De l'usage de la statistique implicative dans les arbres de classification

Gilbert Ritschard

Département d'économétrie, Université de Genève
40, bd du Pont-d'Arve, CH-1211 Genève 11, Suisse
gilbert.ritschard@themes.unige.ch
http://mephisto.unige.ch

Résumé. Cet article met en lumière l'intérêt du concept de statistique implicative dans le contexte des arbres de classification. Dans une première partie nous montrons que les résidus standardisés utilisés en modélisation log-linéaire de tables de contingence sont des alternatives intéressantes de l'indice d'implication de Gras. Nous distinguons ensuite principalement deux usages de ces indices. Le premier, purement descriptif, exploite l'intensité d'implication comme critère d'évaluation a posteriori de la pertinence individuelle des règles de classification associées à chaque feuille de l'arbre. Le second usage, proposé notamment par Zighed et Rakotomalala (2000) utilise l'intensité d'implication pour définir la conclusion de la règle de classification dans chacune des feuilles de l'arbre.

1 Introduction

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, cette idée de force d'implication n'a guère été exploitée dans le contexte de l'apprentissage supervisé, avec l'exception notable cependant de Zighed et Rakotomalala (2000, p. 282-287) sur laquelle nous reviendrons ultérieurement. De même que la force d'implication complémente avantageusement les critères de support et de confiance sur laquelle s'appuient les algorithmes du type *Apriori* de fouille de règles fréquentes, la force d'implication nous semble pouvoir enrichir le classique taux d'erreur utilisé pour valider les règles de classification générées par un arbre, ou par toute autre méthode supervisée.

L'objectif de cet article est d'amorcer une discussion sur la portée et les limites du concept d'implication statistique pour la classification supervisée et plus spécifiquement pour les arbres de classification. Nous commençons à la section 2 par rappeler les concepts d'indice et d'intensité d'implication en illustrant sur un exemple à quoi ils correspondent pour des règles de classification générées par un arbre. Nous discutons l'analogie de l'indice d'implication avec les résidus issus de la modélisation de tables de contingence et l'intérêt de ces derniers comme mesures alternatives de la force d'implication. La section 3 est consacrée à l'évaluation de la qualité individuelle des règles de classification avec l'intensité implicative par comparaison avec le taux d'erreur. Enfin, à la section 4, nous abordons l'utilisation de

l'indice d'implication pour définir la règle de classification associée à chaque feuille de l'arbre. Nous concluons brièvement au point 5.

2 Arbre de classification et indices d'implication

Pour illustrer notre propos, nous considérons un exemple fictif où il s'agit de prédire l'état civil (marié, célibataire, divorcé/veuf) d'individus à partir de la connaissance du genre (NSEX : homme ou femme) et du secteur d'activité (NSACTIV : primaire, secondaire, tertiaire). On dispose pour cela des 273 données récapitulées au tableau 1.

Etat civil	*Sexe*	*Secteur activité*	*Nombre de cas*
marié	homme	primaire	50
marié	homme	secondaire	40
marié	homme	tertiaire	6
marié	femme	primaire	0
marié	femme	secondaire	14
marié	femme	tertiaire	10
célibataire	homme	primaire	5
célibataire	homme	secondaire	5
célibataire	homme	tertiaire	12
célibataire	femme	primaire	50
célibataire	femme	secondaire	30
célibataire	femme	tertiaire	18
divorcé/veuf	homme	primaire	5
divorcé/veuf	homme	secondaire	8
divorcé/veuf	homme	tertiaire	10
divorcé/veuf	femme	primaire	6
divorcé/veuf	femme	secondaire	2
divorcé/veuf	femme	tertiaire	2

TAB. 1 – *Exemple : les données.*

2.1 Arbre et règles

Les arbres de classification sont des outils de classification supervisés. Ils déterminent des règles de classification en deux temps. Dans un premier, une partition de l'espace des prédicteurs est déterminée telle que la distribution de la variable (discrète) à prédire 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.

