Partition BIC optimale de l'espace des prédicteurs

Gilbert Ritschard

*Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Cet article traite du partitionnement optimal de l'espace de prédicteurs catégoriels dans le but de prédire la distribution a posteriori d'une variable réponse elle-même catégorielle. Cette partition optimale doit répondre à un double critère d'ajustement et de simplicité que prennent précisément en compte les critères d'information d'Akaike (AIC) ou bayésien (BIC). Après avoir montré comment ces critères s'appliquent dans notre contexte, on s'intéresse à la recherche de la partition qui minimise le critère retenu. L'article propose une heuristique rudimentaire et démontre son efficacité par une série de simulations qui comparent le quasi optimum trouvé au vrai optimum. Plus que pour la partition elle-même, la connaissance de cet optimum s'avère précieuse pour juger du potentiel d'amélioration d'une partition, notamment celle fournie par un algorithme d'induction d'arbre. Un exemple sur données réelles illustre ce dernier point.

1 Introduction

En apprentissage supervisé, des techniques comme l'analyse discriminante, la régression logistique multinomiale, les modèles bayésiens ou les arbres de décisions induits de données (arbres d'induction) apprennent la distribution a posteriori de la variable à prédire, l'objectif étant d'affecter un cas avec profil $\mathbf{x}$ en termes de prédicteurs à la classe y_i ayant la plus forte probabilité a posteriori $p(Y = y_i|\mathbf{x})$. La régression logistique, l'analyse discriminante et les modèles bayésiens par exemple, modélisent la distribution a posteriori sous forme d'une fonction vectorielle continue de $\mathbf{x}$. Par contraste, les arbres d'induction conduisent à un ensemble fini de distributions, chaque distribution étant associée à une classe d'une partition « apprise » de l'ensemble des profils $\mathbf{x}$ admissibles. Nous nous plaçons dans ce dernier contexte et nous intéressons à la détermination de la partition optimale. Nous examinons tout d'abord les critères d'optimalité qui peuvent s'avérer pertinents. Parmi ceux-ci, nous porterons un intérêt particulier aux critères d'information du type AIC et BIC qui permettent d'arbitrer entre qualité d'ajustement et complexité. Le calcul des AIC et BIC pour une partition quelconque se fait par simple adaptation du principe de l'arbre étendu introduit dans Ritschard et Zighed (2003, 2002) pour le cas des arbres.

La recherche de la partition optimale par exploration exhaustive des partitions étant de complexité non polynomiale, il convient de recourir à des heuristiques. Pour cette première approche du problème, on envisage ici une procédure ascendante dont on examine les performances par une analyse de simulations. L'heuristique est rudi-

mentaire, mais s'avère suffisamment performante pour traiter jusqu'à une centaine de profils différents.

La portée du concept même de partition optimale est quelque peu limitée en raisons de possibles difficultés d'interprétation. Contrairement aux partitions générées par les arbres, la description d'une partition quelconque peut nécessiter en effet des combinaisons souvent difficilement interprétables de conditions. L'optimum fournit cependant et dans tous les cas des indications précieuses sur le potentiel d'amélioration qu'on peut apporter à une partition. Cet intérêt de la partition globalement optimale est illustré par une étude de la réussite des étudiants de première année à la Faculté des sciences économiques de Genève.

L'article est organisé comme suit. La section 2 introduit le cadre formel et les notations. La section 3 précise le concept de partition optimale et définit formellement les critères à optimiser. La section 4 traite de la détermination de l'optimum. On y propose une heuristique dont on analyse empiriquement l'efficacité avec des simulations. La section 5 explicite les limites et l'intérêt du concept de partition optimale qui est illustré à la section 6. Enfin, la section 7 donne quelques pistes de recherche future.

