

REVUE DE STATISTIQUE APPLIQUÉE

E. PÉRINEL

Construire un arbre de discrimination binaire à partir de données imprécises

Revue de statistique appliquée, tome 47, n° 1 (1999), p. 5-30

http://www.numdam.org/item?id=RSA_1999__47_1_5_0

© Société française de statistique, 1999, 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/>

CONSTRUIRE UN ARBRE DE DISCRIMINATION BINAIRE À PARTIR DE DONNÉES IMPRÉCISES

E. Périnel¹

ENSAR-INSFA

65, rue de Saint-Brieuc, 35042 Rennes cedex

RÉSUMÉ

En raison d'un processus de partitionnement parfois rigide, les méthodes de segmentation peuvent offrir des performances peu satisfaisantes lorsque les données étudiées sont recueillies dans un contexte d'incertitude important. Ce travail propose une méthodologie de construction d'arbre de discrimination binaire permettant la prise en compte explicite de phénomènes d'imprécision au niveau de la description des objets étudiés. La prise en compte de données imprécises conduit à un processus de partitionnement plus souple, basé sur l'idée d'affectation probabiliste des objets dans les différents nœuds de l'arbre, à partir d'un critère de coupure utilisant la notion de vraisemblance, la méthode favorise la sélection de divisions plus stables, à l'image des coupures «souples», avec lesquelles nous faisons ici le lien. L'approche est illustrée par un exemple simple et appliqué enfin à un problème médical.

Mots-clés : *segmentation, discrimination par arbre binaire, contexte d'incertitude, données imprécises, partitionnement «souple», modélisation probabiliste, algorithme EM.*

ABSTRACT

Because of a recursive partitioning somewhat inflexible, decision trees may provide unsatisfying performances, especially when the data are observed in the presence of uncertainty. This paper presents a tree-growing methodology that permits to take into account and treat probabilistically imprecise descriptions. Taking into consideration such data leads to the general notion of a soft assignment, based on a probabilistic membership of the objects to the various nodes of the tree. With the help of a likelihood splitting criterion, the approach favours the selection of more robust partitions, just like the soft or fuzzy splits with which we are making the connection. Eventually, the method is illustrated on a simple example and in the context of a medical application.

Keywords : *binary segmentation, discriminant tree-growing, uncertainty, imprecise data, soft assignment, probabilistic assignment, EM algorithm.*

¹ Ce travail a été réalisé alors que l'auteur faisait partie de l'équipe du LISE CEREMADE de l'Université Paris IX-Dauphine et du projet CLOREC de l'INRIA Rocquencourt.

1. Introduction et cadre général

Il y a maintenant plus de trente ans que sont nées les premières méthodes statistiques de segmentation, connues également sous la terminologie de partitionnement récursif ou encore d'arbre de décision binaire. Développées à la fois pour des problématiques d'analyse discriminante (Belson 59), (Friedman 77) et de régression (Morgan *et al.* 63), elles n'ont connu un essor significatif (aussi bien sur le plan théorique que pratique) que près de vingt années plus tard avec les travaux de (Breiman *et al.* 84). Leur ouvrage et logiciel, bien connu désormais sous l'acronyme CART (Classification And Regression Trees), a permis, dans un cadre probabiliste explicite et à travers ses procédures originales dites d'élagage, d'offrir une solution rigoureuse et applicable en pratique au délicat problème de la construction de l'arbre dans une optique inférentielle (on consultera par exemple à ce propos (Breiman *et al.* 84), (Guéguen *et al.* 88) ou encore (Celeux *et al.* 94)).

Actuellement, à la frontière entre les domaines de l'analyse des données-statistique et de l'intelligence artificielle, de nombreux travaux se penchent sur un problème particulièrement sensible en segmentation : celui du traitement d'une information de nature imparfaite, notion générale recouvrant celles d'information imprécise ou incertaine. Cette problématique a donné lieu à des extensions des méthodes usuelles qui font appel bien naturellement au raisonnement probabiliste, mais également de plus en plus souvent à des théories et axiomatiques plus récentes, telles que la théorie des possibilités ou encore la logique floue.

Dans ce travail, nous allons nous intéresser au problème de la construction d'un arbre de **discrimination** binaire lorsque l'on est confronté à un ensemble de données qualifiées d'imprécises, et modélisées de façon probabiliste. Après avoir rappelé brièvement la problématique générale de la segmentation (section 2), nous tenterons d'expliquer pourquoi les phénomènes d'imprécision en général constituent un problème particulièrement sensible pour les méthodes de partitionnement récursif (section 3). Nous décrirons alors plus en détail (section 4, 5 et 6) l'algorithme proposé, basé sur un modèle probabiliste, et où la qualité des partitions «souples» successivement formées est évaluée à partir de la notion de vraisemblance. En raison de l'imprécision entâchant les différentes descriptions des objets étudiés, la recherche d'une meilleure division binaire conduit à un problème de mélange de lois de probabilité que l'on propose de résoudre à l'aide de l'algorithme EM. Enfin, un exemple médical illustrera la méthode proposée (section 7).

2. Rappels sur la segmentation

Formulation du problème. Soit Ω , une population d'objets ω_i ($i = 1, \dots, N$) pour laquelle on connaît *a priori* une partition en K classes. On associe usuellement à cette partition une variable de classe (ou variable à expliquer) notée C , application de Ω dans $\{1, \dots, K\}$, l'ensemble des classes possibles. Chaque objet est par ailleurs décrit par P prédicteurs Y_1, \dots, Y_P (ou variables explicatives) à valeurs dans des domaines O_1, \dots, O_P . Rappelons que les variables Y_j peuvent être indifféremment de nature qualitative, quantitative ou les deux à la fois. Ces notations sont résumées sous la forme d'un tableau à N lignes et $(P + 1)$ colonnes constituant l'ensemble des

données étudiées (tableau 1). Celles-ci forment également ce que (Breiman *et al.* 84) appellent échantillon d'apprentissage, noté $\mathcal{L} = \{(c_i, \mathbf{y}_i)\}_i^2$; nous utiliserons dans la suite la terminologie de *données standard* en référence à ce type de données.

TABLEAU 1
Tableau «standard» des données initiales en segmentation.

	Y_1	...	Y_j	...	Y_P	C
ω_1						
\vdots			\vdots			\vdots
ω_i		...	$Y_j(\omega_i)$...		$C(\omega_i)$
\vdots			\vdots			\vdots
ω_N						

Objectif et principe élémentaire. Comme l'ensemble des techniques statistiques de discrimination, l'objectif d'une segmentation est d'*expliquer la partition* connue *a priori* sur Ω à l'aide des prédicteurs décrivant les objets. Cette explication recouvre deux objectifs : *décrire* les K classes en termes des variables (ou interactions de celles-ci) les plus spécifiques; construire une *règle de décision* dans le but de classer avec le minimum d'erreurs de nouveaux objets à l'une des classes de la partition. Ces deux tâches sont réalisées à travers deux phases successives :

- (1) la phase de segmentation proprement dite, permettant la construction de l'arbre; elle est basée sur la définition *a priori* d'un ensemble de *questions binaires* ainsi que d'un *critère de coupure* permettant d'évaluer leur qualité;

Il s'agit ici de diviser (segmenter) récursivement la population totale, à l'aide des questions binaires, en une partition de plus en plus fine, avec le souci constant d'obtenir des sous-populations présentant une homogénéité maximale vis-à-vis de la variable de classe C .

- (2) L'extraction d'un arbre de «bonne taille», sous-entendu, associé à des qualités prédictives satisfaisantes.

Pour cette phase de nature inférentielle, une majorité des algorithmes actuels procède grosso-modo comme suit : on construit dans un premier temps un «grand arbre»; celui-ci est ensuite élagué pour obtenir l'arbre de «bonne taille», minimisant un risque ou taux d'erreur de classement (estimé par échantillon test ou validation croisée).

Enfin, il est intéressant de rappeler que, contrairement à de nombreuses méthodes usuelles d'analyse des données, la segmentation produit des classes de nature *monothétique* (i.e., caractérisées par des conditions nécessaires et suffisantes), conséquence directe du processus de partitionnement récursif décrit ci-dessus.

² Comme le veut l'usage, les lettres capitales, les minuscules et les lettres grasses désigneront respectivement les variables, leurs réalisations et des vecteurs (de variables ou d'observations).

3. Segmentation et phénomènes d'imprécision

Si la lisibilité des règles de décision fournies par le processus de partitionnement récursif constitue un de ses principaux attraits, ces dernières demeurent en contrepartie particulièrement sensibles à de légères perturbations des données (Lebart *et al.* 95) ou phénomènes d'incertitude pouvant affecter la description des objets (mesures réalisées à partir d'un instrument imprécis, de jugement approximatif émis par un observateur,...).

Un problème d'affectation arbitraire. Face à de telles données, le problème majeur survenant en segmentation est celui de l'affectation arbitraire. Pour illustrer notre propos, considérons tout d'abord la figure 1(a). Celle-ci décrit le problème d'affectation d'un objet dont la valeur du poids, inconnue, est néanmoins supposée comprise dans l'intervalle [55, 60]. En assignant cette description à l'un seul des deux nœuds créés par la coupure (mettons le nœud droit, en supposant que l'on retienne finalement la valeur centrale de l'intervalle, 57.5), on réalise une affectation arbitraire ou incertaine dans la mesure où l'on ignore le poids réel de l'individu. De même pour le schéma 1(b), si l'on estime à 0.8, la probabilité pour qu'un test soit positif, l'affectation au nœud droit n'est pas pour autant à rejeter totalement (ce que l'on ferait en retenant la valeur la plus probable).

