SELECTION DES PREDICTEURS ET ESTIMATION DES TAUX D'ERREUR DE CLASSEMENT EN DISCRIMINATION LINEAIRE

Jean-Christophe TURLOT

STID Université de Pau

Résumé :

On montre que les procédures de sélection d'un sous-ensemble de prédicteurs pertinents pour la discrimination penvent engendrer un biais important dans l'estimation des taux d'erreur de classement par rééchantillonnage (validation croisée, jackknife ou bootstap). Le biais de 'sélection' peut conduire à un choix de prédicteurs en partie illusoire, dépendant des fluctuations d'échantillonnage. Il apparaît que la sélection d'un petit nombre de variables exploratoires, complétant l'information apportée par un ensemble de prédicteurs devant intervenir a priori dans l'élaboration de la règle de décision, constitue une protection contre une sélection trop sujette aux fluctuations d'échantillonnage lorsque la taille du fichier des observations est modérée. En réduisant ainsi le biais de sélection, l' estimation de la qualité de la règle par rééchantillonnage s'en trouve plus précise.

Mots-clés: discrimination, sélection de prédicteurs, erreur de classement.

INTRODUCTION

Dans l'article précédent, portant sur l'estimation de la qualité d'une règle de classement, on a présenté des techniques non-paramétriques (échantillon test,validation croisée et bootstrap) permettant d' évaluer de manière satisfaisante la qualité d'une règle de décision estimée sur un fichier d'apprentissage. Rappelons que l'ensemble des prédicteurs devant intervenir dans l'élaboration de la règle discriminante était supposé connu a priori.
Or en pratique, lorsque les variables observées sont nombreuses, on est le plus souvent amené à en extraire un sous-ensemble de prédicteurs pertinents pour la discrimination. Ce point est rappelé dans le §1; il conduit à la définition d'une hypothèse adaptée au choix des variables à retenir en regard du taux d'erreur de classement en situation gaussienne (§2). Il apparaît que toute procédure de sélection appliquée au fichier d'apprentissage engendre deux biais: un biais par omission et un biais de sélection (§3). Ces deux biais peuvent être importants lorsque la taille du fichier d'apprentissage est modérée. Le biais de sélection à pour effet de réduire la précision de l'estimation des taux d'erreur de classement par rééchantillonnage. Intuitivement, l'élaboration de la règle de décision dépend davantage du fichier d'apprentissage que dans l'article précédent, puisque le choix des prédicteurs et l' estimation des paramètres sont fondés sur un même fichier de données. On en déduit certains principes de sélection visant à réduire les effets des fluctuations d'échantillonnage, ce qui constitue une protection contre une discrimination illusoire: le biais de sélection s'en trouve réduit et l'estimation de la qualité de la règle de classement plus précise (§4). Enfin, on montre comment les taux d'erreur de classement peuvent être estimés par rééchantillonnage, compte tenu de la procédure de sélection d'un sous-ensemble de prédicteurs (§5).

1 NECESSITE DE LA SELECTION D'UN SOUS-ENSEMBLE DE PREDICTEURS

Pour plus de simplicité (et de précision) dans l'exposé, on considère le problème de la sélection des prédicteurs dans le cas de deux populations normales, P1 et P2, de paramètres (μ_1, Σ) et (μ_2, Σ) respectivement.
Lorsque les variables observées sont nombreuses, l'élaboration d'une règle de décision pertinente nécéssite une sélection des prédicteurs. La raison en est la suivante: on peut généralement supposer qu'il existe une partition $(X^{(1)}, X^{(2)})$ de l'ensemble des variables telle que toute l'information utile pour la discrimination soit contenue dans $X^{(1)}$; autrement dit, l' information additionnelle apportée par $X^{(2)}$ sachant $X^{(1)}$ est nulle. De fait, les coefficients de la fonction linéaire discriminante associés à $X^{(2)}$ sont nuls [*Kshirsagar* 1972, *Rao* 1973] , mais leurs estimations sur le fichier d'apprentissage (y,x) ne le sont pas en raison des fluctuations d'échantillonnage. D'un point de vue heuristique, l'introduction dans le modèle du sous-ensemble $X^{(2)}$ contribue à une définition plus imprécise des régions de décision, notées R_1 et R_2 dans l'article précédent, car sur la frontière, le

niveau de la densité est faible pour P1 comme pour P2. La densité est aussi plus diffuse dans l'espace $R^{(p)}$ de l'ensemble des variables, que dans le sous-espace associé à $X^{(1)}$, de plus faible dimension. Une telle frontière, estimée sur le fichier d'apprentissage, peut perdre toute signification dès lors que l'on se place dans un cadre inférentiel. Plus précisément, il est raisonnable de penser qu'il existe un sous-ensemble $X^{(1)}$ de prédicteurs tel que la fonction linéaire discriminante estimée sur le fichier $(y,x^{(1)})$ conduise à un taux d'erreur de classement $Err(y,x^{(1)})$ plus faible que si l'ensemble des variables était pris en compte. Ce point est développé dans le paragraphe qui suit.