2 Cadre formel et notations

On se place dans le cadre de l'apprentissage supervisé avec une variable réponse Y et p prédicteurs $x_1, \ldots, x_p$ et une base de données d'apprentissage de taille n. On considère plus précisément le cas où la variable à prédire Y et les prédicteurs sont tous catégoriels. On note ℓ le nombre de valeurs distinctes de Y, c_k, le nombre de valeurs distinctes du prédicteur x_k, $k = 1, \ldots, p$, et $c \leq \prod_k c_k$ le nombre de profils admissibles $\mathbf{x} = (x_1, \ldots, x_p)$.[1]

Si l'objectif de l'apprentissage est en règle générale la construction d'un classifieur $f(\mathbf{x})$ qui attribue un et un seul état de la variable Y à chaque profil $\mathbf{x}$, les connaissances recherchées peuvent dans certains cas porter sur les probabilités des divers états de Y conditionnellement aux valeurs $\mathbf{x}$ des prédicteurs. Ceci est en particulier le cas dans les sciences de comportement (sociologie, sciences politiques, marketing, histoire, ...) qui cherche à décrire les mécanismes qui régissent les phénomènes étudiés. Nous nous plaçons dans ce contexte est considérons donc le problème de la prédiction de la distribution de probabilité a posteriori de Y, c'est-à-dire de $\mathbf{p}(Y|\mathbf{x}) = (p_1(\mathbf{x}), \ldots, p_\ell(Y\mathbf{x}))$, où l'on note $p_i(\mathbf{x})$ la probabilité $p(Y = y_i|\mathbf{x})$.

Notons que de nombreuses techniques de classification s'appuient de façon plus ou moins explicite sur la distribution a posteriori $\mathbf{p}(Y|\mathbf{x})$, la classification consistant à attribuer les cas à la catégorie la plus probable $\arg\max_i p_i(\mathbf{x})$ compte tenu de $\mathbf{x}$. C'est le cas notamment de la régression logistique, de l'analyse discriminante linéaire ou quadratique, des k plus proches voisins, des arbres et graphes d'induction ou encore des réseaux bayésiens. Cet aspect est en particulier bien mis en évidence dans Hastie et al. (2001).

Dans le cas de variables catégorielles, les données d'apprentissage peuvent être représentées de façon synthétique sous forme d'une table de contingence $\mathbf{T}$ de taille $\ell \times c$

[1] Le croisement de tous les prédicteurs peut donner lieu à des profils non admissibles d'effectif structurellement nul, par exemple (homme, enceinte). Ceci explique l'inégalité utilisée ici.

croisant la variable réponse y avec les profils $\mathbf{x}$. Par exemple, la table 1 présente un jeu de 240 données utilisées pour étudier la distribution du statut marital selon le sexe et le secteur d'activité.

	homme			femme			
marié	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	total
non	50	40	6	0	14	10	120
oui	5	5	12	50	30	18	120
total	55	45	18	50	44	28	240

TAB. 1 – Exemple de table de contingence $\mathbf{T}$

Un élément de la table $\mathbf{T}$ est noté n_{ij} et représente le nombre de cas avec profil $\mathbf{x}_j$ qui dans les données prennent la valeur y_i de la variable réponse. Les totaux des lignes et des colonnes sont respectivement notés $n_{i.}$ et $n_{.j}$. Les estimations du maximum de vraisemblance $\tilde{p}(Y = y_i|\mathbf{x}_j) = n_{ij}/n_{.j}$, c'est-à-dire les distributions colonnes de la table $\mathbf{T}$ donnent l'information la plus fine sur les distributions conditionnelles de la variable réponse au sein de l'ensemble d'apprentissage. Aucun degré de liberté n'est laissé dans ces estimations qui, étant alors entièrement liées à l'échantillon peuvent être très instables lorsqu'on change d'échantillon d'apprentissage. Pour obtenir des estimations plus stables et donc plus pertinente en généralisation, il convient de gagner des degrés de liberté ce qui peut être fait en regroupant des colonnes ou si l'on veut en partitionnant l'ensemble des profils $\mathbf{x}$ admissibles. Se pose alors naturellement la question de l'optimalité de la partition ainsi que de sa pertinence.

3 Le concept de partition optimale

On se propose ici de préciser les propriétés que l'on attend de la partition cherchée est plus formellement d'introduire les critères qu'il s'agira d'optimiser.

Intuitivement, la partition cherchée doit

1. réduire autant que possible l'incertitude quant à la valeur prise par Y dans chaque classe ;
2. avoir le plus petit nombre de classes pour assurer une meilleure stabilité des estimations en laissant le plus de degrés de liberté possibles.

De plus, en particulier lorsqu'on s'intéresse plus à la compréhension des relations de dépendance qu'à la classification, la partition devrait permettre une caractérisation aussi simple que possible des classes pour en facilité l'interprétation. Nous nous concentrons dans un premier temps sur les points énumérés ci-dessus et reviendrons sur l'aspect interprétation à la section 5.