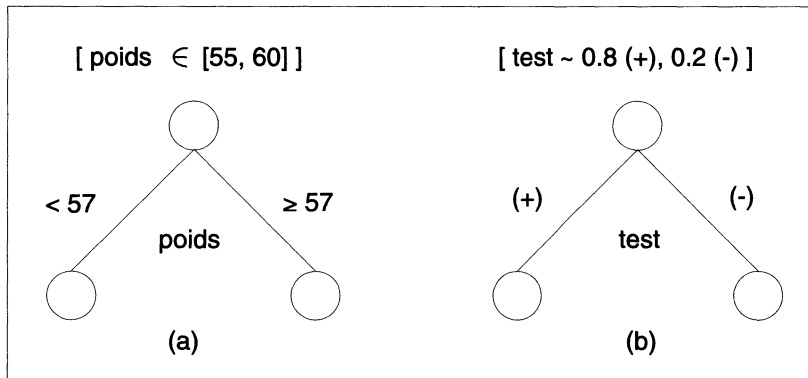


FIGURE 1

*Classement d'observations imprécises :
la segmentation peut conduire à des affectations arbitraires en situation d'incertitude*

Il est important de souligner que ce problème d'affectation arbitraire est susceptible de se produire à la fois pendant la construction de l'arbre et lors de l'utilisation de ce dernier comme règle de classement. Dans la première situation, la sélection d'une question binaire reposant sur un descripteur peu fiable, c'est-à-dire dont les valeurs observées seraient entâchées d'une imprécision importante, conduirait à retenir une division fragile ou instable des objets, susceptible d'engendrer par la suite d'importantes erreurs de classement. En phase de classement, le problème est d'autant plus sensible dans la mesure où la décision est prise à partir d'un petit nombre de descriptions relatives à un seul objet.

Une étude empirique réalisée par (Quinlan 86) met en lumière les constats intuitifs précédents. En perturbant, avec différents niveaux de bruit, les valeurs des descriptions de l'échantillon, l'auteur met en évidence une dégradation des performances décisionnelles de la méthode; ces performances étant d'ailleurs d'autant moins fiables que la procédure tend à sélectionner, en raison du bruit, des variables en réalité peu discriminantes.

Deux approches possibles. On peut voir le problème précédent comme la conséquence de deux choix *a priori* bien distincts. Le premier est lié à la nature des questions binaires «sans nuance» ou «abruptes» retenues dans la majorité des processus de segmentation. Avec ce type de questions, de légères perturbations des données conduisent souvent à une modification sensible de la répartition des individus.

Le second est lié au codage des données et à l'impossibilité en général de prendre en compte et de représenter explicitement une information imprécise. Ces deux constats ont suscité respectivement deux types d'approches que nous détaillons maintenant. Notons que malgré des similitudes évidentes (voir en annexe), ces deux approches découlent néanmoins initialement d'objectifs et de préoccupations différentes.

4. Partitionnement souple ou doux

On trouve différents travaux dans la littérature basés sur la notion d'affectation souple. La plupart des méthodes proposées sont fondées sur l'axiomatique de la théorie des possibilités et celle des ensembles flous (on consultera par exemple (Yuan *et al.* 95) pour une présentation de quelques algorithmes récents sur ce thème). Dans la suite nous nous intéresserons cependant uniquement aux méthodologies fondées sur une approche probabiliste.

4.1. Les coupures douces

Celles-ci ont été essentiellement employées pour des questions binaires basées sur des variables continues. L'idée majeure de ce type de coupures est d'affecter une observation à une région donnée de l'espace de description avec d'autant plus de certitude que celle-ci est éloignée du seuil de coupure. En procédant ainsi, on souhaite substituer à la notion d'affectation déterministe usuelle, un mécanisme plus flexible conduisant à une appartenance probabiliste des objets aux différentes régions de l'espace de description. La figure 2 illustre le principe des deux approches.

Le schéma (a) représente la situation classique en segmentation : l'appartenance à la région définie par la propriété $[Y \geq 53]$ est définie par sa fonction indicatrice $I_{[Y \geq 53]}$, marquant une rupture nette au niveau du seuil de coupure. Dans le schéma de droite, l'affectation à la même région est cette fois-ci déterminée à l'aide d'une fonction continue notée Φ (par exemple, de type sigmoïde) et prenant la valeur 0.5 au niveau de la coupure. Alors que l'observation 53.1 est assignée dans le schéma de gauche avec une totale certitude à la région $[Y \geq 53]$, sa proximité avec le seuil de coupure conduit à l'affecter avec des probabilités sensiblement égales aux deux régions dans la partie droite de la figure.

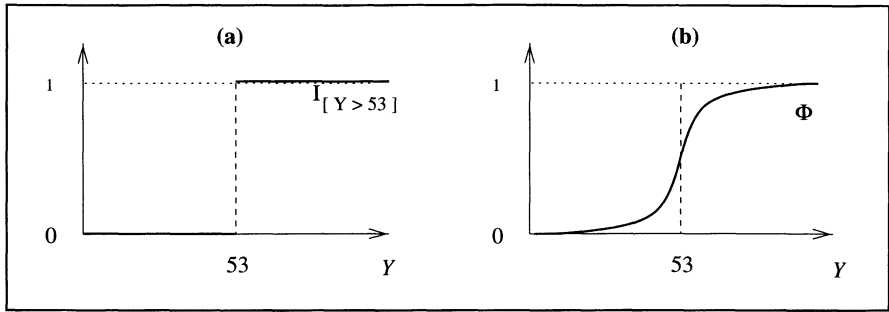


FIGURE 2

Principe d'une coupure souple (b) en regard d'une coupure standard (a)

Ce principe se justifie intuitivement par la remarque suivante : si on considère deux individus d'âge respectif 53 ans plus ou moins un jour, il convient mieux de les traiter de façon quasi-identique plutôt que de les séparer dans deux sous-populations différentes.

4.2. Descriptions imprécises

Lorsque l'on est confronté, lors de la phase de recueil des données, à des observations explicitement imprécises, soumises à une incertitude que l'on est capable au moins approximativement de mesurer, une démarche pour le moins naturelle consisterait à relever et coder cette observation telle qu'on la perçoit dans l'objectif simplement de posséder l'information la plus fidèle possible. Comment de telles observations peuvent-elles être prises en compte et traitées dans un processus de partitionnement binaire usuel? Pour le voir, reprenons les deux exemples de la figure 1. Plutôt qu'un processus d'affectation booléen, il semble ici également plus naturel de réaliser une affectation plus nuancée, en regard du niveau d'imprécision entâchant les descriptions. Par exemple, en supposant que l'attribut *poids* est distribué uniformément entre 55 et 60 kg, on assignera la description aux nœuds gauche et droit avec les probabilités respectives $p = \frac{57 - 55}{60 - 55} = \frac{2}{5}$ et $1 - p = \frac{60 - 57}{60 - 55} = \frac{3}{5}$. Pour l'attribut *test*, l'affectation est encore plus immédiate et conduit bien sûr à des probabilités de 0.8 et 0.2.

Pour une coupure standard déterminée par les deux propriétés $[Y \in A]$ et $[Y \in \bar{A}]$ (notation généralisant les situations Y qualitatif et quantitatif), on détermine ainsi simplement l'appartenance probabiliste aux deux nœuds associés par

$$p_A = \int_A f_Y(y) \cdot dy \quad p_{\bar{A}} = \int_{\bar{A}} f_Y(y) \cdot dy \quad (\text{cas continu})$$

$$p_A = \sum_{y \in A} P_Y(y) \quad p_{\bar{A}} = \sum_{y \in \bar{A}} P_Y(y) \quad (\text{cas discret})$$

où P_Y et f_Y sont les lois et densités de probabilité modélisant l'imprécision d'une observation pour la variable Y .

Formalisation des données imprécises. Formalisons maintenant plus précisément le type de données que nous allons traiter dans notre algorithme, et qui s'inscrivent dans le cadre des travaux développés par (Diday 87) en analyse des données symboliques. Nous supposons donc dorénavant qu'il ne nous est pas toujours possible d'exprimer avec certitude les descriptions des objets pour les variables Y_1, \dots, Y_P ; la classe de ces derniers est par contre toujours supposée parfaitement déterminée. On introduit ainsi pour un objet i particulier, un p -uplet de variables aléatoires $Y_{1,i}, \dots, Y_{P,i}$, définissant autant de fonctions mesurables à valeurs dans O_1, \dots, O_P . Chacune de ces variables aléatoires est par ailleurs associée à une loi de probabilité notée $P_{j,i}$ telle que

$$P_{j,i} : O_j \rightarrow [0, 1]$$

$$v \mapsto P_{j,i}(v) = P[Y_{j,i}^{-1}(v)]$$

La description d'un individu i de la population sera donc dans la suite définie par le $(P + 1)$ -uplet de variables aléatoires

$$(Y_i, C_i) = (Y_{1,i}, Y_{2,i}, \dots, Y_{P,i}, C_i)$$

où C_i est également une variable aléatoire à valeurs dans $\{1, \dots, K\}$. Le tableau 2 représente l'ensemble des données que nous nous proposons de traiter par arbre de discrimination binaire; par abus de langage, nous les qualifierons de *données probabilistes* dans la suite de ce travail.

