierre Valette, 92245 Malakoff cedex
[2] Institut EURECOM, BP 193, F- 06904 Sophia-Antipolis cedex
hayoun@tme.thomson.fr comon@eurecom.fr

On présente d'abord un critère basé sur l'information, permettant de construire un arbre binaire dédié à la classification supervisée. Ensuite, on montre comment ce critère, s'il est remplacé par une estimation, préserve certains caractères d'optimalité. Enfin, cette procédure sera illustrée à l'aide d'un exemple (discrimination entre avions, hélicoptères et missiles).

We present first a discrimination criteria based on the information theory, which can be used to build binary decision trees for classification. We show that the estimated form of this criteria remains, to a certain extent, optimal. At the end of the paper, we show an example (discrimination between flying objects) which demonstrates the efficiency of the criteria.

1. Introduction

Les méthodes de classification par arbres de décision comportent plusieurs avantages sur le plan opérationnel, tels que la simplicité d'implantation, la lisibilité des règles de décision, et la rapidité d'exécution, et cela sans perte d'optimalité.
Classiquement, ces arbres effectuent des coupures successives selon des directions parallèles aux axes (coupures canoniques) [1], de sorte que la profondeur de l'arbre peut être excessive dans certains cas, ce qui compromet les performances précitées. L'idée que nous proposons consiste à envisager des coupures non canoniques, permettant d'obtenir des structures et des temps de classement optimisés.
La détermination de telles coupures peut se faire en utilisant un critère issu de la théorie de l'information, exposé ci-après.

2. Critère d'information

L'arbre permet de résoudre une suite de problèmes de classification en dimension 1 (sur une variable notée X), au lieu d'un seul problème en dimension plus grande.
Notons ω_i l'évènement "appartenir à la classe ω_i " ce qui constitue un abus de langage qui n'est pas bien génant, et W le sous-espace portant la variable X. La variable X caractérisant le mieux la classe ω_i est celle qui maximise l'écart entre les densités $p_x(u)$ et $p_x(u / \omega_i)$. Si on adopte la divergence de Kullback comme mesure de cet écart, on obtient l'information conditionnelle:

$$I(X / \omega_i) = \int_{u \in W} p_x(u / \omega_i) \log \frac{p_x(u / \omega_i)}{p_x(u)} du \quad (1)$$

Cette quantité (déterministe) est toujours positive ou nulle, et s'annulle si et seulement si X et ω_i sont statistiquement liées.

Si on désire ne pas privilégier une classe plutôt qu'une autre, on est conduit à utiliser un critère combinant ces informations conditionnelles. Le critère global s'écrit alors:

$$C(W) = \sum_i P_i \; I(X / \omega_i) \qquad (2)$$

où P_i est la probabilité a priori de la classe ω_i ,estimée par:

$$P_i = \frac{Card(\omega_i)}{\sum_i Card(\omega_i)}$$

L'utilisation de ce critère pour construire des coupures **canoniques** consisterait à chercher laquelle des d composantes x_k rendrait $C(W_k)$ maximale. Similairement, pour des coupures linéaires **non canoniques**, il faudrait trouver la direction w qui maximise:

$$C(w) = \sum_i P_i \int_{z \in R^d} p_x(w^T z / \omega_i) \log \frac{p_x(w^T z / \omega_i)}{p_x(w^T z)} dz$$

La figure 1, en fin d'article, montre un exemple d'application de ce critère.

La mise en oeuvre d'un tel critère nécessite l'utilisation d'un estimateur de densité de probabilité. C'est l'objet du paragraphe suivant.

3. Critère pour des échantillons finis

A l'instar des fonctions non linéaires d'impureté décrites plus haut, le critère proposé à l'origine dans [2] est fondé sur des prédicats, et non sur une approche relevant de la théorie de l'information.
Ce critère est basé sur une quantité, notée ici $\mathcal{H}(v)$, représentant l'information apportée par la variable $X \in W$ pour la

connaissance d'individus de la classe ω_j. Cette grandeur vérifie les quatre prédicats suivants (démonstration dans [2]:

- Si $\mathcal{H}_j(W_1) > \mathcal{H}_j(W_2)$ alors le classement dans ω_j est meilleur avec W_1 qu'avec W_2.

