

Données déséquilibrées, entropie décentrée et indice d'implication

Gilbert Ritschard*, Djamel A. Zighed**, Simon Marcellin**

*Département d'économétrie, Université de Genève

**Laboratoire ERIC, Université de Lyon 2

gilbert.ritschard@unige.ch, {abdelkader.zighed,simon.marcellin}@univ-lyon2.fr

<http://www.unige.ch/ses/metri/>, <http://eric.univ-lyon2.fr>

Résumé. Cet article porte sur l'induction d'arbres de classification pour des données déséquilibrées, c'est-à-dire lorsque certaines catégories de la variable à prédire sont beaucoup plus rares que d'autres. Plus particulièrement nous nous intéressons à deux aspects: d'une part, à définir des critères de construction de l'arbre qui exploitent efficacement la nature déséquilibrée des données, et d'autre part la pertinence de la conclusion à associer aux feuilles de l'arbre. Nous avons récemment abordé cette problématique sous deux angles indépendants: l'un était axé sur le recours à des entropies décentrées, l'autre s'appuyant sur des mesures d'intensités d'implication issues de l'ASI. Nous nous proposons ici de comparer et d'établir les similarités entre ces deux approches.

1 Introduction

Qu'il s'agisse d'induire un arbre, ou d'associer une conclusion à chacune de ses feuilles, les critères utilisés supposent en général implicitement une importance égale des modalités de la variable à prédire. Ainsi, des algorithmes comme CART (Breiman et al., 1984) ou C4.5 (Quinlan, 1993) utilisent comme critère l'amélioration d'une entropie classique, c'est-à-dire centrée sur la distribution uniforme correspondant à l'équiprobabilité des modalités. Le résultat est qu'on obtient ainsi des segmentations en classes dont les distributions tendent à s'écartier le plus possible de la distribution uniforme. De même pour le choix de la conclusion, le critère communément utilisé est simplement la règle majoritaire qui n'a évidemment de sens que si chaque modalité a la même importance. On le voit donc, cette distribution égalitaire des modalités joue le rôle de situation la moins désirable. Mais est-ce vraiment le cas ? Et sinon, de quelles solutions dispose-t-on pour d'une part favoriser les écarts à une distribution non centrée — représentative de la situation la moins désirable — et d'autre part choisir la conclusion la plus pertinente par rapport à cette référence la moins désirable ?

Une première solution nous est fournie par l'indice d'implication dont nous avons montré dans Ritschard (2005) et Pisetta et al. (2007) comment il pouvait s'utiliser avec les arbres de décision. En effet, cet indice est en fait un résidu, soit un écart par rapport à l'indépendance qui est caractérisée dans les arbres par la distribution au nœud initial. Ainsi au lieu de mesurer des écarts par rapport à la distribution uniforme, on mesure des écarts par rapport à cette distribution initiale. Rien n'empêche cependant de considérer des résidus par rapport à d'autres distributions. Voir à ce sujet l'indice d'écart à l'équilibre de Blanchard et al. (2005) et sa généralisation dans Lallich et al. (2005). Une seconde solution consiste à utiliser des entropies décentrées (Marcellin et al., 2006; Zighed et al., 2007) qui généralisent les entropies classiques en les paramétrant par le point où elles prennent leur maximum, laissant ainsi à l'utilisateur la possibilité de déterminer le point d'incertitude maximale.

Nous nous proposons dans ce papier de comparer ces deux approches en discutant leurs avantages respectifs comme critère de construction de l'arbre ainsi que comme critère de choix de la conclusion des règles. Notre discussion nous amènera à proposer une solution hybride où l'on utilise l'entropie décentrée pour induire l'arbre, et l'indice d'implication pour assigner une décision à chaque feuille.

Avant d'entrer dans cette discussion, il convient d'explicitier l'intérêt de la problématique, à savoir les circonstances où la distribution uniforme ne serait pas la moins désirable. C'est l'objet de la section 3, après que nous ayons posé le cadre formel à la section 2. Dans la section 4 nous introduisons un jeu de données qui nous servira d'illustration et rappelons le principe des arbres de décision. A la section 5 nous rappelons les définitions introduites dans Ritschard (2005) sur la notion d'indice d'implication dans le contexte des arbres de décision et examinons

la possibilité de l'utiliser comme critère d'optimalité pour les éclatements successifs lors de la construction de l'arbre. Nous rappelons aussi son intérêt pour l'attribution de la conclusion aux feuilles de l'arbre. La section 6 quant à elle rappelle la forme de l'entropie décentrée introduite dans Marcellin et al. (2006) et Zighed et al. (2007) et commente son usage, en particulier comme critère de développement de l'arbre. La discussion comparative fait l'objet de la section 7. Enfin nous concluons à la section 8.

2 Cadre formel et notations

On se place dans un cadre supervisé où disposant d'une variable dépendante y , dite aussi variable réponse ou à prédire, on cherche à caractériser une fonction $f(x_1, x_2, \dots)$ — un arbre de décision dans notre cas — qui permette de prédire y à partir d'un ensemble x_1, x_2, \dots de variables explicatives (prédicteurs) catégorielles, ordinales ou quantitatives. On s'intéresse ici au cas où la variable réponse est catégorielle avec ℓ modalités y_1, \dots, y_ℓ . Par exemple, s'agissant de diagnostiquer un cancer on aura $y_1 = \text{'a le cancer'}$ et $y_2 = \text{'pas de cancer'}$. Notre propos concerne cependant plus particulièrement les situations où la réponse prend plus de 2 modalités, ce qui est par exemple le cas si l'on retient une catégorie $y_3 = \text{'requiert une analyse supplémentaire'}$ en plus des deux classes précédentes.