TABLEAU 2

*Tableau de données initiales de nature imprécise :
la description de l'individu ω_i pour le prédicteur Y_j est définie
par une variable aléatoire $Y_{j,i}$ distribuée selon la loi de probabilité $P_{j,i}$.*

	Y_1	\dots	Y_j	\dots	Y_P	C
ω_1						
\vdots			\vdots			\vdots
ω_i		\dots	$Y_{j,i} \sim P_{j,i}$	\dots		$C(\omega_i)$
\vdots			\vdots			\vdots
ω_N						

Notons, en regard du cadre statistique usuel, qu'à chaque variable Y_j du tableau 2 ne correspond plus N réalisations $y_{j,1}, \dots, y_{j,N}$ (ou, de manière équivalente, N variables aléatoires $Y_{j,1}, \dots, Y_{j,N}$, ayant une seule réalisation et dont Y_j est une variable parente), mais N variables aléatoires $Y_{j,1}, \dots, Y_{j,N}$ pour lesquelles on peut

réaliser *différentes* hypothèses probabilistes. Dans (Diday *et al.* 97), les données sont qualifiées plus simplement de *multivaluées* lorsque chaque variable $Y_{j,i}$ est renseignée par plusieurs observations, sans hypothèse *a priori* sur leur distribution.

Exemple. Considérons le cas d'une partition *a priori* comportant deux classes ainsi que deux variables aléatoires Y_1, Y_2 , respectivement binaire et continue à valeurs dans $[0, 10]$. On a donc $O_1 = \{0, 1\}$ et $O_2 = [0, 10]$. Considérons ensuite un objet de la classe 1 prenant la valeur 1 avec la probabilité 0.7 pour Y_1 , et dont la valeur pour Y_2 , inconnue, est estimée comprise dans l'intervalle $[3, 5]$. Sa description pour les prédicteurs est définie par le couple de variables aléatoires $(Y_{1,i}, Y_{2,i})$ tel que

$$\begin{aligned} Y_{1,i} &\sim \mathcal{B}(0.7) && \text{où } \mathcal{B}(0.7) \text{ est la loi de Bernoulli de paramètre } 0.7; \\ Y_{2,i} &\sim \mathcal{U}_{[3,5]} && \text{où } \mathcal{U}_{[3,5]} \text{ est la loi uniforme sur } [3,5]. \end{aligned}$$

L'algorithme que nous présentons maintenant est fondé sur le principe d'affectation probabiliste et dans le cas uniquement de données de nature imprécise (les coupures binaires sont classiques). Une présentation de cet algorithme, dans le cadre général de l'analyse des données symboliques peut également être trouvée dans (Ciampi *et al.* 96). La description de cet algorithme nécessite ici l'introduction d'un ensemble de notions probabilistes relatives à la modélisation d'un arbre de décision (ou de prédiction) binaire.

5. Modélisation probabiliste d'un arbre de décision binaire

La présentation décrite ci-dessous s'inspire de celle proposée dans (Ciampi 92) pour l'aspect discrimination de l'algorithme de prédiction par arbre RECPAM. Une des spécificités de la méthode RECPAM est de poser le problème de la construction d'un arbre binaire comme celui de la prédiction d'un paramètre. Plus précisément, la problématique générale d'une segmentation peut être décrite à partir du modèle probabiliste suivant.

En théorie statistique de la décision, la construction d'une règle de classement repose souvent sur l'estimation des densités de probabilité de chacune des classes connues *a priori* (c'est le cas par exemple en discrimination linéaire). La segmentation elle, vise à modéliser directement la loi de probabilité conditionnelle $P(c|y)$, c'est-à-dire à prédire la variable de classe conditionnellement aux observations réalisées sur les prédicteurs.

Atteindre cet objectif à travers un processus de segmentation revient en réalité à faire l'hypothèse suivante : il existe dans l'espace de description un partitionnement en régions (correspondant aux feuilles de l'arbre) où, dans chacune d'elle, les individus sont homogènes du point de vue de la variable à expliquer C . Notons $P(c; \theta_t)$ la loi de probabilité de la variable C , paramétrée par θ_t , au sein de la région t (figure 3).

Nous pouvons alors énoncer l'objectif d'une segmentation comme suit : il s'agit de rechercher un découpage de l'espace des prédicteurs et donc, par suite, un ensemble de modèles probabilistes $(\theta_t)_t$ associé à des régions «maximalement homogènes» vis-à-vis de la variable à expliquer C (c'est-à-dire encore, des régions

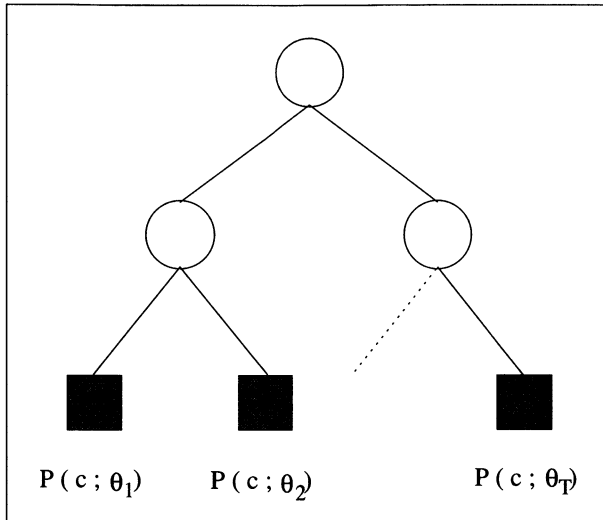


FIGURE 3

*Modélisation probabiliste d'un arbre de décision :
chaque sous-population terminale est modélisée par une loi de probabilité $P(c|\theta_t)$.*

où les lois de probabilité $P(c; \theta_t)$ permettent une bonne prédiction de la variable C). Enfin, on modélisera de manière naturelle un arbre à T nœuds terminaux par la probabilité conditionnelle $P(c|y)$ qui lui est associée et qui, en regard de la figure 3, s'exprime comme suit :

$$P(c|y; \theta) = \sum_{t=1}^T I_t(y) \cdot P(c; \theta_t) \quad (1)$$

où I_t est la fonction indicatrice d'appartenance à la région t (c'est-à-dire, à la feuille t); $\theta = (\theta_1, \dots, \theta_T)$ est le vecteur de paramètres spécifiant chacune des lois de probabilité conditionnelles au sein des T nœuds terminaux.

Application à un problème de discrimination par arbre de deux classes a priori. Chaque région t est décrite ici par une loi de probabilité de Bernoulli de paramètre $\theta_t = P(1|t)$, la probabilité d'observer la classe 1 dans le nœud t . Une bonne segmentation consiste donc ici à former des régions les plus homogènes possible du point de vue de la classe 1 ou 2 (associées donc à des valeurs de θ_t proches de 1 ou 0).

Modélisation dans le cas de données probabilistes. Lorsque les données sont imprécises, nous avons vu qu'il était naturel de procéder à une affectation souple (ici, probabiliste) des individus aux nouveaux nœuds créés par la coupure. Plus généralement, pour une segmentation créant T nœuds terminaux, chaque description

y d'un individu sera désormais assignée à une région donnée t avec la probabilité $P(t|y)$, de sorte que l'équation (1) précédente se réécrit maintenant :

$$P(c|y; \theta) = \sum_{t=1}^T P(t|y) \cdot P(c; \theta_t) \quad (2)$$

Autrement dit, la prédiction de la classe d'un objet est désormais calculée comme une somme pondérée de prédiction «locales», et non plus sur la base de la seule loi de probabilité associée à la région où est affecté l'objet en question.

6. Construction de l'arbre binaire

6.1. Les choix de base

Données étudiées. Nous nous intéressons ici à deux types de descriptions. Pour les variables continues, $Y_{j,i}$ sera supposée distribuée selon une loi uniforme. Pour les variables discrètes (nominales ou ordinales), on considèrera que la loi de $Y_{j,i}$ est de type multinomiale (consistant donc à donner la probabilité associée à chacune des modalités de la variable).

Ensemble des questions binaires. L'ensemble des questions binaires joue un rôle essentiel en segmentation, car c'est celui-ci qui détermine la nature des divisions qui sont effectuées itérativement sur la population étudiée. Il conditionne donc, d'une certaine manière, la forme plus ou moins complexe des régions construites dans l'espace de description, ou encore, il permet de construire des frontières plus ou moins sophistiquées entre les classes de la partition *a priori*. Nous nous limitons dans la suite à un ensemble \mathcal{S} de questions binaires dites *standard*, notées $s = (e, \bar{e})$, de la forme

$$(e = [Y \in V], \bar{e} = [Y \in \bar{V}]) \quad (3)$$

où $\{V, \bar{V}\}$ est une partition du domaine O des valeurs possibles d'un prédicteur Y choisi parmi $\{Y_1, \dots, Y_P\}$.

Qualité d'une partition et Critère de coupure. Réaliser une segmentation consiste, nous le savons, à réitérer sur la population une seule et même opération : la recherche à un nœud donné de la question qui produit la nouvelle «meilleure partition», dans le sens où elle crée des sous-populations qui sont les plus homogènes possibles vis-à-vis de la variable de classe C . L'homogénéité est déterminée ici à partir de la notion d'information statistique ou de vraisemblance (son logarithme, en fait, pour des raisons de calcul), à l'image de la méthodologie RECPAM.