- Si W n'apporte rien au classement d'individus dans la classe ω_j alors $\mathcal{H}_j(W) = 0$.

- Positivité: $\mathcal{H}_j(W) \geq 0, \forall j$.

- Additivité: $\mathcal{H}_j(W_1 \cup W_2) = \mathcal{H}_j(W_1) + \mathcal{H}_j(W_2)$ si $W_1 \cap W_2 = \{0\}$.

Puisque X prend un nombre fini de valeurs (X décrit la projection sur W des individus de la base de données), on peut noter $\mathcal{X}$ l'ensemble des valeurs distinctes prises par X.

En outre, appelons $N_{x,i}$ le nombre d'individus de ω_i dont la projection sur W prend la valeur X, et $N_x = \Sigma_i N_{x,i}$ (N_x est le nombre d'éléments de l'espace total dont la projection sur W est X).

Le critère proposé dans [2] revient alors à utiliser la quantité suivante:

$$\mathcal{H}_j(W) = \sum_{X \in X} \hat{P}(X / \omega_j) \log \frac{\hat{P}(X / \omega_j)}{\hat{P}(X)} \qquad (3)$$

avec

$$\hat{P}(X / \omega_j) = \frac{N_{x,j}}{N_j} \quad , \quad \hat{P}(X) = \frac{N_x}{N}$$

où N_i désigne le nombre d'individus de la base de données issus de la classe ω_i et $N = \Sigma_i N_i$.

De même que dans la section précédente, il faut combiner les mesures d'information pour les différentes classes pour obtenir un critère global:

$$\mathcal{H}(W) = \sum_i P_i \, \mathcal{H}_i(W) \qquad (4)$$

Montrons à présent que le critère (3) peut aussi être obtenu à partir de (1). En effet, si on adopte les estimateurs triviaux de densités:

$$\hat{p}(u / \omega_i) = \frac{1}{N_i} \sum_{x \in \omega_i} \delta(u - x) \qquad (5)$$

alors le calcul du critère (1) nous conduit à prendre:

$$\hat{I}(X / \omega_i) = \int_{u \in W} \hat{p}_x(u / \omega_i) \log \frac{\hat{p}_x(u / \omega_i)}{\hat{p}_x(u)} \, du$$

avec

$$\hat{p}_x(u / \omega_i) = \frac{1}{N_i} \sum_{X \in W, x \in \omega_i} \delta(u - X)$$

soit

$$\hat{p}_x(u / \omega_i) = \sum_{X \in \mathcal{X}} \hat{P}(X / \omega_i) \delta(u - X)$$

et

$$\hat{p}_x(u) = \frac{1}{N} \sum_{X \in W, x \in \bigcup \omega_i} \delta(u - X) = \sum_{X \in \mathcal{X}} \hat{P}(X) \delta(u - X)$$

Or, on a précisément la relation suivante, pour tout triplet de fonctions bornées α, β, γ, en posant $f(u) = \delta(u\text{-}X)$ (la démonstration exploite les propriétés de la fonction de Dirac et est immédiate):

$$\sum_X \gamma(X) f_u \log \frac{\sum_X \alpha(X) f_u}{\sum_X \beta(X) f_u} = \sum_X \gamma(X) \log \frac{\alpha(X)}{\beta(X)} f_u \, , \forall u$$

Autrement dit, on peut remplacer le rapport des sommes par la somme des rapports, ce qui conduit bien à:

$$\hat{I}(X / \omega_i) = \sum_{X \in \mathcal{X}} \hat{P}(X / \omega_i) \log \frac{\hat{P}(X / \omega_i)}{\hat{P}(X)}$$

Ce critère est donc une version empirique du critère d'information conditionnelle. C'est d'ailleurs aussi la façon la moins coûteuse de la mettre en oeuvre.

