Choix des conclusions et validation des règles issues d'arbres de classification

Vincent Pisetta*, Gilbert Ritschard**, Djamel A. Zighed*

*Université Lumière Lyon 2, Laboratoire ERIC
v-pisett@mail.univ-lyon2.fr, abdelkader.zighed@univ-lyon2.fr
http://eric.univ-lyon2.fr
**Université de Genève, Département d'économétrie, Suisse
gilbert.ritschard@unige.ch

Résumé. Cet article traite de la validation de règles dans un contexte de ciblage où il s'agit de déterminer les profils type des différentes valeurs de la variable à prédire. Les concepts de l'analyse statistique implicative fondée sur la différence entre nombre observé de contre-exemples et nombre moyen que produirait le hasard, s'avèrent particulièrement bien adaptés à ce contexte. Le papier montre comment les notions d'indice et d'intensité d'implication de Gras s'appliquent aux règles produites par les arbres de décision et présente des alternatives inspirées de résidus utilisés en modélisation de tables de contingence. Nous discutons ensuite sur un jeu de données réelles deux usages de ces indicateurs de force d'implication pour les règles issues d'arbres. Il s'agit d'une part de l'évaluation individuelle des règles, et d'autre part de leur utilisation comme critère pour le choix de la conclusion de la règle.

1 Introduction

Les arbres de décision utilisés depuis longtemps en statistique (Morgan et Sonquist, 1963; Kass, 1980) sont devenus, suite aux ouvrages de Breiman et al. (1984) et Quinlan (1993) des outils très populaires pour générer des règles de classification et plus généralement des règles de prédictions. On parle ainsi d'arbre de classification lorsque la variable à prédire est catégorielle et que ses valeurs représentent donc des classes. Cependant, et contrairement à ce que l'expression d'arbre de classification peut laisser entendre, la classification n'est pas le seul intérêt des arbres de décisions. Par exemple, en sciences sociales où il s'agit plus de comprendre comment des prédicteurs peuvent affecter les valeurs prises par la variable à prédire que de classer des individus, ils peuvent avoir un intérêt descriptif, ou encore, comme en marketing notamment, on peut les utiliser dans une optique de ciblage. Dans ce dernier cas, plutôt que de prédire la valeur de la réponse, il s'agit de repérer les profils typiques des individus appartenant à chacune des classes de la variable à prédire. On inverse en quelque sorte le problème en cherchant à caractériser les profils propres à la classe, plutôt que la classe à partir du profil.

L'évaluation de la qualité de l'arbre se fonde le plus souvent sur le taux d'erreur de classification. Ce taux de cas mal classés par les règles, qu'il soit calculé sur les données d'apprentissage, des données test ou encore en validation croisée est évidemment pertinent comme

mesure de qualité des règles quand l'objectif est la classification proprement dite. Il ne l'est cependant plus lorsque l'on utilise l'arbre à d'autres fins, et il s'agit alors d'exploiter d'autres mesures mieux adaptées. Dans (Ritschard et Zighed, 2004; Ritschard, 2006), nous avons par exemple proposés des mesures de type déviance qui permettent de juger de la qualité descriptive de l'arbre en évaluant son aptitude à prédire la distribution de la variable réponse pour un profil donné. Ici, nous nous intéressons au cas du ciblage. Quelle information nous donne la règle sur la typicité du profil — la prémisse de la règle — pour sa conclusion ? Nous proposons de mesurer cette typicité à l'aide des concepts de l'analyse statistique implicative.

L'analyse statistique implicative introduite par Régis Gras (Gras, 1979; Gras et Larher, 1992; Gras et al., 1996) comme outil d'analyse de données, a connu ces dernières années un essor remarquable dans le cadre de la fouille de règles d'association du type « si l'on observe A alors on devrait aussi observer B » (Suzuki et Kodratoff, 1998; Gras et al., 2001, 2004). Son principe fondamental consiste à juger de la pertinence d'une relation de dépendance en fonction de la fréquence de ses contre-exemples. Une règle avec peu de contre-exemples est considérée comme plus implicative qu'une règle pour laquelle les contre-exemples sont fréquents. Curieusement, et bien que nous ayons montré (Ritschard, 2005) qu'elle s'appliquait sans difficulté aux règles issues d'arbres, cette idée de force d'implication n'a guère été exploitée dans le contexte de l'apprentissage supervisé. Or, la force d'implication d'une règle évaluée par l'écart entre le nombre observé de contre-exemples et le nombre moyen que générerait le seul hasard correspond précisément à la notion de typicalité du profil pour la conclusion qui nous intéresse ici.