La figure 1 montre l'arbre induit avec la méthode CHAID, un seuil de signification de 5% et une taille minimale des nœuds de 20. Notons que l'on obtient le même arbre avec CART et une valeur de 0.02 pour le gain minimal. L'arbre définit une partition (segmentation) de l'espace des prédicteurs, avec une distribution de la variable à prédire, ici l'état civil, propre à chaque classe de la partition. Pour notre discussion il est commode de représenter ces distributions sous forme d'un tableau croisant la variable à prédire avec les profils définis par les branches (condition des règles). On associe ainsi le tableau 2 à l'arbre de la figure 1.

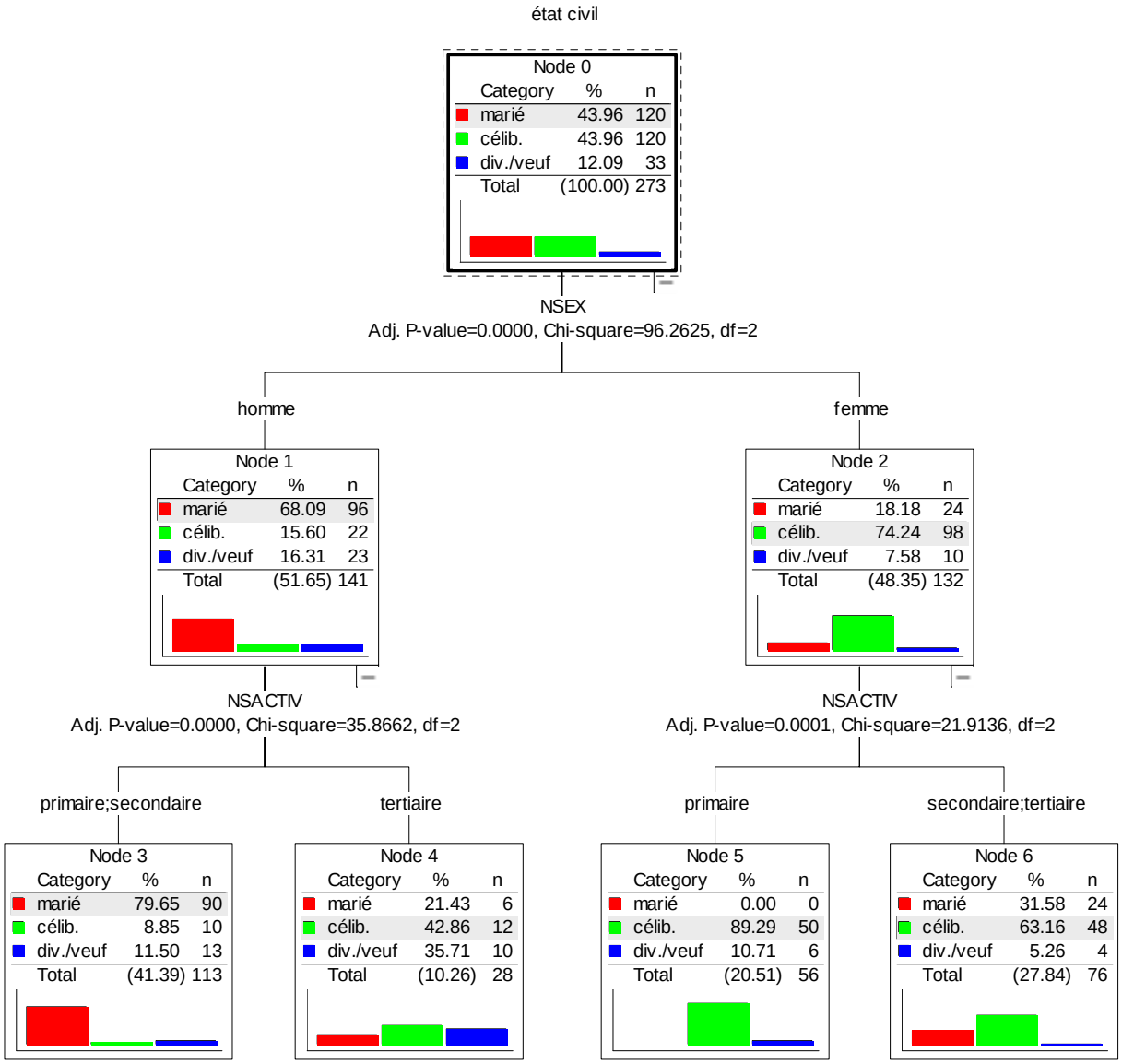

FIG. 1 – *Exemple : arbre induit*

L'approche usuelle pour dériver les règles de classification à partir de l'arbre induit consiste à choisir dans chaque feuille la valeur (classe) la plus fréquente. Par exemple, un

homme travaillant dans le secteur secondaire appartient à la feuille 3 et sera classé parmi les mariés, tandis qu'un homme du secteur tertiaire (feuille 4) sera classé comme célibataire. Dans le tableau 2, les têtes de colonnes définissent les prémisses des règles, tandis que la ligne contenant le maximum donne la conclusion. On note que, dans cet exemple, on ne prédit jamais la catégorie « divorcé ou veuf ».

	Homme		Femme		
Etat civil	primaire, secondaire	tertiaire	primaire	Secondaire, tertiaire	total
Marié	**90**	6	0	24	120
Célibataire	10	**12**	**50**	**48**	120
divorcé/veuf	13	10	6	4	33
Total	113	28	56	76	273

TAB 2 – *Table associée à l'arbre induit.*

L'arbre obtenu définit ainsi les quatre règles suivantes :

1. Homme du secteur primaire ou secondaire ⇒ marié
2. Homme du secteur tertiaire ⇒ célibataire
3. Femme du secteur primaire ⇒ célibataire
4. Femme du secteur secondaire ou tertiaire ⇒ célibataire

Par rapport aux règles d'association cherchées avec des algorithmes du type *Apriori*, les règles de classification présentent les deux caractéristiques suivantes :

1. Les conclusions des règles sont à choisir exclusivement dans l'ensemble des modalités de la variable à prédire.
