Modélisation de tables de contingence par arbres d'induction

Gilbert Ritschard[*] — Djamel A. Zighed[**]

[*] *Département d'économétrie, Université de Genève*
bd du Pont-d'Arve 40, CH-1211 Genève 4
gilbert.ritschard@themes.unige.ch

[**] *Laboratoire ERIC, Université Lyon 2*
Bat. L, C.P. 11, F-69676 Bron Cédex
zighed@univ-lyon2.fr

RÉSUMÉ. *Cet article est consacré à l'évaluation statistique des descriptions de tables de contingence fournies par les arbres d'induction. On se limite au cas particulier de données catégorielles. Trois aspects sont successivement abordés. i) La nature de l'ajustement en apprentissage supervisé, où l'on souligne la distinction entre prédiction de valeurs individuelles et prédiction de leur représentation sous forme de table de contingence. ii) La description de tables fournies par les arbres d'induction que l'on compare notamment à la modélisation log-linéaire utilisée en statistique. iii) L'adaptation au cas des arbres d'induction des mesures et statistiques de qualité d'ajustement utilisées en modélisation log-linéaire. La discussion est complétée par une illustration sur les données du Titanic.*

ABSTRACT. *The paper is concerned with the statistical assessment of the description of contingency tables by induction trees. It focuses on the special case of categorical data. Three topics are successively considered. i) The nature of the fit in supervised learning where we stress the distinction between fitting individual values and fitting their cross-tabulated synthetic representation. ii) The description of contingency tables provided by induction trees which is compared with the log-linear modeling used in statistics. iii) The adaptation of the goodness-of-fit measures and statistics used in log-linear modeling to the case of induction trees. The discussion is completed with an application to the Titanic data set.*

MOTS-CLÉS : *Arbre d'induction, table de contingence, modélisation et tests statistiques, qualité d'ajustement, comparaison de modèles.*

KEYWORDS: *Induction trees, contingency tables, statistical modeling and tests, goodness of fit, models comparison.*

1. Introduction

En apprentissage supervisé, le concept d'ajustement prend généralement une forme particulière. Il consiste à chercher une fonction de prédiction qui, au moyen des attributs prédictifs, permet d'ajuster au mieux l'attribut à prédire. Dans ce cadre, la qualité de l'ajustement est mesurée par le taux de bien classés sur les données d'un échantillon test. Dans certains domaines d'application, comme les sciences humaines et sociales, on s'intéresse plus à savoir comment les prédicteurs influencent la variable réponse qu'aux prédictions individuelles. Ceci conduit à s'intéresser à la distribution de la variable réponse pour les différentes combinaisons de valeur des prédicteurs, c'est-à-dire à la représentation sous forme de table de contingence des données. Dans ce papier nous mettons en exergue ces différentes notions d'ajustement et nous montrons, dans le cas des tables de contingence, comment les arbres d'induction constituent une alternative intéressante aux modèles log-linéaires utilisés en modélisation statistique. Nous proposons différents critères pour apprécier alors la qualité de la description de la table induite par un arbre d'induction

Le papier est organisé comme suit. La section 2 discute de la nature de l'ajustement en apprentissage supervisé en introduisant une distinction entre ajustement des valeurs individuelles et ajustement de leur présentation sous forme de table de contingence. Après avoir souligné l'intérêt de la présentation sous forme de tableau croisé, on aborde à la section 3 le problème de la modélisation de la table, c'est-à-dire de la recherche d'un modèle aussi simple que possible permettant de reconstruire la table de façon satisfaisante. On présente d'abord la modélisation log-linéaire qui est une approche solidement établie statistiquement. On rappelle ensuite la procédure de construction des arbres et explicitons le modèle de reconstruction de la table de contingence qu'ils fournissent. La section 4 montre comment les mesures de qualité d'ajustement utilisées en modélisation log-linéaire peuvent être adaptées pour juger de la description d'une table de contingence déduite d'un arbre induit. A la section 5, nous illustrons la portée des mesures d'ajustement introduites sur un exemple simple.

2. Le concept de l'ajustement en apprentissage supervisé

2.1. *Cadre conceptuel et notations*

On se place dans le cadre de l'apprentissage supervisé consistant à construire une fonction f qui permet de prédire au mieux l'état d'un attribut particulier y au moyen d'un vecteur $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_p)$ de p attributs prédictifs. L'apprentissage se fait sur un échantillon de n individus pour lesquels on connaît $(\mathbf{x}_\alpha, y_\alpha)$, $\alpha = 1, \ldots, n$. Parmi les algorithmes d'apprentissage, on s'intéresse plus particulièrement aux arbres d'induction dits aussi de décision.

