REVUE DE STATISTIQUE APPLIQUÉE

A. GUEGUEN
J. P. NAKACHE

Méthode de discrimination basée sur la construction d'un arbre de décision binaire

Revue de statistique appliquée, tome 36, nº 1 (1988), p. 19-37
<http://www.numdam.org/item?id=RSA_1988__36_1_19_0>

© Société française de statistique, 1988, 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/

Rev. Statistique Appliquée, 1988, **XXXVI** (1), 19-38

MÉTHODE DE DISCRIMINATION BASÉE SUR LA CONSTRUCTION D'UN ARBRE DE DÉCISION BINAIRE

A. GUEGUEN, J.P. NAKACHE

Unité de Recherche INSERM U88,
91 boulevard de l'Hôpital, 75634 Paris Cedex 13

I. Introduction

La méthode de discrimination présentée dans cet article est basée sur la construction d'un arbre de décision binaire. Cette approche est différente des méthodes classiques parmi lesquelles :

— l'analyse discriminante linéaire ou quadratique et la régression logistique qui fournissent des règles de décision sous la forme d'expressions algébriques pas toujours faciles à comprendre, à analyser et à interpréter,
— la méthode de discrimination basée sur l'estimation des densités de probabilité par groupe et la méthode des k plus proches voisins qui sont sensibles au choix de la métrique et qui ne permettent pas de prendre en compte de façon naturelle et simple les variables qualitatives et les données manquantes.

L'utilisation des arbres binaires remonte au programme A.I.D. (Automatic Interaction Detection) proposé par J.A. MORGAN et J.N. SONQUIST [2] dans les années 60. Les importants développements théoriques récents sont dûs à L. BREIMAN *et al.* [1] qui proposent de construire un arbre binaire sans s'imposer de règle d'arrêt de la procédure de division des nœuds. Une procédure d'élagage astucieuse du grand arbre obtenu et l'utilisation d'un échantillon test leur permet d'en extraire le « meilleur » sous-arbre, celui qui correspond à l'estimation la plus fiable du taux d'erreur théorique de classement.

Cette méthode présentée dans le présent article est illustrée par un exemple médical concernant le diagnostic de malignité en présence d'un nodule du foie chez 382 patients divisés en deux groupes : 259 sujets présentant une tumeur maligne et 123 une tumeur bénigne. Le Tableau I fournit la liste des 13 variables qualitatives relevées sur chacun des sujets de l'échantillon.

Un problème de discrimination se pose quand on est en présence d'un tableau de données contenant n sujets répartis en k groupes $G_1, G_2, \ldots, G_k$. Il s'agit alors, d'une part de sélectionner parmi les variables du tableau celles qui sont les plus discriminantes, et d'autre part, de construire une règle de décision permettant d'affecter un nouveau sujet à l'une de ces k classes.

Mots clés : *Discrimination, Arbre de décision binaire, Elagage, Validation croisée, Echantillon test, Données manquantes.*

TABLEAU I

Variables de l'exemple médical

Perte de poids	PPO	1-non	2-oui	
Ascites	ASC	1-non	2-oui	
Douleur (quad. sup. DT)	DOU	1-non	2-oui	
Fièvre	FIE	1-non	2-oui	
Jaunisse	JAU	1-non	2-oui	
Hépatomégalie	HEP	1-non	2-oui	
Hémochromatose	HEM	1-non	2-oui	
Cirrhose	CIR	1-non	2-oui	
Cancer primitif	CPR	1-non	2-oui	
Hépatite antérieure	HAN	1-non	2-oui	
Taux de sédimentation	SED	1-norm.	2-augm.	
Phosph. alcalines	PAL	1-norm.	2-augm.(+)	3-augm.(+)
G.G. transférase	GGT	1-norm.	2-augm.(+)	3-augm.(+)

Dans la méthode présentée, les variables explicatives peuvent être de nature quelconque, mais toute variable quantitative doit être préalablement transformée en une variable discrète ayant suffisamment de modalités de sorte que la perte d'information entraînée par ce codage soit négligeable.

II. Arbre de décision binaire Notions requises et construction

II.1. Arbre de décision binaire

Un tel arbre est construit à l'aide de *divisions successives* de sous-ensembles de l'échantillon en deux descendants. L'idée fondamentale est de sélectionner chaque division d'un nœud de telle sorte que les nœuds *descendants* soient plus « purs » que le nœud *parent.* Autrement dit, que le mélange des groupes soit moins important dans les nœuds descendants que dans le nœud parent.

La figure 1 représente un arbre de décision binaire illustratif où l'on distingue deux types de nœuds :

— les *nœuds intermédiaires* (nœuds non terminaux) entourés d'un cercle; ce sont les nœuds qui fournissent deux descendants *immédiats* : par exemple t_3 qui se divise en t_6 et t_7 ;
— les *nœuds terminaux* (feuilles de l'arbre) entourés d'un carré; ce sont les nœuds qui ne sont plus divisés.

On distingue également :

— les *descendants terminaux* d'un nœud intermédiaire : par exemple les nœuds t_5, t_9, t_{14} et t_{15} sont les descendants terminaux de t_2 ;
— la *branche* B_t d'un nœud t qui est l'ensemble de tous les descendants de t, la *racine* t de la branche n'étant pas incluse;
— un *sous-arbre* de l'arbre T est obtenu à partir de T après avoir *élagué* une ou plusieurs branches.

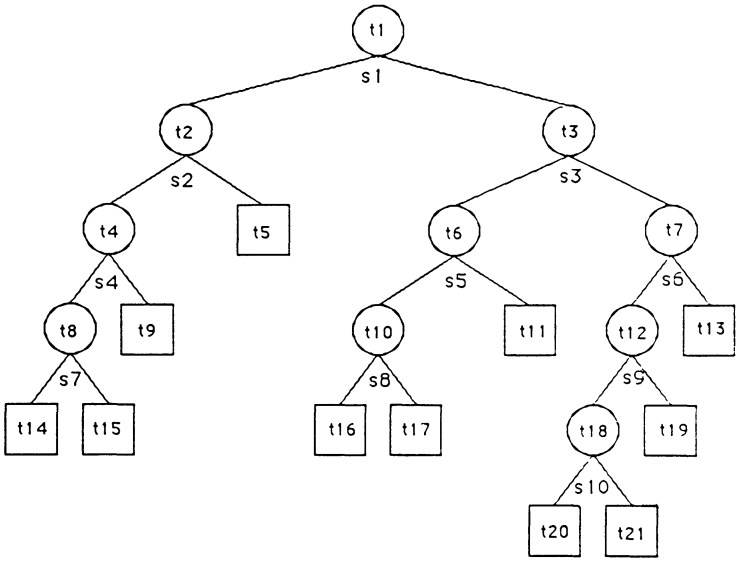


FIGURE 1
Arbre de décision binaire illustratif.