2. Les prémisses des règles sont mutuellement exclusives et définissent une partition de l'espace des prédicteurs.

Il n'en demeure pas moins qu'il s'agit de règles avec une prémisse et une conclusion et qu'on peut leur appliquer en particulier les concepts de support, de confiance et, ce qui nous intéresse tout particulièrement ici, d'indice d'implication.

2.2 Contre-exemples et indice d'implication de Gras

L'*indice d'implication* (voir par exemple Gras et al. 2004, p19) d'une règle se définit à partir des contre-exemples. Dans notre cas il s'agit dans chaque feuille (colonne du tableau 2) 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) de la règle *j* et n_{bj} le maximum de la *j*ème colonne, le nombre de contre-exemples est $n_{\bar{b}j} = n_{.j} - n_{bj}$.

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 $n_{.j}$ cas vérifiant la prémisse de la règle *j* et d'un autre de $n_{\bar{b}.} = n - n_{b.}$ cas qui ne vérifient pas la conclusion de la règle. Sous H_0, et conditionnellement à $n_{b.}$ et $n_{.j}$, le nombre aléatoire $N_{\bar{b}j}$ de contre-exemples est réputé (Lerman et al. 1981) suivre une loi de Poisson de paramètre $n^e_{\bar{b}j} = n_{\bar{b}.} n_{.j} / n$. Ce paramètre $n^e_{\bar{b}j}$ est donc à la fois l'espérance mathématique $\mathrm{E}(N_{\bar{b}j} \mid H_0)$ et la variance $\mathrm{var}(N_{\bar{b}j} \mid H_0)$ 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 $n_{.j}$ cas de *j* selon la distribution marginale, celle du nœud initial de l'arbre (ou marge de droite du tableau 2).

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

$$\mathrm{Imp}(j) = \frac{n_{\bar{b}j} - n^e_{\bar{b}j}}{\sqrt{n^e_{\bar{b}j}}}. \qquad (1)$$

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

$$\mathrm{Imp}(j) = \frac{-(n_{bj} - n^e_{bj})}{\sqrt{n_{.j} - n^e_{bj}}}. \qquad (2)$$

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 3 où la première ligne donne pour chaque règle son nombre $n_{\bar{b}j}$ de contre-exemples et la seconde ligne le nombre n_{bj} de cas vérifiant la règle.