L'article est organisé comme suit. En section 2, nous rappelons les concepts d'indice et d'intensité d'implication et leur utilisation associée à un arbre de classification. Nous discutons ensuite (toujours en section 2) de l'analogie entre indice d'implication et résidus issus de la modélisation de tables de contingence et de l'intérêt de ces résidus comme mesures alternatives de la force d'implication. En section 3 et 4, nous illustrons sur un exemple réel deux utilisations de l'intensité d'implication : évaluer a posteriori les règles issues d'un arbre, et effectuer le choix de la conclusion d'une règle. Enfin, nous présentons des remarques conclusives et des perspectives de développement à la section 5.

2 Arbres et indice d'implication

Les arbres de classification sont des outils de classification supervisés. Ils déterminent des règles de classification en deux temps. Dans une première étape, une partition de l'espace des prédicteurs (x) est déterminée telle que la distribution de la variable (discrète) à prédire (y) diffère le plus possible d'une classe à l'autre de la partition et soit, dans chaque classe, la plus pure possible. La partition se fait successivement selon les valeurs des prédicteurs. On commence par partitionner les données selon les modalités de l'attribut le plus discriminant, puis on répète l'opération localement sur chaque nœud ainsi obtenu jusqu'à la réalisation d'un critère d'arrêt. Dans un second temps, après que l'arbre ait été généré, on dérive les règles de classification en choisissant la valeur de la variable à prédire la plus pertinente, en général simplement la plus fréquente, dans chaque feuille (nœud terminal) de l'arbre.

Pratiquement, on relève dans chaque feuille j, $j = 1, \ldots, \ell$, le nombre n_{ij} de cas qui sont dans l'état y_i. Ainsi, on peut récapituler les distributions au sein des feuilles sous forme d'une table de contingence croisant les états de la variable y avec les feuilles (Tableau 1). On peut

noter que la marge de droite de ce tableau qui donne le total des lignes correspond en fait à la distribution des cas dans le nœud initial de l'arbre.

	feuille 1	...	feuille j	...	feuille ℓ	Total
y_1						
$\vdots$						
y_i			n_{ij}			$n_{i\cdot}$
$\vdots$						
y_k						
Total	$n_{\cdot 1}$		$n_{\cdot j}$		$n_{\cdot \ell}$	n

TAB. 1 – *Table de contingence croisant les états de la réponse y avec les feuilles de l'arbre.*

2.1 Contre-exemples et indice d'implication de Gras

L'indice d'implication (voir par exemple Gras et al., 2004, p 19) d'une règle se définit à partir des contre-exemples. Dans notre cas il s'agit dans chaque feuille (colonne du tableau 1) du nombre de cas qui ne sont pas dans la catégorie majoritaire. Ces cas vérifient en effet la prémisse de la règle, mais pas sa conclusion. En notant b la conclusion (ligne du tableau)[1] de la règle j et n_{bj} le maximum de la $j$ème colonne, le nombre de contre-exemples est $n_{\bar{b}j} = n_{\cdot j} - n_{bj}$. L'indice d'implication est une forme standardisée de l'écart entre ce nombre et le nombre espéré de contre-exemples qui seraient générés en cas de répartition entre valeurs de la réponse indépendante de la condition de la règle.

Formellement, l'hypothèse de répartition indépendante de la condition, que nous notons H_0, postule que le nombre $N_{\bar{b}j}$ de contre-exemples de la règle j résulte du tirage aléatoire et indépendant d'un groupe de $n_{\cdot j}$ cas vérifiant la prémisse de la règle j et d'un autre de $n_{\bar{b}\cdot} = n - n_{b\cdot}$ cas qui ne vérifient pas la conclusion de la règle. Sous H_0 et conditionnellement à $n_{b\cdot}$ et $n_{\cdot j}$, le nombre aléatoire $N_{\bar{b}j}$ de contre-exemples est réputé (Lerman et al., 1981) suivre une loi de Poisson de paramètre $n^e_{\bar{b}j} = n_{\bar{b}\cdot} n_{\cdot j}$. Ce paramètre $n^e_{\bar{b}j}$ est donc à la fois l'espérance mathématique et la variance du nombre de contre-exemples sous H_0. Il correspond au nombre de cas de la feuille j qui seraient des contre-exemples si l'on répartissait les $n_{\cdot j}$ cas de j selon la distribution marginale, celle du nœud initial de l'arbre (ou marge de droite du tableau 1).