Comme y est catégorielle, la prédiction de sa modalité est une *classification*. On assigne un cas j à la classe (modalité) de y que l'on prédit à partir des valeurs x_{j1}, x_{j2}, \dots que prennent les prédicteurs pour ce cas j . Dans ce type de contexte, il est naturel de s'intéresser aux erreurs de classification et à leurs conséquences, c'est-à-dire à leurs coûts. On note c_{ik} le coût de l'erreur réalisée en affectant à la classe i un cas de la classe k , et par $\mathbf{C} = (c_{ik})_{i,k=1,\dots,\ell}$ la matrice $\ell \times \ell$ réunissant l'ensemble de ces coûts. Par exemple, dans le cas de $\ell = 3$ classes, si le coût d'affecter de façon erronée un cas à la première classe est le double des autres erreurs d'affectation et que les classifications correctes n'engendrent pas de frais, on a $c_{12} = c_{13} = 2$ et sinon $c_{ik} = 1$ pour $i \neq k$, et la matrice de coût s'écrit

$$\mathbf{C} = \begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}. \quad (1)$$

Si l'on décide d'assigner à une classe i tous les cas d'un groupe où la distribution effective selon les classes est $\mathbf{p} = (p_1, \dots, p_\ell)'$, l'espérance du coût de l'erreur engendré par une classification en i est $E(\text{coût}_i) = \sum_k p_k c_{ik}$. En notant $\boldsymbol{\iota}$ le vecteur unitaire $(1, \dots, 1)'$, on écrira simplement $\mathbf{p} = \frac{1}{\ell} \boldsymbol{\iota}$ la distribution uniforme de ℓ classes.

3 Coût d'erreur et données déséquilibrées

En apprentissage machine, les *a priori* sont multiples, et leurs conséquences souvent mal contrôlées, conduisent à des conclusions erronées ou des erreurs de prédiction. Nous allons décrire quelques situations typiques où des hypothèses implicites manifestement erronées sont totalement ignorées par l'utilisateur.

Asymétrie des coûts d'erreurs. Un premier élément est le coût des erreurs de décision. En effet, si le coût c_{ik} de l'erreur réalisée en affectant à la classe i un cas de la classe k diffère selon le couple (i, k) , l'incertitude maximale devrait correspondre à la distribution $\mathbf{w} = (w_1, \dots, w_\ell)$ pour laquelle l'espérance du coût $E(\text{coût}_i) = \sum_k w_k c_{ik}$ de chaque décision i serait la même : $E(\text{coût}_i) = \gamma$ pour tout i . Par exemple avec une variable à prédire comprenant trois classes et la matrice de coût (1), la distribution \mathbf{w} la plus incertaine est la solution — non équilibrée — du système $\mathbf{C}\mathbf{w} = \gamma \boldsymbol{\iota}$ avec $\sum_i w_i = 1$, soit

$$\begin{cases} 2w_2 + 2w_3 = \gamma \\ w_1 + w_3 = \gamma \\ w_1 + w_2 = \gamma \end{cases} \Rightarrow \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.2 \\ 0.2 \end{pmatrix} \quad \text{et} \quad \gamma = E(\text{coût}_i) = 0.8.$$

En apprentissage machine, le coût de l'erreur d'affectation n'est pas toujours explicitement pris en compte ni même considéré. Cela conduit, généralement, à admettre implicitement que le coût c_{ik} de l'erreur réalisée en

affectant un cas de la classe k à la classe i est le même que pour l'erreur consistant à affecter un cas de i à k , voire même souvent à admettre qu'il est le même quelle que soit la classe d'affectation. Autrement dit les hypothèses généralement retenues peuvent être résumées ainsi :

1. non négativité : $c_{ik} \geq 0$ pour tout couple de classe i, k .
2. symétrie du coût de l'erreur : $c_{ik} = c_{ki}$ pour tout couple de classes i, k ,
3. coût constant : $c_{ik} = c$ pour tout couple de classes $i, k; i \neq k$,
4. coût nul pour une bonne classification : $c_{ii} = 0$, pour tout i .

De ce fait, la matrice \mathbf{C} des coûts prend une forme particulière, à savoir symétrique et non négative. Ces hypothèses sont bien évidemment lourdes de conséquences et dans la plupart des cas non justifiées comme nous allons le voir à travers un exemple. Considérons pour cela un problème typique issu du monde réel auquel nous pourrions être confrontés en apprentissage. Dans le domaine de la santé et tout particulièrement dans le cancer du sein, annoncer à une patiente qu'elle n'a pas de cancer alors qu'elle en porte un, ce qui représente un faux négatif, pourrait avoir des conséquences désastreuses sur elle et pourrait entraîner son décès. En revanche, un faux positif, qui revient à détecter un cancer chez une patiente alors qu'elle n'en a pas, engendre des conséquences moins graves. Il s'agirait, le plus souvent, d'une multiplication inutile d'exams médicaux, d'un stress chez la personne etc. Il est par conséquent évident dans l'optique du patient que le coût d'un faux positif est inférieur à celui d'un faux négatif. Dans ce cas, les coûts sont asymétriques. D'où cette fois-ci une matrice des coûts non négative et non symétrique

$$c_{ik} \neq c_{ki} \text{ et } c_{ik} \geq 0 \text{ pour tout } i, k$$

comme par exemple la matrice \mathbf{C} donnée en (1).

La difficulté réside évidemment dans la détermination des coûts. De plus on peut relever que le système $\mathbf{C}\mathbf{w} = \gamma\mathbf{1}$ n'admet pas toujours une solution unique non négative. A défaut d'un choix clairement déterminé de la matrice \mathbf{C} , la distribution marginale semble une solution pour le moins plus réaliste que la distribution uniforme, l'erreur réalisée lors de l'affectation d'un cas appartenant à une classe rare portant en règle générale plus à conséquence que celle réalisée avec l'un des nombreux cas d'une classe fréquente.

Le déséquilibre des classes. Dans le domaine de l'analyse des transactions bancaires, la proportion de transactions frauduleuses est habituellement très faible au regard du volume global des transactions. Les taux de fraude ne dépassent guère le 1% et, fort heureusement, restent souvent en deçà de ce seuil. Dans ce cadre, on parle généralement de classes déséquilibrées.