La réduction de l'incertitude ne peut être jugée que par rapport aux données d'apprentissage. Selon le point 1, on attend donc de la partition qu'elle reproduise le mieux possible les distributions colonne de la table $\mathbf{T}$ observée, ce qui est un problème de qualité d'ajustement. Quant au point 2, il a trait à la complexité de la partition qui doit être le plus simple possible pour assurer la meilleure stabilité des distributions

estimées. Ce dernier objectif s'oppose au premier dans la mesure où tout affinement de la partition ne peut que réduire l'incertitude.

Mesure de la qualité d'ajustement

Examinons tout d'abord le problème de l'ajustement. La qualité d'ajustement des distributions observées (les colonnes de la table $\mathbf{T}$) peut se mesurer selon le principe du tableau étendu défini dans Ritschard et Zighed (2003) pour le cas particulier des partitions produites par les arbres. Ce principe est le suivant.

Soit $\mathbf{T}$ la table cible. Pour une partition des profils (colonnes de $\mathbf{T}$) en $q < c$ classes, on génère la table $\mathbf{T}^a$ de dimension $\ell \times q$ qui croise la variable réponse y avec la partition. L'objectif étant de juger de la qualité de l'ajustement de la table cible par cette table réduite, on transforme cette dernière en une table équivalente de même dimension que $\mathbf{T}$. La transformation consiste à ventiler chaque élément n^a_{ik} de $\mathbf{T}^a$ entre les colonnes qui forment la k-ème classe de la partition, la ventilation se faisant proportionnellement aux fréquences marginales des profils concernés. On obtient ainsi la table prédite $\hat{\mathbf{T}}$ d'élément générique

$$\hat{n}_{ij} \quad = \quad \frac{n_{\cdot j}}{\sum_{j' \in J_k} n_{\cdot j'}} n^a_{ik} \tag{1}$$

où J_k est l'ensemble des indices des colonnes de $\mathbf{T}$ qui sont regroupées dans la même classe k de la partition. Pour l'exemple du tableau 1, la partition

$$\{\ \{\text{(hom,pri),(hom,sec)}\},\ \{\text{(hom,ter),(fem,sec),(fem,ter)}\},\ \{\text{(fem,pri)}\}\ \}$$

conduit ainsi à la table prédite $\hat{\mathbf{T}}$ donnée au tableau 2.

	homme			femme			
marié	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	total
non	49.5	40.5	6	0	14.67	9.33	120
oui	5.5	4.5	12	50	29.33	18.67	120
total	55	45	18	50	44	28	240

Tab. 2 – Table prédite $\hat{\mathbf{T}}$

La qualité de l'ajustement de $\mathbf{T}$ peut être évaluée par des statistiques de divergence du khi-2 Cressie et Read (1984), en particulier par les statistiques X^2 de Pearson ou G^2 du rapport de vraisemblance :

$$G^2 = 2\sum_{i=1}^{\ell}\sum_{j=1}^{c} n_{ij}\ \ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right)\ , \qquad\qquad X^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}\ .$$

Sous l'hypothèse que le modèle est correct, c'est-à-dire que la partition modélise correctement les distributions conditionnelles, et sous certaines conditions de régularité, voir par exemple Bishop et al. (1975, chap. 14), ces statistiques suivent une même distribution du khi-2. Nous avons montré dans Ritschard et Zighed (2003) que les degrés de liberté sont dans ce cas $(c-q)(\ell-1)$.

La divergence entre les tableaux $\hat{\mathbf{T}}$ et $\mathbf{T}$ des tableaux 2 et 1 est par exemple de $G^2 = 0.23$ (on a aussi $X^2 = 0.23$), pour 3 degrés de liberté ce qui donne un degré de signification de plus de 97% et indique un ajustement presque parfait.

On peut songer à d'autres indicateurs pour juger de l'écart entre les deux tableaux, par exemple, une différence simple ou normalisée d'entropie entre les deux tables. L'intérêt des statistiques du khi-2 est qu'elles permettent, lorsque les conditions de régularité sont satisfaites, de tester la significativité statistique de la divergence.