On rappelle que, si F est la loi du couple (Y,X) définissant les populations potentielles de l'étude, le 'vrai taux' d'erreur de classement représente le taux d'erreur de classement associé à la règle de décision estimée sur (y,x) et calculé selon la loi F (voir l'article précédent). Ce taux, noté $Err(y,x)$, est conditionnel au fichier d'apprentissage et il est inconnu (puisque la loi F est inconnue). On appelle 'taux apparent' d'erreur, la fréquence des erreurs de classement observées sur le fichier d'apprentissage, selon la terminologie d'Efron [*Efron* 1983, *Turlot* 1989].

2 HYPOTHESE LIEE A LA PROCEDURE DE SELECTION

On se place en situation générale, où aucune structure a priori n'est connue dans l'ensemble $(X^1, \ldots, X^p)$ des prédicteurs. Soit J la famille des sous-ensembles de $\{1,2, \ldots, p\}$: à un élément $j \in J$ est associé le sous-ensemble de prédicteurs correspondant noté $X^{(j)}$. Ainsi à l'élément j correspond la partition $(X^{(j)}, X^{(jc)})$ de X, où $j \cup jc = \{1,2 \ldots p\}$. L'hypothèse de l'existence d'un sous-ensemble $X^{(jo)}$ de prédicteurs préférable à X pour la discrimination des populations P1 et P2 peut se traduire ainsi : on suppose qu'il existe un élement $jo \in J$ tel que les coefficients de la fonction linéaire discriminante soient nuls sur $X^{(joc)}$ et non nuls sur $X^{(jo)}$. L'absence d'information additionnelle apportée par $X^{(joc)}$ est complétée par le caractère minimal de $X^{(jo)}$; on note $H(jo)$ cette hypothèse.

Soit $J1 \in J$ la famille des sous-ensembles de J contenant jo et $J2$ le complémentaire de $J1$ dans J. L'hypothèse $H(jo)$ est caractérisée en termes de distances entre les deux populations:

$$\Delta_j = \Delta \text{ pour tout } j \in J1 \quad \text{et} \quad \Delta_j < \Delta \text{ pour tout } j \in J2,$$

où Δ_j et Δ représentent les distances de Mahalanobis entre P1 et P2 fondées sur $X^{(j)}$ et X respectivement [*Fujikoshi*,1985]. Cette hypothèse peut être justifiée en terme d'erreur de classement dans le cadre asymptotique:

Soit $\{j\}$ un élément de J auquel est associé le sous-ensemble $X^{(j)}$ de prédicteurs; on note $Err(j;y,x^{(j)})$ le vrai taux d'erreur associé à la règle de décision estimée sur $(y,x^{(j)})$. Cette quantité est explicitée dans l'équation (2') de l'article précédent pour $\{j\} = \{1, \ldots, p\}$.

Le vrai taux d'erreur de classement, non conditionnel, associé à $\{j\}$ vaut:

$$(1)\quad R(j) = E[Err(j;Y,X^{(j)})]$$

cette espérance étant calculée selon la loi F_j du vecteur aléatoire $(Y,X^{(j)})$.

On a le résultat suivant [*Fujikoshi* 1985]:

l'hypothèse $H(jo)$ est vraie si et seulement si:

- la limite de $n(R(j) - R(jo))$ est positive (et $< \infty$) pour tout $j \in J1 - \{jo\}$,
- la limite de $R(j) - R(jo)$ est positive pour tout $j \in J2$,

la limite étant entendue au sens où n_1 et n_2 (les effectifs des échantillons de P1 et P2) tendent vers l'infini simultanément (lim $n_1/n_2 = a$, constante non nulle).

Remarque:
Il importe de noter qu'il s'agit d'un résultat 'en moyenne': la quantité minimisée n'est pas $Err(j;y,x^{(j)})$ - le taux d'erreur associé à la règle de décision estimée sur le fichier d'apprentissage $(y,x^{(j)})$, mais son espérance $R(j)$.

