REVUE DE STATISTIQUE APPLIQUÉE

GILLES CELEUX
YVES LECHEVALLIER

Méthodes de segmentation non paramétriques

Revue de statistique appliquée, tome 30, nº 4 (1982), p. 39-53
<http://www.numdam.org/item?id=RSA_1982__30_4_39_0>

© Société française de statistique, 1982, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

METHODES DE SEGMENTATION NON PARAMETRIQUES

Gilles CELEUX, Yves LECHEVALLIER

INRIA

1. INTRODUCTION

L'utilisation de méthodes de segmentation pour des problèmes de discrimination est peu répandue.

Ce manque de succès vient du caractère simple de ces méthodes. En effet, elles construisent des fonctions de décision binaires sur les variables explicatives.

Pourtant, elles présentent des qualités.

Elles sont rapides et surtout très faciles à interpréter. Cette facilité d'interprétation est agréable dans la phase descriptive de l'analyse discriminante où il s'agit d'examiner si les variables explicatives permettent de discriminer les classes définies a priori.

Dans la phase décisionnelle de la discrimination, la clarté des règles de décision rend possible le dialogue homme-machine. Pour cette raison, dans le cas où les taux de reconnaissance sont satisfaisants, ce type de méthode nous paraît préférable à des méthodes donnant des résultats analogues mais plus difficiles à interpréter.

Dans cet article, nous présentons des méthodes de segmentation non paramétriques qui sont basées sur l'approche bayésienne de la discrimination.

Après, avoir exposé la méthode de base dans le cas de deux classes à discriminer, une généralisation permettant de résoudre une discrimination à plus de deux classes est ensuite proposée. Enfin, est présentée une méthode qui peut être utilisée lorsque les classes à reconnaître sont assez mélangées.

2. LA METHODE DE BASE (cas de 2 classes)

Le cadre est celui de la discrimination bayésienne. On dispose d'un échantillon de N individus décrits par p variables quantitatives. Chaque individu i est donc caractérisé par un vecteur de R^p. On note $(x_1, \ldots, x_N)$ cet échantillon.

Par ailleurs, une partition en deux classes W_1 et W_2 est définie a priori sur l'échantillon.

Mots-Clés : **Segmentation, Discrimination Bayésienne, Distance de Kolmogorov-Smirnov.**

On pose card $W_1 = n$, card $W_2 = N - n = m$.

On notera π_1 (resp. π_2) la probabilité a priori pour un élément d'appartenir à la famille W_1 (resp. W_2). On a $\pi_1 + \pi_2 = 1$. On notera ℓ_1 (resp. ℓ_2) le coût de mauvaise classification d'un élément de W_1 (resp. W_2) dans W_2 (resp. W_1).

La méthode que nous présentons dans ce paragraphe est due à J. FRIEDMAN [Fri 77].

2.1. Cas d'une seule variable (p = 1)

Dans ce cas, on note $f_1(x)$ et $f_2(x)$ les densités de probabilité des classes W_1 et W_2 et $F_1(x)$ et $F_2(x)$ les fonctions de répartition correspondantes.

Soit x_{N+1} un nouvel individu. Le principe de la méthode est le suivant : on veut classer x_{N+1} dans W_1 ou W_2 à partir de l'échantillon $x_1, \ldots, x_N$ en utilisant le point coupure c de la manière suivante :

si $x_{N+1} \leqslant c$, il est affecté dans la population "inférieure", W_1 par exemple,

si $x_{N+1} > c$, il est affecté dans la population "supérieure", soit W_2.

Dans ce cadre, on veut choisir c de manière à minimiser le risque de Bayes R de mauvaise classification de cet individu test. On se restreindra ici au cas où $\pi_1 \ell_1 = \pi_2 \ell_2$. Cette hypothèse est vérifiée dans le cas courant où faute de pouvoir mieux faire, on introduit des coûts inversement proportionnels aux probabilités d'apparition des classes.

La valeur du risque de Bayes R(z) de mal classer x_{N+1} en utilisant un point quelconque z comme coupure est :

$$R(z) = \pi_1 \ell_1 (1 - F_1(z)) + \ell_2 \pi_2 F_2(z).$$

On en déduit que sous l'hypothèse $\pi_1 \ell_1 = \pi_2 \ell_2$, on a l'équivalence :

$$\min_z R(z) <=> \max_z (F_1(z) - F_2(z)).$$

La population inférieure et la population supérieure sont définies ainsi : W_1 est la population inférieure si :

$$D(c) = \sup_z |F_1(z) - F_2(z)| = F_1(c) - F_2(c)$$

Sinon, W_1 est la population supérieure.

Comme on ne connaît pas a priori la population inférieure et la population supérieure, le point qui minimise le risque de Bayes est le point c tel que :

$$D(c) = \sup_z D(z) = \sup_z |F_1(z) - F_2(z)|.$$

La quantité D(c) n'est autre que la distance de Kolmogorov-Smirnov entre les deux distributions et est une mesure bien connue de la séparabilité des 2 fonctions de répartition.