La construction d'un tel arbre nécessite la définition d'une règle de division d'un nœud, ou plus précisément, la définition d'un *critère de sélection de la meilleure division d'un nœud.*

II.2. Divisions possibles d'un nœud

Le nombre de divisions d'un nœud t obtenues à l'aide d'une variable dépend de la nature de cette variable :

— une variable *binaire* (0,1) fournit une seule division en deux nœuds t_g et t_d : t_g contient l'ensemble des sujets de t présentant la valeur 0 et t_d l'ensemble des sujets de t pour lesquels la variable prend la valeur 1;
— une variable *qualitative ordinale* à k modalités fournit $(k-1)$ divisions;
— une variable *qualitative nominale* à k modalités donne lieu à $2^{k-1}-1$ divisions. Ce résultat est obtenu aisément en utilisant un raisonnement par récurrence.

Par exemple les divisions possibles d'un nœud t, à l'aide de 3 variables dont une variable binaire (b_1, b_2), une variable ordinale (o_1, o_2, o_3, o_4) et une variable nominale (n_1, n_2, n_3), sont au nombre de 7 :

t_g	t_d
(b_1)	(b_2)
(o_1)	(o_2, o_3, o_4)
(o_1, o_2)	(o_3, o_4)
(o_1, o_2, o_3)	(o_4)
(n_1)	(n_2, n_3)
(n_2)	(n_1, n_3)
(n_3)	(n_1, n_2)

II.3. Division admissible d'un nœud

C'est une division de t en deux nœuds descendants contenant un nombre de sujets supérieur à n_0 qui est à fixer par l'utilisateur. Il est égal à 5 dans le cas de notre exemple.

II.4. Impureté d'un nœud t

A chaque nœud t de l'arbre est associé une mesure de l'impureté i(t) qui représente le degré de mélange des groupes dans t. Dans le cas général de k groupes, i(t) a la forme suivante :

$$i(t) = 2\{\Sigma[P(r/t) \cdot P(s/t);\ r > s = 1, 2, \ldots, k]\}$$

où P(r/t) est la proportion de sujets du groupe G_r dans le nœud t.

En utilisant le résultat :

$$1 = [\Sigma_r P(r/t)]^2 = \Sigma_r[P^2(r/t)] + 2\{\Sigma_{r,s}[P(r/t) \cdot P(s/t);\ r > s]\}$$

il vient :

$$i(t) = 1 - \Sigma_r[P^2(r/t);\ r = 1, 2, \ldots, k]$$

Un nœud est dit « pur » s'il ne contient que des sujets d'un seul groupe : dans ce cas i(t) = 0. Et plus i(t) est élevé, plus important est le mélange des groupes dans t.

Dans le cas particulier de deux groupes (k = 2) :

$$i(t) = 2\,P(1/t) \cdot P(2/t)$$

II.5. Meilleure division d'un nœud

Chaque division d d'un nœud t conduit à une *réduction de l'impureté* dont l'expression est :

$$\Delta i(d, t) = i(t) - p_g\, i(t_g) - p_d\, i(t_d)$$

où p_g et p_d sont les proportions des sujets de t allant respectivement dans les descendants t_g et t_d.

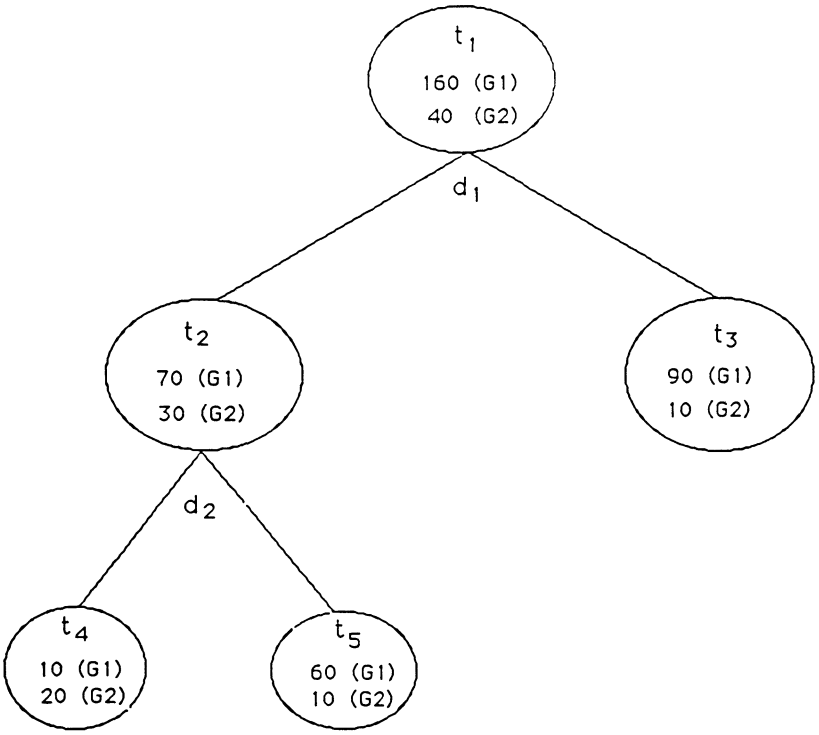
$\Delta i(d, t)$ représente la différence entre l'impureté du nœud parent et la moyenne pondérée des impuretés de ses nœuds descendants immédiats.

La meilleure division d'un nœud t est, parmi toutes les divisions admissibles de t, la division d* qui entraîne la plus grande réduction de l'impureté soit :

$$\Delta i(d^*, t) = \max_d\{\Delta i(d, t)\}$$

Remarques :

1) Ce critère de division, basé sur la réduction de l'impureté, est plus approprié que celui basé sur la réduction du pourcentage de sujets mal classés dans les deux nœuds descendants. En effet, l'examen de l'exemple de la figure 2 montre que, en utilisant un critère basé sur la réduction du nombre de mal classés, le nœud t_1 serait un nœud terminal avec 20 % de sujets mal classés; en utilisant le critère basé sur la réduction de l'impureté, t_1 est divisé successivement par d_1



Réduction de l'impureté	Pourcentage de mal classés	
$\Delta i(d_1, t_1) = 0{,}01$	avant d_1	20%
$\Delta i(d_2, t_2) = 0{,}057$	après d_1	20%
	après d_2	15%

FIGURE 2

(bien que le pourcentage de sujets mal classés reste stable) puis par d_2. A la suite de ces deux divisions le pourcentage de sujets mal classés tombe à 15 %.

2) Dans un nœud t donné, quelle que soit la division d, la réduction de l'impureté $\Delta i(d, t)$ est toujours positive ou nulle.