Classe prédite cpred	Homme		Femme		
	primaire, secondaire	tertiaire	Primaire	secondaire, tertiaire	total
0 (contre-exemple)	23	16	6	28	73
1 (exemple)	**90**	**12**	**50**	**48**	200
Total	113	28	56	76	273

TAB 3 – *Effectifs $n_{\bar{b}j}$ et n_{bj} des contre-exemples et exemples observés.*

De même, le tableau 4 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 $n_{j.}$ cas couverts par la règle selon la distribution marginale. Il est important de noter que ces effectifs attendus ne se déduisent pas des marges du tableau 3. Ils s'obtiennent en répartissant tout d'abord les cas selon la distribution marginale du tableau 2 et en procédant ensuite aux regroupements selon la classe majoritaire observée dans chaque colonne du tableau 2. Ceci explique que les tableaux 3 et 4 n'aient pas la même marge de droite.

Classe prédite	Homme		Femme		
cpred	primaire, secondaire	tertiaire	primaire	secondaire, tertiaire	total
0 (contre- exemple)	63.33	15.69	31.38	42.59	153
1 (exemple)	49.67	12.31	24.62	33.41	120
Total	113	28	56	76	273

TAB 4 – *Effectifs* $n^e_{\bar{b}j}$ *et* n^e_{bj} *de contre-exemples et exemples attendus en cas d'indépendance.*

2.3 Indice d'implication et résidu

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 3 et 4. Ces contributions sont données au tableau 5. Les indices d'implication sont les contributions des éléments de la première ligne. L'indice d'implication de la première règle est ainsi Imp(1) = -5.068. Cette valeur négative indique que le nombre observé de contre-exemples est inférieur au nombre moyen que l'on peut attendre d'une répartition indépendante de la condition et souligne donc l'intérêt de la règle. Pour la règle 2, on a un indice d'implication positif représentatif d'une règle sans intérêt puisque générant plus de contre-exemples que le classement au hasard sans tenir compte de la condition.

Classe prédite	Homme		Femme	
cpred	Primaire, secondaire	tertiaire	Primaire	secondaire, tertiaire
0 (contre- exemple)	**-5.068**	**0.078**	**-4.531**	**-2.236**
1 (exemple)	5.722	-0.088	5.116	2.525

TAB 5 – *Contributions au khi-deux mesurant la « distance » entre les tableaux 3 et 4.*

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 intéressante pour mesurer la force d'implication d'une règle. En particulier on peut citer :

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

Le tableau 6 montre la force d'implication des quatre règles de notre exemple reflétée par chacun de ces résidus. On observe, ce qui n'est pas étonnant, qu'elles sont concordantes.

Résidu		Règle 1	Règle 2	Règle 3	Règle 4
standardisé	res_s	-5.068	0.078	-4.531	-2.236
déviance	res_d	-6.826	0.788	-4.456	-4.847
Freeman-Tukey	res_{FT}	-6.253	0.138	-6.154	-2.414
ajusté	res_a	-9.985	0.124	-7.666	-3.970

TAB 6 – *Les divers résidus.*

Le résidu standardisé qui correspond à l'indice d'implication, est connu pour avoir une variance inférieure à 1. Le problème est que dans la pratique les nombres $n_{b.}$ et $n_{.j}$ 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. Il donne dans notre exemple des valeurs absolues inférieures à celles des trois autres résidus. La seule exception concerne la règle 3, où le nombre 6 de contre-exemples observés est le plus petit.