Plus précisément, notons \mathcal{P}_T , une partition comprenant T nœuds terminaux et associée au modèle probabiliste $P(c|y; \theta)$ (défini donc selon (1) ou (2) suivant que les données sont standard ou probabilistes). L'information portée par la partition \mathcal{P}_T est définie comme suit :

$$\text{Qualité de } \mathcal{P}_T \quad \text{Ginf}(\mathcal{P}_T) = \log L(\theta; c|y) \quad (4)$$

où L est la fonction de vraisemblance du paramètre θ . Intuitivement : plus les sous-populations de \mathcal{P}_T sont homogènes du point de vue de la variable C , plus celle-ci est prévisible ou «vraisemblable» au sein de chacune d'elle, et donc plus la quantité (4) est grande.

Supposons maintenant que l'on recherche pour un nœud donné d'une partition \mathcal{P}_{T-1} à $T - 1$ nœuds terminaux, la question binaire $s = (e, \bar{e}) \in \mathcal{S}$ qui conduit à la nouvelle meilleure partition en T nœuds terminaux. Notons \mathcal{P}_T^s la partition produite par une question s quelconque pour un nœud fixé de \mathcal{P}_{T-1} . Dans ce cas, le problème de la recherche d'une meilleure question binaire revient à maximiser le critère de coupure suivant :

$$\text{Critère de coupure} \quad \max_{s \in \mathcal{S}} \{ \text{Ginf}(\mathcal{P}_T^s) - \text{Ginf}(\mathcal{P}_{T-1}) \} \quad (5)$$

Le second terme de la différence étant constant à tout moment de l'algorithme (il ne dépend pas de s), le problème de la recherche de la meilleure coupure consiste donc à maximiser l'information portée par la seule nouvelle partition \mathcal{P}_T^s .

6.2. Résolution du problème pour des données standard

Le problème de maximisation précédent reste cependant très général sous la forme (5). Nous allons maintenant le préciser pour des données standard avant d'étudier ensuite les modifications qui en découlent dans le cas de données probabilistes.

Soit \mathcal{N} , un nœud quelconque de la partition \mathcal{P}_{T-1} comportant n objets. Une question binaire $s = (e, \bar{e}) \in \mathcal{S}$ crée deux nouveaux nœuds \mathcal{N}_e et $\mathcal{N}_{\bar{e}}$ dont les effectifs n_e et $n_{\bar{e}}$ sont répartis suivant les K classes *a priori* (tableau de contingence TAB. 3). Dans les deux nœuds \mathcal{N}_e et $\mathcal{N}_{\bar{e}}$, la variable C est distribuée selon une loi de probabilité multinomiale de paramètres θ_e et $\theta_{\bar{e}}$; on aura donc par exemple, $\theta_e = (p_e(1), \dots, p_e(K))$ où $p_e(k)$ est la probabilité de la classe k au sein du nœud \mathcal{N}_e . On définit de même $\theta_{\bar{e}}$.

TABLEAU 3

Tableau de contingence croisant les deux nœuds \mathcal{N}_e et $\mathcal{N}_{\bar{e}}$ créés par la coupure avec les classes 1, ..., K de la partition *a priori*.

	1	...	K	
\mathcal{N}_e	$n_e(1)$...	$n_e(K)$	n_e
$\mathcal{N}_{\bar{e}}$	$n_{\bar{e}}(1)$...	$n_{\bar{e}}(K)$	$n_{\bar{e}}$
	$n(1)$...	$n(K)$	n

Comment détermine-t-on dans ce cas la meilleure question binaire au nœud \mathcal{N} ? Il suffit nous l'avons vu de résoudre le problème de maximisation (5), soit encore, d'après les équations (1) et (4), et en supposant de plus l'indépendance entre les

observations i , de maximiser la quantité suivante :

$$\text{Ginf}(\mathcal{P}_T^s) = \log \prod_{i=1}^N \left(\sum_{t=1}^T I_t(\mathbf{y}_i) \cdot P(c_i; \theta_t) \right) \quad (6)$$

Or, lorsque les données sont standard, les objets sont affectés de manière unique à l'un des T nœuds de la partition \mathcal{P}_T . Par suite, il est immédiat de vérifier que la question binaire qui maximise (6) est aussi celle maximisant

$$\text{Ginf}(\mathcal{P}_T^s) = \log \left(\prod_{i \in \mathcal{N}_e} P(c_i; \theta_e) \cdot \prod_{i \in \mathcal{N}_{\bar{e}}} P(c_i; \theta_{\bar{e}}) \right) \quad (7)$$

qui n'est autre que la log-vraisemblance catégorisée, $\log L((\theta_e, \theta_{\bar{e}}); c|\mathbf{y})$, sur les deux nœuds créés par s . En effet, dans ce cas, la recherche d'une nouvelle partition à un nœud donné ne fait intervenir que les seules informations relatives aux deux nœuds fils créés par la coupure (les informations relatives aux autres nœuds étant constantes pour la maximisation).

Le calcul de $\text{Ginf}(\mathcal{P}_T^s)$ nécessite donc l'estimation des paramètres $\theta_e = (p_e(k))_k$, ce qui est possible en résolvant le problème de maximisation (7) : il conduit à retenir de manière naturelle comme estimateurs du maximum de vraisemblance de ces probabilités, les fréquences observées des K classes dans les deux nœuds :

$$\hat{p}_e(k) = \frac{n_e(k)}{n_e}, \quad \hat{p}_{\bar{e}}(k) = \frac{n_{\bar{e}}(k)}{n_{\bar{e}}} \quad (8)$$

Remarque : en substituant ces estimateurs dans l'équation (7), il est aisé de vérifier que la log-vraisemblance se réécrit comme suit :

$$n_e \sum_{k=1}^K \frac{n_e(k)}{n_e} \log \frac{n_e(k)}{n_e} + n_{\bar{e}} \sum_{k=1}^K \frac{n_{\bar{e}}(k)}{n_{\bar{e}}} \log \frac{n_{\bar{e}}(k)}{n_{\bar{e}}} \quad (9)$$

L'expression (9) est donc également, au coefficient $-\frac{1}{n}$ près, une estimation de l'entropie conditionnelle $H[C|Y]$ de la variable C , au sein des deux nœuds créés par la coupure. Ce critère, fréquemment utilisé dans les arbres de décision, par exemple dans (Sethi *et al.* 82) ou (Quinlan 93), permet d'évaluer l'incertitude moyenne de la variable à prédire C , une fois connues les valeurs prises par la variable de coupure Y .

6.3. Extension au cas de données probabilistes

En regard de l'équation (6), le problème de recherche de la meilleure coupure consiste (en utilisant cette fois-ci le modèle (2)) à maximiser la quantité suivante :

$$\text{Ginf}(\mathcal{P}_T^s) = \log \prod_{i=1}^N \left(\sum_{t=1}^T P(t|\mathbf{y}_i) \cdot P(c_i; \theta_t) \right) \quad (10)$$

$$= \log \prod_{i=1}^N (P(e|\mathbf{y}_i) \cdot P(c_i; \theta_e) + P(\bar{e}|\mathbf{y}_i) \cdot P(c_i; \theta_{\bar{e}}) + \lambda) \quad (11)$$

où λ est défini par l'expression $\sum_{t \neq e, \bar{e}} P(t|y_i) \cdot P(c_i; \theta_t)$. Celle-ci est supposée constante

dans la mesure où nous ne remettons pas en question les lois de probabilité (estimées lors des étapes précédentes) des nœuds t , hormis celles relatives aux nœuds fils \mathcal{N}_e et $\mathcal{N}_{\bar{e}}$. Par ailleurs, la probabilité d'appartenance d'une observation i à un nœud t , $p(t|y_i)$, peut être calculée aisément quels que soient i et t (voir leur mode de calcul en 6.4).

Les inconnues du problème sont donc ici la question binaire (e, \bar{e}) ainsi que les paramètres θ_e et $\theta_{\bar{e}}$ qui lui sont associés. Pratiquement, on procède alors comme suit : pour une question binaire s donnée, on calcule une estimation des paramètres θ_e et $\theta_{\bar{e}}$ puis, la valeur de Ginf (\mathcal{P}_T^s) (la qualité de la nouvelle partition) qui en résulte. La meilleure question étant bien sûr celle qui conduit à des estimations de θ_e et $\theta_{\bar{e}}$ maximisant la qualité de la partition. La difficulté du problème réside maintenant dans l'estimation des probabilités conditionnelles représentées par les paramètres θ_e et $\theta_{\bar{e}}$. En effet, les objets étant affectés de façon probabiliste dans les différents nœuds de l'arbre, il ne nous est plus possible ici d'estimer comme avant ces paramètres, «localement» à l'aide des fréquences calculées en (8). Ce phénomène se traduit également au niveau de l'expression de la vraisemblance qui présente une forme «complexe» : contrairement au cas précédent, on ne peut plus isoler les termes relatifs aux deux nœuds créés. La maximisation de celle-ci doit donc être effectuée globalement pour pouvoir estimer θ_e et $\theta_{\bar{e}}$.

Nous avons à résoudre ici ce que l'on appelle usuellement un problème d'estimation de paramètres dans le cadre d'un *mélange de lois de probabilité*. Cette problématique, qui a donné lieu à une abondante publication en classification automatique (voir par exemple (Titterington *et al.* 85)), a par contre été peu étudiée dans une optique prédictive; sur ce thème, on consultera par exemple (Wedel *et al.* 95) ou encore (Jordan *et al.* 93).