En effet l'impureté $i(t) = 2\,P(1/t) \cdot P(2/t)$ est une fonction concave de $P(1/t)$:

$$p_g\, i(t_g) + p_d\, i(t_d) = p_g[2\,P(1/t_g)(1 - P(1/t_g))] + p_d[2\,P(1/t_d)(1 - P(1/t_d))]$$
$$\leq 2\,\{p_g\,P(1/t_g) + p_d\,P(1/t_d)\}\,\{1 - [p_g\,P(1/t_g) + p_d\,P(1/t_d)]\}$$
$$= 2\,P(1/t)\,[1 - P(1/t)] = i(t).$$

Soit m(t) la proportion de mal classés dans le nœud t

$$m(t) = \min\,[P(1/t)\,,\ 1 - P(1/t)]$$

m(t) étant également une fonction concave de $P(1/t)$ la réduction du nombre de mal classés est toujours positive ou nulle, quelle que soit la division d du nœud t.

II.6. Construction de l'arbre, règle d'affectation, taux d'erreur apparent

L'arbre est ainsi construit par divisions successives. La procédure s'arrête quand tous les nœuds sont terminaux; un nœud étant déclaré terminal s'il est pur ou s'il n'y a pas de division(s) admissible(s) pour ce nœud.

La construction de l'arbre se fait, en général, après avoir divisé l'échantillon total en un échantillon de base (E.B.) et un échantillon test (E.T.) dont l'utilisation particulière est fournie en IV.

Dans notre exemple médical, l'échantillon total de 382 patients a permis d'obtenir, par tirage au sort équilibré, un E.B. de 255 sujets représentant 2/3 des patients et un E.T. de 127 sujets correspondant aux 382-255 = 127 patients restants. La figure 3 représente l'arbre T construit à partir de cet E.B.

Cet arbre T possède 16 nœuds terminaux, chacun d'eux étant affecté à l'un des deux groupes bénins ou malins : un nœud est affecté au groupe des tumeurs malignes si la proportion de ces tumeurs est la plus élevée dans ce nœud. On définit ainsi une *règle d'affectation* RA(T) associée à l'arbre T.

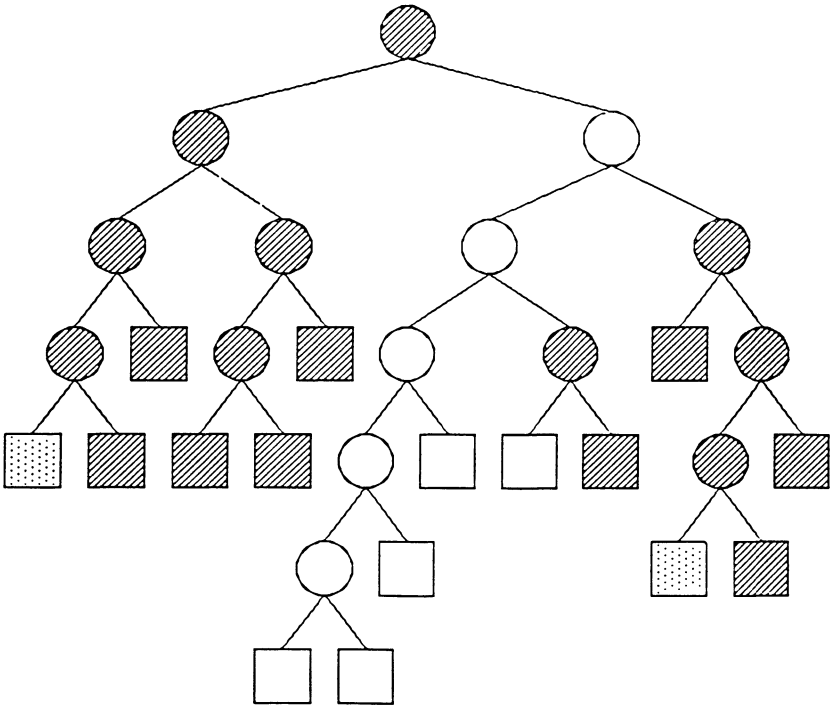


FIGURE 3

Arbre T construit à partir de l'échantillon de 255 patients 2/3 de l'échantillon total (tirage au sort équilibré) : E.B.

(TEAC = 8,63 %)

▨ malins; □ bénins; ⊡ % malins = % bénins

Un *taux d'erreur apparent* de classement (TEAC) est associé à tout arbre de décision binaire. Il représente la proportion de sujets mal classés dans l'ensemble des nœuds terminaux. Dans notre exemple, le TEAC associé à l'arbre T est égal à 8,63 %.

III. Originalité de la méthode présentée

Les méthodes de segmentation existantes sont basées sur la recherche d'une règle, la plus appropriée, d'arrêt de la procédure de division des nœuds, c'est-à-dire sur la recherche d'un critère permettant de déclarer un nœud terminal.

Si le critère de division d'un nœud est basé sur la réduction de l'impureté $\Delta i(d, t)$, une règle d'arrêt consiste à déclarer t nœud terminal si $\Delta i(d, t)$ est inférieur à un seuil β fixé.

Or, un seuil β choisi trop petit conduit à de très nombreuses divisions : on obtient ainsi un arbre très grand avec un taux d'erreur apparent TEAC associé qui donne une estimation trop optimiste du *taux d'erreur théorique* de classement noté dans la suite TETC.

D'autre part, si β est choisi trop grand, l'arbre obtenu est petit et ne tient pas compte de toute l'information discriminante contenue dans l'échantillon; il entraîne un TEAC qui, s'il estime correctement le TETC associé, est trop important : en effet, supposons qu'un nœud t soit déclaré nœud terminal à cause d'une réduction de l'impureté inférieure au seuil β fixé. Un seuil β plus petit aurait pu permettre de diviser t en deux nœuds descendants immédiats susceptibles d'être eux-mêmes divisés de telle manière que la réduction de l'impureté correspondant à ces dernières divisions soit très importante. Ainsi, avec un seuil β trop important on peut être conduit à perdre de bonnes divisions et à ne pas utiliser toute l'information contenue dans l'échantillon.

Par conséquent, l'utilisation d'une règle d'arrêt de la procédure de division des nœuds, principale faiblesse des méthodes de segmentation existantes, ne permet pas d'obtenir un arbre de décision binaire « fiable ».

L'idée fondamentale de la méthode proposée par BREIMAN *et al.* est de construire l'arbre le plus grand possible T_{max} à l'aide d'un échantillon de base en utilisant la procédure décrite en II, ce qui revient en fait à utiliser une règle d'arrêt avec un seuil β fixé à zéro. Un *échantillon test* est ensuite utilisé pour sélectionner le sous-arbre de T_{max} dont le TEAC est le plus petit possible tout en fournissant une estimation correcte du TETC; on notera T* ce « meilleur » sous-arbre.

