Expérimentation de l'entropie décentrée pour le traitement des classes déséquilibrées en induction par arbres

Thanh-Nghi Do*, Nguyen-Khang Pham**
Stéphane Lallich***, Philippe Lenca****

*INRIA Futurs/LRI, Université de Paris-Sud, Orsay, France
dtnghi@lri.fr,
**IRISA, Rennes, France
pnguyenk@irisa.fr
***Université de Lyon, Laboratoire ERIC, Lyon 2, Lyon, France
stephane.lallich@univ-lyon2.fr
****Telecom Bretagne, UMR 2872 TAMCIC, France
philippe.lenca@telecom-bretagne.eu

Résumé. En apprentissage supervisé, les données réelles sont souvent fortement déséquilibrées. Dans le cas des arbres de décision, trois types d'amélioration peuvent être apportés : pour la fonction de segmentation, la règle de décision et la procédure d'élagage. Notre contribution concerne la fonction de segmentation, pour laquelle nous avons proposé une méthode de décentrage des entropies usuelles. Dans ce papier, nous rendons compte des expériences pratiquées sur 25 bases de référence en utilisant C4.5, qui montrent les excellents résultats de l'entropie décentrée face à l'entropie de Shannon, y compris après un *bagging*.

1 Le problème des classes déséquilibrées

En apprentissage supervisé, les classes sont dites déséquilibrées lorsque leurs fréquences *a priori* sont très différentes. Dans le cas d'un problème à 2 classes, la classe la plus fréquente est dite majoritaire, alors que la classe la plus rare est dite minoritaire. Les problèmes réels relèvent très souvent de cette situation, avec une classe minoritaire de fréquence *a priori* inférieure à 0.10 (par exemple en détection de fraudes, *scoring* ou diagnostic médical). Dans un tel cas, les performances des algorithmes de Data Mining sont dégradées, en particulier le taux d'erreur de la classe minoritaire, alors même que cette classe correspond le plus souvent aux exemples positifs dont le coût de mauvaise classification (par exemple, ne pas repérer un fraudeur) est généralement beaucoup plus élevé que celui des exemples de la classe majoritaire (par exemple, déceler à tort un fraudeur).

Ce problème a donné lieu à de nombreux papiers, parmi lesquels on peut citer ceux issus de deux ateliers spécialisés, l'un associé à la conférence AAAI (Japkowicz (2000a)), l'autre à la conférence ICML (Chawla et al. (2003)), ainsi que ceux d'un numéro spécial de SIGKDD (Chawla et al. (2004)). Le traitement des classes déséquilibrées et *cost-sensitive* a été reconnu comme l'un des 10 problèmes les plus importants en Data Mining (Yang et Wu (2006)).

Comme cela est souligné dans les papiers de synthèse de Japkowicz (2000b), Japkowicz et Stephen (2002) et Visa et Ralescu (2005) ou dans les papiers très pédagogiques de Weiss et Provost (2001) et Weiss et Provost (2003), les solutions proposées pour surmonter le déséquilibre des classes peuvent se situer soit au niveau des données, soit au niveau algorithmique.

Les solutions qui interviennent au niveau des données changent la distribution de la variable de classe. Ces solutions comprennent différentes formes de ré-échantillonnage, ainsi le sur-échantillonnage, qui augmente le nombre d'individus de la classe minoritaire (Chawla et al. (2004), Liu et al. (2007)) ou le sous-échantillonnage qui diminue le nombre d'individus de la classe majoritaire (Kubat et Matwin (1997)) de façon aléatoire ou dirigée. Deux méthodes d'apprentissage qui utilisent les exemples de la classe majoritaire ignorés en cas de sous-échantillonnage ont été proposées par Liu et al. (2006). Une étude comparative de Drummond et Holte (2003), qui utilise les arbres de décision, à savoir C4.5 (Quinlan (1993)), a montré la supériorité du sous-échantillonnage sur le sur-échantillonnage.

Au niveau algorithmique, une première solution consiste à rééquilibrer le taux d'erreur en pondérant chaque type d'erreur par le coût correspondant (METACOST de Domingos (1999)). Une étude du bien-fondé de la méthode de rééquilibrage des coûts, pour prendre en compte les coûts de mauvaise classification et traiter les classes déséquilibrées a été proposée par Zhou et Liu (2006). Pour une étude comparative de ces méthodes on peut se reporter à Liu et Zhou (2006) et Weiss et al. (2007). En apprentissage par arbres, les autres solutions algorithmiques consistent à ajuster les estimations de probabilité dans les feuilles de l'arbre et les seuils de décision. Ling et al. (2004) proposent d'utiliser un critère de coût minimal, alors que Du et al. (2007) étudient des stratégies de pré-élagage efficaces pour éviter le sur-ajustement lorsque l'on utilise les méthodes fondées sur les coûts en induction par arbre.