L'indice d'implication de Gras est l'écart $n_{\bar{b}j} - n^e_{\bar{b}j}$ entre les nombres de contre-exemples observés et attendus sous l'hypothèse H_0, standardisé par l'écart type, soit

$$\text{Imp}(j) \quad = \quad \frac{n_{\bar{b}j} - n^e_{\bar{b}j}}{\sqrt{n^e_{\bar{b}j}}} \tag{1}$$

En termes de cas vérifiant la condition, cet indice s'écrit encore

$$\text{Imp}(j) \quad = \quad \frac{-(n_{bj} - n^e_{bj})}{\sqrt{n_{\cdot j} - n^e_{bj}}} \tag{2}$$

[1] Notons que la ligne b contenant le maximum peut évidemment varier selon la colonne.

cpred		feuille 1	...	feuille j	...	feuille ℓ
0	(contre-exemple)	$n_{\bar{b}1}$		$n_{\bar{b}j}$		$n_{\bar{b}\ell}$
1	(exemple)	n_{b1}		n_{bj}		$n_{b\ell}$
Total		$n_{\cdot 1}$		$n_{\cdot j}$		$n_{\cdot \ell}$

TAB. 2 – *Table des nombres observés d'exemples et de contre-exemples.*

cpred		feuille 1	...	feuille j	...	feuille ℓ
0	(contre-exemple)	$n^e_{\bar{b}1}$		$n^e_{\bar{b}j}$		$n^e_{\bar{b}\ell}$
1	(exemple)	n^e_{b1}		n^e_{bj}		$n^e_{b\ell}$
Total		$n_{\cdot 1}$		$n_{\cdot j}$		$n_{\cdot \ell}$

TAB. 3 – *Table des nombres attendus d'exemples et de contre-exemples.*

Pour expliciter le calcul de l'indice, on considère la variable «classe prédite» qui prend la valeur 1 pour chaque cas (exemple) appartenant à la classe majoritaire de sa feuille d'appartenance, et 0 pour les autres (contre-exemples). On note cette variable *cpred*. En croisant cette variable avec les conditions des règles, on obtient le tableau 2 où la première ligne donne pour chaque règle j son nombre $n_{\bar{b}j}$ de contre-exemples et la seconde ligne le nombre n_{bj} de cas vérifiant la règle. De même, le tableau 3 donne les nombres espérés n^e_{bj} d'exemples et $n^e_{\bar{b}j}$ de contre-exemples dans le cas d'une répartition des cas couverts par chaque règle j selon la distribution marginale. Il est important de noter que ces effectifs attendus ne se déduisent pas des marges du tableau 2. Ils s'obtiennent en répartissant tout d'abord les cas selon la distribution marginale du tableau 1 et en procédant ensuite aux regroupements selon la classe majoritaire observée dans chaque colonne du tableau 1.

2.2 Indice d'implication et résidus

Dans sa formulation (1), l'indice d'implication a l'apparence d'un résidu standardisé du type (racine signée de) contribution au khi-deux de Pearson (voir par exemple Agresti, 1990, p.224). Il s'agit en fait de la contribution au khi-deux mesurant la «distance» entre les tableaux 2 et 3. En effet, il suffit de remarquer que le khi-deux ainsi défini peut s'écrire :

$$\chi^2 \quad = \quad \sum_j \frac{(n_{\bar{b}j} - n^e_{\bar{b}j})^2}{n^e_{\bar{b}j}} + \sum_j \frac{(n_{bj} - n^e_{bj})^2}{n^e_{bj}} \tag{3}$$

On reconnaît alors sous le premier signe de sommation dans l'expression (3) le carré de l'indice d'implication de Gras. Cette interprétation de l'indice d'implication en termes de résidu (résidu de l'ajustement du nombre de contre-exemples par le modèle d'indépendance H_0), suggère que d'autres formes de résidus utilisés dans le contexte de la modélisation de tables de contingence puissent également s'avérer intéressantes pour mesurer la force d'implication d'une règle. En particulier on peut citer :

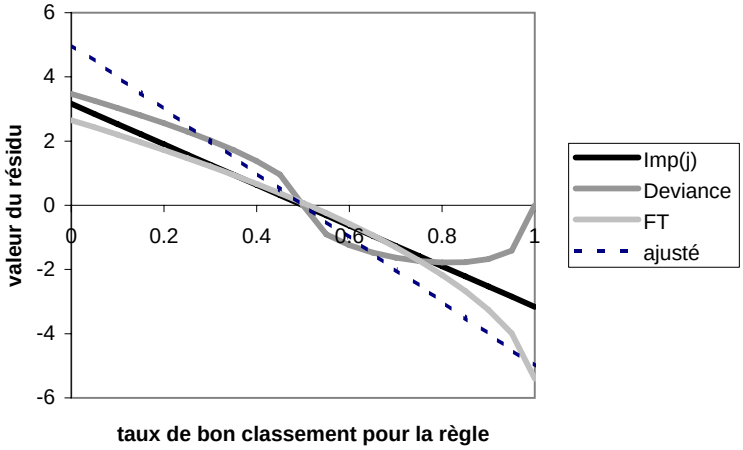

FIG. 1 – *Valeurs des indices selon le taux de biens classés de la règle,* $n_{b\cdot}/n = .5$.