IV. Obtention du meilleur sous-arbre

La solution idéale est la suivante :

Soit H le nombre de nœuds terminaux de T_{max}. Pour toute valeur h telle que $0 \leqslant h \leqslant H - 1$, on considère l'ensemble C_h des sous-arbres de T_{max} ayant $H - h$ nœuds terminaux et on définit T_h tel que :

$$\text{TEAC}(T_h) = \min \{\text{TEAC}(T) \,;\, T \in C_h\}$$

On obtient ainsi une séquence de sous-arbres de T_{max} $\{T_0, T_1, \ldots, T_h, T_{H-1}\}$ dont le premier élément est $T_0 = T_{max}$ et le dernier élément est $T_{H-1} = t_1$ qui est le sous-arbre de T_{max} à un nœud terminal, soit l'échantillon total.

Le TEAC associé à chacun des sous-arbres de cette séquence décroît avec le nombre de nœuds terminaux, soit :

$$\text{TEAC}(T_{h+1}) \geqslant \text{TEAC}(T_h) \quad \text{avec} \quad 0 \leqslant h \leqslant H - 2$$

En effet (démonstration par l'absurde) : supposons qu'il existe h tel que $0 \leqslant h \leqslant H - 2$ et vérifiant l'inégalité TEAC $T_{h+1}) < \text{TEAC}(T_h)$.

On note T^+_{h+1} un des arbres obtenus après avoir divisé un des nœuds terminaux de T_{h+1}. T^+_{h+1} possède $H - h$ nœuds terminaux et il appartient à la classe C_h.

$\text{TEAC}(T^+_{h+1}) \leqslant \text{TEAC}(T_{h+1})$ puisque T^+_{h+1} est un sous-arbre de T_{h+1}.

Or, d'après l'hypothèse, $\text{TEAC}(T_{h+1}) < \text{TEAC}(T_h)$ donc $\text{TEAC}(T^+_{h+1}) < \text{TEAC}(T_h)$ et T_h n'est pas l'arbre qui minimise le TEAC dans la classe C_h.

Le meilleur sous-arbre T^* est ensuite choisi dans la séquence $\{T_0, T_1, \ldots, T_{H-1}\}$ à l'aide d'un *échantillon* test de la façon suivante : pour chaque sous-arbre T_h de la séquence, la règle d'affectation RA (T_h) est appliquée à l'échantillon test et on obtient ainsi un taux d'erreur estimé de classement associé au sous-arbre T_h, noté TEEC (T_h). T^* est alors celui qui rend minimum le TEEC, soit :

$$\text{TEEC}(T^*) = \min\{\text{TEEC}(T_h) \,;\, 0 \leqslant h \leqslant H - 1\}$$

Cependant, même dans le cas d'un arbre T_{max} de taille modérée, il existe un très grand nombre de sous-arbres. Aussi la procédure ci-dessus qui nécessite un algorithme de programmation dynamique compliqué, entraînerait un coût de calcul tellement élevé qu'il la rend prohibitive en pratique.

IV.1. Obtention d'une séquence optimale

En conséquence, Breiman *et al.* proposent d'utiliser une procédure d'*élagage* permettant d'obtenir une sous-séquence de la séquence décrite ci-dessus et appelée *séquence optimale* dans la suite.

Cette séquence optimale se construit, pas à pas, à partir de $T_0 = T_{max}$ en élaguant successivement certaines branches de T_{max}. Ainsi au pas p le sous-arbre $T_{(p)}$ à $t_{(p)}$ nœuds terminaux est obtenu à partir de $T_{(p-1)}$ à $t_{(p-1)}$ nœuds terminaux.

IV.2. Procédure d'élagage

On définit le *critère d'élagage* suivant :

$$\pi(t) = \frac{\text{MC}(t) - \text{MCDT}(t)}{\text{N}(t) \cdot [\text{NDT}(t) - 1]}$$

où MC (t) est le nombre de sujets mal classés du nœud t, MCDT (t) est le nombre de sujets mal classés dans les descendants terminaux de t, N (t) est le nombre de sujets contenus dans le nœud t, et, NDT (t) est le nombre de descendants terminaux de t.

La valeur de ce critère π (t) est calculée pour chaque nœud intermédiaire de $T_{(p-1)}$. Cette valeur est d'autant plus élevée que, d'une part la différence entre la proportion de sujets mal classés dans le nœud t et la proportion de sujets mal classés dans les descendants terminaux du nœud t est grande et que, d'autre part le nombre de descendants terminaux du nœud t est faible.

On élague le sous-arbre $T_{(p-1)}$ en coupant la branche B_{t^*} telle que t^* soit le nœud correspondant à la valeur minimale du critère. On obtient le sous-arbre $T_{(p)}$ à $t_{(p)} = t_{(p-1)} - NDT(t^*) + 1$ nœuds terminaux.

La procédure d'obtention de la séquence optimale $\{T_0, T_{(1)}, \ldots, T_{(p)}, \ldots, T_{(P)}\}$ en pas à pas est très rapide et on montre (BREIMAN *et al.*) que cette séquence est bien une sous-séquence de la séquence idéale $\{T_0, T_1, \ldots, T_{H-1}\}$.

Le meilleur sous-arbre T^* est ensuite sélectionné à l'intérieur de cette séquence optimale en se basant sur l'utilisation d'un échantillon test.

IV.3. Application à l'exemple médical

La figure 4 fournit les valeurs du critère d'élagage pour les différents nœuds intermédiaires du grand arbre $T_0 = T_{max}$ construit à partir de l'échantillon de base de 255 patients. Les branches représentées en traits gras peuvent être élaguées sans accroissement du TEAC. En effet, la valeur du critère associé aux nœuds racines de ces branches est nulle puisque les descendants terminaux de ces nœuds racines sont affectés au même groupe. Ainsi le TEAC associé au sous-arbre $T_{(1)}$ comprenant les 5 nœuds terminaux numérotés 2, 12, 22, 23, 7 est le même que celui de l'arbre $T_0 = T_{max}$.