Aux deux niveaux, dans le cas des arbres de décision avec C4.5, Chawla (2003) a étudié la qualité des estimations probabilistes, l'élagage et le prétraitement des données, trois problèmes habituellement considérés de façon séparée.

Notre contribution se situe, en induction par arbres, au niveau algorithmique pour choisir la mesure à l'aide de laquelle chaque nœud est éclaté. Nous proposons de décentrer l'entropie habituellement utilisée pour choisir la variable assurant le meilleur éclatement (entropie de Shannon dans le cas de C4.5 et entropie quadratique de Gini dans le cas du système CART de Breiman et al. (1984)).

Notre papier est organisé de la façon qui suit. La section 2 est consacrée aux entropies et à leur décentrage, en particulier l'entropie de Shannon. En section 3 les performances des différentes entropies sont comparées sur 25 jeux de données plus ou moins déséquilibrés. Enfin (Section 4), nous concluons et dressons les pistes de recherche futures.

2 De l'entropie de Shannon aux entropies non centrées

Dans cette section, nous présentons brièvement les trois entropies dont nous comparons les performances en section 3. Nous rappelons d'abord les bases de l'entropie de Shannon, ainsi que le principe des coefficients qui lui sont associés, puis nous exposons le principe de notre méthodologie de décentrage et définissons l'entropie de Shannon décentrée issue de cette méthodologie. Nous présentons enfin l'entropie asymétrique, une entropie non centrée particulière proposée dans la littérature. Nous détaillons principalement le cas booléen, situation des expérimentations présentées en section 3.

2.1 Mesures usuelles fondées sur l'entropie de Shannon

En apprentissage supervisé par arbres de classification, de nombreux algorithmes utilisent des coefficients d'association prédictive fondés sur l'entropie de Shannon (1948).

Considérons une variable de classe Y, qui possède q modalités, et un prédicteur catégoriel X qui a k modalités. Le vecteur des fréquences relatives de Y est noté $p = (p_1, \ldots, p_q)$. La distribution de fréquences relatives conjointes du couple (x_i, y_j) est notée $p_{ij}, i = 1, 2, \ldots k; j = 1, 2, \ldots q$. On note $h(Y) = -\sum_{j=1}^{q} p_{.j} \log_2 p_{.j}$ l'entropie de Shannon *a priori* de Y et $h(Y/X) = E(h(Y/X = x_i))$ la moyenne des entropies *a posteriori* de Y conditionnellement à X.

L'entropie de Shannon est une fonction réelle positive de $p = (p_1, \ldots, p_q)$ à valeurs dans $[0 \ldots 1]$, qui vérifie notamment les propriétés suivantes, particulièrement intéressantes dans la perspective de l'apprentissage supervisé (Zighed et Rakotomalala (2000)) :

1. **Invariance par permutation des modalités** : $h(p)$ ne change pas lorsque l'on permute les modalités de Y ;
2. **Maximalité** : $h(p)$ vaut au maximum 1, valeur atteinte lorsque la distribution de Y est uniforme, c'est-à-dire lorsque toutes les modalités de Y ont la même fréquence $1/q$;
3. **Minimalité** : $h(p)$ vaut au minimum 0, valeur atteinte lorsque la distribution de Y est certaine, concentrée sur une seule modalité de Y, qui est de fréquence 1 ;
4. **Concavité stricte** : $h(p)$ est une fonction strictement concave.

Parmi les coefficients d'association usuels fondés sur l'entropie de Shannon, analysés notamment par Wehenkel (1996), Loh et Shih (1997), Shih (1999) et Simovici et Jaroszewicz (2006), nous retiendrons tout particulièrement :

- le gain d'entropie (Quinlan (1975)), qui s'écrit : $h(Y) - h(Y/X)$;
- le coefficient u de Theil (1970), qui normalise le gain d'entropie par l'entropie *a priori* de Y, à savoir $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- le gain-ratio (Quinlan, 1993), défini par $\frac{h(Y)-h(Y/X)}{h(X)}$, qui normalise le gain d'entropie par l'entropie du prédicteur X et non pas par l'entropie *a priori* de Y, afin de ne pas favoriser les prédicteurs ayant un grand nombre de modalités ;
- le coefficient de Kvalseth (1987), défini par $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$, qui normalise le gain d'entropie par la moyenne des entropies de X and Y.