3. CRITERES DE SELECTION ET BIAIS

A. SELECTION D'UN SOUS-ENSEMBLE DE PREDICTEURS

La sélection d'un sous-ensemble $X^{(j)}$ de prédicteurs et l'estimation de la qualité de la règle de classement associée sont deux problèmes statistiques étroitement liés et il est utile de distinguer les trois situations suivantes :

i) une partition de l'ensemble X est donnée a priori : $X = (X^{(1)},X^{(2)})$, $X^{(1)}$ et $X^{(2)}$ étant constitués de $p - r$ et r prédicteurs respectivement;

ii) un ordre naturel dans la sélection des variables $X^1,X^2,\ldots,X^p$ existe a priori ; il s'agit de choisir un sous-ensemble de prédicteurs $X^{(1)} = (X^1,X^2,\ldots,X^{p-r})$ parmi les p sous-ensembles emboîtés $\{X^1\}, \{X^1, X^2\} \ldots \{X^1, X^2, \ldots, X^p\}$;

iii) il n'y a aucun ordre, aucune structure a priori dans l'ensemble des prédicteurs.

Dans la situation (i), il s'agit de choisir entre $X^{(1)}$ et l'ensemble X de tous les prédicteurs; l'aspect minimal de la sélection n'est pas pris en compte et l'hypothèse $H(jo)$ s'identifie à l'absence d'information additionnelle apportée par $X^{(2)}$ sachant $X^{(1)}$ pour la discrimination. Cette hypothèse, notée $H(0)$, peut être testée au moyen des statistiques d'information additionnelle de Rao [*Rao* 1973, *Celeux* 1989] ou d'Akaike; on en rappelle le contexte.

Supposons la loi F de paramètres (μ_1, μ_2, Σ) connue. Le pouvoir discriminant Δ^2 associé à $X = (X^{(1)},X^{(2)})$ se décompose de la manière suivante :

$$(2)\quad \Delta^2 = \Delta_1^2 + \Delta_{2\,1}^2$$

où Δ_1^2 est la distance entre P1 et P2 associée au fichier $X^{(1)}$ et $\Delta_{2\,1}^2$ l'information additionnelle apportée par $X^{(2)}$ sachant $X^{(1)}$. En l'absence d'information addi-

tionnelle (hypothèse H0) on a : H0: $\Delta_1^2 = \Delta^2$, soit : $\Delta_{21}^2 = 0$. De manière équivalente, les coefficients de la fonction linéaire discriminante $\lambda' X$ où $\lambda = \Sigma^{-1}(\mu_1 - \mu_2)$ sont nuls sur $X^{(2)}$ [*Kshirsagar* 1972,*Rao* 1973]. Soit (λ'_1, λ'_2) la décomposition de λ' correspondant à la partition $(X^{(1)},X^{(2)})$ des prédicteurs, l'hypothèse nulle s'écrit encore : $\lambda'_2 = 0$.
Si $X^{(1)}$ et $X^{(2)}$ sont composés de $p - r$ et r variables respectivement, la statistique du test de l'hypothèse d'absence d'information additionnelle de Rao est la suivante :

$$(3)\quad T_{2.1} = ((c1 - p + 1)/r)[(T_p^2 - T_{p-r}^2)/(c1 + T_{p-r}^2)]$$

où $c1 = (n_1 + n_2 - 2)$, T_p^2 et T_{p-r}^2 désignant les statistiques de Hotelling associées à X et $X^{(1)}$ respectivement. Cette statistique a été décrite en détails par Celeux (1989).

Une autre statistique de test de l'hypothèse H0, fondée sur le principe du maximum de vraisemblance, a été proposée par Akaike.

Soit $f(\Theta)$ la fonction de vraisemblance associée au fichier (y,x) et $\hat{\theta}_1$ l'estimateur du maximum de vraisemblance de $\Theta = (\mu_1, \mu_2, \Sigma)$ sous l'hypothèse H0. Le nombre v de paramètres à estimer sous cette hypothèse vaut :
$v(p - r) = 2p - r + p(p + 1)/2$. On considère la statistique AIC_1 définie comme suit :

$$(4)\quad AIC_1 = -2log[f(\hat{\theta}_1)] + 2v(p - r)$$

le nombre $2v(p - r)$ doit être interprété comme une pénalité affectée à la vraisemblance selon le nombre de paramètres du modèle à estimer. Cette statistique est notée AIC lorsque le modèle est complet ; $\hat{\theta}$ est alors l'estimateur du maximum de vraisemblance usuel de (μ_1, μ_2, Σ).
Le test de l'hypothèse H0 peut être fondé sur la statistique $A_{2.1}$:

$$(4')\quad A_{2.1} = AIC_1 - AIC$$

On peut montrer que les deux statistiques $T_{2.1}$ et $A_{2.1}$ sont équivalentes dans la situation (i) ($A_{2.1}$ est une fonction de $T_{2.1}$ et de $p - r$). Sous les hypothèses du §1, $T_{2.1}$ suit une loi de Fisher à r et $(c1 - p + 1)$ degrés de liberté, centrée si H0 est vraie.