Il est admis que le principe d'affectation d'un individu à une classe donnée repose sur la règle de la minimisation du risque du nombre d'erreurs. Ainsi, on affecte généralement un individu à la classe majoritaire. Du fait du déséquilibre des classes, la stricte application de la règle de minimisation du nombre d'erreurs conduit à considérer que toutes les transactions sont non frauduleuses, ce qui évidemment n'est pas très utile dans une optique de détection des fraudes. Conscient de cela, les utilisateurs adoptent le plus souvent l'une des positions suivantes :

1. Soit ils définissent explicitement le coût de rater une fraude et celui d'en prédire une à tort et cherchent à minimiser le coût total des erreurs au lieu de leur nombre.
2. Soit ils contournent la règle de minimisation du nombre d'erreurs en considérant que dès lors que la probabilité d'avoir une transaction frauduleuse atteint un certain seuil δ ($< 50\%$) fixé par l'utilisateur, alors l'affectation est faite sur la classe frauduleuse. Ils prennent ainsi implicitement en compte une fonction de coût non symétrique dans laquelle le coût de rater une transaction frauduleuse est plus élevé.
3. Soit, ce qui est une autre façon d'introduire implicitement des coûts d'erreurs de classification, ils rééquilibrent les classes dans les données d'apprentissage de sorte à travailler sur un échantillon où les deux classes sont identiquement représentées et où la règle de minimisation du nombre d'erreurs n'est plus transgressée.

La taille des échantillons. En plus de la nécessité de mieux exploiter la nature déséquilibrée des données, un autre aspect important est également le plus souvent ignoré par les critères fondés sur les entropies classiques. Il s'agit des effectifs concernés par les distributions. En apprentissage nous travaillons sur un échantillon de taille fini. Les règles de décision induites par un arbre de décision ou une règle d'association par exemple, couvrent

des portions différentes d'individus. Utiliser une mesure d'entropie à l'état brut pour apprécier la qualité d'une règle s'avère inconsistant. En effet, si nous disposons de deux règles qui conduisent à la même distribution de probabilité des classes, elles auront la même valeur d'entropie quel que soit l'effectif couvert par l'une ou l'autre. Or, les probabilités à la base du calcul des entropies sont estimées par les fréquences empiriques au niveau de l'échantillon couvert par la règle. L'entropie qui traduit l'incertitude sur la décision ne devrait pas être la même sous prétexte que les vecteurs de distribution des probabilités sont identiques. Il faut également introduire la taille de l'échantillon car à distributions équivalentes, notre incertitude en matière de décision devrait être plus faible sur un effectif plus important. Il importe donc de distinguer l'entropie théorique de l'entropie empirique, et de tenir compte des propriétés statistiques de cette dernière en tant qu'estimation de la première.

Les trois situations que nous venons de décrire expriment clairement la nécessité de mieux réfléchir aux hypothèses sous-jacente lors de la mise en œuvre d'un algorithme d'apprentissage. Dans les méthodes qui génèrent des règles de décision du type «SI *condition* ALORS *conclusion*» comme les arbres de décision (Breiman et al., 1984; Quinlan, 1993), ou les graphes d'induction (Zighed et Rakotomalala, 2000), les mesures d'entropie sont fréquemment utilisées. Or, justement, celles-ci reposent sur des hypothèses implicites peu compatibles avec les situations où la distribution équilibrée n'est pas la moins désirable.

Solutions proposées dans la littérature. Plusieurs approches ont été proposées pour traiter l'asymétrie des coûts et le déséquilibre des classes. On peut distinguer quatre orientations (Barandela et al., 2003). La première introduit des approches explicitement sensibles au coût des erreurs. Cela permet soit d'instancier la matrice des coûts, soit de rééquilibrer artificiellement les effectifs des classes. Pour la seconde, une stratégie de (ré-)échantillonnage permet de sous-représenter la classe majoritaire ou sur-représenter la classe minoritaire (Provost, 2000; Witten et Frank, 2005). Dans la troisième, des méthodes comme METACOST produisent plusieurs instances d'un même classifieur par *bootstrap*, réétiquettent chaque exemple par un critère de vote à la majorité des classifieurs et construisent ensuite un nouveau modèle sur l'échantillon réétiqueté (Domingos, 1999). Enfin, la quatrième approche remplace la fonction d'entropie par un critère incluant explicitement la notion de coût (Chen et al., 2004). Les solutions envisagées dans cet article s'apparentent à cette dernière dans la mesure où elles conduisent également d'une part à modifier le critère de développement de l'arbre et d'autre part à remplacer le principe du vote majoritaire simple pour le choix de la conclusion.

4 Données illustratives et principe des arbres de décision

Pour illustrer notre propos, nous reprenons les données fictives utilisées dans Ritschard (2005) et récapitulées au tableau 1. La variable à prédire est l'état civil, le sexe et le secteur d'activité étant les prédicteurs disponibles.

état civil	homme			femme			total
	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	
marié	50	40	6	0	14	10	120
célibataire	5	5	12	50	30	18	120
divorcé/veuf	5	8	10	6	2	2	33
total	60	53	28	56	46	30	273

TAB. 1 – *Données illustratives.*

Les arbres de classification sont des outils 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 , l'état civil dans notre exemple) diffère le plus possible d'une classe à l'autre de la partition. 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 dans