Ces coefficients correspondent à différentes normalisations de l'entropie de Shannon, qui atteint son maximum lorsque la distribution de Y est uniforme. Même si c'est le gain d'entropie par rapport à l'entropie *a priori* de Y qui est effectivement normalisé, il n'en reste pas moins que les entropies de Y et $Y/X = x_i$ qui interviennent dans le calcul du gain d'entropie sont évaluées sur une échelle dont la valeur de référence (entropie maximale) correspond à la distribution uniforme des classes, ce qui est inadapté pour des classes déséquilibrées. Il serait plus logique d'évaluer directement le gain d'entropie en utilisant une échelle pour laquelle la valeur de référence correspondrait à la distribution *a priori* des classes dans le nœud considéré. Cette critique des coefficients fondés sur le gain d'entropie prend toute son importance lorsque les classes sont très déséquilibrées ou lorsque les coûts de mauvaise classification diffèrent largement, et elle fonde la méthode de décentrage que nous proposons pour le traitement de telles données.

2.2 L'entropie de Shannon décentrée

La stratégie de construction d'une entropie de Shannon décentrée, dans le cas booléen, est esquissée dans Lallich et al. (2005) puis approfondie dans Lallich et al. (2007b). Ce travail portait sur la paramétrisation de différentes mesures statistiques de l'intérêt des règles d'association, où l'on compare suivant différentes normalisations la confiance de la règle $\mathtt{A} \rightarrow \mathtt{B}$ à un paramètre θ choisi par l'utilisateur plutôt qu'à la fréquence *a priori* de B. C'est ainsi que nous avons été amenés à décentrer l'entropie de Shannon de $\mathtt{B}/\mathtt{A}$ pour faire en sorte que celle-ci soit maximale lorsque $p_{b/a}$, la confiance de $\mathtt{A} \rightarrow \mathtt{B}$, est égale à θ. Par la suite, constatant l'intérêt de cette entropie décentrée pour l'apprentissage supervisé de classes déséquilibrées, nous avons entrepris l'étude approfondie du décentrage des entropies généralisées (Lallich et al., 2007a,c), tant dans le cas booléen que dans celui des variables catégorielles. Nous présentons en détail le cas booléen, puis nous donnerons les formules du cas général.

Le cas booléen. Considérons une variable de classe Y qui comporte $q = 2$ modalités. La distribution de fréquences relatives de Y pour les valeurs 0 et 1 est notée $(1-p, p)$, où p désigne la fréquence de $Y = 1$ et son entropie de Shannon est notée $h(p)$. Nous souhaitons associer à cette distribution une entropie décentrée notée $\eta_\theta(p)$, qui est maximale lorsque $p = \theta$, θ étant fixé par l'utilisateur et pouvant prendre n'importe quelle valeur entre 0 et 1. Pour définir l'entropie décentrée, suivant la démarche décrite dans Lallich et al. (2005), nous proposons de transformer la distribution $(1-p, p)$ en $(1-\pi, \pi)$, de telle sorte que :

- π augmente de 0 à $1/2$, lorsque p augmente de 0 à θ ;
- π augmente de $1/2$ à 1, lorsque p augmente de θ à 1.

En recherchant une transformation du type $\pi = \frac{p-b}{a}$, sur chacun des intervalles $0 \leq p \leq \theta$ et $\theta \leq p \leq 1$, on obtient :

$$\pi = \frac{p}{2\theta} \text{ si } 0 \leq p \leq \theta, \ \pi = \frac{p+1-2\theta}{2(1-\theta)} \text{ si } \theta \leq p \leq 1$$