2.4 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 $p(N_{\bar{b}j} \geq n_{\bar{b}j} \,|\, H_0)$. 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 $P(n^e_{\bar{b}j})$. 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 $n^e_{\bar{b}j}$ = 100. Ainsi, en notant φ(.) la fonction de distribution d'une normale standardisée, on a :

$$p(N_{\bar{b}j} \leq n_{\bar{b}j} \mid H_0) \cong \varphi\left((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j}) / \sqrt{n^e_{\bar{b}j}} \right). \quad (3)$$

On appelle *intensité d'implication* (Gras, 1996) le complémentaire à 1 de cette *p*-valeur. Gras et al. (2004) la définissent en termes de l'approximation normale (3), mais sans la correction pour la continuité. Pour notre part nous la calculerons comme :

$$\text{Intens}(j) = 1 - \varphi\left((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j}) / \sqrt{n^e_{\bar{b}j}} \right). \quad (4)$$

Dans tous les cas, cette intensité s'interprète comme la probabilité d'obtenir, sous l'hypothèse H_0, un nombre de contre-exemples inférieur à celui observé pour la règle *j*. Le tableau 7 donne ces intensités pour les quatre règles de notre exemple. A titre de comparaison, nous y reportons également le complémentaire à 1 des *p*-valeurs des résidus déviance, de Freeman-Tukey et ajusté, calculés avec la correction pour la continuité, c'est-à-dire en ajoutant 0.5 aux nombres observés de contre-exemples.

Résidu		Règle 1	Règle 2	Règle 3	Règle 4
standardisé	res_s	1.000	0.419	1.000	0.985
déviance	res_d	1.000	0.099	1.000	1.000
Freeman-Tukey	res_{FT}	1.000	0.350	1.000	0.988
ajusté	res_a	1.000	0.373	1.000	1.000

TAB 7 – *Variantes d'intensité d'implication (avec correction pour continuité).*

3 Pertinence individuelle des règles

L'intensité 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. Par exemple, parmi les 4 règles issues de notre arbre illustratif, les règles 1, 3 et 4 sont clairement pertinentes, tandis que la deuxième avec une intensité implicative inférieure à 50%, signifiant qu'elle génère plus de contre-exemples que l'indépendance, est à rejeter.

La question est alors que faire des cas couverts par la condition de la règle à rejeter. Deux solutions nous semblent pouvoir être envisagées : 1) la fusion avec les cas couverts par une autre règle et 2) changer de conclusion. Ce dernier cas est discuté à la section 4. Pour ce qui est de la fusion, comme on est dans le contexte des arbres, il est naturel de songer à regrouper la feuille correspondante avec l'une de ses feuilles sœur, ce qui revient à élaguer la branche non pertinente de l'arbre. Dans notre exemple, ceci conduit à fusionner les règles 1 et 2 en une

nouvelle règle « Homme $\Rightarrow$ marié ». Les résidus sur les contre-exemples de cette nouvelle règle sont respectivement res_s = -3.8, res_d = -7.1, res_{FT} = -8.3 et res_a = -4.3. Excepté le résidu déviance, ils indiquent une légère détérioration par rapport à la qualité implicative de la règle 1.

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 $\text{err}(j) = n_{\bar{b}j} / n_{.j} = 1 - n_{bj} / n_{.j}$, 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 classification indépendante de toute condition. Par ailleurs, le taux d'erreur est étroitement lié au choix de la règle majoritaire pour le choix de la conclusion des feuilles. A titre indicatif, le taux d'erreur est respectivement pour nos quatre règles de 0.20, 0.57, 0.11 et 0.36. La 2ème règle est donc aussi la moins bonne de ce point de vue. La comparaison avec l'erreur de la règle majoritaire au nœud initial, qui vaut 0.56, montre que ce taux de 0.57 est mauvais. Ainsi, pour être vraiment informatif sur la pertinence d'une règle, le taux d'erreur doit être comparé avec celui d'une règle naïve de référence, ce qu'inclut précisément l'indice d'implication. Sur notre exemple, la comparaison des taux d'erreur avec celui de la règle naïve conduit à la même appréciation que l'indice d'implication.

Notons encore qu'en pratique pour pouvoir déceler le sur-apprentissage, on considère le taux d'erreur en généralisation ou en validation croisée. Rien n'empêche évidemment de faire de même avec la qualité implicative en calculant des intensités implicatives en généralisation.