1. Le résidu «déviance», $res_{dev}(j) = \text{signe}(n_{\bar{b}j} - n^e_{\bar{b}j})\sqrt{|2n_{\bar{b}j}\log(n_{\bar{b}j}/n^e_{\bar{b}j})|}$, qui est la racine signée de la contribution (en valeur absolue) au khi-deux du rapport de vraisemblance (Bishop et al., 1975, pp.136-137).
2. Le résidu ajusté d'Haberman, $res_{adj}(j) = (n_{\bar{b}j} - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}(n_{b\cdot}/n)(1 - n_{\cdot j}/n)}$, qui est le résidu standardisé de Pearson divisé par son erreur standard (Agresti, 1990, p.224).
3. Le résidu de Freeman-Tukey, $res_{FT} = \sqrt{n_{\bar{b}j}} + \sqrt{1 + n_{\bar{b}j}} + \sqrt{1 + 4n^e_{\bar{b}j}}$, qui résulte d'une transformation de stabilisation de la variance (Bishop et al., 1975, p.137).

Le résidu standardisé, qui correspond à l'indice d'implication de Gras, est connu pour avoir une variance inférieure à 1. Le problème est que dans la pratique les nombres $n_{b\cdot}$ et $n_{\cdot j}$ dépendent de l'échantillon considéré et sont donc eux-mêmes aléatoires. Ainsi $n^e_{\bar{b}j}$ n'est qu'une estimation du paramètre de la loi de Poisson. On doit alors tenir compte du fait que dans la formule (1), le dénominateur n'est qu'une estimation de l'écart type. Les résidus déviance, de Freeman-Tukey et ajusté sont mieux adaptés à cette situation et sont réputés avoir dans la pratique une distribution plus proche de la normale $N(0, 1)$ que le simple résidu standardisé. Ce dernier, et par conséquent l'indice d'implication de Gras, tend à sous-estimer la force d'implication.

La figure 1 montre les valeurs des résidus (en ordonnée) en fonction de la proportion $n_{bj}/n_{\cdot j}$ (en abscisse) de cas qui dans la feuille j vérifient la conclusion b de la règle. Les courbes sont représentées pour $n = 100$, une règle j associée à une feuille couvrant 20% des cas, et une proportion marginale $n_{b\cdot}/n$ de cas vérifiant la conclusion b de 50%. A gauche de ce seuil, les indices prennent tous une valeur positive indiquant que la règle fait moins bien que le hasard. On peut relever le comportement curieux du résidu déviance dont la valeur tend vers 0 lorsque le nombre de contre-exemples tend vers 0. Cela suggère que la règle devient non implicative quand le nombre de contre-exemples devient nul, ce qui n'est évidemment pas sastisfaisant. L'indice de Gras et le résidu ajusté évoluent de manière linéaire avec la proportion

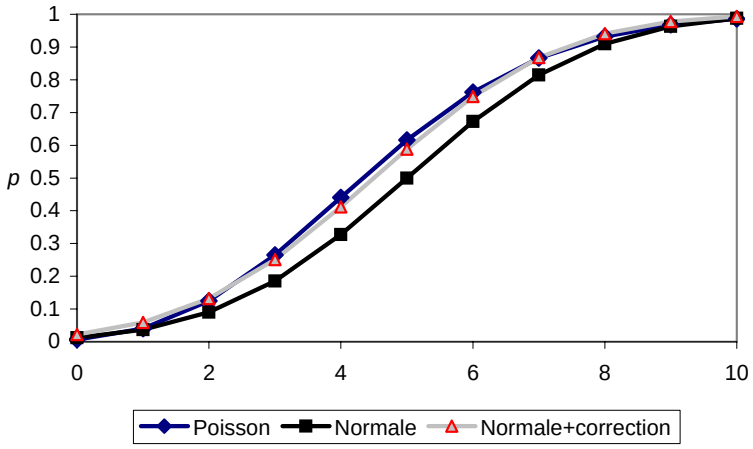

FIG. 2 – *Distributions normale avec et sans correction de continuité, et de Poisson ($n^e_{\bar{b}j} = 5$).*

$n_{bj}/n_{.j}$, le résidu ajusté prennant ses valeurs sur une étendue plus importante. Quant au résidu de Freeman-Tukey, on relève que sa variation s'accélère lorsque le taux de biens classés de la règle approche de 1.