Arbitrage avec la complexité

La partition optimale du seul point de vue de la qualité de l'ajustement est évidemment la partition la plus fine qui prédit exactement la table cible. Comme mentionné plus haut, il convient de tenir également compte de la taille de la partition. Une stratégie peut ainsi consister à chercher la partition de taille minimale qui assure une déviance non statistiquement significative. Le problème avec cette approche est double comme le souligne en particulier Raftery (1995). D'une part, lorsque n devient grand le moindre écart devient statistiquement significatif. D'autre part, de multiples modèles (il faut entendre partitions dans notre cas) ajustent de façon satisfaisante la table cible. On se trouve dès lors confronté à une incertitude quant au bon modèle. Kass et Raftery (1995) ont montré que la minimisation du critère BIC permet dans une approche bayésienne de minimiser l'incertitude liée au modèle pour les données observées. Ce critère initialement introduit par Schwarz (1978), s'exprime dans notre cas comme la combinaison de la déviance G^2 et d'une pénalisation pour la complexité mesurée par $(q\ell - q + c)$ (voir Ritschard et Zighed, 2003). Le critère AIC de Akaike (1983) est une variante qui pénalise la complexité moins fortement et indépendamment de n.

$$\text{AIC} = G^2 + 2(q\ell - q + c) \qquad \text{et} \qquad \text{BIC} = G^2 + (q\ell - q + c)\log(n) \ .$$

La minimisation de l'un ou l'autre de ces critères répond parfaitement à notre objectif d'optimalité de la partition.[2]

4 Détermination de la partition optimale

L'exploration exhaustive de toutes les partitions est de complexité non polynomiale (np-complet). Le nombre $B(c)$ de partitions des c profils est donné par la formule itérative de Bell (1938) $B(c) = \sum_{k=0}^{c-1} \binom{c-1}{k} B(k)$. La figure 1 montre clairement que ce nombre explose totalement au delà d'une dizaine de profils.

Limite de l'approche par les arbres

Parmi les diverses méthodes d'apprentissage, les arbres d'induction ont la particularité de travailler avec un nombre fini de distributions a posteriori $p(Y|\mathbf{x})$ par opposition avec des approches comme la régression logistique ou l'analyse discriminante qui

[2]Dans les situations classiques, le critère AIC est connu pour être biaisé. Par exemple, dans le cadre des modèles linéaires, AIC asymptotiquement sélectionne des modèles plus complexes que le vrai modèle.

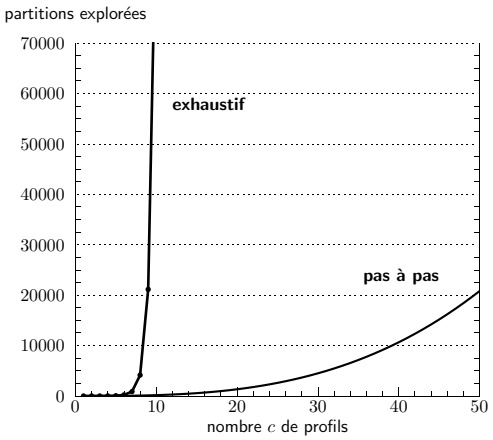

FIG. 1 – Nombre de partitions explorées par la procédure exhaustive et borne supérieure de ce nombre pour l'approche pas à pas

impliquent des fonctions continues de $\mathbf{x}$. Les arbres sont donc bien adaptés à notre contexte. Ils construisent les partitions de façon descendante. En partant du nœud initial constitué de tous les profils, ils procèdent par éclatements successifs des nœuds jusqu'à ce qu'un critère d'arrêt soit atteint. Les éclatements successifs, c'est-à-dire pour chaque nœud le choix d'un prédicteur et le partitionnement du nœud selon les modalités de ce prédicteur, se font par optimisation d'un critère local, par exemple la significativité d'un khi-2 dans CHAID Kass (1980) ou le ratio de gain dans C4.5 Quinlan (1993).