L'objectif est, en apprentissage, d'utiliser ensuite la fonction de prédiction $f(\mathbf{x})$ afin de prédire les valeurs particulières de y pour des individus dont nous ne connais-

sons que les valeurs des attributs prédictifs $\mathbf{x}$. La fiabilité de la prédiction est, dans une large mesure, liée à la qualité de l'ajustement de y par $f(\mathbf{x})$.

De façon générale, la qualité d'ajustement d'un modèle se réfère à sa capacité à reproduire les données. Pour la prédiction d'une variable quantitative y, par exemple dans le cas de la régression linéaire, l'objectif est clair. Il s'agit d'obtenir des valeurs prédites $\hat{y}_\alpha$ qui s'ajustent le mieux possible aux valeurs observées y_α, pour $\alpha = 1, \ldots, n$. Dans les tables de contingence qui nous occupent, les données sont groupées. Les tables en donnent une présentation synthétique. On peut alors chercher à prédire la présentation synthétique, c'est-à-dire la distribution des cas.

2.2. *Données d'illustration*

Afin d'illustrer les principaux concepts que nous manipulons, nous utilisons le jeu de 18 données fictif du tableau 1. On s'intéresse à prédire l'***activité*** (*salarié, formation, chômeur*) à l'aide des trois attributs prédictifs dichotomisés ***sexe***, ***âge*** et ***statut***.

i	***sexe***	***âge***	***statut***	***activité***
1	homme	jeune	seul	salarié
2	homme	jeune	couple	formation
3	homme	jeune	couple	formation
4	homme	adulte	seul	salarié
5	homme	adulte	seul	chômage
6	homme	adulte	seul	chômage
7	homme	adulte	seul	chômage
8	homme	adulte	couple	salarié
9	homme	adulte	couple	salarié
10	homme	adulte	couple	chômage
11	homme	adulte	couple	chômage
12	homme	adulte	couple	chômage
13	homme	adulte	couple	chômage
14	femme	jeune	seul	salarié
15	femme	jeune	seul	chômage
16	femme	jeune	couple	formation
17	femme	adulte	seul	salarié
18	femme	adulte	seul	salarié

Tableau 1. *Données illustratives*

2.3. *Classification et ajustement de tables de contingence*

La classification avec des variables catégorielles nominales conduit à exploiter l'information sur les modalités des variables prédictives $x_1, \ldots, x_p$ pour prédire la catégorie de la variable réponse y. Les données peuvent dans ce cas être représentées synthétiquement sous la forme d'une table de contingence à $p + 1$ dimensions croisant toutes les variables prédictives et à prédire. De façon équivalente, ce tableau multidimensionnel peut être représenté par une table de contingence $\mathbf{T}$ à deux dimensions croisant la variable à prédire y avec la variable composite vectorielle résultant

	homme				femme			
	jeune		adulte		jeune		adulte	
Activité	seul	couple	seul	couple	seul	couple	seul	couple
salarié	1	0	1	2	1	0	2	0
formation	0	2	0	0	0	1	0	0
chômeur	0	0	3	4	1	0	0	0

Tableau 2. *Table de contingence* $\mathbf{T}$ *des données illustratives*

du croisement de tous les prédicteurs. La variable y a ℓ modalités et la variable prédictive composite $c = \prod_{j=1}^{p} c_j$, où c_j est le nombre de modalités de l'attribut x_j, $j = 1, \ldots, p$. La table $\mathbf{T}$ est donc de taille (ℓ, c). Par exemple, les données du tableau 1 sont représentées par la table de contingence $\mathbf{T}$ du tableau 2 avec $\ell = 3$ lignes et $c = 2 \cdot 2 \cdot 2 = 8$ colonnes.

L'objectif de la classification est de prédire pour chaque cas la classe d'appartenance y (la ligne du tableau) compte tenu de l'information sur son profil $\mathbf{x}$, c'est-à-dire connaissant la colonne où il se trouve. Plusieurs techniques de classification dont la régression logistique et les arbres d'induction procèdent en deux étapes : 1) Modéliser la distribution de probabilité $\mathbf{p} = \big(p(Y = y_1,) \ldots, p(Y = y_\ell)\big)$ de la variable à prédire en fonction du profil $\mathbf{x}$, c'est-à-dire trouver une fonction vectorielle $\mathbf{p}(\mathbf{x})$ pour prédire $\mathbf{p}$. 2) Classer selon la règle majoritaire : $\hat{y} = f(\mathbf{x}) = g(\mathbf{p}(\mathbf{x})) = \arg\max_i \hat{p}_i(\mathbf{x})$.