Dans la situation (ii) on est amené à effectuer p tests simultanément. Si le sous-ensemble $X^{(1)}$ est acceptable pour la discrimination, alors tout sous-ensemble de X contenant $X^{(1)}$ doit également l'être. Un test répondant à ce principe de cohérence est obtenu par l'utilisation simultanée de la statistique d'information additionnelle et du principe d'union-intersection de Roy. Cette procédure sera présentée plus loin et dans un cadre plus général. On peut aussi choisir $X^{(1)}$ au vu de l'évolution du critère $A_{2.1}$ en fonction du nombre de variables introduites dans le modèle. Le rôle joué par la pénalité $2v(p - r)$ prend ici tout son sens : lorsque l'introduction d'une nouvelle variable dans le modèle n'en améliore pas sensiblement la vraisemblance, la pénalité a pour effet de rejeter ce modèle au profit d'un modèle de plus petite dimension.

On peut aussi utiliser le critère d'information additionnelle de Rao pour la recherche de $\{jo\}$; cependant, le caractère minimal de ce sous-ensemble nécéssite l'introduction d'une procédure d'arrêt dans la sélection des prédicteurs. On no-

tera que, dans la situation présente, les critères de Rao et d'Akaike ne sont pas équivalents et ne vérifient pas nécéssairement le principe de cohérence.
L'utilisation simultanée du critère d'information additionnelle et du principe d'union-intersection de Roy est aussi plus rigoureuse en ce sens que l'on contrôle le risque d'erreur de sélection (tests emboités); en contre-partie, comme pour tout test d'hypothèse nulle multiple, cette procédure est conservatrice: trop de sous-ensembles sont déclarés admissibles. En d'autres termes, le sous-ensemble minimal retenu ne contient pas en général toute l'information utile pour la discrimination.

La situation (iii) est la plus complexe. L'élément jo doit être choisi parmi tous les sous-ensembles de J. Lorsque le nombre p de candidats prédicteurs est important, la question qui se pose est celle-ci: peut-on trouver un sous-ensemble de variables acceptable pour la discrimination, lorsque le choix doit se faire parmi un très grand nombre de sous-ensembles candidats? A cette question se rattache évidemment celle de l'estimation de la qualité de la règle de décision associée au sous-ensemble de prédicteurs finalement retenu. Pour éclairer la nature du problème, il est utile d'exhiber les deux biais inhérents à toute procédure de sélection.

B. BIAIS PAR OMISSION ET BIAIS DE SELECTION

L'élément $j_e(y,x)$ de J optimisant le critère choisi (réalisant, par exemple, le minimum d'un estimateur de $R(j)$), dépend évidemment du fichier d'apprentissage (y,x), réalisation du couple aléatoire (Y,X) de loi F. Dans le cadre non-conditionnel, tout critère de sélection conduit à une distribution de probabilité sur J dépendant de F (que la loi de F soit connue ou non). En conséquence des résultats du §2, s'il existe un élément jo vérifiant l'hypothèse $H(jo)$, alors l'espérance de l'erreur vraie de classement associée à la procédure de sélection est supérieure à $R(jo)$. De même, la distance de Mahalanobis correspondante est inférieure en espérance à $\Delta = \Delta_{jo}$. Ce biais s'écrit :

$$(5)\quad b_o = E[R(j_e(Y,X))] - R(jo)$$

il est appelé 'biais par omission' ; il vaut $Err(j_e(y,x)) - Err(jo)$ sur le fichier d'apprentissage. Ce biais n'est faible que si la distribution de probabilité sur J est concentrée sur des éléments $\{j\}$ tels que $R(j)$ soit très proche de $R(jo)$. Cela dépend, bien entendu, de la loi F.

Un sous-ensemble $X^{(j)}$ étant sélectionné, il s'agit ensuite d'estimer $Err(j_e(y,x))$, le vrai taux d'erreur d'affectation associé à la règle estimée sur le fichier $(y,x^{(j_e)})$. Pour mieux décrire la difficulté qui apparaît ici, on fixe a priori le nombre de prédicteurs à retenir; autrement dit, l'élément jo est supposé appartenir à la sous-famille $J(p-r)$ décrivant tous les sous-ensembles composés de $p-r$ prédicteurs. Dans cette situation, celà a un sens de prendre comme critère de sélection un estimateur sans biais de la distance $\Delta^2(j)$ de Mahalanobis entre P1 et P2 associée à $X^{(j)}, j \in J(p-r)$; soit $D^2(j)$ cet estimateur [*Tomassone&al.*1988]. Le problème qui apparaît est alors le suivant: supposons que dans $J(p-r)$, plusieurs sous-ensembles $\{j\}$ soient en compétition, c'est à dire tels que les distances $\Delta^2(j)$ qui leur sont associées soient voisines; la quantité $D^2(j_e)$ maximisant $D^2(j)$ sur $J(p-r)$, n'est pas un estimateur sans biais de la distance de Mahalanobis entre les deux populations P1 et P2 décrites par le vecteur $X^{(j_e)}$ de prédicteurs. Cette distance est surestimée et le biais positif peut être important lorsque plusieurs sous-ensembles $X^{(j)}$ sont en compétition. Ce biais optimiste est appelé 'biais de sélection' ; son amplitude est de l'ordre suivant [*Schaafsma*1982] :