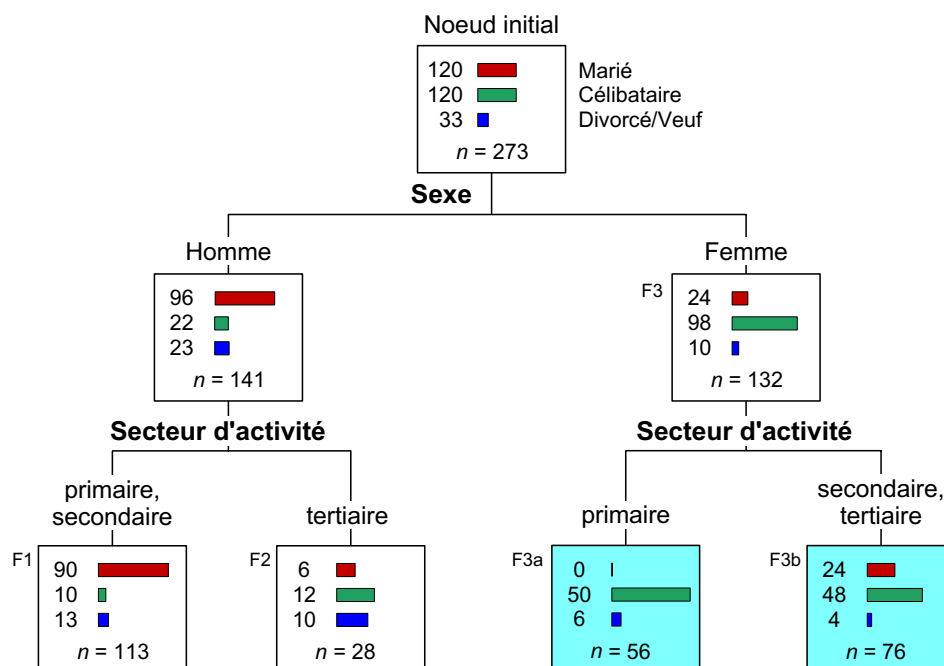


FIG. 1 – Arbre induit. Feuilles F1, F2, F3 avec indice implication, F1, F2, F3a, F3b avec entropie décentrée.

chaque feuille (nœud terminal) de l’arbre. On retient classiquement pour cela la valeur la plus fréquente, mais nous reviendrons précisément sur ce point.

Pratiquement, on relève dans chaque feuille j , $j = 1, \dots, q$, 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 2). On peut noter que la marge de droite de ce tableau qui donne le total $n_{i\cdot}$ des lignes correspond en fait à la distribution des cas dans le nœud initial de l’arbre. Les $n_{\cdot j}$ désignent les totaux des colonnes.

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

TAB. 2 – Table de contingence croisant les états de la réponse y avec les feuilles de l’arbre.

5 Indice d’implication

L’indice d’implication (voir par exemple Gras et al., 2004, p. 19) d’une règle se définit à partir des contre-exemples. Dans le cas des arbres de classification il s’agit dans chaque feuille (colonne du tableau 2) du nombre de cas qui ne sont pas dans la catégorie qui lui a été attribuée. 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 nombre de cas qui vérifient cette conclusion dans 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 $n_{\bar{b}j}^e = n_{\bar{b}.} n_{.j}$. Ce paramètre $n_{\bar{b}j}^e$ 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 $n_{.j}$ 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 $n_{\bar{b}j} - n_{\bar{b}j}^e$ entre les nombres de contre-exemples observés et attendus sous l'hypothèse H_0 , standardisé par l'écart type, soit, en ajoutant la correction pour la continuité en vue de la comparaison avec la loi normale

$$\text{Imp}(j) = \frac{n_{\bar{b}j} - n_{\bar{b}j}^e + .5}{\sqrt{n_{\bar{b}j}^e}} \quad (2)$$

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

$$\text{Imp}(j) = \frac{-(n_{bj} - n_{bj}^e) + .5}{\sqrt{n_{.j} - n_{\bar{b}j}^e}} \quad (3)$$

Une valeur positive de l'indice indique que la règle fait moins bien que le hasard et n'apporte donc aucune information implicative. Seules les valeurs négatives ont donc un intérêt. Plus l'indice — l'écart par rapport au hasard — est grand (en valeur absolue), plus la force implicative de la règle est forte.

Dans (Ritschard, 2005), nous avons proposé des variantes inspirées des résidus utilisés en modélisation de tables de contingence multidimensionnelles. Il s'agit du résidu déviance, du résidu ajusté d'Haberman et du résidu de Freeman-Tukey qui ont une variance plus proche de 1 que le résidu standardisé utilisé par Gras. Le premier a cependant un comportement tendant vers 0 quand le nombre de contre-exemples s'approche de 0 qui le disqualifie (Pisetta et al., 2007). Les deux autres évoluent de façon similaire à l'indice de Gras tout au moins du point de vue qui nous intéresse ici de l'ordre de préférence des conclusions que suggèrent les valeurs de l'indice. Nous nous contentons donc ci-après de discuter l'usage de l'indice de Gras.

5.1 Gain d'implication comme critère d'optimalité des éclatements

L'indice permet de mesurer la force implicative de la règle. On peut alors songer à l'exploiter comme critère de développement de l'arbre. L'idée est de rechercher à chaque nœud l'éclatement qui produirait le meilleur gain en termes de force implicative des règles, en admettant évidemment qu'on retienne à chaque nœud la conclusion qui maximise l'intensité d'implication. On se heurte cependant ici à une difficulté d'agrégation. En effet, s'il est aisé de calculer l'indice d'implication avant l'éclatement, on se retrouve après l'éclatement avec plusieurs nœuds et donc un ensemble de valeurs d'indices d'implication qu'il nous faut synthétiser en une seule valeur qui puisse être comparée avec l'indice d'implication avant l'éclatement. Une possibilité est de prendre simplement une moyenne pondérée par les effectifs des nœuds concernés. Une autre solution, qui ferait sens si l'on est intéressé en priorité à obtenir quelques règles très fortes tout en s'accommodant de règles peu implicatives, est de retenir le maximum des intensités obtenues. Pour rester dans la logique de l'indice d'implication, une troisième solution d'indice d'implication pour l'ensemble S de sommets résultant de l'éclatement est (en incluant la correction pour la continuité)

$$\text{ImpT}(S) = \frac{\sum_{j \in S} n_{\bar{b}j} - \sum_{j \in S} n_{\bar{b}j}^e + .5}{\sqrt{\sum_{j \in S} n_{\bar{b}j}^e}} = \frac{\sum_{j \in S} (n_{\bar{b}j} - n_{\bar{b}j}^e) + .5}{\sqrt{\sum_{j \in S} n_{\bar{b}j}^e}} \quad (4)$$

¹Notons que b peut évidemment varier d'une colonne à l'autre.