Ainsi, le modèle de classification $f(\mathbf{x}) = g(\mathbf{p}(\mathbf{x}))$ repose sur le modèle descriptif de la distribution $\mathbf{p}(\mathbf{x})$. Ce dernier modèle vise à décrire comment les prédicteurs influencent la distribution de Y. Il fournit en cela des connaissances précieuses, en particulier en sciences sociales, où l'objectif est souvent la compréhension de phénomènes plutôt que la prédiction et la classification.

Pour une taille d'échantillon n donnée et une répartition $n_{\cdot 1}, \ldots, n_{\cdot c}$ fixée entre colonnes, c'est-à-dire entre valeurs du vecteur $\mathbf{x}$ des attributs prédictifs, le modèle descriptif $\mathbf{p}(\mathbf{x})$ est équivalent à un modèle de prédiction de la table de contingence $\mathbf{T}$. En effet, chacune des c colonnes de la table correspond à un vecteur $\mathbf{x}_j$, $j = 1, \ldots, c$ différent. Ainsi, $n_{\cdot j}\mathbf{p}(\mathbf{x}_j)$ donne une prédiction de la j-ème colonne de $\mathbf{T}$. Dès lors, il est légitime de s'intéresser à l'ajustement de la table $\mathbf{T}$ par le modèle descriptif. En notant $\hat{n}_{ij}$ les effectifs de la table induite $\hat{\mathbf{T}}$, il s'agit d'évaluer globalement les écarts entre ces effectifs et les effectifs observés n_{ij} du tableau $\mathbf{T}$.

3. Méthodes d'ajustement de tables de contingence

3.1. *Le modèle log-linéaire*

Pour simplifier la présentation, nous considérons des tableaux à trois dimensions croisant trois variables $x_1 = A$, $x_2 = B$ et $x_3 = C$. La généralisation à plusieurs dimensions est immédiate. Soit n_{ijk} les effectifs observés de la table. La modélisa-

tion log-linéaire vise à exprimer le logarithme de ces valeurs par une somme d'effets propres et d'interactions d'ordre 2 ou supérieur des variables. Le modèle saturé qui reproduit parfaitement les effectifs observés s'écrit par exemple

$$\log \hat{n}_{ijk} \quad = \quad \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

où λ_i^A représente l'effet propre de A, λ_{ij}^{AB} l'interaction d'ordre 2 entre A et B et λ_{ijk}^{ABC} l'interaction d'ordre 3. On impose aux paramètres des contraintes de normalisation pour que le nombre de paramètres indépendants du modèle saturé n'excèdent pas le nombre de cellules du tableau. L'objectif de la modélisation est de déterminer par suppression d'interactions, ou par d'autres contraintes sur les paramètres, le modèle le plus parcimonieux qui reproduit de façon satisfaisante le tableau observé $\{n_{ijk}\}$.

Notons que cette approche ne fait pas de distinction entre variables à prédire et prédicteurs. On peut néanmoins tenir compte de cette distinction a posteriori en ne s'intéressant qu'aux interactions entre variables à prédire et prédicteurs.

Les paramètres d'un modèle log-linéaire peuvent être estimés par le maximum de vraisemblance. Les estimations s'obtiennent en résolvant le système d'équations définissant les conditions du premier ordre (voir [AGR 90] p. 187) avec un algorithme du type Newton-Raphson. On obtient les mêmes estimateurs en postulant un processus multinomial (n fixé a priori) ou un processus de Poisson (n aléatoire).

La sélection du modèle s'appuie sur la statistique du rapport de vraisemblance ou déviance G^2 qui dans le cas de 3 variables s'écrit : $G^2 = 2\sum_{i,j,k} n_{ijk} \ln(n_{ijk}/\hat{n}_{ijk})$. Elle consiste en une procédure pas à pas qui, dans le cas d'une procédure *backward* par exemple, élimine successivement les interactions qui impliquent l'accroissement le moins significatif du G^2. La procédure s'arrête lorsqu'il ne reste que des interactions dont la suppression entraîne une variation statistiquement significative du G^2.

3.2. *Les Arbres d'induction*

3.2.1. *Objectifs et principes*

Les arbres d'induction sont, parmi les techniques d'apprentissage, les plus utilisées. Ce succès est essentiellement dû à leur simplicité dans la mise en œuvre comme dans l'interprétation des résultats. Le principe de leur construction est fort simple. Au moyen des attributs prédictifs $x_1, \ldots, x_p$, ils construisent une succession de partitions sur l'ensemble d'apprentissage.