En pratique, F_1 et F_2 ne sont pas connues et on les estime par les fonctions de répartition empiriques définies ainsi :

$$\hat{F}_1(x) = \begin{cases} 0 \text{ si } x < x_1^1 \\ \dfrac{k}{n} \text{ si } x_k^1 \leqslant x < x_{k+1}^1 \\ 1 \text{ si } x_n^1 \leqslant x \end{cases} \qquad \hat{F}_2(x) = \begin{cases} 0 \text{ si } x < x_1^2 \\ \dfrac{k}{m} \text{ si } x_k^2 \leqslant x < x_{k+1}^2 \\ 1 \text{ si } x_m^2 \leqslant x \end{cases}$$

avec $\qquad n = \text{card } W_1 \quad \text{et} \quad m = N - n = \text{card } W_2$

formules dans lesquelles x_k^i est le $k^{\text{ième}}$ point de la classe W_i $(i = 1, 2)$ les points étant rangés par ordre croissant.

On estime $D(c)$ par $\hat{D}(\hat{c}) = \sup_z |\hat{F}_1(z) - \hat{F}_2(z)|$.

Le risque effectif de mauvaise classification résultant de cette procédure est $R(\hat{c})$. Le risque "estimé" de mauvaise classification est

$$\hat{R}(\hat{c}) = \ell_1 \pi_1 (1 - \hat{F}_1(\hat{c})) + \ell_2 \pi_2 \hat{F}_2(\hat{c})).$$

On a les propositions suivantes [Sto 54] :

Proposition 1 :

$\hat{D}(\hat{c})$ converge en probabilité vers $D(c)$.

Proposition 2 :

$D(\hat{c})$ converge en probabilité vers $D(c)$.

Sous l'hypothèse $\pi_1 \ell_1 = \pi_2 \ell_2$, la proposition 1 implique que le risque "estimé" $\hat{R}(\hat{c})$ converge en probabilité vers le risque de Bayes $R(c)$ et la proposition 2 implique que le risque effectif $R(\hat{c})$ converge en probabilité vers le risque de Bayes $R(c)$.

2.2. Extension de la procédure à plusieurs variables

Dans le cas de plusieurs variables, on effectue la première coupure sur la variable qui fournit la plus grande distance de Kolmogorov-Smirnov entre les deux classes a priori.

Par cette procédure, l'échantillon est découpé en deux sous échantillons portés par deux segments. Dans la suite, on identifiera chaque sous échantillon au segment qui le porte.

Pour chaque segment ainsi construit, on réitère la procédure. Le découpage des segments s'arrête lorsqu'un test d'arrêt est vérifié.

Plus précisément, l'algorithme peut se résumer ainsi :

On calcule pour toutes les variables la quantité :

$D(c_j) = \sup_z |\hat{F}_1^j(z) - \hat{F}_2^j(z)|$ ($\hat{F}_1^j$ (resp. $\hat{F}_2^j$) représentant la fonction de répar-

tition empirique de la classe W_1 (resp. W_2) pour la variable j) et l'on effectue la coupure pour la variable j* telle que :

$$D(c_{j*}) = \sup(D(c_j)).$$

La coupure se fait au point c_{j*}.

Si l'un des deux segments satisfait au test d'arrêt, il est affecté à l'une des deux classes (celle qui est majoritaire dans le sous échantillon) et on obtient ainsi un segment terminal. Sinon, on reprend la procédure à partir de ce segment. En particulier, on recalcule $D(c_j)$ même pour la variable j choisie au (x) pas précédent(s), en effet dans le cas où $f_1(x)$ et/ou $f_2(x)$ sont multimodes, il se peut qu'une seule coupure ne fournisse pas une bonne discrimination.

2.2.1. Le test d'arrêt

Il reste à définir le test d'arrêt. L'affectation à une des classes est faite sur la base de l'estimation du rapport $f_1(x)/f_2(x)$. Le cardinal de chaque classe dans les segments doit être assez grand pour permettre une estimation raisonnable du rapport des densités. Ainsi le partitionnement s'arrêtera chaque fois que le partitionnement suivant n'assurera pas des échantillons de taille minimum pour chaque classe.

Le choix de la taille t minimale doit être déterminé par l'utilisateur et dépend du problème posé. t doit croître avec n, mais plus lentement que n. En effet, Gordon et Ohlsen montrent [GoOh 78] que la procédure décrite ici est asymptotiquement efficace au sens de Bayes si

$$\lim_n \frac{t(n)}{n} = 0 \quad \text{et} \quad \lim_n \frac{t(n)}{\sqrt{n}} = +\infty.$$

Autrement dit, toujours sous la restriction $\ell_1 \pi_1 = \ell_2 \pi_2$, lorsque la taille des deux classes devient grande, le risque de Bayes de classement d'un nouvel individu approche, avec une probabilité arbitrairement proche de 1, le risque de Bayes basé sur une procédure de Bayes construite à partir d'une connaissance complète des distributions F_1 et F_2. Il s'agit en fait de la généralisation au cas multivariable de la proposition 2.

2.2.2. Affectation de nouveaux individus

Le partitionnement par cette procédure conduit à un arbre de décision binaire. Un segment à chaque étape est représenté par un nœud de l'arbre. Le sommet de l'arbre redonne l'échantillon tout entier. Les deux successeurs de chaque nœud non terminal représentent les 2 segments définis par partitionnement. Les nœuds terminaux représentent les segments terminaux.