Attribut utilisé	nbre sommets	moyenne pondérée	maximum	ImpT
sexe	2	4.17	4.59	5.94
secteur	3	0.82	1.34	1.50
primaire	2	0.31	0.44	0.46
tertiaire	2	0.79	0.82	1.09

TAB. 3 – Gains de force implicative pour les éclatements possibles au premier niveau.

soit l'écart standardisé entre le nombre total de contre-exemples observés des règles et le total attendu.

Pour notre exemple, nous donnons au tableau 3 le gain de force implicative apporté par les différents éclatements possibles au premier niveau. Le gain est la différence entre la valeur de l'indice au nœud que l'on veut éclater (soit 0 au nœud initial) et l'indice synthétique pour les nœuds résultant de l'éclatement. Le sexe s'impose clairement comme meilleur attribut prédictif. Il est intéressant de relever que le gain mesuré avec l'indice total est en règle générale plus fort que l'écart par rapport au maximum.

Attribut utilisé	nbre sommets	moyenne pondérée	maximum	ImpT
Sommet : Homme				
secteur	3	-0.71	0.22	1.18
primaire	2	-1.30	-0.10	0
tertiaire	2	0.48	1.23	1.18
Sommet : Femme				
secteur	3	-1.83	-0.15	0
primaire	2	-1.46	-0.15	0
tertiaire	2	-0.81	-0.01	0

TAB. 4 – Gains de force implicative pour les éclatements possibles au deuxième niveau.

On procède donc à l'éclatement selon le sexe, et l'on donne au tableau 4 les gains possibles au niveau 2 pour chacun des sommets "Homme" et "Femme". Pour les femmes, aucun gain de force implicative n'est possible avec la seule variable qui nous reste à savoir le secteur d'activité. La raison en est simplement que quelque soit l'éclatement, la catégorie pour laquelle on a l'implication la plus forte reste la même (célibataire) dans tous les nœuds qu'on obtient. Pour les hommes, il en est de même si l'on segmente entre le secteur primaire et le reste. Par contre, une segmentation en deux, tertiaire contre le reste ou en trois, permet un gain égal en termes d'implication totale. Le partage en deux paraît cependant plus intéressant puisqu'il se traduit, contrairement à l'éclatement en 3, par un gain positif également en termes d'implication moyenne.

5.2 Choix de la conclusion des règles

Chaque feuille (nœud terminal) de l'arbre caractérise une règle dont la prémisse est définie par les conditions d'embranchement le long du chemin menant du nœud initial à la feuille, la conclusion de la règle correspondant à la modalité assignée à la feuille. Comme déjà mentionné, le choix se porte de façon classique sur la modalité la plus fréquente. Dans certaines circonstances, il est plus pertinent de retenir la modalité assurant la plus forte implication. Il en est en particulier ainsi dans le contexte du ciblage où il s'agit de déterminer les profils types de chaque modalité de la variable cible y , et non pas, comme en classification, de prévoir la modalité que prendra un individu avec un profil donné.

Notons que l'usage de l'indice d'implication pour le développement de l'arbre suppose implicitement que la conclusion attribuée est dans chaque feuille la modalité qui assure la plus forte valeur négative de l'indice d'implication. La conclusion est ainsi dans ce cas automatiquement déterminée. A titre d'exemple, en induisant l'arbre avec l'indice d'implication on obtient les trois règles du tableau 5 qui correspondent aux feuilles F1, F2

et F3 dans la figure 1. On notera en particulier que la conclusion attribuée à la 2ème règle n'est pas la modalité majoritaire.

Le recours à l'indice d'implication pour le choix des conclusions reste cependant également possible pour des arbres induits selon d'autres critères.

Règle	Condition	Conclusion
R1	Homme et secteur primaire ou secondaire	→ marié
R2	Homme et secteur tertiaire	→ divorcé
R3	Femme	→ célibataire

TAB. 5 – Meilleures règles en termes de force implicative.

6 Entropie décentrée

Les mesures d'entropie ont été définies mathématiquement par un ensemble d'axiomes en dehors du contexte de l'apprentissage machine. On peut trouver des travaux détaillés dans Rényi (1960) et Aczél et Daróczy (1975). Leur transfert vers l'apprentissage s'est fait de manière hâtive et sans prêter trop attention à la pertinence de leurs axiomes fondateurs. Ainsi, nous avons souligné dans Zighed et al. (2007) l'intérêt de relâcher l'axiome exigeant que l'entropie soit maximale à la distribution uniforme, et par suite évidemment l'axiome de symétrie stipulant que l'entropie doit être insensible à l'ordre des probabilités constituant la distribution. En nous fondant sur une axiomatique plus générale, nous avons proposé une entropie décentrée d'une distribution (p_1, \dots, p_ℓ) généralisant l'entropie quadratique dans le cas où $\ell = 2$. Sa forme théorique, standardisée pour que sa valeur maximale soit égale à 1, est :

$$h_w(p_1, p_2, \dots, p_\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{p_i(1-p_i)}{(-2w_i+1)p_i + w_i^2} \quad (5)$$

où $\mathbf{w} = (w_1, \dots, w_\ell)$ est un vecteur de paramètres caractérisant la distribution d'incertitude maximale. On obtient une version empirique en remplaçant les p_i par leurs estimations de Laplace $\hat{p} = (n_i + 1)/(n + \ell)$. D'autres formes d'entropies décentrées ont également été proposées par Lallich et al. (2007). En lieu et place d'indicateurs fondés sur des entropies, on pourrait également envisager par exemple des mesures de divergences telles que celles discutées dans Jaroszewicz et Simovici (2001).