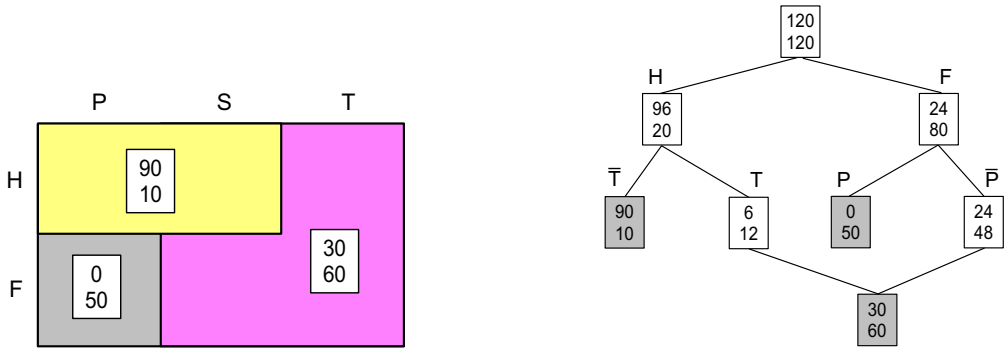

FIG. 2 – Prédiction de la distribution du statut marital : une partition de l'espace des prédicteurs non réalisable avec un arbre

Dans l'optique de déterminer la partition globalement optimale, on peut songer à remplacer le critère local par un critère global du type AIC ou BIC. Les arbres présentent cependant l'inconvénient de ne pas pouvoir générer toutes les partitions et donc de rater éventuellement la partition globalement optimale. La figure 2 par exemple, illustre une partition qui ne peut pas être obtenue avec un arbre, tandis que la figure 3

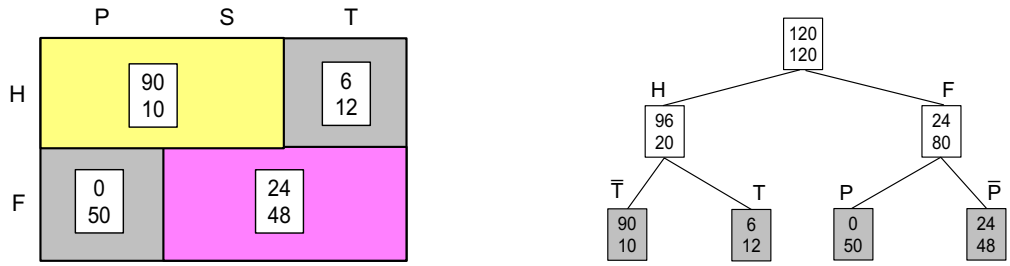

FIG. 3 – Prédiction de la distribution du statut marital : une partition de l'espace des prédicteurs réalisable avec un arbre

caractérise le type de partition produite par un arbre. On notera que la partition de la figure 2 est précisément celle qui donne lieu à la table prédite $\hat{\mathbf{T}}$ du tableau 2. Les valeurs de AIC et BIC des deux partitions illustrées sont respectivement de 18.2 et 49.6 pour la partition de la figure 2 contre 20.2 et 55 pour celle de la figure 3. On a donc ici un exemple de situation où l'on ne peut pas atteindre la partition optimale avec un arbre.

Procédure pas à pas arrière

Une solution alternative est de procéder par regroupements successifs deux à deux des classes en partant de la partition la plus fine jusqu'à ce que le critère AIC ou BIC ne puisse plus être réduit. Il s'agit d'une version simplifiée de l'heuristique proposée dans Ritschard et al. (2001) et étudiée dans Ritschard (2001) où l'on explorait simultanément les regroupements en lignes et en colonnes. Les regroupements se font ici uniquement sur les colonnes (profils) de la table cible $\mathbf{T}$. En effet, des regroupements sur la variable à prédire modifieraient la nature des distributions à prédire et rendrait caduque la comparaison des AIC ou BIC.

Performance de l'heuristique

La borne supérieure du nombre de cas explorés par l'heuristique est $1+\sum_{i=2}^{c}\binom{i}{2} = 1+\frac{c(c^2-1)}{6}$ et sa complexité est donc polynomiale en $O(c^3)$. C'est évidemment plus complexe que les arbres, mais suffisant tant que le nombre de partitions admissibles c n'excède pas une centaine, contre 8 ou 9 pour la procédure exhaustive. On peut voir l'évolution de cette borne supérieure avec c dans la figure 1.

Pour juger de la capacité de l'heuristique à trouver la partition globalement optimale, nous avons procédé par simulation. Une séries de 200 tables de contingence de $n = 10'000$ cas ont été générées aléatoirement. On a généré des tables de $\ell = 4$ lignes et $c = 7$ colonnes, le nombre 7 de colonnes étant le plus grand qui permette d'appliquer la procédure exhaustive sur les 200 tables en un temps raisonnable. Pour chaque table nous avons comparé la valeur du critère (AIC ou BIC) quasi-optimale trouvée par l'heuristique à la valeur optimale déterminée avec la procédure exhaustive.