La résolution d'un problème de maximisation tel que (11) fait généralement appel à deux types de méthodes : les procédures d'estimation standard comme l'algorithme de Newton-Raphson, ou celles fondées sur l'algorithme EM (Dempster *et al.* 77). Nous présentons ici ce second type de méthodes qui semblent avoir actuellement la préférence d'un plus grand nombre d'auteurs (Wedel *et al.* 95). Nous ne donnerons dans ce travail qu'un bref aperçu du principe de cet algorithme et des résultats qu'il fournit, et renvoyons le lecteur à (Périnel 96) pour un exposé plus détaillé.

6.4. Estimation des paramètres par l'algorithme EM

Principe de l'algorithme EM. L'idée de base de l'algorithme EM (en référence aux deux étapes *Expectation – Maximization*) est de supposer que la fonction de vraisemblance, «complexe» à maximiser, résulte d'une situation de données incomplètes. Dans notre cas, il s'agit de la non connaissance de l'affectation des objets aux différents nœuds de l'arbre. On introduit alors des variables, dites cachées, apportant cette information manquante. Grâce à cet «artifice» de calcul, on construit alors une vraisemblance simplifiée, dont la maximisation itérative fournit, à la convergence, les estimateurs du maximum de vraisemblance du problème complexe.

L'expression de la log-vraisemblance simplifiée, pour des données dites complètes, est définie ici par

$$l_c(\theta; c|y) = \log \prod_{i=1}^N \prod_{t=1}^T (P(t|y_i) \cdot P(c_i; \theta_t))^{\delta_{it}}. \quad (12)$$

où δ_{it} est la variable indicatrice de l'appartenance d'un objet i au nœud t (c'est-à-dire, $(\delta_{it} = 1) \Leftrightarrow (i \in \mathcal{N}_t)$). Chacune des itérations de l'algorithme EM comporte deux étapes :

- L'étape **Espérance** : elle consiste à calculer $E[l_c]$, la valeur espérée de l_c , conditionnellement aux valeurs observées pour les prédicteurs;
- L'étape **Maximisation** : elle fournit les estimateurs du maximum de vraisemblance de θ_e et $\theta_{\bar{e}}$, solutions du problème de maximisation de $E[l_c]$.

Nous nous contenterons ici de donner l'expression des estimateurs des paramètres θ_e et $\theta_{\bar{e}}$ obtenus à une étape (r) quelconque de l'algorithme :

$$\hat{p}_e^{(r)}(k) = \frac{\sum_i c_{ik} \cdot E^{(r)}(\delta_{ie})}{\sum_i E^{(r)}(\delta_{ie})}, \quad \hat{p}_{\bar{e}}^{(r)}(k) = \frac{\sum_i c_{ik} \cdot E^{(r)}(\delta_{i\bar{e}})}{\sum_i E^{(r)}(\delta_{i\bar{e}})} \quad (13)$$

avec : $E^{(r)}$, l'opérateur d'espérance mathématique à l'étape (r); c_{ik} et δ_{ie} , les variables indicatrices de l'appartenance de l'observation i à la classe k et au nœud e .

Sur le plan informatique, il est important de signaler que l'algorithme EM est parfois contraint d'effectuer un nombre important d'itérations avant de converger. Les performances de certains algorithmes récents (l'algorithme SEM (Celeux *et al.* 92), qui incorpore une étape stochastique; des méthodes basées sur l'accélération d'Aitken (Lindstrom *et al.* 88) ou du gradient conjugué (Jamshidian *et al.* 93); l'algorithme ECM qui substitue à l'étape de maximisation usuelle, une succession d'étapes CM, où l'utilisation d'une vraisemblance *conditionnelle* permet de simplifier sensiblement les calculs (Meng *et al.* 93)), laissent néanmoins espérer une amélioration sensible des performances relatives à cette phase numérique.

Remarque sur la sélection d'un arbre de «bonne taille». Cette phase prédictive s'écartant de la problématique centrale de ce travail, nous n'y accorderons ici qu'une place limitée. Notons tout d'abord que la construction même d'un arbre et les procédures d'élagage usuelles constituent des phases relativement bien séparées et indépendantes (voir par exemple (Mingers 89)). De ce fait, il est possible de «greffer» à la procédure de construction d'arbre décrite ci-dessus un des nombreux algorithmes d'élagage développés dans la littérature (voir par exemple (Esposito 97)).

Cependant, la construction qui est proposée ci-dessus peut suggérer d'adopter l'approche suivante (Ciampi *et al.* 96) : en divisant à chaque étape le nœud terminal

généralisant l'accroissement d'information maximal, on crée ainsi une suite de sous-arbres emboîtés, chacun d'eux étant le «meilleur»³ pour un nombre de nœuds terminaux donné. Cette sous-suite constitue en quelque sorte le pendant de la sous-suite optimale obtenue par la méthodologie CART durant la phase d'élagage et non au cours la construction de l'arbre. Dans un second temps, on pourra alors choisir l'arbre associé à une estimation optimale du taux d'erreur de classement (en utilisant par exemple un échantillon test).

En résumé : la phase de recherche de la meilleure question (coupure) binaire à une étape T quelconque de l'algorithme de segmentation (i.e., visant à déterminer la meilleure partition comportant T nœuds terminaux) est réalisée à travers les étapes suivantes :

1. On fixe un nœud t parmi les $T - 1$ nœuds terminaux de la partition courante \mathcal{P}_{T-1} ;
2. Pour chaque question binaire $s = (e, \bar{e})$ au nœud t , on construit la partition induite \mathcal{P}_T^s à T nœuds terminaux;
3. Pour déterminer les probabilités conditionnelles de la variable C dans les deux nouveaux nœuds, on résout alors le problème de maximisation suivant :

$$\max_{\theta_e, \theta_{\bar{e}}} \text{Ginf}(\mathcal{P}_T^s)$$

4. L'estimation des paramètres θ_e et $\theta_{\bar{e}}$ permet de calculer la valeur de l'information portée par la partition induite par (e, \bar{e}) ;
5. On retient enfin le nœud t et la question s conduisant à la partition dont la qualité mesurée par Ginf est maximale.

Calcul des probabilités $P(t|y_i)$. De manière générale, si l'appartenance au nœud t est conditionnée par le fait de satisfaire à R propriétés e_1^t, \dots, e_R^t , alors la probabilité qu'un objet i décrit par y_i appartienne au nœud t est donnée par

$$P(t|y_i) = P(e_1^t \cap \dots \cap e_R^t | y_i)$$

où, on le rappelle, e_r^t est une condition du type $[Y \in V]$. Par suite, si l'on admet que les aléas affectant chacune des observations réalisées pour les prédicteurs d'un individu i donné sont indépendants⁴ (sous l'hypothèse donc d'indépendance des

³ Étant donné bien sûr la partition précédente, puisqu'il s'agit d'un algorithme dit «glouton», et donc non globalement optimal.

⁴ Ce serait par exemple faire l'hypothèse que ce qui conduit un médecin à estimer le poids d'un patient donné dans l'intervalle $[62, 65]$ n'est pas influencé par le fait qu'il estime que la réaction de ce patient à un test soit positive avec la probabilité 0.8.

variables Y_{ij} , et non des Y_j), on a alors

$$P(t|\mathbf{y}_i) = \prod_r P(e_r^t | \mathbf{y}_i) \quad (14)$$

où chacun des termes du produit correspond au calcul d'une affectation «souple» présentée au paragraphe 4.2. Les \mathbf{y}_i étant non observées, on peut néanmoins estimer les différentes probabilités (14) sur la base des lois de probabilité des variables Y_{ij} (qui elles sont connues), comme l'illustre l'exemple simple suivant. Reprenons la description imprécise de l'individu (i) dans l'exemple de la section 4.2. En se limitant aux seuls prédicteurs, nous l'avons noté $(\mathcal{B}(0.7), \mathcal{U}_{[3,5]})$. Considérons ensuite l'arbre de la figure 4.

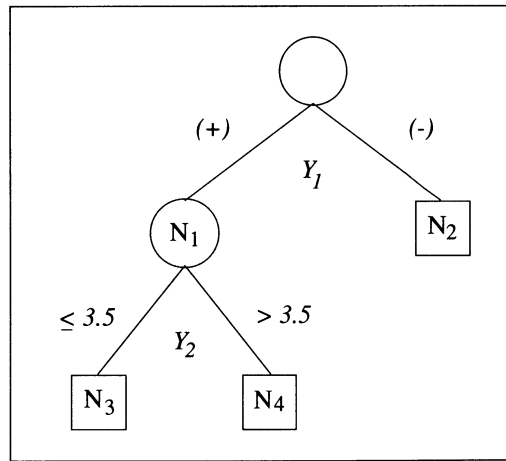


FIGURE 4

Exemple d'arbre illustrant le calcul des probabilités $P(t|\mathbf{y}_i)$

Pour cet individu particulier, les probabilités d'appartenir aux nœuds N_1 et N_2 , c'est-à-dire de prendre les valeurs (+) et (-) sont bien sûr calculées par $P(1|\mathbf{y}_i) = 0.7$ et $P(2|\mathbf{y}_i) = 0.3$. Pour déterminer ensuite les probabilités d'appartenance à N_3 et N_4 , nous raisonnons sous l'hypothèse d'indépendance des variables aléatoires Y_{i1} et Y_{i2} ; on obtient alors :

$$\begin{aligned} P(3|\mathbf{y}_i) &= P([Y_{i1} = (+)] \cap [Y_{i2} \leq 3.5]) = P[Y_{i1} = (+)] \times P[Y_{i2} \leq 3.5] \\ &= 0.7 \times \frac{3.5 - 3}{2} = 0.175 \\ P(4|\mathbf{y}_i) &= 0.7 \times \frac{5 - 3.5}{2} = 0.525 \end{aligned}$$

6.5. Un exemple illustratif

Afin d'illustrer au mieux le problème de la recherche d'une meilleure coupure binaire (choix d'une question, estimation des paramètres puis calcul de l'information associée à la question), considérons l'exemple suivant : deux variables prédictrices binaires ont été observées (avec imprécision) sur 10 objets répartis en deux groupes *a priori* ($C = 1$ ou 2). Le tableau 4 résume l'ensemble des données initiales. Nous nous sommes posés ici le problème très simple de la recherche de la première coupure, c'est-à-dire ici de connaître la variable expliquant le mieux dans un premier temps la variable C . La procédure de segmentation a fourni les résultats présentés dans le tableau 5.