La règle de classification d'un nouvel individu est donc simple. On l'affecte à la classe caractérisant le segment dans lequel il tombe. Partant du sommet, l'individu descend l'arbre jusqu'à ce qu'il arrive à un segment terminal. Si celui-ci représente une classe unique, il est affecté à cette classe. Si le segment représente un mélange de classes, il est affecté à la classe majoritaire. En descendant l'arbre, la décision d'aller à gauche ou à droite est prise ainsi : si $x_{j*} \leqslant c_{j*}$, on va au successeur de gauche, sinon on va au successeur de droite. Ici j* et c_{j*} représentent la variable et la coupure correspondantes à ce nœud.

2.2.3. Remarques

La méthode garantit la meilleure segmentation en deux classes au sens du critère. Mais elle ne passe pas en revue tous les choix de séquences de coupures

possibles et donc rien n'assure que la segmentation finale soit la meilleure possible au sens du risque de Bayes. Aussi, l'examen de toutes les suites de coupures n'est pas envisageable même pour de petits échantillons et pour un petit nombre de variables.

Dans ce qui précède, nous avons supposé que $\pi_1 \ell_1 = \pi_2 \ell_2$, ce qui garantit pour la procédure la propriété d'efficacité asymptotique au sens de Bayes. Ceci étant, le critère choisi, ici la distance de Kolmogorov-Smirnov entre les 2 distributions, se justifie en dehors des considérations de convergence, par son bon pouvoir de séparation entre 2 distributions.

Si l'on désire conserver la propriété de convergence du risque de Bayes et si l'hypothèse $\pi_1 \ell_1 = \pi_2 \ell_2$ n'apparaît pas satisfaisante, on peut définir le critère à partir d'une estimation de $\pi_1, \ell_1, \pi_2, \ell_2$.

On doit alors à chaque pas de l'algorithme trouver c qui minimise le risque de Bayes $R(z) = \ell_1 \pi_1 (1 - F_1(z)) + \ell_2 \pi_2 F_2(z)$ et en pratique on doit considérer $R(z) = \hat{\ell}_1 \hat{\pi}_1 (1 - \hat{F}_1(z)) + \hat{\ell}_2 \hat{\pi}_2 \hat{F}_2(z)$ où $\hat{\pi}_1$, $\hat{\ell}_2$, $\hat{\pi}_1$, $\hat{\ell}_2$ représentent les estimations des valeurs théoriques. La modification introduite dans l'algorithme garantit l'efficacité asymptotique au sens du risque de Bayes [Go0h 78].

3. UNE METHODE DANS LE CAS MULTI-CLASSES

Une extension possible de cette procédure à des problèmes à plus de 2 classes consiste à les considérer comme une succession de problèmes à 2 classes [Fri77].

Si on a K classes ($K > 2$) à discriminer, on construit les K arbres de décision correspondant à chaque classe (contre toutes les autres).

En fait, cette approche n'est pas satisfaisante.

D'une part, la procédure d'affectation de nouveaux individus perd en rapidité.

D'autre part, les décisions offertes par cette méthode peuvent être contradictoires si l'on considère chaque arbre de décision pris séparément :

Soit par exemple, trois classes A, B, C. Au vu de l'arbre de décision de la classe A, un individu peut être affecté à la classe A et au vu de l'arbre de la classe B, être affecté à la classe B, etc.

Aussi, nous avons développé une méthode permettant la segmentation en ne construisant qu'un seul arbre de décision.

On retrouve cette préoccupation dans [GGM80]. Cependant les auteurs de cet article utilisent une approche différente de celle que nous présentons ci-dessous. En particulier l'agrégation des classes se fait en amont de la construction de l'arbre de décision.

3.1. Cas d'une seule variable

Le problème est de construire un seul arbre de décision associé à un ensemble de K classes. Nous notons $W_1, \ldots W_K$ les K classes et $\Pi_1, \ldots, \Pi_K$ les probabilités a priori associées à ces classes et $F_1, \ldots F_K$ leurs fonctions de répartitions

théoriques. Soit $W = \{ W_1, \ldots, W_K \}$ alors pour $A \in \mathcal{P}(W)$ la fonction de répartition théorique est :

$$F_A(x) = \frac{1}{\Pi_A} \sum_{W_i \in A} \Pi_i \, . \, F_i(x)$$

avec

$$\Pi_A = \sum_{W_i \in A} \Pi_i \, .$$

Ainsi pour un ensemble A de classes a priori le risque de Bayes au point z est égal à :

$$R(z) = \pi_A \, \ell_A \, (1 - F_A(z)) + \ell_{\bar{A}} \, \pi_{\bar{A}} \, F_{\bar{A}}(z)$$

avec

$$\bar{A} = W - A$$

Généralement on fixe les coûts ℓ_A et $\ell_{\bar{A}}$ inversement proportionnels aux probabilités, d'où $\pi_A \, \ell_A = \pi_{\bar{A}} \, \ell_{\bar{A}}$.

On a

$$R(z) = \pi_A \, \ell_A \, (F_{\bar{A}}(z) - F_A(z)) + \pi_A \, \ell_A \, .$$

Comme on ne connaît pas a priori la population inférieure et la population supérieure le point c qui minimise le risque de Bayes est :

$$D(c) = \operatorname*{Sup}_z \, |F_A(z) - F_{\bar{A}}(z)| \, .$$

On estime D(c) par $\hat{D}(c) = \operatorname*{Sup}_z \, |\hat{F}_A(z) - \hat{F}_{\bar{A}}(z)| \, .$

Proposition 3 :

$D(\hat{c})$ converge en probabilité vers D(c) ce qui implique que le risque effectif $R(\hat{c})$ converge vers le risque R(c) sous l'hypothèse $\pi_A \, \ell_A = \pi_{\bar{A}} \, \ell_{\bar{A}}$.