4. Application

4.1 *Les arbres de décision*

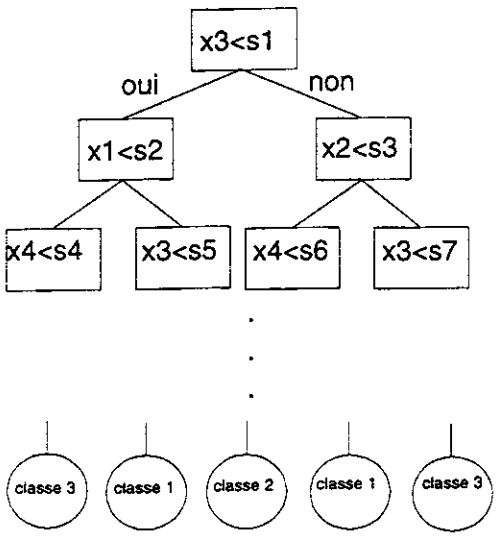
On rappelle brièvement le principe des arbres de décision.

Partant d'une base d'apprentissage (collection d'objets dont les classes d'appartenance sont connues), on détermine pour chaque variable, la dichotomie optimale au sens d'un certain critère lié à la séparation de la population en deux groupes, puis on choisit la variable qui induit la meilleure séparation.

On procède de la même façon pour les deux populations filles et ainsi de suite jusqu'à l'obtention de populations suffisamment « pures ».

On obtient, au terme de cette procédure, un structure arborescente qui permet de classer un individu inconnu à l'aide de questions binaires en cascade, portant sur les variables descriptives de l'objet à reconnaître.

On obtient une structure du type suivant:



Les avantages généraux offerts par les arbres de décision sont les suivants:

- Simplicité de la procédure de classement (structure séquentielle analogue au raisonnement humain) et rapidité de classement.

- Lisibilité des règles de décision, permettant une bonne compréhension des problèmes.

- Robustesse au bruit et aux données aberrantes.

- Facilité de prise en compte de critères non paramétriques et introduction possible d'informations a priori.

4.1 *Exemple*

On présente des résultats illustrant le critère (6). Ce critère a été employé pour construire un discriminateur de type arbre de décision. Le détail de la méthode ne sera pas exposée ici, mais le lecteur pourra se référer à [2].

Le problème consiste à discriminer 3 classes d'objet volant décrit par 4 variables descriptives correspondant à des mesures capteur.

Les classes considérées sont: Avion (C1), Hélicoptère (C2) et Missile (C3).

Les variables descriptives sont:

- Variable X: température équivalente (Watts/stéradian)
- Variable Y: surface équivalente radar (m^2)
- Variable Z: vitesse radiale (m/s)
- Variable W: altitude de vol (m)

On dispose d'une base de données synthétiques comprenant 9000 éléments. Cette base a été scindée en deux parties, l'une dédiée à l'apprentissage et l'autre à l'estimation des performances.
Ces données ont été générées à partir de distributions classiques: loi uniforme, loi de gauss, loi de Rayleigh, ...

Les résultats présentés ont été obtenus avec un logiciel créé à Thomson-CSF (divisions RCM et TME), qui permet de construire des arbres de décision basés sur différents critères d'optimalité (critère du Khi2, critère de Kolmogorov, critère d'impureté,...).

L'arbre obtenu, construit à partir du critère (6), **comprend 122 noeuds et affiche un taux de reconnaissance de 90.1%** (taux estimé sur une base distincte de la base d'apprentissage).
On précise que les coupures effectuées sont parallèles aux axes et qu'il n'a pas été introduit de coûts d'erreurs d'affectation.

Ces performances ont été comparées avec celles d'autres méthodes de classification. A l'aide du logiciel NEUROCLASS™ de Thomson-CSF [3], d'autres classifieurs ont été construits à partir de la même base.
Les méthodes testées sont la discrimination linéaire (DL) [2,3], la discrimination quadratique (DQ) [2.3], les « plus proches voisins » (KPPV) [2,3], la méthode de la pseudo-inverse (PI) [3] et la méthodes des nuées dynamiques supervisées (NDS) [3].