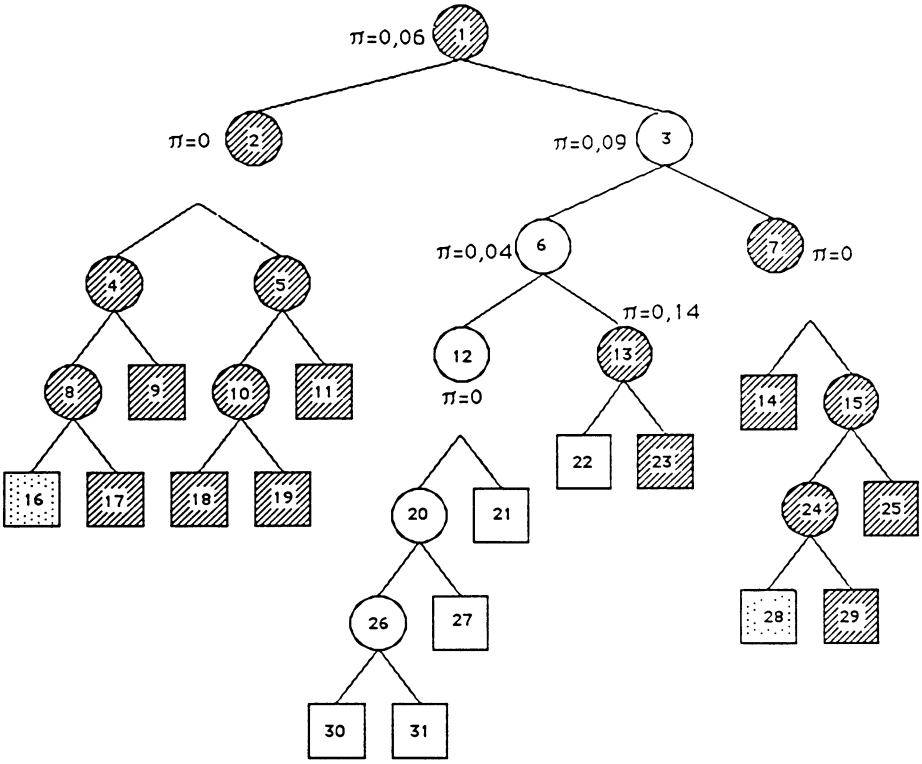
La procédure d'élagage est alors effectuée à partir de $T_{(1)}$ et le Tableau II fournit les résultats de cette procédure appliquée à notre échantillon de base.

La règle d'affectation associée à chacun des sous-arbres de la séquence est appliquée à l'échantillon test de 127 sujets représentant le tiers des patients de notre exemple médical. Les résultats obtenus sont résumés dans le Tableau III. L'arbre sélectionné est $T_{(1)}$ qui correspond à une estimation du TETC égale à 7,87 %.

Remarque : Il est possible de fournir un intervalle de confiance associé à cette estimation, en effet :

Soient n_1 et n_2 les tailles respectives de l'E.B. et de l'E.T. qui sont indépendants et qui suivent la même distribution. L'estimation du TECT est égale à la proportion p_t de sujets mal classés dans l'E.T.; et nous avons à résoudre le problème simple de l'estimation d'une proportion p_t et de sa variance qui est donc égale ici à $p_t(1 - p_t)/n_2$.

Dans notre exemple l'estimation du TETC, soit p_t est égale à 7,87 %; l'estimation de son écart-type est : $\sqrt{0{,}0787\,(1 - 0{,}0787)/127} = 0{,}024$. L'intervalle de confiance à 95 % du TETC est donc de 7,87 % ± 4,68 %.



Malins ; Bénins ; % Mal. = % Bén.

TEAC de l'arbre $T_0 = T_{max}$ 8,63%

TEAC de l'arbre $T_{(1)}$ 8,63%

FIGURE 4

TABLEAU II

Procédure d'élagage appliquée à l'arbre T_{max} obtenu à partir de l'échantillon de base (255 patients)

T_{max} : 16 nœuds terminaux — TEAC = 8.63 %
$T_{(1)}$: 5 nœuds terminaux — TEAC = 8.63 %

Obtention de $T_{(2)}$

t	N (t)	Bénins	Malins	MC (t)	MCDT (t)	NDT (t)	π (t)
1	255	82	173	82	22	5	.0588
3	131	77	54	54	17	4	.0941
6	88	70	18	18	10	3	.0451
13	14	4	10	4	2	2	.1428

$T_{(2)}$: 3 nœuds terminaux (2,6,7) — TEAC = 11.765 %

Obtention de $T_{(3)}$

1	255	82	173	82	30	3	.1019
3	131	77	54	54	25	2	.2213

$T_{(3)}$: 1 nœud terminal — TEAC = 32.15 %

Séquence optimale $\{T_{(1)}, T_{(2)}, T_{(3)}\}$

TABLEAU III

Sélection du meilleur sous-arbre à l'aide de l'échantillon test
($T_{(1)}$ correspondant à une estimation du TETC égale à 7,87 % ± 4,68 %)
NNT : nombre de nœuds terminaux
π : critère d'élagage
TEAC : taux d'erreur apparent de classement
TETC : taux d'erreur théorique de classement

Echantillon de base (2/3 des sujets) 255				Ech. test (1/3 des sujets) 127
Séquence optimale				
Arbres	NNT	π	TEAC	Estimation du TETC
T_{max}	16	0,0	8,63	7,87
$T_{(1)}$	5	0,0	8,63	7,87
$T_{(2)}$	3	0,045	11,77	11,81
$T_{(3)}$	1	0,102	32,15	32,28

V. Validation croisée

L'utilisation d'un échantillon test n'est intéressante que dans le cas de grands échantillons puisque la précision de l'estimation du TETC dépend de la taille de l'échantillon test. Si l'échantillon est de taille moyenne, ce qui est le cas dans notre exemple médical, il est conseillé d'utiliser une méthode de validation croisée. Cette méthode permet de prendre en compte tous les sujets de l'échantillon à la fois pour construire (échantillon de base) et pour tester (échantillon test) l'arbre.

L'échantillon total L est utilisé pour construire l'arbre T_{max} et la séquence optimale $S = \{T_{(j)}, J = 1, \ldots, P\}$ de laquelle sera extrait le meilleur sous-arbre. Cependant, on ne dispose pas d'échantillon test pour le sélectionner; aussi procède-t-on de la manière suivante :

— l'échantillon total L est divisé de manière aléatoire en k sous-ensembles disjoints $L_1, L_2, \ldots, L_k$ de même taille. Ainsi à chaque échantillon L_i est associé l'échantillon L^i complémentaire de L_i dans L.
— en utilisant la procédure décrite en IV et pour chaque i avec $i = 1, \ldots, k$, l'échantillon de base L^i fournit un arbre T^i_{max} et une séquence optimale $S_i = \{T_{(j),i}, j = 1, \ldots, P_i\}$.
— les règles d'affectation $RA(T_{(j),i})$ sont alors appliquées à l'échantillon test L_i et on note $n_{j,i}$ le nombre de sujets mal classés par la règle $RA(T_{(j),i})$ dans l'échantillon test L_i.

Les Tableaux IV et IV′ résument les résultats de cette procédure appliquée à notre exemple médical dans lequel nous avons fixé k à 10.

Ces calculs vont nous permettre d'estimer le TETC de chaque sous-arbre $T_{(p)}$ de la séquence S.