En toute rigueur, la distribution transformée devrait être notée $(1-\pi_\theta, \pi_\theta)$. Nous la noterons plus simplement $(1-\pi, \pi)$, pour ne pas alourdir les formules. Il s'agit bien d'une distribution de fréquences, puisque $0 \leq \pi \leq 1$. L'entropie de Shannon décentrée de $(1-p, p)$ est alors définie comme l'entropie de Shannon de $(1-\pi, \pi)$, notée $\eta_\theta(p)$:

$$\eta_\theta(p) = h(\pi) = -\pi \log_2 \pi - (1-\pi) \log_2 (1-\pi)$$

A proprement parler, il est clair que par rapport à $(1-p, p)$, $\eta_\theta(p)$ n'est pas une entropie. Ses propriétés doivent être analysées en tenant compte du fait que $\eta_\theta(p)$ est l'entropie de la distribution transformée $(1-\pi, \pi)$. C'est ainsi que parmi les propriétés signalées précédemment, l'entropie décentrée préserve les propriétés 3 (minimalité) et 4 (concavité stricte) et que l'on doit adapter la propriété 2 (maximalité), le maximum ayant lieu dorénavant pour $\pi = 0.5$, i.e. pour $p = \theta$. En revanche la propriété 1 (invariance par permutation) est volontairement abandonnée. Le détail des démonstrations est donné dans Lallich et al. (2007a).

Le cas des variables catégorielles. La construction de l'entropie de Shannon décentrée est étendue au cas où Y et X sont catégorielles, à q et k modalités respectivement, en suivant un raisonnement analogue à celui tenu dans le cas booléen (Lallich et al., 2007a,c). Pour une variable Y ayant q modalités, la distribution uniforme correspond au cas où les fréquences des différentes modalités de Y sont égales à $\frac{1}{q}$, ce qui amène à définir l'entropie décentrée

par $\eta_\theta(p) = h(\pi^*)$, où p désigne la distribution de fréquences de Y et π^* la distribution transformée définie comme suit (pour $q = 2$ on retrouve la définition donnée dans le cas booléen) :

- $\pi_j = \frac{p_j}{q\theta_j}$ si $0 \leq p_j \leq \theta_j$, et $\pi_j = \frac{q(p_j-\theta_j)+1-p_j}{q(1-\theta_j)}$ si $\theta_j \leq p_j \leq 1$
- $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^q \pi_j}$, d'où $0 \leq \pi_j \leq 1$ et $\sum_{j=1}^q \pi_j = 1$

2.3 Décentrage des entropies généralisées

L'entropie de Shannon n'est pas la seule fonction de diversité que l'on puisse utiliser pour construire des coefficients d'association prédictive. Une présentation unifiée des trois principaux coefficients que sont le λ de Guttman, le u de Theil et le τ de Goodman et Kruskal, a été proposée par Goodman et Kruskal (1954), sous la dénomination de coefficients PRE (*Proportional Reduction in Error)*. De façon plus générale, nous avons élargi cette définition pour proposer les coefficients *PRD* (*Proportional Reduction in Diversity*) qui sont l'analogue d'un gain normalisé d'entropie de Shannon, lorsque l'entropie de Shannon est remplacée par n'importe quelle fonction de diversité concave (Lallich (2002)).

La particularité de notre approche est de proposer une méthode de décentrage qui s'adapte à n'importe quelle entropie (Lallich et al., 2007c), aussi bien l'entropie de Shannon que plus généralement une entropie de type bêta (Daroczy, 1970) ou une entropie de rangs (Lallich, 2002).

2.4 L'entropie asymétrique

Dans une optique de construction d'une mesure d'association prédictive, particulièrement dans le cas des arbres de décision, Marcellin et al. (2006a) ont proposé l'entropie asymétrique, $h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p+\theta^2}$, dans le cas d'une variable de classe booléenne. Cette mesure est asymétrique au sens où l'on peut choisir la distribution pour laquelle elle atteint son maximum. Par rapport aux propriétés classiques des entropies, les propriétés 3 (minimalité) et 4 (stricte concavité) sont conservées, mais la propriété 2 (maximalité) est modifiée de telle sorte que la mesure soit maximale pour une distribution $(1-\theta, \theta)$ fixée par l'utilisateur. On remarquera que dans le cas $\theta = 0.5$, l'entropie asymétrique correspond à l'entropie quadratique de Gini.