6.1 Utilisation de l'entropie décentrée

Par rapport à l'indice d'implication qui compare le nœud obtenu au nœud initial en termes de distribution entre exemples et contre-exemples, l'entropie décentrée ne privilégie pas de catégorie particulière et compare l'ensemble de la distribution. Elle ne préjuge donc pas de la catégorie qui sera assignée au nœud.

Attribut utilisé	nbre sommets	Gini	entropie décentrée	
			théorique	empirique
sexe	2	0.150	0.210	0.201
secteur	3	0.016	0.024	0.023
primaire	2	0.001	0.003	0.003
tertiaire	2	0.011	0.017	0.016

TAB. 6 – Gains d'entropie pour les éclatements possibles au premier niveau.

Nous donnons au tableau 6 les gains d'entropie pour les divers éclatements possibles au premier niveau. A titre de comparaison nous donnons les gains obtenus en termes de l'indice de Gini, qui est l'entropie quadratique

classique, et de l'entropie décentrée théorique et empirique. La version théorique est obtenue en remplaçant dans la formule (5) les p_i par les fréquences observées, et la version empirique en les remplaçant par les estimations de Laplace.

Pour ce premier éclatement, les entropies classiques et décentrées conduisent au même résultat. Le sexe est la variable à retenir, tout comme il l'était avec le gain d'implication. On peut relever par ailleurs pour les entropies décentrées que le gain tend à être moins fort en termes de mesure empirique que théorique.

Attribut utilisé	nbre sommets	Gini	entropie décentrée	
			théorique	empirique
Sommet : Homme				
secteur	3	0.084	0.111	0.089
primaire	2	0.020	0.025	0.012
tertiaire	2	0.082	0.106	0.098
Sommet : Femme				
secteur	3	0.042	0.075	0.048
primaire	2	0.042	0.073	0.052
tertiaire	2	0.013	0.019	0.012

TAB. 7 – Gains d'entropie pour les éclatements possibles au deuxième niveau.

Le tableau 7 propose la même comparaison pour les éclatements possibles au second niveau. Les résultats divergent ici selon le type d'entropie utilisé. Pour ce qui est du sommet "Homme", Gini et l'entropie décentrée théorique sélectionnerait l'éclatement en trois, tandis que la version empirique de l'entropie décentrée privilégie l'éclatement qui oppose le secteur tertiaire aux deux autres secteurs. Il n'y a donc que ce dernier indice qui donne un résultat concordant avec l'optique implication discutée précédemment.

Pour le sommet "Femme", il y a également divergence, l'entropie empirique favorisant à nouveau un éclatement en deux plutôt qu'en trois, soit le secteur primaire contre les deux autres.

Notons qu'à nouveau les gains sont plus faibles avec la version empirique, les écarts étant ici plus importants en raison des effectifs plus faibles des nœuds. C'est la sensibilité aux effectifs que nous souhaitons.

6.2 Choix de la conclusion selon la contribution à l'entropie

L'entropie qui mesure l'écart entre deux distributions ne se prête pas en tant que telle à la mesure de l'intérêt de chaque modalité dans la feuille. La contribution de chaque modalité à cette entropie nous donne par contre une information utile de ce point de vue. La seule valeur de cette contribution n'est cependant pas suffisante. Il nous faut tenir compte également du signe de l'écart. En effet, une faible contribution à l'entropie indique une classe qui se démarque fortement de sa proportion marginale, mais cet écart est pertinent seulement si l'effectif observé dépasse l'effectif attendu en cas d'indépendance. On propose alors de sélectionner dans chaque feuille j la modalité qui maximise le critère

$$\max_i \eta_{w,ij} = \text{signe}(n_{ij} - n_{ij}^e) (1 - h_{w,ij}), \quad j = 1, \dots, q \quad (6)$$

où $h_{w,ij}$ est la contribution effective de la modalité i à l'entropie de la feuille j , n_{ij} le nombre observé de cas de modalité i dans la feuille j , et n_{ij}^e le nombre attendu sous l'hypothèse d'indépendance. On retient ainsi la modalité dont la contribution $h_{w,ij}$ est la plus faible parmi celles dont on observe plus de cas qu'attendus par hasard. Notons que, l'écart $n_{ij} - n_{ij}^e$ étant nécessairement non négatif pour au moins un i , la fonction "signe" pourrait tout aussi bien être remplacée par la fonction logique $(n_{ij} - n_{ij}^e > 0)$ qui prend la valeur 1 lorsque l'écart est positif et 0 sinon.

Le tableau 8 donne les valeurs de $\eta_{w,ij}$ pour les 4 feuilles de l'arbre de la figure 1. On note que les conclusions sélectionnées concordent avec celles d'implication maximale.

	F1	F2	F3a	F3b
marié	0.59	-0.79	-0.09	-0.93
célibataire	-0.42	-1.00	0.39	0.88
divorcé/veuf	-1.00	0.81	-1.00	-0.95

TAB. 8 – Contributions à l'entropie décentrée des feuilles.

7 Discussion

Nous avons vu que tant l'indice d'implication que l'entropie décentrée pouvaient servir de critère de développement de l'arbre. Les deux approches fournissent également des éléments permettant d'attribuer aux feuilles une conclusion appropriée dans un contexte de données déséquilibrées où le rappel de catégories faiblement représentées est plus important que le taux total d'erreurs de classification. Les deux approches ne sont pas pour autant équivalentes. L'indice d'implication oppose la classe pour laquelle on a l'implication maximale aux autres, tandis que l'entropie asymétrique prend en compte tout le détail de la distribution. Quels sont alors les avantages et inconvénients respectifs ?