Démonstration : il suffit d'appliquer les propositions 1 et 2 aux classes A et $\bar{A}$.

3.2. Détermination de la classe A

Cette détermination peut se faire au départ [GGM80] ou au fur et à mesure de la construction de l'arbre.

On note $\mathcal{A}$ l'ensemble des partitions en 2 classes de $\{W_1, \ldots, W_k\}$. Dans ce cas, la classe A choisie est la réunion de classes a priori qui rend maximum le critère suivant :

$$\sup_{A \in \mathcal{A}} D(c) = \sup_z \sup_{A \in \mathcal{A}} |F_A(z) - F_{\bar{A}}(z)|$$

Remarque : Pour z fixé, la recherche de la classe A ne nécessite pas l'énumération complète de tous les cas possibles d'un regroupement en deux classes d'un ensemble de k classes a priori, mais seulement la construction de (k − 1) regroupements candidats (cf. [CeLe.80]).

3.3. Extension de la procédure à plusieurs variables

Cette extension est identique à celle du paragraphe 2.2. Ainsi, on calcule pour toutes les variables la quantité :

$$D(c_j) = \sup_{A \in \mathscr{A}} \sup |F_{\bar{A}}(x) - F_A(x)|$$

et l'on effectue la coupure pour la variable j* tel que :

$$D(c_{j*}) = \sup_j \{D(c_j)\} .$$

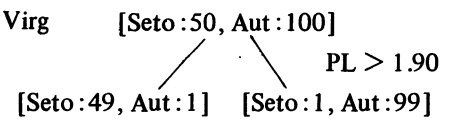
3.4. Exemple

Pour illustrer cette méthode, nous présenterons son application sur les données classiques de Fisher [Fis 36] concernant 3 populations d'Iris : Setosa (Seto), Versicolor (Vers), Virginica (Virg). On dispose d'un échantillon de 50 individus pour chaque type d'Iris. Chaque Iris est caractérisé par quatre variables : longueur du sépale (SL), largeur du sépale (SW), longueur du pétale (PL), largeur du pétale (PW).

Nous présentons tout d'abord les trois arbres de décision obtenus pour la séparation d'une classe contre les deux autres.

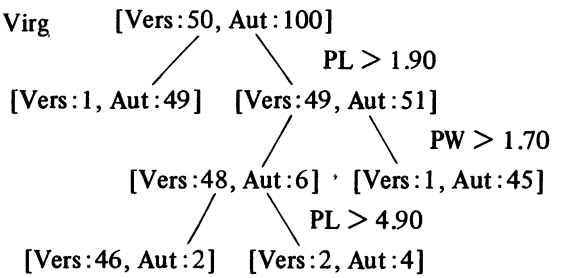
Setosa :

Aut = Vers + Virg



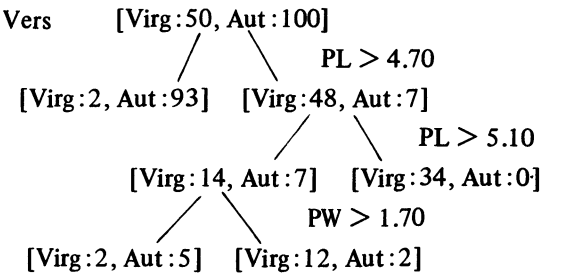
Versicolor :

Aut = Seto + Virg



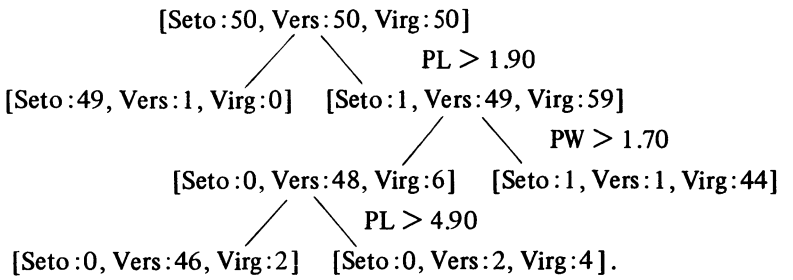
Virginica :

Aut = Seto + Vers



Finalement, en affectant les points par la règle majoritaire, on obtient 9 individus mal classés : 1 setosa, 4 versicolor et 4 virginica.

L'unique arbre de décision obtenu par la méthode décrite ci-dessus est le suivant :



Cet arbre donne 7 individus mal classés : 1 sétosa, 4 versicolor, 2 virginica.

De plus, il nécessite un nombre moindre de comparaisons que l'examen de tous les arbres de chaque type.

4. UNE VARIANTE : ARBRES DE DECISION TERNAIRES

L'algorithme que nous avons présenté vaut par sa rapidité de mise en œuvre et la facilité d'interprétation des résultats.

La principale critique que l'on peut lui faire est son manque de finesse. S'il est très efficace et agréable pour reconnaître des classes relativement bien séparées, il est moins performant pour reconnaître des classes "proches" l'une de l'autre.

Dans cette partie, nous présentons une variante dont le but est de fournir une discrimination plus fine tout en conservant la facilité d'interprétation des résultats de la méthode précédente et qui reste performante au point de vue de la rapidité.