Dans Zighed et al. (2007), considérant que la distribution de Y n'est connue qu'à travers sa distribution empirique issue d'un échantillon de taille n, les mêmes auteurs souhaitent que pour une même distribution empirique, la valeur de l'entropie décroisse lorsque n augmente, définissant ainsi la consistance (propriété 5). C'est ainsi qu'ils sont conduits à transformer la propriété 3 (minimalité) en une propriété 3' (minimalité asymptotique) où l'entropie d'une variable certaine doit seulement tendre vers 0 lorsque $n \to \infty$. Pour obtenir ces propriétés, ils proposent de remplacer les fréquences empiriques p_j par les estimateurs de Laplace des fréquences théoriques $\widetilde{p_j} = \frac{np_j+1}{n+q}$. En outre, ils étendent leur approche au cas où la variable de classe possède q modalités, $q > 2$, et proposent en définitive une entropie asymétrique consistante définie par :

$$h_\theta(p) = \sum_{j=1}^q \frac{\widetilde{p_j}(1-\widetilde{p_j})}{(1-2\theta_j)\widetilde{p_j}+\theta_j^2}$$

3 Expérimentations

Dans cette section nous comparons les résultats obtenus en induction par arbres sur 25 bases de référence, suivant que l'on utilise notre entropie décentrée (notée OCE, comme *off-centered entropy*), l'entropie asymétrique (AE) ou l'entropie de Shannon usuelle (SE). Pour ces comparaisons, nous avons intégré OCE et AE à l'algorithme C4.5 de Quinlan (1993).

Les comparaisons ont été faites à partir de 25 jeux d'essais plus ou moins déséquilibrés décrits dans le tableau 1. Les colonnes 2, 3 et 4 indiquent le nom du jeu d'essai, le nombre de cas et le nombre d'attributs. Les 17 premières bases proviennent du site de l'UCI (Blake et Merz, 1998), les 6 suivantes du site de Statlog (Michie et al., 1994), la 24e du projet Delve (http ://www.cs.toronto.edu/∼delve/), la 25e provenant de Jinyan et Huiqing (2002).

n°	Base	Nb. cas	Nb. dim	Classe min	Classe max	Validation
1	Opticdigits	5620	64	10%(0)	90%(rest)	trn-tst
2	Tictactoe	958	9	35%(1)	65%(2)	10-fold
3	Wine	178	13	27%(3)	73%(rest)	loo
4	Adult	48842	14	24%(1)	76%(2)	trn-tst
5	20-newsgrp	20000	500	5%(1)	95%(rest)	3-fold
6	Breast	569	30	35%(M)	65%(B)	10-fold
7	Letters	20000	16	4%(A)	96%(rest)	3-fold
8	Yeast	1484	8	31%(CYT)	69%(rest)	10-fold
9	Connect-4	67557	42	10%(draw)	90%(rest)	3-fold
10	Glass	214	9	33%(1)	67%(rest)	loo
11	Spambase	4601	57	40%(spam)	60%(rest)	10-fold
12	Ecoli	336	7	15%(pp)	85%(rest)	10-fold
13	Abalone	4177	8	9%(15-29)	91%(rest)	10-fold
14	Pendigits	10992	16	10%(9)	90%(rest)	trn-tst
15	Car	1728	6	8%(g, vg)	92%(rest)	10-fold
16	Bupa	345	6	42%(1)	58%(2)	10-fold
17	Page blocks	5473	10	10%(rest)	90%(text)	10-fold
18	Pima	768	8	35%(1)	65%(2)	10-fold
19	German	1000	20	30%(1)	70%(2)	10-fold
20	Shuttle	58000	9	20%(rest)	80%(1)	trn-tst
21	Segment	2310	19	14%(1)	86%(rest)	10-fold
22	Satimage	6435	36	24%(1)	90%(rest)	trn-tst
23	Vehicle	846	18	24%(van)	76%(rest)	10-fold
24	Splice	3190	60	25%(EI)	75%(rest)	10-fold
25	ALL-AML	72	7129	35% (AML)	65%(ALL)	loo

TAB. 1 – *Description des bases*

Lorsqu'une base comportait plus de 2 classes, nous nous sommes ramenés par regroupement de classes au cas de données booléennes déséquilibrées. Les colonnes 5 et 6 montrent comment nous avons opéré le regroupement. Par exemple, dans le cas de la base *OpticDigits*, le chiffre 0 est considéré comme la classe minoritaire (10%), alors que le regroupement des autres chiffres constitue la classe majoritaire (90%). Dans le cas de la base *20-newsgroup*, utili-

sée en catégorisation de textes, nous avons utilisé une procédure de sélection de mots pertinents fondée sur l'information mutuelle pour extraire un tableau attributs-valeurs à 500 dimensions.