Le passage d'une partition à la suivante se fait en optimisant un critère d'évaluation. Il s'agit de comparer la valeur de ce critère entre la partition courante et la nouvelle. Si la nouvelle est meilleure, elle est alors conservée et le processus est réitéré à partir de la nouvelle partition. Dans le cas des arbres d'induction, comme la méthode CART [BRE 84] ou C4.5 [QUI 93], les partitions sont emboîtées et de plus en plus fines.

	homme				femme	
	jeune		adulte			
Activité	seul	couple	seul	couple		Totaux
salarié	1	0	1	2	3	7
formation	0	2	0	0	1	3
chômeur	0	0	3	4	1	8
Totaux	1	2	4	6	5	18

Tableau 3. *Table de contingence* $\mathbf{T}^a$ *associée à la partition finale*

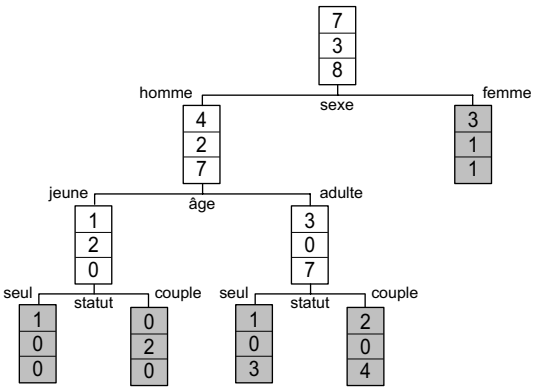

Figure 1. *Arbre induit pour les données illustratives*

A chaque partition, hormis la partition grossière qui se situe à la racine de l'arbre, est associé un tableau de contingence $\mathbf{T}^a$ de taille (ℓ, q) dont les lignes sont les différentes valeurs de y et les colonnes, les q différents sommets terminaux de l'arbre qui forment la partition. Sur l'exemple de la figure 1, la partition engendrée par les attributs ***sexe***, ***âge*** et ***statut*** fournit le tableau de contingence 3.

Le terme générique du tableau $\mathbf{T}^a$ de taille (ℓ, q) sera noté n_{ik} où i fait référence à la i-ème valeur de y et k au k-ème sommet de l'arbre. Comme il s'agit de partitions construites sur les mêmes données, nous aurons toujours, quelle que soit la partition courante, $n_{i\cdot} = \sum_k n_{ik}$ et $n = \sum_{i,k} n_{ik}$.

Pour évaluer le passage d'une partition S_i à la suivante S_{i+1}, on utilise généralement des mesures de gain informationnel comme l'indice de Gini, l'entropie de Shannon ou le ratio du gain, pour ne citer que ceux là. Par souci de simplification des notations, on notera $G(S_{i+1})$ le gain informationnel en passant de la partition S_i à la partition S_{i+1} ou simplement $G(S)$ s'il n'y a pas d'ambiguïté. A titre illustratif, on donne ci-dessous l'expression du ratio du gain. Pour d'autres critères voir par exemple [ZIG 00].

$$G_R(S) = \frac{-\sum_{i=1}^{\ell} \frac{n_{i\cdot}}{n} \log_2 \frac{n_{i\cdot}}{n} + \sum_{j \in S} \frac{n_{\cdot j}}{n} \sum_{i=1}^{\ell} \frac{n_{ij}}{n_{\cdot j}} \log_2 \frac{n_{ij}}{n_{\cdot j}}}{-\sum_{j \in S} \frac{n_{\cdot j}}{n} \log_2 \frac{n_{\cdot j}}{n}} \quad .$$

Une grande partie des mesures utilisées dans la construction d'un arbre de décision possèdent une propriété commune dite propriété de conservation des flux. Elle signifie que $G(S)$ ne croît jamais quelle que soit la partition engendrée : $G(S) \geq 0$ pour tout S. Autrement dit, $G(S)$ est une fonction non croissante par raffinement d'un arbre.