En allouant les 10′000 cas entre les $4 \cdot 7 = 28$ cases du tableau selon une distribution uniforme, on génère des tables caractérisées par une absence totale de structure pour lesquelles tout regroupement de profils donne lieu à un déficit d'ajustement trop important pour pouvoir être compensé par la réduction de taille de la partition. De façon donc non surprenante, l'heuristique a trouvé les BIC et AIC optimaux dans les 200 cas, la partition optimale étant dans presque tous les cas la partition la plus fine, la valeur moyenne des 200 critères optimaux étant de 254.8 contre 257.8 pour la partition la plus fine.

Pour introduire un peu de structure, nous avons procédé à une seconde séries de simulation en générant les tables selon des distributions uniformes conditionnelles emboîtées : un pourcentage aléatoire entre 0 et 100% des cas est alloué à la première ligne, puis un pourcentage aléatoire des cas restant à la seconde ligne et ainsi de suite jusqu'à la dernière ligne, le total de chaque ligne étant ensuite réparti de la même façon dans la ligne. On génère ainsi des tables où les gros effectifs ont tendance à se concentrer en haut et à gauche.

		AIC	BIC
Ecarts non nuls seulement :			
	proportion	6%	12.5%
	maximum	1.55	8.79
	moyenne	0.62	1.62
	écart type	0.44	1.73
	asymétrie	0.48	0.979
Ensemble des écarts nuls et non nuls :			
	moyenne	0.04	0.20
	écart type	0.18	0.95
	asymétrie	5.72	6.21
Ecarts relatifs :	maximum	0.035	0.048
	moyenne	0.013	0.009
Valeur initiale du critère		56	257.9
Moyenne des optima globaux		50.98	213.5

Tab. 3 – Simulations : écarts entre optima et quasi-optima du AIC et du BIC

Le tableau 3 résume les résultats de ces simulations. On constate que si la proportion d'optima manqués est significative avec respectivement 6% et 12.5%, les écarts entre la solution quasi-optimale de l'heuristique et l'optimum global reste faible dans tous les cas avec un écart relatif maximal de 1.3% pour l'AIC et de 4.8% pour le BIC. De même, les écarts moyens restent faibles en regard de la réduction moyenne (respectivement 5.02 et 44.4) de la valeur initiale du critère. Nous avons obtenus des résultats très similaires, en fait même légèrement moins bons, avec des tables 4×6. Ceci est plutôt encourageant, dans la mesure où cela indique que ces performances ne semblent pas devoir se détériorer lorsque c augmente.

5 Portée et limites du concept de partition optimale

On a examiné jusqu'ici comment définir formellement la partition optimale et les possibilités et difficultés que soulèvent sa détermination. Nous nous proposons maintenant de revenir sur l'aspect interprétation en discutant brièvement les enseignements que nous apporte la connaissance de cet optimum global.

En premier lieu, il importe de préciser que dans notre approche la complexité est mesurée en termes de taille de la partition. Ceci n'assure pas la simplicité de description des classes comme l'illustrent notamment les exemples des figures 2 et 3, la partition en quatre de la seconde figure étant plus facile à décrire que celle en trois de la première. La difficulté dans ce dernier cas tient au fait que l'une des classes est définie par une alternative de conditions et non uniquement en termes de conjonctions de conditions comme dans le cas des arbres.

Si cette difficulté de décrire simplement la partition optimale limite la portée du concept de partition optimale, la connaissance de la valeur optimale du critère (AIC ou BIC) fournit par contre une indication précieuse pour juger de la qualité d'une solution produite par un arbre d'induction. En effet, une solution proche de la valeur optimale nous indiquera qu'on a peu de chances de pouvoir améliorer les choses en jouant sur les paramètres de contrôle, tandis qu'un grand écart par rapport à la valeur optimale pourra justifier des efforts dans ce sens. Dans cette optique nous suggérons de calculer les deux valeurs optimales AIC et BIC ainsi que la taille des partitions correspondantes.