Les trois entropies sont comparées selon la taille de l'arbre (TS), la précision globale (complément à 1 du taux d'erreur, notée *Acc*), la précision sur la classe minoritaire notée *Amin* et la précision sur la classe majoritaire notée *Amax*. La comparaison est faîte sans et avec *bagging*. Rappelons que le *bagging* consiste à agréger les arbres obtenus à partir de différents échantillons *bootstrap* de l'échantillon initial (Breiman, 1996). Le protocole de test est présenté dans la colonne 7 du tableau 1. Dans certains cas, la base est déjà divisée en ensemble d'apprentissage (trn) et ensemble test (tst). Sinon, nous avons procédé par validation croisée. Le *leave-one-out* (loo) est utilisé lorsque la base comporte moins de 100 cas. Autrement, nous avons utilisé la validation croisée à k segments, avec $k = 3$ ou $k = 10$, suivant la taille de la plus petite classe. La synthèse des comparaisons 2 à 2 des entropies est présentée dans les tableaux 2, 3 et 4. Prenons l'exemple du tableau 2 : pour comparer les performances de OCE et SE, on construit d'abord le test de conformité à zéro de la moyenne de OCE-SE (ligne 1 à 4). Cette comparaison est re-doublée par le test non paramétrique du signe (ligne 5 à 8) qui a l'avantage d'être indépendant des distributions sous-jacentes. On note *** la signification à 1/1000, ** à 1/100, * à 5/100. Pour ces comparaisons, la règle de prédiction adoptée est la règle majoritaire, en dépit du fait qu'elle n'est pas adaptée aux classes déséquilibrées. En effet, lorsque dans une feuille, la classe minoritaire passe de 0.05 *a priori* à 0.40, c'est un beau résultat, pourtant celui-ci ne modifie pas la décision de la règle majoritaire. La recherche d'une règle mieux adaptée est l'une des améliorations envisagées.

D'après le tableau 2, face à SE, OCE améliore 23 fois contre 1 (***) la précision sur la classe minoritaire, pour un gain moyen de 0.020 (***), tout en améliorant 21 fois contre 3 (***) la précision globale, pour un gain moyen de 0.008, qui est à la limite de la signification. En cas de *bagging*, les améliorations apportées par OCE sont très hautement significatives, tant en ce qui concerne la fréquence d'amélioration de la précision, qui est améliorée 21 fois contre 2 (***) sur la classe minoritaire, 17 fois (*) contre 6 sur la classe majoritaire et 23 fois contre 0 (***) au niveau global, qu'en ce qui concerne le gain de précision qui est de 0.015 sur la classe minoritaire (***) et de 0.008 (**) au niveau global. Seul le gain de précision sur la classe majoritaire qui vaut 0.007 n'est pas significatif en raison d'une forte variabilité des résultats.

Comparée à SE (tableau 3), AE l'emporte de façon significative, 20 fois contre 4 (**) pour *Amin*, avec une amélioration moyenne de 0.014 (***), 18 fois contre 6 (*) pour *Acc*, sans que le gain moyen de 0.003 soit significatif et 10 fois contre 9 avec un gain moyen de 0.003 pour *Amax*, ce qui n'est pas significatif. En cas de *bagging*, l'amélioration de la précision sur la classe minoritaire qui est de 0.005 n'est plus significative ; en revanche, l'amélioration de la précision globale qui est de 0.004 devient significative.

Entre OCE et AE, les deux entropies non centrées, dans le cadre du protocole appliqué, OCE a un léger avantage, mais celui-ci n'est significatif qu'en cas de *bagging* (tableau 4). En effet, sans *bagging* préalable, si l'on ne tient pas compte des exaequo, les scores entre OCE et AE sont de *13-9* pour *Amin*, *14-6* pour *Acc* et *12-9* pour *Amax*. Parallèlement, les gains moyens sont de 0.005, 0.004 et 0.003. En cas de *bagging*, l'avantage d'OCE devient significatif pour *Amin* et *Acc*, les scores passant respectivement à *16-5* (*) et *17-3* (**), avec des gains de 0.010 (*) et 0.004 (**), mais reste non significatif pour *Amax* avec un score de *11-8*.