2.3 Intensité d'implication et p-valeur

Il est naturel de s'intéresser à la p-valeur, ou degré de signification, des indices d'implication observés. Cette p-valeur correspond à la probabilité $p(N_{\bar{b}j} \leq n_{\bar{b}j}|H_0)$. Quand $n^e_{\bar{b}j}$ est petit, le calcul peut se faire, conditionnellement à $n_{b.}$ et $n_{.j}$, avec la loi de Poisson de paramètre $n^e_{\bar{b}j}$. Pour $n^e_{\bar{b}j}$ grand (≥ 5), la loi normale donne une bonne approximation, à condition toutefois de procéder à la correction pour la continuité, la différence pouvant atteindre encore 2.6 points de pourcentage pour $n = 100$. La figure 2 montre les fonctions de répartition de la loi de Poisson et de la loi normale avec et sans correction de continuité pour $n^e_{\bar{b}j} = 5$. On peut relever que l'approximation par la loi normale, en particulier avec la correction pour la continuité, reste bonne même pour $n^e_{\bar{b}j}$ relativement petit. Ainsi, en notant $\phi(\cdot)$ la fonction de distribution d'une normale standardisée, on a

$$p(N_{\bar{b}j} \leq n_{\bar{b}j} \mid H_0) \quad \simeq \quad \phi\Big((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}}\Big) \ . \tag{4}$$

On appelle *intensité d'implication* (Gras et al., 1996) le complémentaire à 1 de cette p-valeur. Gras et al. (2004) la définissent en termes de l'approximation normale (4), mais sans la correction pour la continuité. Pour notre part, nous la calculerons pour une règle j comme

$$\text{IntImp}(j) \quad = \quad 1 - \phi\Big((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}}\Big) \ . \tag{5}$$

Dans tous les cas, cette intensité s'interprète comme la probabilité d'obtenir, sous l'hypothèse H_0, un nombre de contre-exemples supérieur à celui observé pour la règle j.

3 Validation des règles

L'indice d'implication et ses variantes proposées à la section précédente s'avèrent tout naturellement utiles pour juger de la pertinence individuelle des règles de classification. Cette information vient enrichir les critères usuels d'évaluation globale du classifieur. Nous nous proposons d'illustrer ici cet usage sur un jeu de données réelles.

Nous considérons des données collectées dans le cadre de l'étude STULONG (Tomečková et al., 2002) sur les effets pathologiques de la consommation d'alcool qui a été réalisée conjointement par l'Université Charles et l'Hôpital universitaire Charles de Prague. Les données portent sur 1341 patients dont on s'intéresse ici à prédire les habitudes de consommation (*jamais*, *occasionnellement*, *régulièrement*) à partir des cinq prédicteurs quantitatifs répertoriés au tableau 4.

Code	Nom
syst1	pression artérielle systolique (mesure 1)
syst2	pression artérielle systolique (mesure 2)
BMI	indice de masse corporel
chlst	taux de cholestérol
Nb cigarettes	nombre de cigarettes / jour

TAB. 4 – *Les différents prédicteurs.*

La figure 3 montre l'arbre obtenu avec la méthode CHAID (avec seuil de significativité de 5% et une taille minimale des nœuds de 10). On en induit sept règles de classification. Celles-ci correspondent aux feuilles (nœuds terminaux), les conclusions étant données par la classe