(6) $bs = c(p-r)[Var(D^2_{je})]^{\frac{1}{2}}$ où $c(p-r) \sim 3$ ou 4 si $p - r = p/2$.

L'importance du biais de sélection rend difficile le choix entre deux sous-ensembles de prédicteurs appartenant à $J(p-r)$ et $J(p-r')$ lorsque la taille du fichier est modérée, car l'introduction d'une pénalité -fonction du nombre de paramètres à estimer- ne suffit pas à rendre comparables les deux biais de sélection.

4. PRATIQUE DE LA SELECTION DES PREDICTEURS ET TAUX D'ERREUR DE CLASSEMENT

Sous l'éclairage du paragraphe précédent, nous décrivons brièvement trois types de procédures de sélection en situation générale: méthodes exhaustives, méthodes pas à pas et méthodes de choix partiel.

Une technique naturelle de sélection des prédicteurs consiste à rechercher l'élément $\{j_e\}$ minimisant l'estimation de $R(j)$ sur J, puisque $\{jo\}$ réalise le minimum de $R(j)$ sur J sous l'hypothèse $H(jo)$ (§2).
Parmi les estimateurs de R(j) -j étant fixé- l'estimateur $L(j)$ de Mc Lachlan [*McLachlan* 1976,1980] possède la propriété de converger assez rapidement vers $R(j)$ avec n. Cet estimateur est fonction de $[n_1, n_2, k(j), D^2(j)]$, où $k(j)$ représente le nombre de prédicteurs et $D^2(j)$ l'estimateur usuel de la distance de Mahalanobis entre P1 et P2, fondée sur l'observation du sous-ensemble $X^{(j)}$ de variables (voir l'article précédent). L'expression de la statistique $L(j)$ ainsi que ses propriétés asymptotiques sont décrites en détails dans [*McLachlan* 1976] , [*Anderson* 1984] , [*Fujikoshi* 1985].

On a le résultat asymptotique suivant [*Fujikoshi* 1985] :
La minimisation sur J du critère d'Akaike $AIC(j)$ et du critère $L(j)$ conduisent à des distributions de probabilité *sur* J asymptotiquement identiques (n_1 et n_2 tendant vers l'infini simultanément). autrement dit, ces deux critères de sélection exhaustive sont asymptotiquement équivalents (sous les hypothèses de normalité et d'identité des matrices des covariances). Cette distribution $\{p(j), j \in J\}$ n'est pas concentrée sur l'élément $\{jo\}$ lorsque la taille n du fichier (y,x) tend vers l'infini: $p_n(j)$ tend vers 0 sur $J2$, mais vers des quantités non nulles, explicitées dans [*Fujikoshi* 1985] , pour les éléments de $J1$. Le biais par omission (équations 6 et 6') tend vers 0 en $O(1/n)$.

Si de telles procédures exhaustives sont ainsi justifiées lorsque n est grand, en réalité le nombre de variables observées est souvent élevé, de sorte que leur mise en oeuvre est trop coûteuse. On a alors recours à des méthodes sous-optimales: les méthodes pas à pas [*Celeux*,1989; *Hand*,1981]. Une difficulté nouvelle apparaît ici: elles ne conduisent pas, en général, à l'optimum sur J du critère choisi. Ce sont néanmoins les techniques les plus fréquemment utilisées en pratique dès que le nombre de variables est supérieur à 10. Elles sont le plus souvent fondées sur le critère d'information additionnelle de Rao, avec procédure d'arrêt dans la sélection [*Costanza et Affifi*, 1979] ,ou sur le critère d'Akaike.

Lorsque la taille du fichier (y,x) est modérée, celà ne signifie pas qu'une recherche exhaustive conduise toujours à un 'meilleur' sous-ensemble de prédicteurs que celui obtenu par une méthode pas à pas fondée sur le même critère. En effet, les procédures pas à pas ont pour effet de réduire -d'une certaine manière- le nombre de sous-ensembles candidats à comparer: en d'autres termes, le biais de sélection -qui peut fortement perturber la pertinence des prédicteurs choisis- semble s'en trouver, à chaque étape, réduit. Pour illustrer celà, on reprend certains des résultats de simulation de Murray [1977].