6 Illustration

Afin d'illustrer les enseignements apportés par les AIC et BIC optimaux, nous donnons quelques résultats obtenus avec des données relatives aux 762 étudiants qui ont commencé leur première année d'études à la Faculté des sciences économiques et sociales de Genève en 1998. Il s'agit de données administratives réunies par Petroff et al. (2001). L'objectif était d'évaluer les chances de respectivement réussir, redoubler ou être éliminé à la fin de la première année d'études. La figure 4 montre l'arbre obtenu avec la procédure CHAID (Kass, 1980) implémentée dans Answer Tree (SPSS, 2001). Parmi une trentaine de prédicteurs potentiels, CHAID en a sélectionné 5 dont deux quantitatifs, l'année d'immatriculation à l'université et l'âge à l'obtention du diplôme de l'école secondaire. Les cinq variables avec les discrétisations et regroupements de modalités proposés par CHAID sont le type de diplôme secondaire (3 modalités), l'âge de son obtention (4), la date d'immatriculation (2), le tronc commun choisi (2) et la nationalité (2). La table cible définie par ces variables contient 88 colonnes. Elle a 3 lignes correspondant aux 3 situations possibles de l'étudiant après un an d'étude.

La valeur du AIC et du BIC pour la partition définie par l'arbre est donnée dans le tableau 4. On donne également la taille de la partition, la déviance G^2 avec ses degrés de liberté et son degré de signification, ainsi que le pseudo R^2 ajusté. Ces valeurs peuvent être comparées à celles de deux variantes, CHAID2 qui est CHAID sans l'éclatement du sommet 4 ($nationa \notin \{$GE, hors Europe$\}$) et CHAID3 sans l'éclatement des sommets 4 et 5 ($nationa \in \{$GE, hors Europe$\}$). Le modèle saturé correspond à la partition la plus fine et le modèle d'indépendance au cas où tous les profils sont regroupés en seul groupe.

Patition BIC optimale

Modèle	q	d	G^2	sig(G^2)	pseudo R^2_{ajust}	AIC	BIC
Saturé	88	0	0	1	1	528	1751.9
Meilleur AIC	14	148	17.4	1	.941	249.4	787.2
CHAID	9	158	177.9	0.133	.336	390.0	881.3
CHAID2	8	160	187.4	0.068	.309	395.4	877.5
CHAID3	7	162	195.2	0.038	.289	399.2	872.1
Meilleur BIC	6	164	75.2	1	.745	275.2	738.8
Indépendance	1	174	295.1	0.000	0	475.8	892.3

TAB. 4 – SES 98 : qualités d'ajustement d'un choix de modèles

On constate tout d'abord que CHAID et ses deux variantes ont des AIC ou BIC très voisins, sensiblement meilleurs que ceux du modèle saturé et dans une moindre mesure que ceux du modèle d'indépendance. A priori il est difficile de dire si pour améliorer la partition il vaut mieux l'affiner ou au contraire réduire sa taille. C'est ici que les valeurs des AIC et BIC optimaux s'avèrent utiles. Elles indiquent par exemple que les deux options peuvent être envisagées. Le meilleur AIC correspond à une partition plus fine de 14 classes qui montre notamment un potentiel important d'amélioration de la qualité d'ajustement. Le meilleur BIC indique que l'on peut atteindre une qualité comparable d'ajustement avec une partition plus sommaire de 6 classes seulement. On a dans tous les cas une indication forte qu'il est possible d'améliorer sensiblement la qualité de la partition produite par CHAID. Cet exemple semble aussi confirmer une