TABLEAU 4

Exemple de données imprécises :
pour les deux prédicteurs, seule la probabilité de la modalité 1 a été indiquée ;
par exemple, pour ω_7 , on a en fait $P(Y_{1,7} = 1) = 0.2$
et $P(Y_{1,7} = 0) = 0.8$; ou encore, $Y_{1,7} \sim \text{Bernoulli}(0.2)$.

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
C	1	1	1	1	1	1	2	2	2	2
Y_1	0.8	1	1	0.2	0.7	0.9	0.2	0.1	0.2	0
Y_2	0.6	0.6	0.7	0.6	0.6	0.6	0.4	0.4	0.4	0.3

TABLEAU 5

Résultats de la recherche de la première meilleure segmentation

Question binaire		Prob. condit. estimées		Information (Ginf)
e	\bar{e}	$(p_e(1), p_e(2))$	$(p_{\bar{e}}(1), p_{\bar{e}}(2))$	
$[Y_1 = 1]$	$[Y_1 = 0]$	(1, 0)	(0.044, 0.956)	-2.83
$[Y_2 = 1]$	$[Y_2 = 0]$	(1, 0)	(0, 1)	-4.80

- **Questions binaires.** Elles sont dans cet exemple très simples et on en dénombre autant que de variables prédictrices (soit 2).
- **Probabilités estimées.** Pour les deux questions binaires, la probabilité de la classe 1 est égale à 1 conditionnellement à la modalité 1 ($p_e(1)=1$). Elle est par contre respectivement très faible (4.4%) et nulle conditionnellement à la modalité 0 pour les deux coupures.
- **Information.** Si les probabilités conditionnelles estimées sont peu différentes pour les deux questions binaires, il n'en va pas de même en ce qui concerne la qualité des deux partitions induites, la première étant nettement plus favorable que la seconde. Ceci s'explique par le fait que l'incertitude sur les valeurs de la variable Y_1 est bien

moins importante que celle affectant Y_2 et donc, en d'autres termes, que la répartition des objets est «plus stable» pour Y_1 que pour Y_2 .

Pourtant il est ici intéressant de noter que la décision inverse aurait été prise si l'on avait choisi dans le tableau 4 d'affecter comme description à un objet, la valeur la plus probable pour chacun des descripteurs (par exemple, respectivement 1 et 1 comme descriptions de ω_1 pour Y_1 et Y_2). En effet, dans cette situation (tableau 6), la seconde question aurait conduit à une répartition «idéale» des individus (des nœuds dits purs) et donc à une log-vraisemblance nulle, c'est-à-dire optimale; pour la question basée sur le premier prédicteur, l'information est ici peu différente de celle résultant de la prise en compte de l'imprécision (celle-ci étant relativement faible pour Y_1).

TABLEAU 6
Résultats de la recherche de la première meilleure segmentation
pour les données précises associées au tableau 4

Question binaire		Prob. condit. estimées		Information (Ginf)
e	\bar{e}	$(p_e(1), p_e(2))$	$(p_{\bar{e}}(1), p_{\bar{e}}(2))$	
$[Y_1 = 1]$	$[Y_1 = 0]$	(1, 0)	(0.2, 0.8)	-2.50
$[Y_2 = 1]$	$[Y_2 = 0]$	(1, 0)	(0, 1)	0

Calcul des effectifs. Comme dans le cas d'une segmentation standard, on peut tenter d'associer à chaque nouveau nœud son effectif en termes des classes à expliquer. Les effectifs «réels» n'étant pas connus en raison de l'affectation probabiliste, on peut néanmoins, à partir des probabilités $p(t|y_i)$, construire les quantités suivantes, qui se ramènent aux effectifs usuels en cas de données précises :

$$n_t = \sum_{k=1}^K n_t(k) \quad \text{avec} \quad n_t(k) = \sum_{i \in C_k} p(t|y_i) \quad (15)$$

où $\{i \in C_k\}$ désigne l'ensemble des objets ω_i appartenant à la classe k . Les nombres n_t et $n_t(k)$ désignent donc respectivement l'effectif du nœud t et celui de la classe k dans le nœud t . Dans l'exemple précédent, la coupure basée sur Y_1 aboutit aux effectifs suivants :

$$n_e = n_e(1) + n_e(2) = 4.6 + 0.5 = 5.1$$

$$n_{\bar{e}} = n_{\bar{e}}(1) + n_{\bar{e}}(2) = 1.4 + 3.5 = 4.9$$

Notons finalement que les probabilités estimées par la méthode du maximum de vraisemblance diffèrent ici de celles que l'on obtiendrait en retenant comme estimateurs $\hat{p}_t(k) = \frac{n_t(k)}{n_t}$ (stratégie retenue par exemple dans (Arraya 95)).

7. Application

Problématique. En oncologie, le problème de l'identification des tumeurs de nature neuro-endocrine⁵ constitue actuellement une question complexe et essentielle (en vue tout particulièrement d'attribuer à ces tumeurs un traitement thérapeutique adéquat). Pour aborder ce problème, médecins et chercheurs disposent aujourd'hui de nouveaux moyens d'investigation dont, en particulier, l'utilisation de marqueurs biologiques de différenciation (Caillou *et al.* 92). Ces techniques visent à mettre en évidence au niveau d'un type cellulaire donné (par exemple, la cellule *C* de la thyroïde) un ensemble d'anticorps particuliers (la calcitonine, la neuro-specific enolase,...). L'objectif de cette étude est de tenter de mieux cerner la «frontière» du système neuro-endocrinien (NE) sur la base des résultats produits par ces marqueurs.

Description des données. On dispose pour cela d'une population de types cellulaires appartenant ou non au système NE. La variable à expliquer vaut donc ici 1 (ou 0) selon qu'un type cellulaire appartient (ou non) au système NE.

Chaque type cellulaire est décrit par le médecin sur la base des résultats produits par un ensemble de marqueurs de différenciation. Pour chaque marqueur, deux type de descriptions sont fournies par le médecin : la *localisation* de la réaction au niveau de la cellule; l'*intensité* de la réaction. Les modalités possibles pour chacun de ces deux descripteurs (respectivement qualitatif et quantitatif) sont les suivantes :

- **localisation**, variable nominale à 5 modalités :
(CD) : Diffus dans le Cytoplasme; (CG) : dans les Grains de sécrétion Cytoplasmique; (CP) : à la Périphérie du Cytoplasme; (M) : Membrane; (EC) : Extérieur Cellule;
- **intensité**, variable discrète ordonnée :
Elle varie sur l'échelle $\llbracket 0...4 \rrbracket$, désignant tous les entiers compris entre 0 (faible intensité) et 4 (intensité très marquée).

La description d'un type cellulaire pour un marqueur donné correspond en fait à la *connaissance* qu'a le médecin de la réaction provoquée en général par ce marqueur sur ce type cellulaire (il s'agit également, selon une terminologie employée par exemple dans (Lebbe 91), d'une description de concept médical). De ce fait, elle comporte de manière naturelle des imprécisions ou incertitudes plus ou moins importantes comme l'illustre l'exemple suivant. Il s'agit de la description du type cellulaire *cellule vésiculaire de la thyroïde* dans son état *normal*, en opposition à l'état tumoral (la représentation logique employée ici, correspondant à une *assertion*, est empruntée au formalisme de l'analyse des données symboliques) :

$$\begin{aligned}
 [classe = \overline{NE}] \quad & \wedge [intensité (CALCI.) = 0] \\
 & \wedge [intensité (HNK1) = \llbracket 0, 3 \rrbracket] \quad \wedge [localisation (HNK1) = \{CD, EC\}] \\
 & \wedge [intensité (PS1) = \llbracket 0, 2 \rrbracket] \quad \wedge [localisation (PS1) = CD] \quad \dots
 \end{aligned}$$

⁵ D'un point de vue fonctionnel, le système neuro-endocrinien est un mécanisme de régulation qui relie les fonctions nerveuses et des cellules endocrines à travers tout le corps; certains types cellulaires de la thyroïde ou du pancréas par exemple, font partie de ce système.