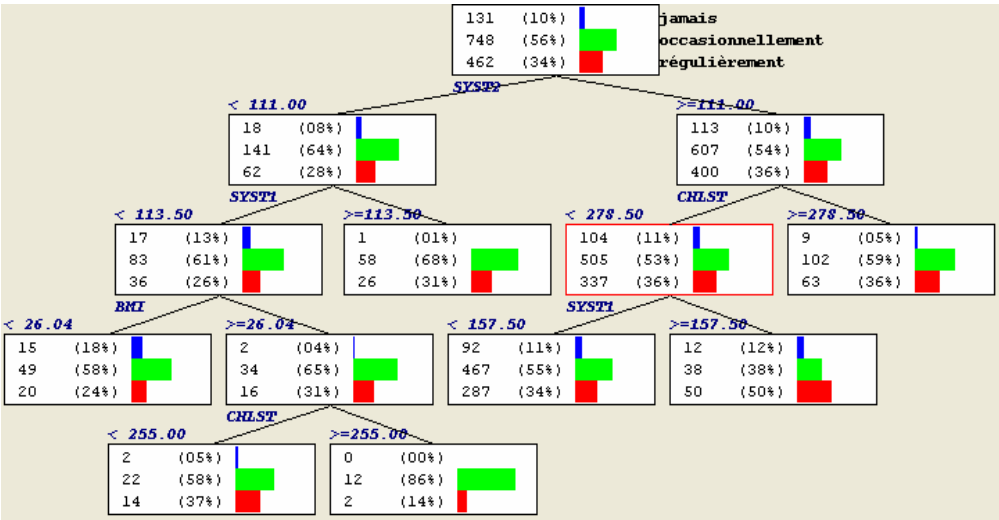

FIG. 3 – *Arbre induit sur les données STULONG, méthode CHAID.*

R1	(syst2 $<$ 111) et (syst1 $\geq$ 113.5)	$\Rightarrow$	buveur occasionnel
R2	(syst2 $<$ 111) et (syst1 $<$ 113.5) et (BMI $<$ 26.04)	$\Rightarrow$	buveur occasionnel
R3	(syst2 $<$ 111) et (syst1 $<$ 113.5) et (BMI $\geq$ 26.04) et (chlst $<$ 255)	$\Rightarrow$	buveur occasionnel
R4	(syst2 $<$ 111) et (syst1 $<$ 113.5) et (BMI $\geq$ 26.04) et (chlst $\geq$ 255)	$\Rightarrow$	buveur occasionnel
R5	(syst2 $\geq$ 111) et (chlst $\geq$ 278.5)	$\Rightarrow$	buveur occasionnel
R6	(syst2 $\geq$ 111) et (chlst $<$ 278.5) et (syst1 $<$ 157.5)	$\Rightarrow$	buveur occasionnel
R7	(syst2 $\geq$ 111) et (chlst $<$ 278.5) et (syst1 $\geq$ 157.5)	$\Rightarrow$	buveur régulier

TAB. 5 – *Les sept règles induites.*

majoritaire et leurs prémisses définies par les chemins, mutuellement exclusifs, qui mènent aux feuilles. Les sept règles sont explicitées au tableau 5. Remarquons en premier lieu que toutes les règles sauf une concluent à «buveur occasionnel», modalité majoritaire au nœud initial de l'arbre. Ceci est typique de situations de déséquilibre (répartition des modalités de y éloignée de la situation d'équiprobabilité), où les algorithmes d'apprentissage ont parfois du mal à discriminer les différentes classes. On relève également, qu'aucune règle ne conclut à «jamais». La faible représentativité au nœud initial de l'arbre de cet état induit une difficulté à trouver des règles isolant ces individus. Le tableau 6 présente la classique matrice de confusion associée à cet arbre. Les défauts cités jusqu'ici y apparaissent de façon plus flagrantes, en particulier l'erreur associée à la modalité «jamais».

	prédiction		
état réel	jamais	occasionnellement	régulièrement
jamais	0	119	12
occasionnellement	0	710	38
régulièrement	0	412	50

TAB. 6 – *Matrice de confusion, règle majoritaire, taux d'erreur* $= 43\%$.

Les valeurs des résidus définis en section 2.1 sont présentées pour chacune des règles dans le tableau 7. Les valeurs négatives indiquent que le nombre de contre-exemples observé est inférieur à celui attendu sous la condition d'indépendance entre la prémisse et la conclusion de la règle. Dès lors, les valeurs négatives sont synonymes de «qualité». La règle R6 ((syst2 $\geq$ 111) et (chlst $<$ 278.5) et (syst1 $<$ 157.5) $\Rightarrow$ buveur occasionnel) pour laquelle les résidus sont positifs, est une règle qui fait moins bien que l'indépendance au sens que le nombre de contre-exemples observé est supérieur au nombre moyen que générerait le hasard. La règle peut donc être considérée comme non pertinente.

Il est intéressant ici de faire une comparaison de la qualité implicative avec le taux d'erreur communément utilisé pour l'évaluation de règles de classification. Le nombre de contre-exemples considérés est précisément le nombre d'erreurs produites par la règle sur l'échantillon d'apprentissage. Le taux d'erreur correspond ainsi au pourcentage de contre-exemples parmi les cas couverts par la règle, soit $n_{\bar{b}j}/n_{.j}$ pour la règle j, ce qui est encore le complémentaire à 1 de la confiance. Le taux d'erreur souffre donc des mêmes inconvénients que la confiance. En particulier, il ne nous dit rien sur ce que la règle apporte de plus qu'une clas-

	Standardisé	Deviance	Freeman-Tukey	Ajusté
R1	-1.73	-2.79	-1.81	-2.39
R2	-0.35	-1.34	-0.31	-0.49
R3	-0.20	-0.83	-0.14	-0.27
R4	-1.68	-1.40	-1.93	-2.27
R5	-0.56	-2.04	-0.54	-0.81
R6	0.25	2.07	0.26	0.56
R7	-1.92	-3.43	-2.01	-3.40

TAB. 7 – *Valeurs des résidus pour chacune des règles de l'arbre.*

sification indépendante de toute condition. Pour notre règle R6 par exemple, la confiance est de 55% contre 56% pour le classifieur naïf consistant à classer tout le monde comme « buveur occasionnel », classe la plus fréquente au nœud initial.