Dans la construction d'un arbre d'induction, on s'arrête généralement avant d'avoir atteint la partition la plus fine. Cet arrêt, conduit implicitement à admettre que le résultat de l'heuristique est quasi optimal, c'est-à-dire que l'information que nous fournirait tout sur-arbre serait négligeable. C'est cette notion de quasi-équivalence entre la partition courante et toutes les partitions plus fines qui en sont issues qui justifie notre approche de l'ajustement d'une table de contingence par un arbre d'induction. La question est alors de savoir si le tableau $\mathbf{T}^a$ associé à un arbre induit constitue un bon ajustement du tableau $\mathbf{T}$ issu du croisement de tous les attributs qui interviennent dans la construction. Sur l'exemple de la figure 1, cela revient à s'interroger sur la qualité de l'ajustement de la table de contingence 2, par la table 3. Le tableau 2 correspond à la partition la plus fine engendrée par les attributs qui apparaissent dans l'arbre qui a conduit au tableau 3. Du point de vue du gain informationnel la réponse est oui. Ce résultat devrait donc se confirmer par un test statistique d'ajustement de table de contingence.

3.2.2. *Extension d'un arbre et arbre saturé*

Pour mesurer la qualité de l'ajustement du tableau $\mathbf{T}$ de taille (ℓ, c), le tableau 2 de notre exemple, par le tableau $\mathbf{T}^a$ de taille (ℓ, q), avec $q \leq c$, découlant de l'arbre, le tableau 3 de notre exemple, on se heurte évidemment au problème du nombre différent de colonnes des deux tableaux. Nous proposons alors de transformer le tableau $\mathbf{T}^a$ défini par l'arbre en une forme $\hat{\mathbf{T}}$ étendue équivalente qui possède le même nombre de colonnes que $\mathbf{T}$. Il s'agit du tableau $\hat{\mathbf{T}}$ associé à l'extension maximale de l'arbre induit :

Définition 1 (Extension maximale de l'arbre induit) *Pour des variables prédictives catégorielles, on appelle* extension maximale de l'arbre induit *ou* arbre induit étendu, *l'arbre qui résulte de tous les éclatements successifs possibles de ses sommets terminaux au moyen des attributs retenus. On applique aux feuilles de l'extension la distribution $\mathbf{p}^a_{|k}$ du nœud terminal parent de l'arbre initial. On note $\hat{\mathbf{p}}_{|j}$ les distributions conditionnelles des feuilles de l'extension maximale de l'arbre.*

Par exemple, la figure 2, illustre l'extension maximale de l'arbre induit. L'effectif des sommets ajoutés est ventilé selon la distribution du sommet terminal de l'arbre induit dont ils sont issus. La table $\hat{\mathbf{T}}$ associée est donnée au tableau 4.

Il est évident que d'un point de vue informationnel, le tableau induit et le tableau de son extension ont exactement la même valeur. On propose alors d'évaluer l'ajustement d'un arbre par une mesure de divergence entre le tableau $\hat{\mathbf{T}}$ de taille (ℓ, c), généré par l'extension maximale de l'arbre, et le tableau $\mathbf{T}$.

	homme				femme			
	jeune		adulte		jeune		adulte	
Activité	seul	couple	seul	couple	seul	couple	seul	couple
salarié	1	0	1	2	1.2	0.6	1.2	0
formation	0	2	0	0	0.4	0.2	0.4	0
chômeur	0	0	3	4	0.4	0.2	0.4	0

Tableau 4. *Table* $\hat{\mathbf{T}}$ *associée à l'extension maximale de l'arbre induit*

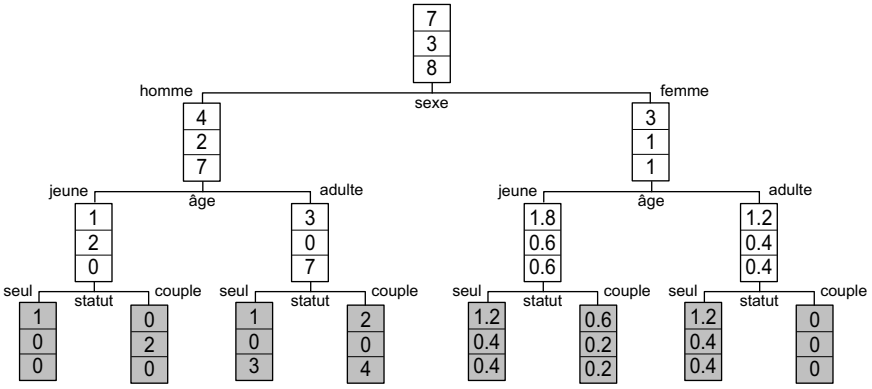

Figure 2. *Extension maximale de l'arbre induit*

Par analogie avec la modélisation log-linéaire où le modèle saturé reproduit exactement la table $\mathbf{T}$, on définit la notion d'arbre saturé :