Cette description s'interprète comme suit : le type cellulaire vésiculaire de la thyroïde, de classe non NE (ou \overline{NE}), réagit négativement au marqueur de la calcitonine; positivement au marqueur HNK1 (mais entre des niveaux pouvant varier de 0 à 3) de façon diffuse dans le cytoplasme ou bien à l'extérieur de la cellule; pour le marqueur PS1, l'intensité de la réaction peut varier de 0 à 2 et apparaît de façon diffuse dans le cytoplasme; etc. Nous avons adopté dans la suite un codage de type probabiliste uniforme consistant à associer des probabilités égales à chacun des états exprimant l'imprécision. Par exemple :

$$\llbracket 2, 3 \rrbracket \longrightarrow P(2) = P(3) = \frac{1}{2}$$

Cette hypothèse, certes simplificatrice, répond néanmoins au principe raisonnable et souvent employé dans les applications du *maximum d'entropie*; à savoir, l'attribution de probabilités égales à des événements pour lesquels on ne possède pas d'information précise quant à leur degré d'incertitude.

Questions binaires. Des questions binaires standard ont été employées dans le cadre de cette application. Rappelons que celles-ci sont déterminées par un couple de propriétés $[Y \in V]/[Y \in \overline{V}]$ où $\{V, \overline{V}\}$ forment ici une partition du domaine des intensités possibles $\llbracket 0...4 \rrbracket$ ou de celui des localisations possibles $\{CD, CG, M, EC, CP\}$.

Résultats et commentaires. Notons tout d'abord que si l'intérêt médical des résultats de cette application est pour l'instant limité, cette dernière s'inscrit néanmoins dans le cadre d'une démarche méthodologique dont l'importance est liée à l'utilisation actuellement grandissante de ces techniques de marquage en immunohistochimie. D'un point de vue pratique, certains marqueurs très spécifiques du système neuroendocrinien permettent, pour notre jeu de données, une caractérisation immédiate de la variable à expliquer (ceux-ci produisent donc des arbres «triviaux» – mais attendus – à une seule coupure binaire). Ces marqueurs ont donc par la suite été écartés du jeu de données afin de construire des arbres permettant la mise en évidence d'interactions entre les marqueurs restants. C'est un de ces derniers arbres que nous produisons maintenant (voir figure 5) et que nous allons commenter en regard des éléments théoriques introduits dans les sections précédentes.

Nous avons stoppé la construction de l'arbre, soit en présence de nœuds purs, soit encore en procédant à un «élagage expert», c'est-à-dire en admettant qu'une coupure supplémentaire n'apportait plus d'amélioration sensible pour le problème posé. L'arbre de la figure 5 est ici très simple et conduit à trois sous-populations de types cellulaires (trois feuilles) dont deux peuvent être étiquetées \overline{NE} (les probabilités estimées des deux classes sont indiquées à l'intérieur de chaque nœud). Ainsi, le nœud terminal situé à l'extrême droite décrit des types cellulaires dont la réaction aux marqueurs HNK1 et PS1 est de niveau au moins supérieur à la valeur 1. Dans la même optique que l'exemple du paragraphe 6.4, la procédure conduit ici à sélectionner des questions binaires basées sur des prédicteurs (marqueurs) produisant moins d'incertitude au niveau de la réaction des différents types cellulaires. Par exemple, pour deux marqueurs qui auraient conduit à des partitions de qualité égales

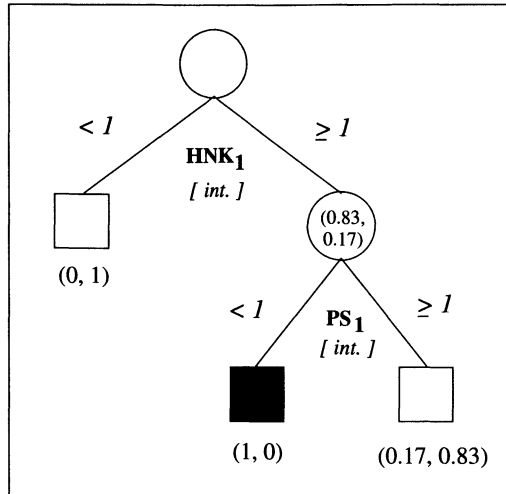


FIGURE 5

Exemple d'arbre binaire permettant la caractérisation de certains types cellulaires du système neuro-endocrinien.

sur la vue de l'intensité, pour un codage standard (par exemple (+) ou (-)), la méthode sélectionnera ici celui dont les plages de variabilité de l'intensité sont moins importantes.

Utilisation de l'arbre comme règle de classement. Comme tout arbre de segmentation, on peut souhaiter dans un second temps employer l'arbre construit dans une optique décisionnelle. Considérons par exemple, le type cellulaire décrit plus haut (cellule vésiculaire, notée cv), en ignorant sa classe d'appartenance \overline{NE} . De manière générale, la probabilité qu'un objet i soit affecté à la classe k est déterminée, selon un calcul de probabilités conditionnelles usuel, par :

$$P(k|i) = \sum_{t=1}^T p(t|i) \cdot P_t(k) \quad (16)$$

où la probabilité que i appartienne au nœud t , $p(t|i)$, est calculée selon le principe décrit par l'équation (14). Par exemple, la probabilité d'appartenance de l'objet cv à la feuille de droite est calculée par

$$p[int.(HNK1) \geq 1 | cv] \times p[int.(PS1) \geq 1 | cv] = \frac{3}{4} \times \frac{2}{3} = 0.5$$

On rappelle en effet, qu'à l'intervalle $[0, 2]$ est supposé correspondre une probabilité de $\frac{1}{3}$ pour chacun des états 0, 1 et 2. Par suite, en notant p la probabilité

que le type cellulaire précédent appartienne à la classe \overline{NE} (sa classe *a priori*), on obtient donc :

$$p = \frac{1}{4} \times 1 + \left(\frac{3}{4} \times \frac{1}{3}\right) \times 0 + \left(\frac{3}{4} \times \frac{2}{3}\right) \times 0.83 \simeq 0.67$$

D'autre part, un type cellulaire qui réagirait toujours négativement à ces deux marqueurs serait affecté à la classe \overline{NE} avec une probabilité de 1.

8. Conclusion

Dans l'optique (davantage paramétrique) de la méthodologie RECPAM, nous avons proposé un algorithme de construction d'arbre de discrimination binaire, permettant le traitement de données de nature imprécise. Il est important de souligner tout d'abord que, partant d'un ensemble de descriptions – éventuellement entâchées d'une forte imprécision – cette approche parvient à la construction d'une règle de classement usuelle, c'est-à-dire exprimée en termes de propriétés précises portant sur les prédicteurs. Déjà proposé par quelques auteurs, le principe d'affectation probabiliste employé dans notre approche a toutefois la particularité : d'une part, d'être employé *dès la phase de construction de l'arbre* (et non uniquement lors d'une phase décisionnelle, c'est-à-dire une fois l'arbre construit à partir de données standard, comme c'est le cas par exemple dans (Quinlan 90)); d'autre part, de permettre une estimation rigoureuse des probabilités conditionnelles de la variable à expliquer dans chacun des nœuds de l'arbre. Il est aussi intéressant de noter, du point de vue de l'interprétation des résultats, que ces probabilités sont exprimées selon une sémantique identique à celle représentant l'imprécision des différentes descriptions. Ceci n'est pas toujours le cas pour certains arbres de décision, où à des descriptions imprécises de type floue correspondent des probabilités particulières appelées probabilités conditionnelles floues. Enfin, sur le plan des perspectives, ce travail devrait connaître des extensions sur deux points en priorité : la possibilité de prise en compte *a priori* d'une imprécision quant à l'appartenance d'un objet à l'une des classes de la partition (problème fréquemment rencontré en pratique); le développement d'une méthodologie identique, mais dans une optique de régression.

Remerciements

Pour leur très précieuse collaboration dans le cadre de l'application médicale, j'aimerais remercier chaleureusement le Docteur Bernard Caillou du service d'histopathologie A de l'Institut Gustave Roussy à Villejuif, ainsi que Mireille Gettler-Summa, Maître de conférences à l'Université Paris IX-Dauphine.

Références bibliographiques

ARRAYA R., (1995). Induction of decision trees when examples are describes with noisy measurements and with fuzzy class membership. In *Seminaire du projet CLOREC*, INRIA Rocquencourt, juin.

- BELSON W.A., (1959). Matching and prediction on the principle of biological classification, *Applied Statistics*, vol. VIII.
- BREIMAN L., FRIEDMAN J.H., OLSHEN R.A. and STONE C.J., (1984). Classification and regression trees. Belmont : Wadsworth.
- CAILLOU B., TARTOUR E. et SCHLUMBERGER M., (1992). Les tumeurs neuro-endocrines, *Revue Prat.*, vol.42, 7.
- CELEUX G. et NAKACHE J.P., (1994). *Analyse discriminante sur variables qualitatives*, PolyTechnica.
- CELEUX G. and GOVAERT G., (1992). A Classification EM algorithm for clustering and two stochastic versions, *Computational Statistics & Data Analysis*, vol.14, 315–332.
- CIAMPI A., (1992). Constructing prediction trees from data : the RECPAM approach. *Proceedings from the Prague 1991 University Summer School on computational aspects of model choice*, 105–152, Physica Verlag, Heidelberg.
- CIAMPI A., DIDAY E., LEBBE J., PÉRINEL E. and VIGNES R., (1996). Recursive partition with probabilistically imprecise data. In : *Ordinal and Symbolic Data Analysis*, 201-212, Diday E. et al. editors, Springer-Verlag.
- DEMPSTER A., LAIRD N. and RUBIN D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, B39, 1–38.
- DIDAY E., (1987). Des objets de l'analyse des données à ceux de l'analyse des connaissances. In *Induction symbolique et numérique à partir de données*, E. Diday et Y. Kodratoff (eds.), Cépaduès.
- DIDAY E. et ÉMILION, R., (1997). Treillis de Galois maximaux et Capacités de Choquet. *C.R. Acad. Sci. Paris, Analyse mathématique*, t. 324, série 1.
- ESPOSITO F., MALERBA D. and SEMERARO G., (1997). A comparative analysis of methods for pruning decision trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol.19, n°5, 476-492.
- FRIEDMAN J.H., (1977). A Recursive Partitioning Decision Rule for Nonparametric Classification, *IEEE Transactions on Computers*, april, 404–408.
- GUEGUEN A. et NAKACHE J.P., (1988). Méthode de discrimination basée sur la construction d'un arbre de décision binaire, *Revue de Statistique Appliquée*, vol.XXXVI, 1, 19–38.
- JAMSHIDIAN M., JENNRICH R.I., (1993). Conjugate gradient acceleration of the EM algorithm, *Journal of the American Statistical Association*, march, vol. 88, 421.
- JORDAN M.I. and JACOBS R.A., (1993). Hierarchical mixtures of experts and the EM algorithm, août 93, soumis à *Neural Computation*.
- LEBART L., MORINEAU A. et PIRON M., (1995). *Statistique exploratoire multidimensionnelle*, Dunod
- LEBBE J., (1991). Représentation des concepts en biologie et en médecine, *Thèse de l'Université Pierre et Marie Curie, Paris VI-Jussieu*.