La question est évidemment de savoir quoi faire d'une règle non pertinente d'un point vue implicatif. On peut soit décider de la conserver si le but est la qualité globale de classification. Dans le cas où, au contraire, l'on veut privilégier la force implicative de chaque règle, deux solutions sont envisageables :

- fusionner la règle avec une de ses règles sœurs ;
- changer la conclusion de la règle.

En fusionnant les règles R6 et R7, ce qui revient à élaguer la branche non pertinente de l'arbre, on obtient une nouvelle règle ((syst2 $>= 111$) et (chlst < 278.5) $\Rightarrow$ buveur occasionnel). Les valeurs des résidus pour cette nouvelle règle sont : $res_{std} = 1.11$, $res_{dev} = 4.5$, $res_{FT} = 1.11$, $res_{adj} = 2.73$. Ils sont positifs et indiquent clairement une détérioration par rapport à la situation précédente. En fait, on peut observer sur l'arbre de la figure 3 qu'en remontant la branche à partir de la feuille correspondant à la règle R6, on ne rencontre que des nœuds où la classe majoritaire « occasionnellement » a une fréquence inférieure à celle relevée au nœud initial. La fusion ne peut donc pas être une solution dans ce cas particulier tant que l'on garde le principe de la classe majoritaire pour le choix de la conclusion. Ceci nous amène donc à discuter l'autre solution consistant à changer la conclusion de la règle en choisissant la modalité qui maximise l'intensité d'implication.

4 Choix de la conclusion des règles

Si l'objectif est de maximiser l'intensité d'implication des règles, dans le but en particulier de déterminer les profils les plus caractéristiques de chaque état de la variable à prédire, il semble naturel de choisir la conclusion de la règle qui maximise cette intensité plutôt que la classe majoritaire. L'idée de choisir ainsi la classe maximisant l'intensité d'implication (i.e. minimisant le résidu) a notamment déjà été exploitée par Zighed et Rakotomalala (2000, pp.282-287). A titre d'exemple, nous donnons dans le tableau 8 la conclusion sélectionnée par cette procédure pour chacune des règles et selon le résidu utilisé comme critère de choix. On observe que si les conclusions restent celles de la classe majoritaire pour les règle R1, R4 et R5, le principe de la maximisation de l'implication donne des conclusions différentes

	Imp(j)	Déviance	Freeman-Tukey	Ajusté	Majorité
R1	2	2	2	2	2
R2	1	1	1	1	2
R3	2	3	2	3	2
R4	2	2	2	2	2
R5	2	2	2	2	2
R6	1	1	1	1	2
R7	3	3	3	3	3

$1 =$ jamais, $2 =$ occasionnellement, $3 =$ régulièrement

TAB. 8 – *Conclusion selon le résidu utilisé comme critère.*

pour les quatre autres règles. Pour la règle R3, la conclusion varie entre «occasionnellement» et «fréquemment» selon le critère implicatif retenu. Pour les règles R2, R6 et R7 les quatre indices d'implication conduisent au même résultat. Il est intéressant de relever également, qu'avec ce critère implicatif, chacune des trois modalités de la variable à prédire est retenue comme conclusion pour au moins une règle. De plus, on peut souligner que, dans tous les cas, les indices d'implication — dont les valeurs ne sont pas montrées ici — restent négatifs.

Une expérience intéressante consiste à recalculer la matrice de confusion nouvellement obtenue. Le tableau 9 montre cette dernière pour le cas où l'on utilise le résidu standardisé, soit l'indice de Gras. Le taux d'erreur global est évidemment plus élevé qu'au tableau 6 ce qui n'est pas surprenant puisqu'on ne vise plus ici à minimiser l'erreur de classification. Le tableau fournit cependant des enseignements utiles sur deux plans. Premièrement, on peut observer que la maximisation de l'intensité d'implication améliore considérablement la valeur des mesures de rappel intra-classe pour les modalités faiblement représentées. On a également confirmation qu'il n'y pas ici, et contrairement au tableau 6, d'état de y qui ne puisse être prédit par au moins une règle.