Alternativement, on peut songer, dans l'esprit des critères BIC ou du MDL (minimum description length), à prendre en compte la complexité de la condition. Comme on cherche ici à minimiser l'indice d'implication, il s'agit en fait de pénaliser l'indice d'implication par la longueur k de la branche, l'idée étant que plus la condition est simple, moins on risque de se tromper dans l'assignation de cas à la feuille correspondante. Comme première proposition d'indice d'implication pénalisé, nous suggérons la forme suivante inspirée du critère BIC (Raftery 1995) :

$$\text{Imp}_{\text{pen}}(j) = res_d(j) + \ln(n)k \,. \qquad (5)$$

Règle	res_d	k	Imp_{pen}
1	-6.826	2	4.39
2	0.788	2	12.01
3	-4.456	2	6.76
4	-4.847	2	6.37
Homme $\Rightarrow$ marié	-7.119	1	**-1.51**
Femme $\Rightarrow$ célib.	-7.271	1	**-1.66**

TAB 8 – *Indice d'implication pénalisé par la complexité de la condition.*

Pour notre exemple, la pénalité pour chaque condition supplémentaire est ln(273) = 5.6 et les valeurs de l'indice pénalisé sont données au tableau 8.

Cette forme pénalisée n'apporte pas d'information nouvelle pour ce qui est de la comparaison des quatre règles initiales de notre exemple qui ont toutes la même longueur $k = 2$. Par contre elle s'avère utile pour évaluer la pertinence de la fusion des deux règles 1 et 2. Le tableau 8 fait ressortir que la règle fusionnée « homme ⇒ marié » est très supérieure aux autres, notamment aux règles 1 et 2. Elle donne un signal clair en faveur de cette fusion.

Au nœud initial, le résidu est nul ainsi que le nombre de conditions *k*. L'indice pénalisé associé à cette situation vaut donc 0. Ainsi, un indicateur pénalisé positif, indiquerait qu'on ne peut guère attendre de la règle qu'elle se comporte mieux en généralisation qu'un classement indépendant de toute condition. Pour notre exemple, aucune des quatre règles ne parait satisfaisante de ce point de vue. Les règles plus simples définies par le premier niveau de l'arbre devraient être plus performantes puisqu'elles donnent lieu à des valeurs négatives.

4 Choix de la conclusion dans les feuilles

Un dernier aspect que nous souhaitons aborder porte sur le recours à la statistique implicative pour le choix de la conclusion. L'idée, qui a notamment été exploitée par Zighed et Rakotomalala (2000, p 282-287), est de choisir dans chaque feuille la classe pour laquelle on maximise l'intensité d'implication.

A titre d'exemple, nous donnons au tableau 9 les variantes d'indices d'implication et les intensités correspondantes (avec correction pour la continuité) pour les trois conclusions possibles pour la règle 2. Les colonnes « célibataires » correspondent au choix de la classe majoritaire considérée jusqu'ici. Les quatre indices font ressortir la modalité « divorcé ou veuf » comme meilleure conclusion pour la deuxième règle, avec une intensité d'implication qui va de 89,1% pour l'indice de Gras (résidu standardisé) à 99,9% pour la déviance. On constate ici à nouveau, que l'indice de Gras semble sous-estimer légèrement l'intensité.

Résidu		*Indices*			*intensité*		
		marié	célib.	div./v	marié	célib	div./v
Standardisé	res_s	1.592	0.078	**-1.333**	0.043	0.419	**0.891**
Déviance	res_d	3.856	0.788	**-3.357**	0.000	0.099	**0.999**
Freeman-Tukey	res_{FT}	1.501	0.138	**-1.372**	0.054	0.398	**0.895**
Ajusté	res_a	2.402	0.117	**-2.011**	0.005	0.379	**0.968**
Imp_{pen}		15.07	12.01	**7.86**			

TAB. 9 – *Indices et intensités d'implication de la règle 2 selon la conclusion choisie.*

La dernière ligne du tableau donne l'indice pénalisé. Du point de vue de cet indice, la règle 2 optimale a une performance de 7.86 proche de celle des règles 3 et 4 (voir tableau 8). Elle reste par contre positive et ne remet donc pas en cause la fusion des règles 1 et 2.