- LINDSTROM M.J. and BATES D.M., (1988). Newton Raphson and EM algorithms for linear mixed-effects models for repeated-measured data, *Journal of the American Statistical Association*, vol.83, 1014–1022.
- MENG X.L. and RUBIN D.B., (1993). Maximum likelihood estimation via the ECM algorithm : a general framework, *Biometrika*, vol.80, 2, 267–278.
- MINGERS J., (1989). An empirical comparison of pruning methods for decision-tree induction, *Machine Learning*, vol. 4, 2, 227–243.
- MORGAN J.N. and SONQUIST J.A., (1963). Problems in the analysis of survey data, and a proposal, *J.A.S.A.*, vol.58, 302.
- PÉRINEL E., (1996). Méthodes de segmentation et analyse des données symboliques. Le cas de données probabilistes imprécises. *Thèse de l'Université Paris IX-Dauphine*.
- QUINLAN J.R., (1986). The effect of noise on concept learning. In Michalski, Carbonel & Mitchell (Eds.), *Machine Learning : an artificial intelligence approach*. San Mateo, CA : Morgan Kaufman.
- QUINLAN J.R., (1990). Probabilistic decision trees. In : *Machine Learning III*, Kodratoff Y., Michalski R. (eds.), 140–152.
- QUINLAN J.R., (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufman, San Mateo, California.
- SETHI I.K. and SARVARAYUDU G.P.R., (1982). Hierarchical classifier design using mutual information, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, 441–445.
- TITTERINGTON D.M., SMITH A.F.M. and MAKOV U.E., (1985). *Statistical analysis of finite mixture distributions*, Wiley & Sons (Eds.), Wiley series in probability and mathematical statistics.
- WEDEL M. and DE SARBO W.S., (1995). A mixture likelihood approach for generalized linear models, *Journal of Classification*, vol.12, 21–55.
- YUAN Y. and SHAW M.J., (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69, 125–139.

Annexe : lien entre coupure souple et observations imprécises

- Considérons n observations (y_1, \dots, y_n) d'une variable réelle Y . Ces n observations sont dites précises ou standard en accord avec la terminologie introduite dans ce travail.
- Soit F_s , une fonction définie sur \mathbb{R} représentant une coupure souple associée à la question binaire : « $[Y \leq s]$ ou $[Y > s]$? ». La coupure souple F_s n'est en fait rien d'autre qu'une fonction de répartition possédant de plus la propriété $F_s(s) = 0.5$, traduisant simplement une incertitude maximale au niveau du seuil de coupure. Les fonctions F_s et $1 - F_s$ déterminent donc respectivement l'appartenance aux régions $[Y \leq s]$ et $[Y > s]$. On supposera dans la suite que $F_s(y) = F(y - s)$ où F est une fonction de répartition telle que $F(0) = 0.5$.

- On construit sur la base du n -uplet (y_1, \dots, y_n) , n observations imprécises définies par autant de densités de probabilité f_{Y_1}, \dots, f_{Y_n} respectivement symétriques en y_1, \dots, y_n , et issues d'une même famille de lois de probabilité que l'on notera f .
- Soit I_s , la fonction caractéristique définissant la coupure standard (ou coupure nette) en s , c'est-à-dire telle que $I_s(y) = 1$ si $y > s$ (0 sinon). I_s et $1 - I_s$ définissent donc l'appartenance aux régions $[Y > s]$ et $[Y \leq s]$.
- Enfin, on note $\delta_i = y_i - s$, l'écart observé entre la description de Y pour l'individu i et le seuil de coupure s . L'ensemble des notions précédentes se trouve résumé dans le cadre des deux graphiques de la figure 6.

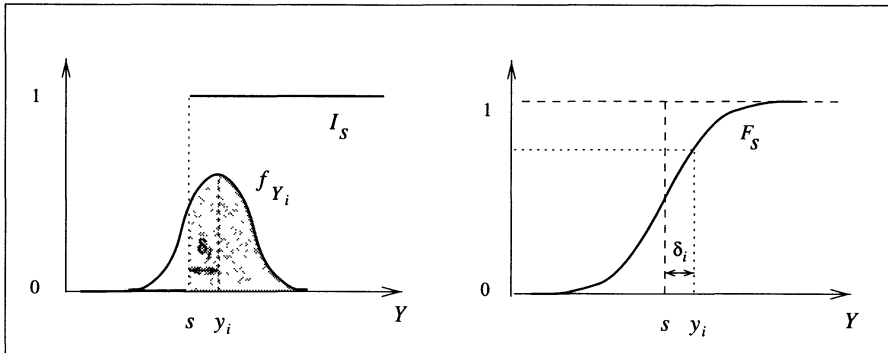


FIGURE 6

*Coupure douce et donnée précise (à droite);
Coupure nette et donnée imprécise (à gauche)*

On a alors le résultat immédiat suivant :

Pour toute observation y_i ($i = 1, \dots, n$), la probabilité d'appartenance à la région $[Y > s]$ est déterminée par :

$$P[Y_i > s] = F_s(y_i) = \int_s^{+\infty} f_{Y_i}(y).dy \quad (17)$$

sous la condition que l'imprécision des observations et la coupure souple soient reliés par $f = F'$.

Autrement dit, l'utilisation d'une coupure souple F_s sur un jeu de données précises $(y_i)_i$ est analogue à l'emploi d'une coupure nette usuelle sur des données imprécises modélisées par f_{Y_i} , dont la forme est définie à partir de la modélisation de la coupure F_s .

Le résultat est naturel et découle des hypothèses relatives à la modélisation des différentes imprécisions. On a tout d'abord, puisque F a été définie comme la

fonction de répartition associée à f :

$$\int_s^{+\infty} f_{Y_i}(y).dy = F_{Y_i}(+\infty) - F_{Y_i}(s) = 1 - F_{Y_i}(s)$$

Par ailleurs, f étant symétrique en Y_i , on a

$$f_{Y_i}(y_i + h) = f_{Y_i}(y_i - h)$$

Par suite, en intégrant entre δ_i et $+\infty$, et puisque $y_i - \delta_i = s$ et $F_{Y_i}(u) = F(u - y_i)$,

$$\begin{aligned} \int_{\delta_i}^{+\infty} f_{Y_i}(y_i + h).dh &= \int_{\delta_i}^{+\infty} f_{Y_i}(y_i - h).dh \\ \Leftrightarrow 1 - F_{Y_i}(y_i + \delta_i) &= F_{Y_i}(y_i - \delta_i) \\ \Leftrightarrow 1 - F_{Y_i}(s) &= F_{Y_i}(y_i + \delta_i) \\ \text{d'où } 1 - F_{Y_i}(s) &= F(\delta_i) = F(y_i - s) = F_s(y_i) \end{aligned}$$

Interprétation. Lorsque l'on emploie des coupures douces, cela revient donc au même que de supposer que toutes les grandeurs observées sont affectées d'une imprécision identique, en présence d'une coupure qui elle est classique. Ceci permet de donner l'interprétation suivante d'une coupure souple, en termes de stabilité, rejoignant sous certains aspects, celle trouvée dans (Quinlan 90). Supposons en effet que l'on augmente sensiblement le degré d'inclinaison d'une coupure au seuil s sans pour autant modifier (ou, en modifiant très faiblement) la qualité de la partition associée. Cela signifie en d'autres termes que des perturbations de même grandeur autour des valeurs observées ne modifierait pas la qualité de la coupure et par suite que cette dernière est stable face à des phénomènes de bruit affectant les données.

De manière plus générale, il est intéressant de noter qu'un processus d'affectation peut être modélisé à travers la notion de loi produit de convolution. Si l'on note en effet $\Delta = Y - S$, l'écart aléatoire entre une observation modélisée par f_Y et une coupure modélisée par une fonction de répartition G_S , la probabilité d'appartenance d'une observation à la région $[Y > s]$ est calculée par l'expression suivante

$$P[\Delta \geq 0] = \int f_Y(y).G_S(y).dy = \int g_S(s).(1 - F_Y(s)).ds$$

qui généralise l'équation (17) au cas d'une coupure et d'une observation pouvant être simultanément imprécises.