Ensuite, et c'est ici l'intérêt principal du choix de la conclusion selon le principe de la maximisation de l'intensité, la matrice fait ressortir que les règles sont ici plus discriminantes par rapport à la répartition au nœud initial de l'arbre. Ainsi, l'arbre généré peut être vu comme la représentation d'une typologie des modalités de la variable à prédire y. L'interprétation des règles, qui ne sont plus alors des règles de classification, doit elle être revue en terme de typicité de la condition pour la conclusion choisie. Ainsi, les personnes ayant une pression artérielle systolique 2 inférieure à 111, une pression artérielle systolique 1 inférieure à 113.5 et un indice de masse corporelle inférieur à 26.04 sont caractéristiques des «jamais buveurs». Au contraire,

		prédiction	
état réel	jamais	occasionnellement	régulièrement
jamais	107	12	12
occasionnellement	516	194	38
régulièrement	307	105	50

TAB. 9 – *Matrice de confusion, maximisation intensité implicative, taux d'erreur* $= 72\%$.

les buveurs réguliers sont caractérisés par une pression artérielle systolique 2 élevée (≥ 111) et un taux de cholestérol également élevé. Enfin, les buveurs occasionnels ne sont pas clairement caractérisés, bien qu'il existe des circonstances typiques comme un taux de cholestérol élevé allié à un BMI également élevé, sans toutefois connaître de problème au niveau de la pression artérielle. La difficulté à discriminer les buveurs occasionnels des autres peut également venir du fait qu'il y a différents types de buveurs, la définition « occasionnellement » étant elle-même subjective.

5 Conclusion

Nous avons montré dans ce papier l'intérêt de la statistique implicative pour les arbres de classification. Nous avons tout d'abord défini et rappelé le concept de force d'implication et réalisé un parallèle avec les résidus utilisés en modélisation des tables de contingence. L'utilisation de ces concepts a deux objectifs intéressants. D'une part, la statistique implicative fournit des outils pour juger de manière pertinente l'intérêt des règles de classification dans une optique de ciblage, c'est-à-dire lorsqu'on cherche à caractériser les profils type de chaque valeur de la variable réponse. D'autre part, elle offre des critères utiles pour sélectionner les conclusions des règles dans ce contexte de ciblage. Les règles obtenues selon ces critères de force d'implication, ne sont plus des règles de classification, mais permettent d'effectuer une typologie des différentes modalités de la variable réponse y.

Une perspective intéressante dans ce contexte de ciblage est alors d'utiliser l'intensité d'implication pour la construction de l'arbre et non pas pour évaluer uniquement a posteriori les règles. On peut penser à utiliser des critères réalisant un compromis entre classement pur (obtenir le meilleur taux d'erreur possible) et validation statistique (intensité d'implication) en utilisant des mesures comme l'intensité d'implication entropique. Etant de nature statistique, elle permet d'effectuer directement un pré-élagage de l'arbre et offre un critère d'arrêt tout à fait naturel. Ces questions et réflexions méritent toutefois une étude plus approfondie.

Références

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

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

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques*. Thèse d'état, Université de Rennes 1, France.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Laher, M. Polo, H. Ratsimba-Rajohn, et A. Totohasina (1996). *L'implication statistique : Nouvelle méthode exploratoire de données*. Recherches en didactique des mathématiques. Grenoble : La pensée sauvage.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, 3–30.

Gras, R., P. Kuntz, R. Couturier, et F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage ECA 1*(1-2), 69–80.

Gras, R. et A. Larher (1992). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématique, Informatique et Sciences Humaines* (120), 5–31.

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

Lerman, I. C., R. Gras, et H. Rostam (1981). Elaboration d'un indice d'implication pour données binaires I. *Mathématiques et sciences humaines* (74), 5–35.

Morgan, J. N. et J. A. Sonquist (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*, 415–434.

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

Ritschard, G. (2005). De l'usage de la statistique implicative dans les arbres de classification. In R. Gras, F. Spagnolo, et J. David (Eds.), *Actes des Troisièmes Rencontres Internationale ASI Analyse Statistique Implicative*, Volume Secondo supplemento al N.15 of *Quaderni di Ricerca in Didattica*, pp. 305–314. Palermo : Università degli Studi di Palermo.

Ritschard, G. (2006). Computing and using the deviance with classification trees. In A. Rizzi et M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 55–66. Berlin: Springer.

Ritschard, G. et D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information RNTI E-1*, 45–67.

Suzuki, E. et Y. Kodratoff (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow et M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, Proceedings*, pp. 10–18. Berlin : Springer.

Tomečková, M., J. Rauch, et P. Berka (2002). Data from longitudinal study of atherosclerosis risk factors. In P. Berka (Ed.), *Discovery Challenge Workshop Notes. ECML/PKDD-2002*. Helsinki.

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

Summary

This paper deals with rule validation in a targeting framework where the goal is to characterize typical profiles for each of the outcome classes. Implicative statistic analysis which is founded on the difference between the observed number of counter-examples and the mean number of them we may expect from hazard, is well suited for this issue. We show how notions such as Gras' implication index and intensities can be applied to rules derived from trees. We propose alternatives to Gras' index based on residuals used in the modeling of contingency tables. Then, using a real world data set, we discuss two usages of these measures of rule implication strength. The first one is individual validation of rules, and the second one concerns its use as a criterion for selecting the conclusion of the rule.