Les résultats sont les suivants:

Méthode	DL	DQ	KPPV (K=5)	PI	NDS
Taux (%)	82.6	90.3	90.4	84.4	86.9

On constate donc que la méthode proposée présente, sur cet exemple, d'excellentes performances, comparables à celles de la discrimination quadratique et des plus proches voisins qui sont des méthodes très coûteuses en temps de calcul.
Le temps de classement d'un individu inconnu par l'arbre est nettement inférieur à celui des autres méthodes (simples comparaisons à des seuils, entre 5 et 10 en pratique). Par exemple, sur un processeur spécialisé tel que le TMS320C40 , le temps de classement est inférieur à 2 microsecondes.
Le tableau suivant donne une évaluation du temps de classement (TMS320C40) d'un individu inconnu pour d'autres algorithmes sur cette application:

ALGORITHME	Temps de classement
Arbre de décision	1 µs
DL	10 µs
DQ	20 µs
KPPV (K=5)	15 ms
PI	400 µs

CONCLUSION

Dans le cadre de la reconnaissance statistique de formes, nous avons proposé un critère d'optimalité issu de la théorie de l'information et qu'il est possible d'estimer pratiquement. Ce critère permet de sélectionner la variable ou la combinaison de variables la plus discriminante. Il peut être utilisé pour construire des classifieurs de type arbre de décision, où pour chaque noeud à segmenter, le problème du choix de la meilleure variable se pose.
Un algorithme de construction d'arbres utilisant ce critère a été mis au point et on montre sur un exemple que les performances sont très bonnes puisqu'elles sont équivalentes à celles des meilleures algorithmes connus. De surcroît, on bénéficie des avantages des arbres de décision: simplicité, lisibilité, rapidité de classement.

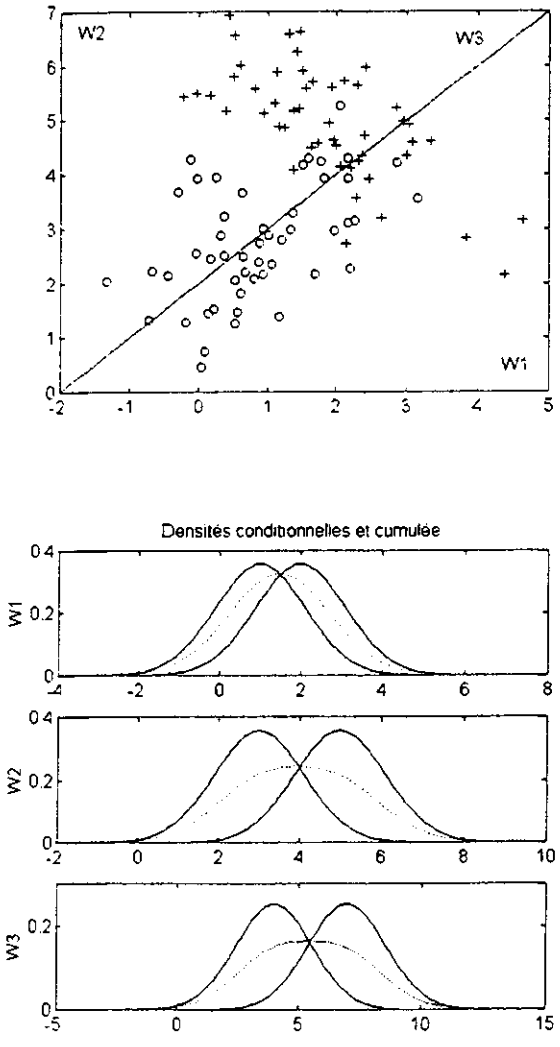


Figure 1: *En haut, exemple de problème à 2 classes et 2 variables; on cherche le meilleur des 3 espaces projetés W1, W2, W3. En bas, le critère C(W) mesure sur l'espace projeté W, l'écart entre les lois conditionnelles (en trait plein) et la loi cumulée. Le meilleur de ces 3 sous-espaces est W3, qui conduit à la valeur maximale de C(W).*

Références

[1] L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN, C.J. STONE, *Classification and Regression Trees*, Wadsworthand Brooks, 1984.

[2] F. NIVELLE, *Optimisation et robustesse des algorithmes de reconnaissance de formes: Application aux traitements des signaux radar*, Doctorat de l'ENST, Octobre 1994.

[3] E. PERNOT, *Choix d'un classifieur en discrimination*, Doctorat de l'université Paris-Dauphine, Avril 1994.