Pour cela, il faut donner une réponse à la question : quel est, pour chaque i ($i = 1, \ldots, k$), l'arbre de la séquence S_i le plus proche de $T_{(p)}$? BREIMAN *et al.* proposent d'utiliser les valeurs des critères d'élagage : pour chaque arbre $T_{(p)}$ de la séquence S on définit π'_p qui est la moyenne géométrique $(\sqrt{\pi_p \pi_{p+1}})$ de l'intervalle $[\pi_p, \pi_{p+1}[$. Dans la séquence S_i l'arbre $T_{(j),i}$ le plus proche de $T_{(p)}$ est celui dont la valeur du critère d'élagage π_{ji} est inférieure ou égale à π'_p tout en étant la plus grande possible. Par exemple, dans les séquences S_i ($i = 1, \ldots, 10$), les sous-arbres les plus proches de $T_{(2)}$ sont $T_{(3),1}$, $T_{(2),2}$, $T_{(1),3}$, $T_{(2),4}$, $T_{(2),5}$, $T_{(1),6}$, $T_{(1),7}$, $T_{(1),8}$, $T_{(2),9}$, $T_{(1),10}$. Les nombres de sujets mal classés par ces 10 sous-arbres dans les échantillons test sont $n_{3,1} = 2$, $n_{2,2} = 2$, $n_{1,3} = 6$, $n_{2,4} = 7$, $n_{2,5} = 3$, $n_{1,6} = 5$, $n_{1,7} = 4$, $n_{1,8} = 3$, $n_{2,9} = 1$ et $n_{1,10} = 4$, ce qui entraîne un total de 37 mal classés et une estimation du TETC de $T_{(2)}$ égale à $37/382 = 9.69\,\%$.

Ainsi à chaque sous-arbre de la séquence S (Tableau V) correspond un TEEC. Le meilleur sous-arbre est alors celui qui rend minimum le TEEC; soit, dans notre cas, l'arbre $T_{(3)}$ à 5 nœuds terminaux qui correspond à une estimation du TETC égale à 8,64 %.

Remarque : il est beaucoup plus difficile d'obtenir un écart-type de l'estimation du TETC quand la validation croisée est utilisée : en effet il n'y a pas indé-

TABLEAU IV

Séquences optimales des échantillons $L^1, \ldots, L^5$

L^1	$T_{(1),1}$	$T_{(2),1}$	$T_{(3),1}$	$T_{(4),1}$	$T_{(5),1}$	$T_{(6),1}$
$\pi.10^3$	0,0	2,01	4,81	9,43	44,00	100,58
mal classés dans L_1	1	1	2	2	4	13
L^2	$T_{(1),2}$	$T_{(2),2}$	$T_{(3),2}$	$T_{(4),2}$		
$\pi.10^3$	0,0	2,07	52,00	97,67		
mal classés dans L_2	3	2	2	13		
L^3	$T_{(1),3}$	$T_{(2),3}$	$T_{(3),3}$			
$\pi.10^3$	0,0	36,39	110,47			
mal classés dans L_3	6	10	12			
L^4	$T_{(1),4}$	$T_{(2),4}$	$T_{(3),4}$	$T_{(4),4}$	$T_{(5),4}$	
$\pi.10^3$	0,0	2,00	9,52	44,72	109,01	
mal classés dans L_4	7	7	7	9	12	
L^5	$T_{(1),5}$	$T_{(2),5}$	$T_{(3),5}$	$T_{(4),5}$	$T_{(5),5}$	$T_{(6),5}$
$\pi.10^3$	0,0	2,00	6,54	18,69	43,65	109,01
mal classés dans L_5	4	3	1	0	2	12

pendance entre les k échantillons de base et les k échantillons test. Cependant, si on accepte en pratique de ne pas tenir compte de cette non indépendance comme le font BREIMAN *et al.*, on obtient une formule semblable à celle obtenue quand la méthode de l'échantillon test est utilisée. Si p_c est l'estimation du TETC obtenue à partir des n sujets de l'échantillon total, l'estimation de l'écart-type du TETC est $\sqrt{p_c (1 - p_c)/n}$ soit : $\sqrt{0{,}0864 (1 - 0{,}0864)/382} = 0{,}0144$. D'où l'intervalle de confiance à 95 % du TETC égal à 8,64 % ± 2,82 %.

Application à l'exemple médical.

La figure 5 représente le meilleur sous-arbre sélectionné en utilisant la méthode validation croisée appliquée à notre exemple médical. La règle de décision est donnée sous la forme de cinq chemins dont 2 en faveur du groupe G_1 (tumeurs bénignes) et 3 en faveur du groupe G_2 (tumeurs malignes).

TABLEAU IV'

Séquences optimales des échantillons $L^6, \ldots, L^{10}$

L^6	$T_{(1),6}$	$T_{(2),6}$	$T_{(3),6}$	$T_{(4),6}$	
$\pi.10^3$	0,0	13,89	50,78	104,65	
mal classés dans L_6	5	6	6	12	
L^7	$T_{(1),7}$	$T_{(2),7}$	$T_{(3),7}$	$T_{(4),7}$	$T_{(5),7}$
$\pi.10^3$	0,0	9,26	27,52	47,24	100,29
mal classés dans L_7	4	4	2	3	12
L^8	$T_{(1),8}$	$T_{(2),8}$	$T_{(3),8}$	$T_{(4),8}$	$T_{(5),8}$
$\pi.10^3$	0,0	9,43	18,52	50,39	100,29
mal classés dans L_8	3	3	3	3	12
L^9	$T_{(1),9}$	$T_{(2),9}$	$T_{(3),9}$	$T_{(4),9}$	$T_{(5),9}$
$\pi.10^3$	0,0	1,98	18,69	44,00	100,29
mal classés dans L_9	1	1	1	3	12
L^{10}	$T_{(1),10}$	$T_{(2),10}$	$T_{(3),10}$		
$\pi.10^3$	0,0	46,61	107,56		
mal classés dans L_{10}	4	6	13		

Par exemple, un sujet avec des phosphatases alcalines (PAL) normales sans hépatite (HEP) et sans perte de poids (PPO) est affecté à G_1 ; un sujet avec des phosphatases alcalines normales et présentant une hépatite est affecté au groupe G_2.

Dans la figure 5, les pourcentages indiqués sous chaque nœud terminal représentent les fréquences respectives des sujets bénins et malins dans le nœud.