4.1. Principe de la méthode dans le cas de 2 classes

L'idée est la suivante : au lieu de construire des arbres de décision binaires, on va construire des arbres de décision à 3 branches, la branche du milieu étant une branche d'indécision.

Le formalisme de cette variante est tout à fait analogue à celui de la méthode de base. Aussi nous l'exposons en suivant le même plan que précédemment.

4.1.1. Cas d'une seule variable ($p = 1$)

Les définitions sont les mêmes qu'au début du § 2. Mais en plus des coûts ℓ_1 (resp. ℓ_2) de mauvaise classification pour un élément de W_1 (resp. W_2), on introduit le coût de non décision ℓ'_1 (resp. ℓ'_2) pour un élément de W_1 (resp. W_2).

Le problème s'énnonce alors ainsi :

Au vu de l'échantillon $x_1, \ldots, x_N$, on veut classer un individu supplémentaire x_{N+1} dans W_1 ou W_2 en utilisant *deux* points coupure c_1 et c_2 ($c_1 \leqslant c_2$) de la manière suivante :

si $x_{N+1} \leqslant c_1$, il est affecté dans la population "inférieure", W_1 par exemple.

si $x_{N+1} > c_2$, il est affecté dans la population "supérieure", soit W_2,

si $c_1 < x_{N+1} \leqslant c_2$, on ne prend pas de décision, ce qui signifie que la variable considérée ne permet pas de décider de l'affectation de x_{N+1} .

Dans ce cadre, on veut choisir c_1 et c_2 de manière à minimiser le risque de Bayes $R(z_1, z_2)$ de mal classer un individu en utilisant les deux points coupures z_1 et z_2 $(z_1 \leqslant z_2)$.

Là encore, on se restreindra au cas $\pi_1 \ell_1 = \pi_2 \ell_2$. Et de manière cohérente avec cette dernière restriction, on supposera de plus que $\pi_1 \ell'_1 = \pi_2 \ell'_2$.

Enfin, on posera $\ell_1 = a\ell'_1$ avec $a \geqslant 1$, ce qui entraîne que $\ell_2 = a\ell'_2$, car de $\pi_1 \ell_1 = \pi_2 \ell_2$ on tire $\pi_1 a\ell'_1 = \pi_2 \ell_2 = a\pi_2 \ell'_2$ puisque $\pi_1 \ell'_1 = \pi_2 \ell'_2$.

Notons que prendre $a \geqslant 1$ est naturel. Il semble raisonnable de convenir que le coût de mal classer un élément de W_1 soit plus grand que le coût de ne pas classer cet individu à l'aide de la variable considérée.

Le risque de Bayes s'écrit :

$$R(z_1, z_2) = \pi_1 \ell_1 (1 - F_1(z_2)) + \pi_2 \ell_2 F_2(z_1) + \pi_1 \ell'_1 (F_1(z_2) - F_1(z_1)) + \\ + \pi_2 \ell'_2 (F_2(z_2) - F_2(z_1)) .$$

Dans le cas où $\pi_1 \ell_1 = \pi_2 \ell_2 = a\pi_1 \ell'_1 = a\pi_2 \ell'_2$, il vient :

$$R(z_1, z_2) = \pi_1 \ell'_1 [a + F_1(z_2)(a-1) - F_1(z_1) + F_2(z_1)(a-1) + F_2(z_2)] .$$

On a alors l'équivalence :

$$\min_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} R(z_1, z_2) \Longleftrightarrow \max_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} [F_1(z_2)(a-1) + F_1(z_1) - \\ F_2(z_1)(a-1) - F_2(z_2)]$$

En fait, on ne connaît pas a priori la population inférieure. Mais l'algorithme détermine au début la population inférieure de la manière suivante : W_1 est la population inférieure si

$$D(c) = \sup_z |F_1(z) - F_2(z)| = F_1(c) - F_2(c)$$

Dans la suite, on supposera donc que W_1 est la population inférieure.

La variante consiste donc à chercher les deux points coupures c_1 et c_2 qui rendent maximum le critère :

$$D(z_1, z_2) = [F_1(z_2)(a-1) + F_1(z_1) - F_2(z_1)(a-1) - F_2(z_2)].$$

En pratique, on rend maximum

$$\hat{D}(z_1, z_2) = [\hat{F}_1(z_2)(a-1) + \hat{F}_1(z_1) - \hat{F}_2(z_1)(a-1) - \hat{F}_2(z_2)]$$

et donc on estime

$$D(c_1, c_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} D(z_1, z_2)$$

par

$$\hat{D}(\hat{c}_1, \hat{c}_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} \hat{D}(z_1, z_2) .$$

De manière analogue au cas binaire, on a les propriétés de convergence suivantes (voir démonstrations en annexe) :

Proposition 4

$\hat{D}(\hat{c}_1, \hat{c}_2)$ converge vers $D(c_1, c_2)$ en probabilité.

Proposition 5

$D(\hat{c}_1, \hat{c}_2)$ converge en probabilité vers $D(c_1, c_2)$.

On en déduit le résultat suivant :

Le risque effectif de mauvaise classification $R(\hat{c}_1, \hat{c}_2)$ converge en probabilité vers le risque de Bayes $R(c_1, c_2)$.

4.1.2. Extension de la méthode à plusieurs variables

L'algorithme est alors le même que dans le cas binaire.