Définition 2 (Arbre saturé) *Pour des variables prédictives catégorielles, on appelle* arbre saturé, *un arbre qui résulte de tous les éclatements successifs possibles selon les modalités des variables prédictives.*

4. Qualité d'ajustement des arbres d'induction

Les critères usuels pour juger de la qualité de l'ajustement d'une table $\mathbf{T}$ par la prédiction $\hat{\mathbf{T}}$ sont les statistiques de divergence du khi-2, tels que le G^2 du rapport de vraisemblance, appelé également déviance, et le X^2 de Pearson :

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

Sous l'hypothèse que le modèle est correct et sous certaines conditions de régularité, voir par exemple [BIS 75] chap. 14, ces statistiques suivent une même distribution du khi-2 avec pour degrés de liberté le nombre de cellules de la table de contingence moins le nombre de paramètres indépendants du modèle de prédiction des n_{ij}.

Ces statistiques du khi-2 permettent de tester la significativité de la divergence. On en déduit également des indicateurs normalisés de qualité d'ajustement, en particulier le pseudo R^2 qui mesure la proportion de réduction du G^2 que permet le modèle par rapport au modèle d'indépendance où l'on ne tient pas compte de l'information donnée par les prédicteurs, soit $R^2 = \big(G^2(I) - G^2(M)\big)/G^2(I) = 1 - G^2(M)/G^2(I)$, où $G^2(I)$ est le G^2 du modèle d'indépendance et $G^2(M)$ celui du modèle ajusté. On préfère souvent à cet indicateur sa forme ajustée pour les degrés de liberté

$$R^2_{\text{ajust}} = 1 - \frac{G^2(M)/d_M}{G^2(I)/d_I}$$

où d_I et d_M sont respectivement les degrés de libertés du modèle d'indépendance et du modèle ajusté. Pour la comparaison de modèles de complexité différente on recourt également aux critères d'information AIC d'Akaike ou au critère bayésien BIC qui, pour rendre compte de l'incertitude liée au choix du modèle, pénalisent le G^2 pour la complexité mesurée en terme de nombre de paramètres indépendants.

4.1. *Paramètres du modèle et degrés de liberté*

Afin de déterminer les degrés de liberté des statistiques du khi-2, on doit tout d'abord préciser les paramètres du modèle. Formellement, le modèle de reconstruction de la table $\mathbf{T}$ s'écrit en notant $\mathbf{T}_j$ la j-ème colonne de $\mathbf{T}$:

$$\hat{\mathbf{T}}_j \quad = \quad n\, a_j \mathbf{p}_{|j}, \quad j = 1, \ldots, c \tag{1}$$

Ses paramètres sont le nombre total n de cas, les proportions a_j de cas par colonne $j = 1, \ldots, c$, et les c vecteurs de probabilités $\mathbf{p}_{|j} = \mathbf{p}(Y|j)$ correspondant à la distribution de Y dans chaque colonne j de la table.

Un arbre induit non saturé définit une partition de l'ensemble $\mathcal{X}$ des profils $\mathbf{x}$ possibles. Chacun de ses q sommets terminaux correspond donc à un sous-ensemble $\mathcal{X}_k \subseteq \mathcal{X}$, $k = 1, \ldots, q$ de profils x_j pour lequel on impose la contrainte

$$\mathbf{p}_{|j} = \mathbf{p}^a_{|k} \quad \text{pour tout } \mathbf{x}_j \in \mathcal{X}_k \qquad k = 1, \ldots, q \tag{2}$$

où $\mathbf{p}^a_{|k}$ désigne la distribution dans le sommet k de l'arbre induit.

Les degrés de liberté d_M du modèle sont donnés par le nombre de contraintes (2), soit $d_M = (c - q)(\ell - 1)$.

Sous réserves des conditions de régularité, les statistiques X^2 et G^2 de tables associées à des arbres suivent donc, lorsque le modèle est correct, une distribution du khi carré avec $(c - q)(\ell - 1)$ degrés de liberté.

La pénalisation pour la complexité dont tiennent compte les critères AIC et BIC sont fonctions du nombre de paramètres indépendants qui vaut dans notre cas $q\ell - q + c$. On a ainsi :

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

5. Illustration

Afin d'illustrer les concepts introduits, on considère les données sur le Titanic où l'on dispose de deux variables binaires : le genre (*sex* = male,female) et l'âge (*age* = adult,child), et d'une variable nominale : la classe (*class* = $c1, c2, c3$,crew) pour discriminer entre survivants et décédés (*living* = yes,no).