VI. Conclusions

Les idées qui font de la méthode présentée une méthode décisionnelle intéressante et performante sont les suivantes :

— elle est valable pour discriminer entre deux ou plus de deux groupes à l'aide

TABLEAU V

Détermination du meilleur sous-arbre

S	$T_{(1)}$	$n_{1,1}$	$T_{(2)}$	$n_{2,1}$	$T_{(3)}$	$n_{3,1}$	$T_{(4)}$	$n_{4,1}$	$T_{(5)}$	$n_{5,1}$
nb de n.term.	10		7		5		3		1	
$\pi.10^3$	0		1,82		16,81		46,43		102,09	
$\pi'.10^3$	0		5,53		27,94		68,85		> 68,85	
S_1	$T_{1,1}$	1	$T_{3,1}$	2	$T_{4,1}$	2	$T_{5,1}$	4	$T_{6,1}$	13
S_2	$T_{1,2}$	3	$T_{2,2}$	2	$T_{2,2}$	2	$T_{3,2}$	2	$T_{4,2}$	13
S_3	$T_{1,3}$	6	$T_{1,3}$	6	$T_{1,3}$	6	$T_{2,3}$	10	$T_{3,3}$	12
S_4	$T_{1,4}$	7	$T_{2,4}$	7	$T_{3,4}$	7	$T_{4,4}$	9	$T_{5,4}$	12
S_5	$T_{1,5}$	4	$T_{2,5}$	3	$T_{4,5}$	0	$T_{5,5}$	2	$T_{6,5}$	12
S_6	$T_{1,6}$	5	$T_{1,6}$	5	$T_{2,6}$	6	$T_{3,6}$	6	$T_{4,6}$	12
S_7	$T_{1,7}$	4	$T_{1,7}$	4	$T_{3,7}$	2	$T_{4,7}$	3	$T_{5,7}$	12
S_8	$T_{1,8}$	3	$T_{1,8}$	3	$T_{3,8}$	3	$T_{4,8}$	3	$T_{5,8}$	12
S_9	$T_{1,9}$	1	$T_{2,9}$	1	$T_{3,9}$	1	$T_{4,9}$	3	$T_{5,9}$	12
S_{10}	$T_{1,10}$	4	$T_{1,10}$	4	$T_{1,10}$	4	$T_{2,10}$	6	$T_{3,10}$	13
nb tot. m. c.		(38)		(37)		(33)		(48)		(123)
TEEC		9.95 %		9.69 %		8.64 %		12.57 %		32.20 %

de variables qualitatives, ce qui nécessite le découpage en classes des variables quantitatives s'il en existe;

— elle fournit une règle de décision simple proche en médecine, par exemple, du raisonnement médical;
— elle est robuste vis à vis de données aberrantes ou de sujets mal étiquetés;
— elle tient compte naturellement des interactions qui peuvent exister dans les données;
— le critère de division qui utilise la notion d'impureté d'un nœud est plus approprié que le critère basé sur le pourcentage de bien classés;
— elle n'utilise pas de règle d'arrêt de la procédure de division basée sur des tests statistiques comme c'est le cas, en général, dans les méthodes de segmentation ou les méthodes de discrimination;
— l'échantillon test utilisé dans cette méthode sert à la fois pour déterminer le meilleur sous-arbre et pour obtenir l'estimation la plus précise du taux théorique d'erreur de classement;
— elle permet d'affecter à une classe un sujet présentant des données manquantes même si l'arbre a été construit à partir de données complètes, et ce grâce aux *divisions suppléantes* (Annexe 1);
— elle permet des divisions d'un nœud suivant des *combinaisons booléennes de variables* (Annexe 2);
— il est enfin possible de prendre en compte, dans la règle d'affectation, des coûts de mauvaise classification et des probabilités a priori d'appartenance aux groupes.

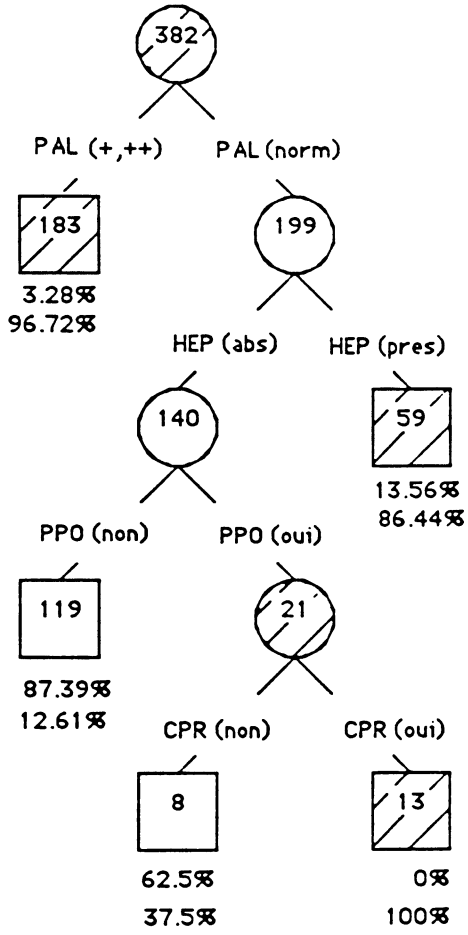


FIGURE 5
Meilleur sous-arbre obtenu par la méthode de validation croisée
□ nœud en faveur de G_1 (tumeurs bénignes)
▨ nœud en faveur de G_2 (tumeurs malignes)
Estimation du TETC = (8,64 % ± 2,82 %)

Références

[1] L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN and C.J. STONE (1984). — *Classification and Regression Trees.* Wadsworth International Group.

[2] J.A. SONQUIST and J.N. MORGAN (1960). — The Detection of Interaction Effects.

Annexe 1

Prise en compte des données manquantes dans la construction de l'arbre de décision binaire et dans l'affectation d'un nouveau sujet à une classe.

Divisions suppléantes

L'idée est la suivante : on définit une mesure d'association entre deux divisions quelconques d et d' d'un nœud t. Si la meilleure division d^* du nœud t est obtenue à partir de la variable V_j, on définit pour chaque variable V_i avec $i = 1, \ldots, V$ (où V est le nombre de variables) et $i \neq j$, la division $\tilde{d}_i$ obtenue à partir de la variable V_i la plus semblable à d^*; $\tilde{d}_i$ est la division suppléante de d^*.

La meilleure division suppléante est, parmi les $V - 1$ divisions suppléantes $\tilde{d}_i$ avec $i \neq j$, la division suppléante la plus semblable à d^*.

De la même manière, on peut définir la seconde meilleure division suppléante, la troisième, etc.

Si pour un sujet la valeur de la variable V_j est manquante, on décide de son affectation à la branche gauche ou droite en utilisant la meilleure division suppléante de d^*. Si, d'autre part, la valeur du sujet est aussi manquante pour la variable sur laquelle la meilleure division suppléante est définie, on utilise la variable de la deuxième meilleure division suppléante etc...

Détermination des divisions suppléantes

Soit V_j la variable ayant permis de définir d^* comme la meilleure division du nœud t; on note A_j et $\overline{A}_j$ les deux sous-ensembles complémentaires des modalités de V_j qui caractérisent respectivement les branches gauche et droite de d^*.