A chaque pas on sélectionne la variable qui maximise le critère :

$$\hat{D}(\hat{c}_1, \hat{c}_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} [(a-1)\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\hat{F}_2(z_1) - \hat{F}_2(z_2)]$$

et on effectue les deux coupures aux points $\hat{c}_1$ et $\hat{c}_2$.

On réitère la procédure sur chacun des segments ainsi obtenus. Le test d'arrêt est le même que dans le cas binaire.

Le partitionnement par cette procédure conduit à un arbre de décision où de chaque sommet partent deux ou trois branches selon les cas. Il part deux branches si au sommet considéré les points coupures $\hat{c}_1$ et $\hat{c}_2$ sont confondus, trois branches si $\hat{c}_1 < \hat{c}_2$.

Ce dernier cas se produit lorsque la variable j sélectionnée au sommet considéré ne permet pas de discriminer entre les deux classes tous les individus dont la coordonnée x_j pour cette variable vérifie $\hat{c}_1 < x_j \leqslant \hat{c}_2$.

L'introduction de cette branche du "milieu" permet ainsi d'éviter une affectation peu sûre d'individus à l'une des deux classes définies a priori.

Lorsque la branche du milieu disparaît, c'est-à-dire lorsque pour le sommet considéré et la variable sélectionnée correspondante on a $\hat{c}_1 = \hat{c}_2$, le critère s'écrit :

$$\hat{D}(\hat{c}_1, \hat{c}_1) = a \sup_z |\hat{F}_1(z) - \hat{F}_2(z)|.$$

L'unique point coupure est alors le même que l'on obtient dans le cas binaire.

Ainsi, si les classes a priori sont bien séparées, cette variante donnera les mêmes résultats que l'algorithme de base : de chaque sommet de l'arbre de décision, il ne partira que deux branches.

4.2. Cas multi-classes ($K > 2$)

Il est clair que, quelle que soit la manière d'envisager le problème multi-classes, les algorithmes que nous avons présentés se généralisent sans peine pour cette variante.

4.3. Le choix de a

Dans la variante proposée, le nombre $a \geqslant 1$ tel que $\ell_1 = a\,\ell_1'$ est un paramètre de l'algorithme qui doit être défini par l'utilisateur.

Si a est trop grand, l'arbre de décision aura de nombreuses branches et sera difficile à interpréter. Si a est trop petit, l'arbre sera peu différent d'un arbre de décision binaire.

Nous pensons que prendre $a = 3$, soit $\ell_1 = 3\,\ell_1'$ est une assez bonne solution.

En pratique, l'utilisateur pourra faire varier a au cours de différents essais de manière à obtenir un arbre de décision qui le satisfasse en regard du problème posé.

4.4. Considérations numériques

Dans ce paragraphe, nous allons voir qu'il n'est pas nécessaire de passer en revue tous les couples (z_1, z_2) pour optimiser le critère

$$\hat{D}(\hat{c}_1, \hat{c}_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} \hat{D}(z_1, z_2).$$

Proposition [CeLe80] : Pour $a > 2$, le couple (c_1, c_2) vérifiant

$$\hat{D}(c_1, c_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} \hat{D}(z_1, z_2)$$

est tel que :

$$\hat{D}(c_1, c) = \sup_{z_1 \leqslant c} \hat{D}(z_1, c)$$

et

$$\hat{D}(c, c_2) = \sup_{z_2 \geqslant c} \hat{D}(c, z_2)$$

où c vérifie

$$\sup |\hat{F}_1(z) - \hat{F}_2(z)| = |\hat{F}_1(c) - \hat{F}_2(c)|.$$

Remarque : faire l'hypothèse $a > 2$ est raisonnable car a ne doit pas être trop voisin de 1 et cela permet d'accélérer l'algorithme.

En effet, la proposition permet de simplifier la recherche du couple (c_1, c_2) optimal.

On cherche d'abord le point c tel que :

$$\sup_z |\hat{F}_1(z) - \hat{F}_2(z)| = |\hat{F}_1(c) - \hat{F}_2(c)|.$$

Cela nous permet de déterminer les populations inférieure et supérieure.

c_1 est alors obtenu par maximisation de $D(z_1, c)$ et c_2 par maximisation de $D(c, z_2)$.

Cet algorithme nécessite 2N investigations (N étant la taille de l'échantillon) : N investigations pour la recherche de c, N investigations pour la recherche de c_1 et de c_2.

Or s'il avait fallu examiner tous les couples (c_1, c_2) avec $c_1 \leqslant c_2$ pour optimiser le critère, cela aurait nécéssité $\frac{N^2 + N}{2}$ investigations.

4.5. Exemples d'application

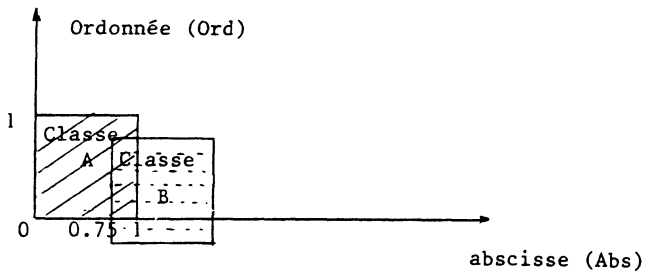
Nous présentons deux applications sur des données simulées. Ces exemples permettent de bien illustrer l'intérêt de la méthode.

On trouvera des applications de cette méthode à des problèmes médicaux dans [CeLe80], [CLL82].