7.1 Avantages et limites

Considérons tout d'abord l'optique du développement de l'arbre. De ce point de vue, l'indice d'implication a quelque analogie avec le critère 'Twoing' de CART (Breiman et al., 1984) qui pour chaque éclatement possible cherche la partition en deux des valeurs de la variable cible qui maximise l'indice de Gini. L'avantage est qu'on a ainsi un critère qui devient plus robuste en se fondant sur des effectifs moins dispersés qui le rendent notamment moins sensible aux variations à l'intérieur de chacune des deux classes. Le même argument vaut pour l'indice d'implication bien que dans ce cas la première classe n'ait toujours qu'une seule catégorie.

Utiliser l'indice d'implication comme critère d'éclatement présuppose que la catégorie maximisant l'implication sera assignée au nœud. Ceci assure évidemment une cohérence à la procédure, mais limite évidemment aussi l'usage de l'arbre obtenu au contexte où ce choix de la conclusion selon la force implicative s'avère pertinent.

Pour ce qui est de l'entropie décentrée, elle mesure la proximité à la distribution de référence, proximité que l'on cherche à minimiser de sorte à obtenir des distributions aussi différentes que possible de la référence. On peut ici faire l'analogie avec le critère du khi-deux utilisé par l'algorithme CHAID (Kass, 1980) qui conduit également à choisir la segmentation pour laquelle les distributions s'écartent le plus possible de celle d'indépendance. La différence est que dans CHAID le référentiel change à chaque nœud puisque le critère consiste à s'éloigner le plus possible de la distribution du nœud qu'on éclate, tandis qu'avec le gain d'entropie décentré on cherche à se démarquer de la distribution du nœud initial qui reste la même à toutes les étapes du calcul.

Comme l'illustre en particulier notre exemple, les deux approches conduisent à des arbres relativement semblables même s'il l'on peut imaginer des situations peu claires où les éclatements proposés peuvent différer. Remarquons tout de même que l'indice d'implication ne propose pas d'éclatement lorsque les conclusions prévues pour les nœuds qui en résulteraient sont les mêmes. Ainsi, dans la figure 1, le développement s'arrête à la feuille F3 avec l'indice d'implication, alors même qu'on réalise un gain d'entropie décentrée en éclatant le nœud en F3a et F3b. On peut donc s'attendre à obtenir des arbres moins complexes avec l'indice d'implication qu'avec l'entropie décentrée.

Sur le plan de la complexité de calcul, la mise en œuvre de l'entropie décentrée semble un peu plus immédiate, l'indice d'implication nécessitant de tester à chaque fois les différentes possibilités d'opposer une catégorie aux autres. Ceci n'affecte la complexité de l'algorithme que par un facteur multiplicatif c correspondant au nombre de catégories de la variable cible.

Enfin, on a relevé à la section 3 l'importance de disposer de critères qui soient sensibles à la taille des effectifs. L'entropie décentrée l'est dans sa forme empirique, la sensibilité à l'effectif découlant de l'utilisation des estimations de Laplace des probabilités. Quant à l'indice d'implication, qui peut être vu comme un résidu standardisé, il est calculé à partir des effectifs et non des proportions et est donc sensible aux effectifs par construction.

Si l'on considère à présent l'attribution de la conclusion aux feuilles de l'arbre, l'indice d'implication présente l'avantage d'avoir une interprétation claire abondamment discutée dans la littérature : l'implication statistique est d'autant plus forte que la règle admet étonnamment peu de contre-exemples.

Le critère $\eta_{w,ij}$, complémentaire à un de la contribution à l'entropie décentrée, est moins intuitif. Il mesure en quelque sorte l'importance de l'écart entre la fréquence de la catégorie dans la feuille et la proportion avec laquelle cette même catégorie est observée dans l'ensemble de la population. Contrairement à l'indice d'implication, il se fonde sur la fréquence même de la catégorie, et non sur ses contre-exemples. Le critère $\eta_{w,ij}$ conduit ainsi à privilégier la catégorie dont la fréquence domine relativement le plus fortement sa proportion marginale.

Les deux critères trouvent leur justification dans une perspective de ciblage où l'on s'intéresse à savoir pour quelle valeur de la variable cible le profil décrit par la condition de la règle est la plus typique. Par exemple, un médecin sera intéressé en priorité à savoir quelle est la population la plus exposée au risque de développer un cancer. De même, il est naturel d'axer en priorité des actions de marketing, de prévention ou de contrôle sur les groupes de population qui seront les plus réceptifs même lorsque ceux-ci ne sont pas majoritairement concernés par les actions envisagées.

On peut noter que les deux indicateurs sélectionnent la même catégorie lorsqu'une seule fréquence de la feuille dépasse la proportion marginale. Les choix peuvent cependant diverger dans le cas contraire. A titre d'exemple, nous donnons au tableau 9 la distribution d'une feuille j pour laquelle on obtient des conclusions non concordantes. On observe que l'écart entre effectifs observés (les n_{ij}) et attendus selon la distribution marginale est le même, -6 , pour les catégories A et B. Relativement, l'écart est plus important pour la catégorie A que privilégie la contribution à l'entropie. L'indice d'implication privilégie par contre B, pour laquelle on a moins de contre-exemples.

catégorie	en tout	distribution dans feuille j			effectifs attendus	contre-exemples		Indice implication	Contrib. $\eta_{w,ij}$ à l'entropie
	w_i	n_{ij}	f_{ij}	\hat{p}_{ij}		observés	attendus		
A	10%	12	.2	0.206	6	48	54	-0.75	0.064
B	20%	18	.3	0.302	12	42	48	-0.79	0.047
C	70%	30	.5	0.492	42	30	18	2.95	-0.147
Total	100%	60	1	60	1	-	-	-	-

TAB. 9 – Illustration de la différence entre indice d'implication et contribution à l'entropie décentrée.

Un avantage de l'indice d'implication est qu'il peut être comparé avec une distribution normale, ce qui justifie d'ailleurs la correction pour la continuité que nous lui avons apporté. Ceci permet de dire par exemple dans le cas du tableau 9 que le choix de la conclusion n'est pas solidement établi statistiquement puisque la valeur de l'indice et en deçà du seuil critique de -1.645 pour un risque de 5%. Notons que pour la comparaison avec la loi normale il serait préférable d'utiliser l'une des variantes proposées dans Ritschard (2005), l'indice de Gras tendant à avoir une variance inférieure à 1.