V_i étant une variable différente de V_j, pour chaque division d_{il} où B_{il} et $\overline{B}_{il}$ caractérisent les modalités de V_i qui définissent les branches gauche et droite, on construit le tableau suivant :

		d^*			
		A_j	$\overline{A}_j$	DM	total
d_{il}	B_{il}	n_{gg}	n_{gd}	n_{gm}	$n_{g.}$
	$\overline{B}_{il}$	n_{dg}	n_{dd}	n_{dm}	$n_{d.}$
	DM	n_{mg}	n_{md}	n_{mm}	$n_{m.}$
	total	$n_{.g}$	$n_{.d}$	$n_{.m}$	$n_{..}$

où :

n_{gg} et n_{dd} sont les nombres de sujets classés respectivement à gauche et à droite à la fois par d^* et d_{il},

n_{dg} est le nombre de sujets classés à droite par d_{il} et à gauche par d^*, tandis que n_{gd} est le nombre de sujets classés à gauche par d_{il} et à droite par d^*,

n_{gm} et n_{dm} sont les nombres de sujets non classés par d^* et classés respectivement à gauche et à droite par d_{il}, tandis que n_{mg} et n_{md} sont les nombres de sujets non classés par d_{il} et classés respectivement à gauche et à droite par d^*,

n_{mm} est le nombre de sujets non classés ni par d^* ni par d_{il}.

Ce tableau permet d'estimer la probabilité p_{il} pour que la division d_{il} prédise correctement d^* :

$$p_{il} = (n_{gg} + n_{dd})/n_{..}$$

La division suppléante de d^*, obtenue à partir de V_i est $\tilde{d}_i = d_{iq}$ avec :

$$p_{iq} = \max(p_{il} ; l = 1, 2, \ldots, L)$$

où L est le cardinal de l'ensemble des divisions possibles obtenues à partir de la variable V_i.

Remarque :

Le nombre de divisions obtenues à partir d'une variable dépend du type de la variable; il y a $k - 1$ ou $2^{k-1} - 1$ divisions suivant que la variable est qualitative à k modalités ordonnées ou non ordonnées.

Dans le cas présent de la détermination des divisions suppléantes, il y a deux fois plus de divisions à considérer; en effet A_j peut définir soit la branche gauche, soit la branche droite de d^* de façon arbitraire, et, par la suite les branches gauche et droite ne sont pas équivalentes. Si B_{il} caractérise la branche gauche de la division d_{il}, on note $\overline{d}_{il}$ la division dont la branche gauche est définie par $\overline{B}_{il}$; à chaque fois qu'une division d_{il} est prise en compte, il faut également considérer la division $\overline{d}_{il}$.

Annexe 2

Divisions d'un nœud à l'aide de combinaisons booléennes de variables.

Les combinaisons booléennes possibles à partir de plusieurs variables peuvent être très nombreuses. Pour des raisons de temps calcul et aussi pour ne pas compliquer la structure de l'arbre et son interprétation, on peut se limiter à la détermination des combinaisons du type :

$$d_{j1} \cap d_{j2} \cap \ldots \cap d_{jl}$$

avec : t_g définie par $V_{j1} = \{A_{j1}\}$ et $V_{j2} = \{A_{j2}\}$ et ... et $V_{jl} = \{A_{jl}\}$

et t_d définie par $V_{j1} = \{\overline{A}_{j1}\}$ ou $V_{j2} = \{\overline{A}_{j2}\}$ ou ... ou $V_{jl} = \{\overline{A}_{jl}\}$

où :

— A_{j1} représente un sous-ensemble des modalités de la variable V_{j1} et $\overline{A}_{j1}$ son complémentaire.

— A_{j2} représente un sous-ensemble des modalités de la variable V_{j2} et $\overline{A}_{j2}$ son complémentaire.

— etc.

Division admissible

Une division du nœud t est dite admissible si les nombres de sujets des descendants immédiats t_g et t_d de t sont supérieurs à n_0 fixé.

Sélection de la meilleure division d'un nœud

Le choix de la meilleure division d'un nœud se fait par une procédure en pas à pas ascendant.

On note D_0 l'ensemble des divisions admissibles, et, au premier pas on choisit $d^*_{j1} \in D_0$ telle que :

$$\Delta i\,(d^*_{j1}, t) = \max\{\Delta i\,(d_j, t)\ ;\ d_j \in D_0\}$$

avec t_g définie par $V_{j1} = \{A_{j1}\}$ et t_d définie par $V_{j1} = \{\overline{A}_{j1}\}$

où A_{j1} représente un sous-ensemble des modalités de la variable V_{j1} et $\overline{A}_{j1}$ son complémentaire.

On considère D_1 l'ensemble des divisions admissibles obtenues à l'aide des variables V_j telles que $j \neq j1$. Pour chaque division d_j appartenant à D_1 on définit la division $d^*_{j1} \cap d_j$ qui fournit les branches

$$t_g : V_{j1} = \{A_{j1}\} \text{ et } V_j = \{A_j\}$$
$$t_d : V_{j1} = \{A_{j1}\} \text{ ou } V_j = \{\overline{A}_j\}$$

où A_j représente un sous-ensemble des modalités de V_j et $\overline{A}_j$ son complémentaire.

Au deuxième pas on retient, parmi les divisions $d^*_{j1} \cap d_j$ avec $d_j \in D_1$ la division $d^*_{j1} \cap d^*_{j2}$ telle que :

$d^*_{j1} \cap d^*_{j2}$ soit admissible

et $$\Delta i\,(d^*_{j1} \cap d^*_{j2}, t) = \max\{\Delta i\,(d^*_{j1} \cap d_j, t)\ ;\ d_j \in D_1\}$$

et $$\frac{\Delta i\,(d^*_{j1} \cap d^*_{j2}, t)}{\Delta i\,(d^*_{j1}, t)} > 1 + \beta$$

c'est-à-dire, telle que le rapport des deux critères au pas 2 et au pas 1 soit supérieur à un seuil $1 + \beta$ fixé.

On considère l'ensemble D_2 des divisions admissibles obtenues à l'aide des variables V_j avec $j \neq j1$ et $j \neq j2$, et on définit comme précédemment les divisions $d^*_{j1} \cap d^*_{j2} \cap d_j$ avec $d_j \in D_2$.

Au troisième pas on choisit d^*_{j3} telle que :

$d^*_{j1} \cap d^*_{j2} \cap d^*_{j3}$ soit admissible

et $$\Delta i\,(d^*_{j1} \cap d^*_{j2} \cap d^*_{j3}, t) = \max\{\Delta i\,(d^*_{j1} \cap d^*_{j2} \cap d_j, t)\ ;\ d_j \in D_2\}$$

et $$\frac{\Delta i\,(d^*_{j1} \cap d^*_{j2} \cap d^*_{j3}, t)}{\Delta i\,(d^*_{j1} \cap d^*_{j2}, t)} > 1 + \beta$$

— ainsi de suite...

— la procédure s'arrête quand il n'y a plus de divisions admissibles susceptibles de produire une augmentation suffisante du critère.