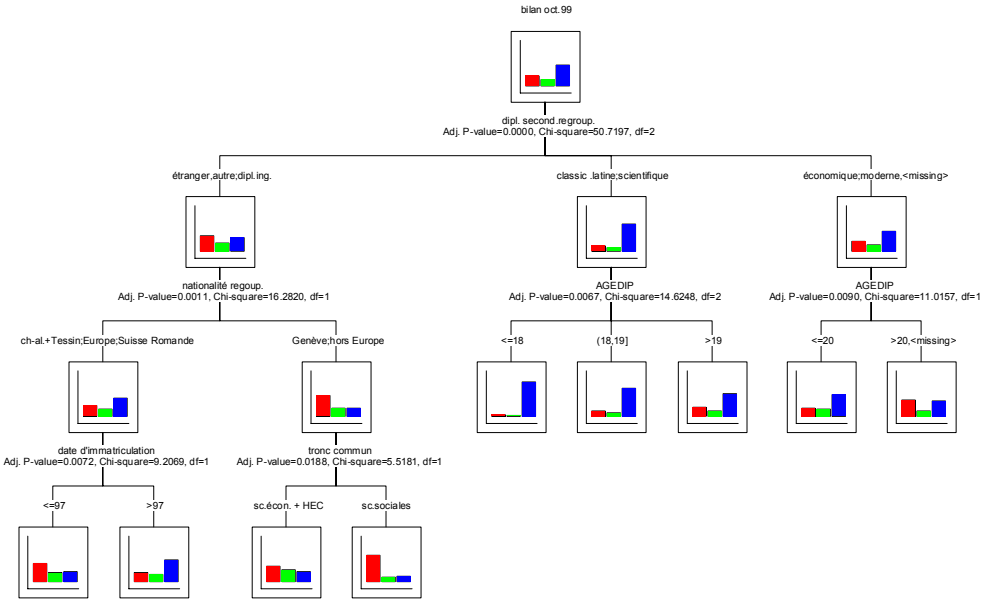

FIG. 4 – Bilan après une année en SES : arbre CHAID

tendance du AIC à sélectionner un modèle trop complexe.

7 Conclusion

Cet article présente une première approche assez rustre pour déterminer la partition optimale dans une optique de prédiction de la distribution a posteriori de la variable réponse. L'heuristique proposée doit certainement pouvoir être améliorée et d'autres approches du type descendant notamment mériteraient également d'être explorées. Il convient ici de mentionner en particulier les approches qui à l'instar de Sipina (Zighed et Rakotomalala, 2000) génèrent des graphes d'induction plutôt que des arbres en autorisant des fusions comme celles illustrées à la figure 2. On peut s'attendre à ce que le couplage de la stratégie graphe d'induction avec des critères du type AIC ou BIC donnent des résultats performants tant du point de vue du temps de calcul que des possibilités d'approcher la solution optimale.

Références

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institue 50*, 277–290.

Bell, E. T. (1938). The iterated exponential numbers. *Ann. Math. 39*, 539–557.

Bishop, Y. M. M., S. E. Fienberg, et P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge MA : MIT Press.

Cressie, N. et T. R. Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society 46*, 440–464.

Hastie, T., R. Tibshirani, et J. Friedman (2001). *The Elements of Statistical Learning*. New York : Springer.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Kass, R. E. et A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Petroff, C., A.-M. Bettex, et A. Korffy (2001, Juin). Itinéraires d'étudiants à la faculté des sciences économiques et sociales : le premier cycle. Technical report, Université de Genève, Faculté SES.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo : Morgan Kaufmann.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC : The American Sociological Association.

Ritschard, G. (2001). Performance d'une heuristique d'agrégation optimale bidimensionnelle. *Extraction des connaissances et apprentissage 1*(4), 185–196.

Ritschard, G. et D. A. Zighed (2002). Qualité d'ajustement d'arbres d'induction. Technical report, Groupe Gafo Qualité, CNRS, Paris. 16p.

Ritschard, G. et D. A. Zighed (2003). Modélisation de tables de contingences par arbres d'induction. *Revue des sciences et technologies de l'information — ECA 17*(1–3), 381–392.

Ritschard, G., D. A. Zighed, et N. Nicoloyannis (2001). Maximisation de l'association par regroupement de lignes ou colonnes d'un tableau croisé. *Revue Mathématiques Sciences Humaines 39*(154/155), 81–97.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago : SPSS Inc.

Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'induction : apprentissage et data mining*. Paris : Hermes Science Publications.

Summary

This paper is concerned with the partitioning of the predictor space best suited to generate reliable estimates of class posteriors. This optimal partition has to face a double goodness-of-fit and simplicity criteria. We thus focus on the Akaike (AIC) and Bayesian (BIC) information criteria that specifically take these two aspects into account. We show how they apply in our framework and then investigate how to reach their optimal value. The paper provides a crude heuristic and studies its efficiency by comparing the quasi optimum with the true optimum on a series of simulated tables. The optimum provides useful insight on how much a partition, for instance the partition provided by an induction tree algorithm, could be improved. This point is illustrated with a real dataset.