Le modèle est le suivant: X est composé de 10 prédicteurs non corrélés entre eux, de moyennes 1/4 et -1/4 sur P1 et P2 respectivement, et de variances égales à 1. Tous les prédicteurs ont ainsi le même pouvoir discriminant et $jo = \{1, \ldots, p\}$. Le critère de sélection choisi est le taux apparent d'erreur: *err*. Au moyen d'un grand nombre de fichiers (y,x) simulés, Murray a trouvé qu'en moyenne 5.5 prédicteurs étaient sélectionnés par recherche exhaustive et 5.9 par la méthode pas à pas (avec comme critère d'arrêt la non augmentation de *err*); les taux apparents d'erreur sont de 13.7% et 15.2% en moyenne, alors que les vrais taux d'erreur (non conditionnels) sont de 28.2% et 27.6% respectivement. On peut en déduire le biais par omission, puisque si $p=10$, $Err = 21.5\%$: il vaut 6.7% par recherche exhaustive et 6.1% par la méthode pas à pas. Les biais de sélection sont en moyenne de 14.5% et 12.4% respectivement.
Bien entendu, la situation décrite par le modèle choisi n'est pas agréable sous l'angle du problème de la sélection de prédicteurs; de même, le critère choisi *err* est très sensible aux fluctuations d'échantillonnage, de sorte que les biais sont importants. Cependant, outre le fait que cet exemple permet d'exhiber les deux biais de manière concrète, il montre aussi qu'une méthode exhaustive peut conduire à la sélection d'un plus petit nombre de prédicteurs qu'une méthode pas à pas. Notons aussi, que lorsque la taille du fichier passe de 50 à 100, les résultats ne sont que très peu modifiés [*Murray*,1977]. Ces résultats de simulation peuvent susciter d'autres commentaires [*Turlot*,1989].

D'un point de vue pratique, on a intérêt à jouer sur une réduction naturelle des sous-ensembles de prédicteurs à comparer, par la prise en compte de connaissances a priori: supposons que des études antérieures aient montré la pertinence d'un ensemble de variables pour la discrimination; il s'agit alors de voir si de nouvelles variables -reflètant par exemple un autre aspect du problème ou apportant un complément d'information- peuvent sensiblement améliorer la qualité de la règle de décision. Si l'information additionnelle contenue dans cet ensemble de variables exploratoires est importante, on peut adjoindre pas à pas et de manière ascendante, un petit nombre d'entre elles au système initial de prédicteurs. Par l'usage de variables dites 'forcées', le programme SELDIS de la bibliothèque MODULAD permet une telle approche visant à réduire le biais par omission, ce qui constitue une protection contre un choix de prédicteurs en partie artificiel.

5. ESTIMATION DES TAUX D'ERREUR DE CLASSEMENT

Lorsque la taille du fichier est modérée, les techniques d'estimation de $Err(y, x^{(je)};F)$ par rééchantillonnage proposées dans les programmes DISC (validation croisée) et DIS2G (bootstrap) ne prennent pas en compte le fait que $\{j_e\}$ est sélectionné au vu du fichier d'apprentissage au moyen des programmes FUWIL ou SELDIS.

Soit $Err(j_e) = Err(y,x^{(je)};F)$ le vrai taux d'erreur de classement associé au fichier $(y,x^{(je)})$; on note $op(j_e) = op(y,x^{(je)},F)$ l'optimisme de l'estimateur $err(j_e) = err(y,x^{(je)})$.

Le biais:

(7) $b(j_e) = E(Err(j_e) - err(j_e)) = E(op(j_e))$

est plus important que dans la situation de l'article précédent où j était supposé connu, de sorte que les mêmes techniques conduisent à des estimations des taux d'erreur de classement trop optimistes comme l'ont montré Snappin et Knoke [1989] au moyen de simulations.

Pour obtenir un estimateur plus précis de $b(j_e)$, on peut intégrer les techniques de rééchantillonnage dans la procédure de sélection.
Les prédicteurs étant sélectionnés (par exemple au moyen d'une procédure pas à pas) et la fonction discriminante estimée sur le fichier d'apprentissage, un estimateur bootstrap du taux d'erreur de classement associé à la règle de décision peut être construit de la manière suivante:

1. calcul de $err(j_e)$
2. tirage d'un échantillon bootstrap $(y,x)_\alpha$, puis sélection d'un sous-ensemble de prédicteurs et estimation de la règle de décision r_α (la même procédure de sélection étant appliquée au fichier d'apprentissage et au fichier bootstrap);
3. calculs de $err(\alpha)$, de $Err(\alpha)$ et de l'écart entre ces deux quantités définies dans l'article précédent ($Err(\alpha)$ ne représente pas le vrai taux d'erreur de classement associé à r_α, mais le taux apparent d'erreur observé sur le fichier (y,x));
4. répétition des étapes 2 et 3 sur un grand nombre A d'échantillons bootstrap;
5. calcul de:

$$\frac{1}{A}\sum_{\alpha}(Err(\alpha) - err(\alpha))$$

cette quantité définissant l'estimateur bootstrap du biais, compte-tenu de la procédure de sélection.