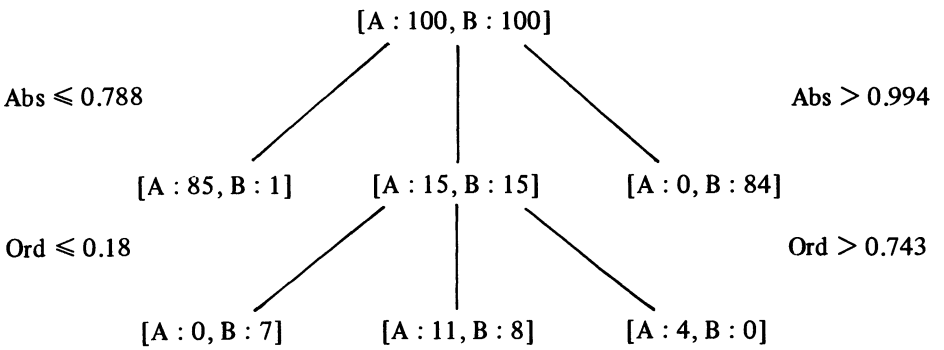
Application 1

On a tiré au hasard 100 points de R^2 suivant un loi uniforme sur le cube $[0,1] \times [0,1]$ (classe A) et 100 points de R^2 suivant une loi uniforme sur le cube (classe B) $[0.75, 1.75] \times [-0.25, 0.75]$.

Les deux classes A et B ainsi définies admettent donc à peu près la représentation suivante :



L'arbre obtenu est le suivant :



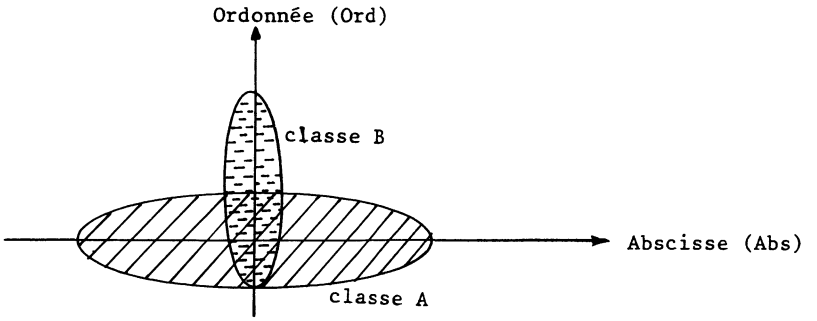
Cet arbre permet bien de retrouver la zone de recouvrement des classes A et B. Elle correspond à la classe d'indécision I suivante :

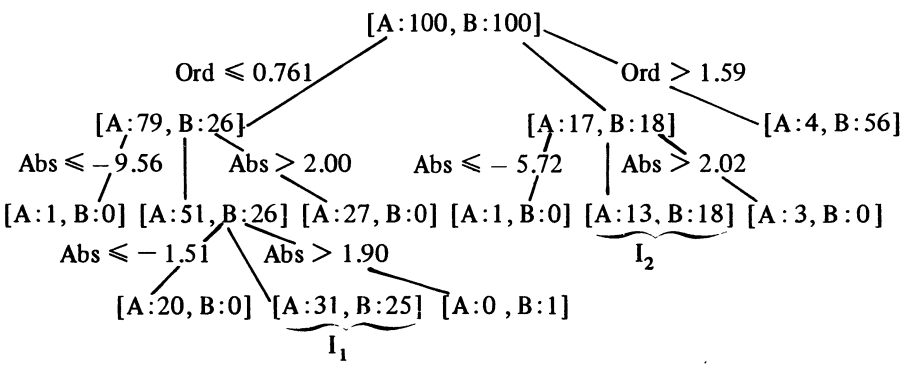
$$I = \{(x, y) \in R^2 / 0.788 < x \leqslant 0.994, 0.128 < y \leqslant 0.743\}.$$

Application 2

On a tiré au hasard 100 points de R^2 suivant une loi normale de vecteur moyenne $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ et de matrice de variances covariances $\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ et 100 points de R^2 suivant une loi normale de vecteur moyenne $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ et de matrice de variances covariances $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

Les deux classes A et B ainsi définies admettent les ellipses d'inertie à 95 % :





Cet arbre donne 4 mal classés, tous issus de la classe A, et 87 non classés situés dans les segments I_1 et I_2.

En effet, le découpage des segments I_1 et I_2 est illusoire : à chaque nouvelle segmentation tentée, il n'y a pas plus d'un ou deux points plus petits que la coupure inférieure ou plus grands que la coupure supérieure.

On voit que $I_1 \cup I_2$ recouvre à peu près la zone de recouvrement des deux classes A et B :

$$I_1 = \{(x , y) \in R^2 / -1.51 < x \leqslant 1.90 , y < 0.761\}$$

$$I_2 = \{(x , y) \in R^2 / -5.72 < x \leqslant 2.02 , 0.761 < y \leqslant 1.59\}.$$

Le programme, écrit en FORTRAN, de ces algorithmes est disponible auprès des auteurs à l'INRIA. Domaine de Voluceau. Rocquencourt. Le Chesnay 78150.

ANNEXE

Démonstration de la Proposition 4

$\forall (z_1, z_2), \hat{D}(z_1, z_2)$ converge en probabilité vers $D(z_1, z_2)$ car la fonction de répartition empirique converge en probabilité vers la fonction de répartition.