7.2 Vers une approche hybride

Au vu des remarques précédentes nous proposons d'exploiter de préférence l'entropie décentrée pour le développement de l'arbre et l'indice d'implication pour l'attribution des conclusions aux feuilles.

Pour le développement de l'arbre, l'entropie décentrée est un peu plus simple à mettre en œuvre, mais nous semble surtout être un instrument plus général, l'indice d'implication étant trop étroitement lié à la procédure de choix de la conclusion et donc à la seule élaboration de règles. Par exemple, le non éclatement lorsque les règles obtenues prennent la même conclusion peut être un handicap dans la mesure où cela empêche de repérer des sous-groupes pour lesquels la règle serait plus fiable que pour d'autres. L'entropie décentrée nous semble de ce point de vue permettre plus de nuances.

Une fois l'arbre construit par contre, l'indice d'implication nous semble mieux indiqué pour le choix de la conclusion, de par l'importance accordée aux contre-exemples et la possibilité qu'il offre de juger de la signification statistique du lien entre prémisse et conclusion de la règle.

8 Conclusion et perspectives

Nous nous sommes dans cet article intéressés à la problématique des données déséquilibrées dans le contexte des arbres de décision. Nous avons présenté et discuté avantages et inconvénients de deux approches, l'une fondée sur l'indice d'implication et l'autre sur une entropie décentrée. Il apparaît que ces deux approches conduisent à des solutions semblables bien qu'obéissant à des logiques totalement différentes. L'entropie décentrée semble être un critère plus naturel pour le développement de l'arbre tandis que l'indice d'implication présente des avantages certains pour sélectionner la catégorie à attribuer aux feuilles. Ces appréciations relèvent cependant essentiellement de la conjecture et doivent encore être validées empiriquement. Nous travaillons actuellement à la mise au point d'un protocole d'expérimentation qui devrait permettre de caractériser les différences de comportement à l'échelle des deux types d'indicateur. En particulier, il s'agira d'évaluer la performance des règles obtenues du point de vue du rappel des valeurs peu fréquentes, mais aussi d'évaluer l'information descriptive des arbres générés. Enfin, il nous faudra encore populariser ces procédures en les implémentant dans des plateformes aisément accessibles.

Références

- Aczél, J. et Z. Daróczy (1975). *On measures of information and their characterizations*. New York: Academic Press.
- Barandela, R., J. S. Sánchez, V. García, et E. Rangel (2003). Strategies for learning in class imbalance problems. *Pattern Recognition* 36(3), 849–851.
- Blanchard, J., F. Guillet, H. Briand, et R. Gras (2005). Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In Gras et al. (2005), pp. 131–138.
- Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.
- Chen, C., A. Liaw, et L. Breiman (2004). Using random forest to learn imbalanced data. Technical report, Berkeley, Department of Statistics, University of California.
- Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pp. 155–164.
- 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., F. Spagnolo, et J. David (Eds.) (2005). *Actes des Troisièmes Rencontres Internationale ASI Analyse Statistique Implicative*, Volume Secondo supplemento al N.15 of *Quaderni di Ricerca in Didattica*, Palermo. Università degli Studi di Palermo.
- Jaroszewicz, S. et D. A. Simovici (2001). A general measure of rule interestingness. In L. De Raedt et A. Siebes (Eds.), *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001, Freiburg, Germany, September 3-5, 2001, Proceedings*, Volume 2168 of *LNAI*, pp. 253–265. Berlin: Springer.
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2), 119–127.
- Lallich, S., P. Lenca, et B. Vaillant (2005). Variation autour de l'intensité d'implication. In Gras et al. (2005), pp. 237–246.
- Lallich, S., P. Lenca, et B. Vaillant (2007). Construction d'une entropie décentrée pour l'apprentissage supervisé. In *QDC 2007, Actes du 3ème atelier Qualités des données et connaissances, EGC janvier 2007, Namur*, pp. 45–54.
- 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.
- Marcellin, S., D. A. Zighed, et G. Ritschard (2006). Detection of breast cancer using an asymmetric entropy measure. In A. Rizzi et M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 975–982. Berlin: Springer. (on CD).

- Pisetta, V., G. Ritschard, et D. A. Zighed (2007). Choix des conclusions et validation des règles issues d'arbres de classification. In M. Noirhomme et G. Venturini (Eds.), *Extraction et Gestion des Connaissances (EGC 2007)*, Volume E-9 of *Revue des nouvelles technologies de l'information RNTI*, pp. 485–496. Cépaduès.
- Provost, F. (2000). Learning with imbalanced data sets. In *Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets*.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
- Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1, Berkeley, pp. 547–561. University of California Press.
- Ritschard, G. (2005). De l'usage de la statistique implicative dans les arbres de classification. In Gras et al. (2005), pp. 305–314.
- Witten, I. H. et E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Amsterdam: Morgan Kaufman (Elsevier).
- Zighed, D. A., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In M. Noirhomme et G. Venturini (Eds.), *Extraction et Gestion des Connaissances (EGC 2007)*, Volume E-9 of *Revue des nouvelles technologies de l'information RNTI*, pp. 81–86. Cépaduès.
- Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'induction : apprentissage et data mining*. Paris : Hermes Science Publications.

Summary

This paper is concerned with the induction of classification trees for imbalanced data, i.e. for the case where some categories of the target variable are much less frequent than other ones. More specifically, we address two aspects. On the one hand, we look for growing criteria that efficiently take into account the specific imbalanced nature of the data. On the other hand, we deal with the relevance of the conclusion that should be assigned to the leaves of a grown tree. We have recently considered two independent ways for dealing with these issues. The first one consisted in defining and using out centered entropies, and the second one on relying on measures of implication strength derived from implicative statistics. The aim of this paper is to compare and establish the relationship between these two approaches.