Des résultats de simulation [*Snappin et Knoke*,1989] montrent que le biais est approximé de manière satisfaisante par cet estimateur. Cependant, une telle démarche suggère plusieurs remarques:

- La quantité estimée n'est pas $b(j_e) = E(op(j_e))$, mais $b = \sum b(j)p(j)$. Il s'agit donc d'une correction moyenne calculée sur l'ensemble des systèmes de prédicteurs $X^{(j)}, j \in J$, selon la distribution de probabilité $\{p(j), j \in J\}$ qui dépend à la fois de la loi F du vecteur aléatoire (Y,X) et de la procédure de sélection.

- A chaque échantillon bootstrap correspond un sous-ensemble de prédicteurs dont la sélection ne prend pas en compte le biais par omission. Ceci signifie que, si la distribution $\{p'(j), j \in J\}$ associée à la loi empirique F' est fortement concentrée sur un élément $\{j\}$, il y a stabilité dans le choix des prédicteurs, mais pas nécéssairement absence de biais par omission.

- Le temps de calculs est important, que l'on utilise la technique bootstrap ou la validation croisée (dont les résultats sont comparables).

La pratique de l'échantillon test décrite dans l'article précédent demeure valable: $err(j_e)$ calculé sur le fichier d'épreuve constitue un estimateur sans biais de $Err(y,x^{(je)},F)$.
Plus récemment ont été proposées de nouvelles techniques d'estimation des taux d'erreur fondées sur des méthodes de lissage des probabilités de classement a posteriori [*Snappin et Knoke*, 1989] dont la mise en oeuvre est moins lourde. Il semble, au vu de simulations, qu'un estimateur du taux d'erreur d'affectation correctement lissé donne des résultats comparables à ceux obtenus par rééchantillonnage selon la démarche décrite ci-dessus.

CONCLUSION

Le choix des prédicteurs est une opération délicate: retenir l'ensemble des variables conduit à l'élaboration d'une fonction de classement non optimale si l'information additionnelle apportée par un sous-ensemble d'entre elles est nulle; effectuer une sélection de prédicteurs peut conduire à une fonction linéaire discriminante non satisfaisante en regard du potentiel de prévision inclus dans les données. Le cadre classique de deux populations normales ne différant que par leur moyenne permet d'exhiber clairement le biais de sélection et d'en évaluer l'importance tant dans le choix des prédicteurs (biais par omission) que dans l'estimation du taux d'erreur de classement par rééchantillonnage.
On peut intégrer le biais de sélection dans la recherche des sous-ensembles de prédicteurs admissibles, sous forme d'une hypothèse nulle multiple: $H(Jo)$, Jo décrivant la sous-famille des systèmes de prédicteurs $\{j\}$ vérifiant $\Delta_j^2 = \Delta^2$. Le critère d'information additionnelle de Rao, combiné avec le principe d'union-intersection de Roy permet de trouver à un niveau de probabilité donné, l'ensemble Jo. Cette procédure [*McKay* 1976,1977] vérifie le principe de cohérence: si $\{j\}$ est admissible, alors tout sous-ensembe $\{j'\}$ contenant $\{j\}$ l'est aussi; si $\{j\}$ n'appartient pas à Jo, aucun des sous-ensembles de $\{j\}$ n'est admissible; cette propriété est en accord avec la caractérisation de l'hypothèse $H(jo)$ (§2). En contre-partie, comme toute procédure multiple, elle est conservatrice: en d'autres termes, les solutions minimales admissibles ne contiennent en général pas toute l'information utile pour la discrimination. Ceci peut être interprété comme un effet du biais de sélection. On notera néanmoins que, si cette approche ne permet pas le choix du sous-ensemble optimal $\{jo\}$, elle exhibe les principaux systèmes pertinents de prédicteurs. En ce sens, on peut dire qu'il s'agit là d'une méthode d'analyse des données dans le cadre de la sélection des prédicteurs.