Le croisement des trois variables exogènes donne lieu à $2 \cdot 2 \cdot 4 = 16$ cellules dont les 2 correspondant aux enfants filles et garçons membres de l'équipage sont structurellement vides. L'arbre théorique maximal donne ainsi lieu à $c = 14$ feuilles. Le tableau 5 donne la répartition observée des données dans ces 14 feuilles qui est aussi la répartition générée par le modèle saturé (l'arbre maximal théorique). Cette ventilation des effectifs constitue la table de contingence $\mathbf{T}$ des données qui est présentée ici sous forme transposée pour des raisons de mise en page.

La figure 3 montre l'arbre induit obtenu avec la procédure CHAID de Answer Tree [SPS 01] en fixant à 10 la taille minimale des nœuds. L'arbre induit compte $q = 9$ feuilles terminales. La variable endogène étant binaire, on a $\ell = 2$. Ainsi, avec cet arbre, on dispose de $(c-q)(\ell-1) = 14-9 = 5$ degrés de liberté. Le tableau 5 donne les effectifs des 14 feuilles de l'extension maximale de l'arbre induit. Les feuilles de l'arbre étendu sont numérotées selon j, la valeur de k repérant la feuille parente de l'arbre induit.

On trouve à la ligne « CHAID » du tableau 6 les valeurs des statistiques du khi-2 du rapport de vraisemblance et de Pearson qui mesurent la divergence entre les effectifs observés et ceux générés par l'arbre. Ces valeurs sont faibles et indiquent, avec des

feuille					observé		selon arbre		
					living				
j	k	*sex*	*age*	*class*	yes	no	yes	no	Total
1	1	male	adult	c1	57	118	57	118	175
2	2			c2	14	154	14	154	168
3	3			c3	75	387	75	387	462
4	4			crew	192	670	192	670	862
5	5		child	c1	5	0	5	0	5
6	5			c2	11	0	11	0	11
7	6			c3	13	35	13	35	48
8	7	female	adult	c1	140	4	140.03	3.97	144
9	8			c2	80	13	81.47	11.53	93
10	9			c3	76	89	75.77	89.23	165
11	8			crew	20	3	20.15	2.85	23
12	7		child	c1	1	0	0.97	0.03	1
13	8			c2	13	0	11.39	1.61	13
14	9			c3	14	17	14.23	16.77	31
Total					711	1490	711	1490	2201

Tableau 5. *Titanic : effectifs observés et déduits de l'arbre CHAID*

Modèle	d	G^2	sig(G^2)	X^2	sig(X^2)	pseudo R^2_{ajust}	AIC	BIC
CHAID	5	3.72	0.590	2.10	0.835	.986	49.7	180.7
Indépendance	13	671.96	0.000	650.09	0.000	0	702.0	787.4
Saturé	0	0	1	0	1	1	56	215.5
CHAID2	6	35.81	0.000	27.85	0.000	.885	79.8	205.1
CHAID3	6	10.68	0.098	8.44	0.208	.966	54.7	180.0
CART	4	0.08	0.999	0.05	0.999	.999	48.1	184.8
C4.5	6	43.32	0.000	40.10	0.000	.860	87.3	212.6
Sipina	7	5.15	0.642	3.16	0.870	.986	47.2	166.8
Meilleur BIC	8	9.08	0.335	7.82	0.452	.978	49.1	163.0

Tableau 6. *Titanic : qualités d'ajustement d'un choix de modèles*

degrés de signification de plus de 50%, que l'arbre ajuste très bien le tableau observé. Dans ce même tableau, on donne les valeurs des critères d'ajustement pour le modèle d'indépendance, c'est-à-dire l'arbre constitué du seul nœud initial. L'indépendance est clairement rejetée et le pseudo R^2 indique que l'arbre CHAID explique 98.6% de la déviance du modèle d'indépendance. Les indicateurs BIC et AIC montrent que le déficit d'ajustement de l'arbre induit par rapport à l'arbre saturé est largement compensé par la réduction de la complexité. De même ils indiquent que l'accroissement