En particulier $\hat{D}(c_1, c_2)$ converge en probabilité vers $D(c_1, c_2)$. On peut alors écrire :

$$\forall \epsilon > 0, \forall \eta > 0, \exists N, \text{tq pour tout } n > N_1$$

$$P[\hat{D}(c_1, c_2) < D(c_1, c_2) - \epsilon] < \eta$$

Or, par construction, $\hat{D}(\hat{c}_1, \hat{c}_2) > \hat{D}(c_1, c_2)$ d'où

$$P[\hat{D}(\hat{c}_1, \hat{c}_2) < D(c_1, c_2) - \epsilon] < \eta \tag{1}$$

pour tout $n > N_1$, D'autre part :

$$\hat{D}(\hat{c}_1, \hat{c}_2) - D(c_1, c_2) = \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} [(a-1)\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\hat{F}_2(z_1) - \hat{F}_2(z_2)] - \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} [(a-1)F_1(z_2) + F_1(z_1) - F_2(z_1)(a-1) - F_2(z_2)]$$

$$\hat{D}(\hat{c}_1, \hat{c}_2) - D(c_1, c_2) \leqslant \sup_{\substack{z_1, z_2 \\ z_1 \leqslant z_2}} \{[(a-1)\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\hat{F}_2(z_1) - \hat{F}_2(z_2)] - [(a-1)F_1(z_2) + F_1(z_1) - (a-1)F_2(z_1) - F_2(z_2)]\}.$$

La quantité sous le sup s'écrit encore :

$$[(a-1)(\hat{F}_1(z_2) - F_1(z_2)) + (\hat{F}_1(z_1) - F_1(z_1)) - (a-1)(\hat{F}_2(z_1) - F_2(z_1)) - \hat{F}_2(z_2) - F_2(z_2))]$$

On en déduit :

$$D(\hat{c}_1, \hat{c}_2) - D(c_1, c_2) \leqslant \sup_{z_2} (a-1)|\hat{F}_1(z_2) - F_1(z_2)| + \sup_{z_2} |\hat{F}_1(z_1) - F_1(z_1)| + \sup (a-1)|\hat{F}_2(z_1) - F_2(z_1)| + \sup_{z_2} |\hat{F}_2(z_2) - F_2(z_2)|.$$

Et d'après le théorème de Glyvenko-Cantelli, pour $i = 1, 2$ $\sup_z |\hat{F}_i(z) - F_i(z)|$ tend vers 0 presque sûrement donc a fortiori en probabilité.

On a donc :

$$\forall \epsilon > 0, \forall \eta > 0, \exists N_2 \text{ tq pour tout } n > N_2,$$

$$P[\hat{D}(\hat{c}_1, \hat{c}_2) - D(c_1, c_2)) > \epsilon] \tag{2}$$

Finalement, on déduit de (1) et (2) que :

$$\forall \epsilon > 0 , \forall \eta > 0 , \exists N_3 = \sup (N_1 , N_2) \text{ tq pour tout } n > N_3$$

$$P[\,|\hat{D}(\hat{c}_1 , \hat{c}_2) - D(c_1 , c_2)| > \epsilon] < 2\eta .$$

Démonstration de la Proposition 5

On a :

$$|D(\hat{c}_1 , \hat{c}_2) - D(c_1 , c_2)| \leqslant |D(\hat{c}_1 , \hat{c}_2) - \hat{D}(\hat{c}_1 , \hat{c}_2)| + |\hat{D}(\hat{c}_1 , \hat{c}_2) - D(c_1 , c_2)| .$$

D'après la proposition précédente $\hat{D}(\hat{c}_1 , \hat{c}_2)$ converge en probabilité vers $D(c_1 , c_2)$ et $\hat{D}(\hat{c}_1 , \hat{c}_2)$ converge en probabilité vers $D(\hat{c}_1 , \hat{c}_2)$ puisque la fonction de répartition empirique converge en probabilité vers la fonction de répartition. On en déduit aisément le résultat annoncé.

BIBLIOGRAPHIE

[CeLe 80] G. CELEUX, Y. LECHEVALLIER. – *Méthodes de discrimination non paramétriques asymptotiquement efficaces au sens de Bayes.* Rapport de Recherche n° 52. INRIA, 1980.

[CLL 82] G. CELEUX, N. LAURO, Y. LECHEVALLIER. – Contrilenti dell' analizi multidimensionale nello studio di gruppi cliniqui a priori mal définiti. *Rivista di statistica applicata* (Italie), 1982.

[Fis 36] R.A. FISHER. – The use of multiple measurements in taxinomia problems. *Ann. of Enginics, 7.*

[Fri 77] J.H. FRIEDMAN. – A recursive partitionary decision rule for non parametric classification. *IEEE Trans. Comput,* pp. 404-408 (Avril 1977).

[Go0h 78] L. GORDON, R.A. OHLSEN. – Asymptotically efficient solutions to the classification problem. *Annals of Statistics.* Vol. 6, n° 3, 1978.

[GGM 80] D.E. GUSTAFSON, S. GELFAND, S.K. MITTER. – *A non parametric multiclass partitioning method for classification.* Proceed. of 5th International Conference on Pattern Recognition.

[Sto 54] D.S. STOLLER. – *Univariate two-population distribution free discrimination. J. Amer. Statist. Assoc.* Vol. 49, pp. 770-775, 1954.