En somme, si l'on considère que l'objet de l'analyse discriminante est l'élaboration d'une règle de décision permettant de classer au mieux les observations futures, définissant les populations potentielles de l'étude, la statistique classique -sous la forme d'un modèle simple- peut éclairer la nature du problème posé. La qualité de la règle de classement dépend de la méthode de sélection des prédicteurs; or, il apparaît que la pertinence d'un critère -qu'il s'agisse d'une optimisation sur J ou d'une méthode sub-optimale- dépend de la structure même des données. Murray (1977), Schaafsma (1982) et Copas (1983) dans un article général sur la sélection de prédicteurs (dont le commentaire de Miller a inspiré en partie cette présentation), vont plus loin: ils mettent en doute la possibilité de sélection de prédicteurs pertinents au vu du fichier d'apprentissage. Toutefois, lorsque le nombre d'unités statistiques est important en regard du nombre de variables, les résultats de Fujikoshi (1985) montrent que ces propos doivent être nuancés.
En pratique, si l'on dispose d'un fichier de taille modérée, la sélection d'un sous-ensemble de prédicteurs effectuée sur la base de connaissances a priori (variables forcées), puis la sélection d'un petit nombre de prédicteurs parmi les variables exploratoires -si l'information additionnelle apportée par celles-ci est conséquente- constitue une protection contre une sélection en partie artificielle, le biais de sélection s'en trouvant réduit. De fait, l'estimation par rééchantillonnage des taux d'erreur de classement présente un plus faible biais.

BIBLIOGRAPHIE

Anderson (1984) : An Introduction to Multivariate Statistical Analysis ; Wiley.

Campbell (1978) : The Influence Function as an Aid in Outlier Detection in Discriminant Analysis ; Appl.Statist.,27, pp 251-258.

Celeux (1989) : Discrimination sur variables quantitatives; support de cours de l'école Modulad (Strasbourg 1989), INRIA.

Celeux & Turlot (1989) : Estimation de la qualité d'une règle discriminante; revue Modulad,4.

Copas (1983) : Regression, Prediction and Shrinkage; JRSS,B;45, pp 311- 354 (with discussion).

Costanza & Affifi (1979) : Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis ; JASA,74, pp777-785.

Daudin (1989) : Discrimination logistique; support de cours de l'école Modulad (Strasbourg 1989), INRIA.

Efron (1982) : The Jackknife,the Bootstrap and Others Resampling Plans ; SIAM,Monograph 38.

Efron (1983) : Estimating the Error Rate in a Predictive Rule : Improvement on Cross-Validation ; JASA,78, pp 316-331.

Fujikoshi (1985) : Selection Variables in Two-group Discriminant Analysis by Error Rate and Akaike's Information Criteria ; Journal of Multivariate Analysis,17, pp 27-37.

Gong (1986) : Cross-validation,the Jackknife, and the Bootstrap : Excess Error Estimation in Forward Logistic Regression. JASA, 81, pp 108-113.

Hand (1981) : Discrimination and Classification. Wiley.

Lachenbruch et Mickey (1968) : Estimation of Error Rates in Discriminant Analysis ; Technometics,10, pp1-11.

Mc.Kay (1976) : Simultaneous Procedures in Discriminant Analysis Involving Two Groups ; Technometrics,18, pp 47-53.

Mc.Kay (1977) : A Graphical Aid to Selection of Variables in Two-group Discriminant Analysis ; Appl.Statist.,27, pp 259-263.

Kshirsagar (1972) : Multivariate Analysis, Marcel Dekker,New-York.

Mc Lachlan (1976) : A Criterion for Selecting Variables for the Linear Discriminant Function. Biometrics, 32, pp 529-534.

Mc Lachlan (1980) : On the Relationship between the F Test and the Overall Error Rate for Variable Selection in Two-group Discriminant Analysis. Biometrics, 36, pp 501-510.

Murray (1977) : A Cautionnary Note on Selection of Variables in Discriminant Analysis ; Appl.Statist.,26, pp 246-250.

Rao (1973) : Linear Statistical Inference and its Applications. Wiley.

Schaafsma (1982) : Selecting Variables in Discriminant Analysis for Improving upon Classical Procedures ; Handbook of Statistics,Vol 2.

Snapinn & Knoke (1985) : An Evaluation of Smoothed Classification Error Rate Estimators ; Technometrics,27, pp 199-206.

Snapinn & Knoke (1989) : Estimation of Error Rates in Discriminant Analysis with Selection of Variables ; Biometrics,45, pp 289-299.

Stone (1977) : An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's criterion ; JRSS,B,39, pp 44-47.

Tomassone, Danzart, Daudin et Masson (1988) : Discrimination et Classement. Masson.

Turlot (1989) : Validité de la règle de classement en discrimination; support de cours de l'école Modulad (Strasbourg 1989), INRIA.