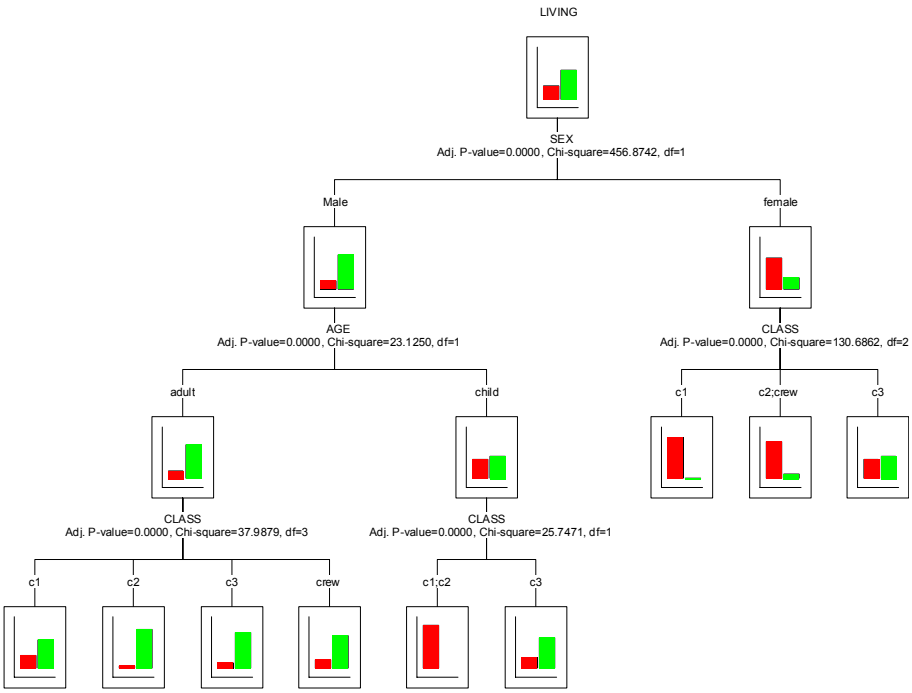

Figure 3. *Arbre induit avec la procédure CHAID*

de complexité de l'arbre par rapport à l'indépendance est nettement compensé par le gain en ajustement.

La feuille $k = 5$ (enfant de sexe masculin en 1ère ou 2ème classe) ne contient que 16 observations. On peut se demander s'il est pertinent de distinguer ces cas des autres enfants de sexe masculin. En renonçant à cet éclatement (modèle CHAID2), on constate que l'ajustement se détériore fortement. Le G^2 s'accroît de $\Delta G^2 = 32.1$ pour un gain de 1 degré de liberté ce qui indique que l'éclatement est statistiquement significatif. Les valeurs AIC et BIC montrent que le modèles CHAID2 est moins satisfaisant du point de vue du compromis entre ajustement et complexité. Par contre, si l'on fusionne les feuille $k = 2, 3$ (hommes adultes en 2ème et 3ème classe), on obtient le modèle CHAID3 qui, bien que dégradant significativement la qualité d'ajustement de l'arbre CHAID ($\Delta G^2 = 6.96$ pour 1 degré de liberté,) génère des effectifs qui ne s'écartent pas significativement des observations. Le BIC de CHAID3 est légèrement meilleur que celui de CHAID. Finalement, nous donnons à titre indicatif les statistiques d'ajustement pour les partitions induites par CART dans Answer Tree, par C4.5 et Sipina dans Sipina for Windows [Sip 00], ainsi que pour la meilleure partition possible en terme de BIC.

6. Conclusion

Cet article aborde la question de la qualité de l'ajustement d'une table de contingence par des arbres d'induction. Il s'agit d'un aspect peu discuté dans la littérature sur l'extraction de connaissances alors même que la qualité d'ajustement fait partie des outils classiques d'évaluation de modèles en statistique. La qualité d'ajustement fournit des indications complémentaires aux indicateurs de qualité traditionnellement utilisés pour les arbres d'induction que sont le taux d'erreur de prédiction, la qualité des partitions, le degré de complexité. En particulier, elle permet d'évaluer la pertinence statistique d'un arbre induit.

7. Bibliographie

[AGR 90] AGRESTI A., *Categorical Data Analysis*, Wiley, New York, 1990.

[BIS 75] BISHOP Y. M. M., FIENBERG S. E., HOLLAND P. W., *Discrete Multivariate Analysis*, MIT Press, Cambridge MA, 1975.

[BRE 84] BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J., *Classification And Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.

[QUI 93] QUINLAN J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, San Mateo, 1993.

[Sip 00] SIPINA FOR WINDOWS V2.5, http ://eric.univ-lyon2.fr, 2000, Logiciel.

[SPS 01] SPSS, Ed., *Answer Tree 3.0 User's Guide*, SPSS Inc., Chicago, 2001.

[ZIG 00] ZIGHED D. A., RAKOTOMALALA R., *Graphes d'induction : apprentissage et data mining*, Hermes Science Publications, Paris, 2000.