5 Conclusion

Notre motivation dans cet article était de souligner l'intérêt de l'analyse implicative pour les arbres de classification. La discussion du concept d'indice d'implication nous a amené à montrer que les résidus utilisés dans le contexte de la modélisation de tables de contingence offraient des alternatives intéressantes à l'indice de Gras. La comparaison avec ces alternatives fait apparaître que l'indice de Gras tend à sous-estimer la force d'implication des règles. Nous avons ensuite discuté l'utilisation des indices d'implication pour l'évaluation individuelle des règles de classification. L'avantage principal par rapport au taux d'erreur et que ces indices incluent une comparaison avec une règle naïve indépendante de la condition. Dans l'optique d'évaluer la performance en généralisation, nous avons fait une première proposition d'indice pénalisé par la complexité de la condition qui devrait s'avérer intéressante pour évaluer la performance de la règle en généralisation a priori sur la base des seules données d'apprentissage. Enfin, nous avons brièvement discuté le recours à l'implication statistique pour le choix de la conclusion la plus pertinente pour une règle. En particulier, nous avons mis en évidence que la modalité la plus fréquente n'est pas nécessairement le meilleur choix.

Cet article n'est qu'une amorce de discussion. Les variantes de l'indice de Gras, les corrections pour la continuité, l'indice pénalisé sont autant d'éléments qui méritent une étude plus approfondie tant sur le plan théorique qu'empirique. Par ailleurs, si nous avons à la section 4 souligné l'intérêt des indices d'implication pour le choix de la conclusion optimale dans une feuille de l'arbre, nous pensons également étudier l'intérêt de ces indices, peut-être dans leur forme pénalisée, pour l'induction proprement dite de l'arbre. L'étude du lien entre la performance individuelle des règles et la performance globale de l'arbre de classification est également un sujet que nous envisageons d'approfondir.

Références

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

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

Blanchard, J., P. Kuntz, F. Guillet, et R. Gras (2004). Mesure de la qualité de règles d'association par l'intensité d'implication entropique. *Revue des nouvelles technologies de l'information RNTI E-1*, 33-43.

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

Gras, R., P. Kuntz, et H. Briand (2001). Les fondements de l'analyse statistique implicative et leur prolongement pour la fouille de données. *Mathématique et Sciences Humaines 120* (154-155), 9-29.

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.

Gras, R. et H. Ratsima-Rajohn (1996). L'implication statistique, une nouvelle méthode d'analyse de données. *RAIRO Recherche Opérationnelle 30* (3), 217-232.

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. (1996). *L'implication statistique: Nouvelle méthode exploratoire de données*. Recherches en didactique des mathématiques. Grenoble: La pensée sauvage.

Gras, R. (2000). Les fondements de l'analyse statistique implicative. In R. Gras et M. Bailleul (Eds.), *La fouille dans les données par la méthode d'analyse statistique implicative*, pp. 11-32. Ecole polytechnique de l'Université de Nantes, IRIN et IUFM Caen, ISBN 2-9516505-0-7.

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

Lerman, I. C., R. Gras, and H. Rostam (1981b). Elaboration d'un indice d'implication pour données binaires II. *Mathématiques et sciences humaines* (75), 5-47.

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.

Suzuki, E. and Y. Kodratoff (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow and M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, 1998, Proceedings*, LNAI 1510, pp. 10-18. Berlin: Springer.

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

Summary

This paper highlights the interest of implicative statistics for classification trees. In a first part we show that residuals used in the modeling of contingency tables provide interesting alternatives to Gras' index of implication. We then consider two main usages of these indexes. The first is purely descriptive and concerns the a posteriori individual evaluation of the classification rules. The second usage, considered for instance by Zighed and Rakotomalala (2000), relies upon the intensity of implication to define the conclusion in each leaf of the induced tree.