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	-7,28	0,76	1,98	0,58	0,75	1,47	0,65
écart-type	26,55	1,91	2,13	2,32	1,22	1,58	2,93
t-ratio	-1,37	2,00	4,65	1,26	3,11	4,66	1,11
p-value	0,1831	0,0574	0,0001	0,2215	0,0048	0,0001	0,2786
OCE	6	21	23	12	23	21	17
=	4	1	1	5	2	2	2
SE	15	3	1	8	0	2	6
p-value	0,0784	0,0003	0,0000	0,5034	0,0000	0,0001	0,0347

TAB. 2 – OCE *vs.* SE

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	0,44	0,34	1,44	0,29	0,39	0,51	0,50
écart-type	31,81	0,90	1,87	1,84	0,79	1,59	1,58
t-ratio	0,07	1,92	3,85	0,79	2,45	1,61	1,59
p-value	0,9454	0,0673	0,0008	0,4360	0,0218	0,1214	0,1249
AE	12	18	20	10	16	14	13
=	2	1	1	6	4	3	4
SE	11	6	4	9	5	8	8
p-value	1,0000	0,0227	0,0015	1,0000	0,0266	0,2863	0,3833

TAB. 3 – AE *vs.* SE

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	-7,72	0,42	0,54	0,29	0,37	0,96	0,15
écart-type	21,91	1,56	2,18	1,92	0,72	1,47	1,68
t-ratio	-1,76	1,34	1,25	0,76	2,55	3,26	0,44
p-value	0,0908	0,1917	0,2243	0,4551	0,0174	0,0033	0,6625
OCE	9	14	13	12	17	16	11
=	6	5	3	4	5	4	6
AE	10	6	9	9	3	5	8
p-value	1,0000	0,1153	0,5235	0,6636	0,0026	0,0266	0,6476

TAB. 4 – OCE *vs.* AE

4 Conclusion et travaux futurs

Pour adapter les fonctions de segmentation utilisées en apprentissage par arbres au cas des classes déséquilibrées, nous avons proposé une stratégie de décentrage des entropies usuelles qui permet d'évaluer la qualité d'un éclatement par référence directe à la distribution *a priori* de la variable de classe dans le noeud considéré et non pas en fonction de l'écart à l'équirépartition. Dans ce papier, à partir d'une expérimentation menée sur 25 bases booléennes plus ou moins déséquilibrées, avec l'algorithme C4.5 où seule était modifiée la fonction de segmentation, nous montrons que les entropies non centrées, AE de Marcellin et al. (2006b) et surtout notre entropie OCE obtenue par décentrage de l'entropie de Shannon, améliorent de façon quasi systématique la prédiction sur la classe minoritaire pour une précision globale au moins aussi grande. L'utilisation du *bagging* conforte ces résultats, particulièrement pour OCE.

Outre ces résultats, le principal avantage de notre démarche est que nous proposons une méthode de décentrage qui s'applique à n'importe quel type d'entropie, qu'il s'agisse de l'entropie de Shannon testée ici, de l'entropie quadratique de Gini utilisée dans l'algorithme CART (Breiman et al., 1984), ou de toute autre entropie. Il reste à proposer un critère d'élagage pour déterminer la taille finale de l'arbre et surtout à imaginer une règle de décision adaptée aux classes déséquilibrées. Il serait aussi intéressant de prendre en compte les matrices *cost-sensitive*.

Références

Blake, C. L. et C. J. Merz (1998). UCI repository of machine learning databases. http ://www.ics.uci.edu/∼mlearn/MLRepository.html.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. Stone (1984). *Classification and Regression Trees*. Wadsworth International,.

Chawla, N. (2003). C4.5 and imbalanced datasets : Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In *ICML'Workshop on Learning from Imbalanced Data Sets*.

Chawla, N., N. Japkowicz, et A. Kolcz (Eds.) (2003). *ICML'Workshop on Learning from Imbalanced Data Sets*.

Chawla, N., N. Japkowicz, et A. Kolcz (Eds.) (2004). *Special Issue on Class Imbalances*, Volume 6 of *SIGKDD Explorations*.

Daroczy, A. (1970). Generalized information functions. *Information and Control* (16), 36–51.

Domingos, P. (1999). Metacost : A general method for making classifiers cost sensitive. In *Int. Conf. on Knowledge Discovery and Data Mining*, pp. 155–164.

Drummond, C. et R. Holte (2003). C4.5, class imbalance, and cost sensitivity : Why under-sampling beats over-sampling. In *ICML'Workshop on Learning from Imbalanced Data Sets*.

Du, J., Z. Cai, et C. X. Ling (2007). Cost-sensitive decision trees with pre-pruning. In *Canadian Conf. on Artificial Intelligence*, Volume 4509 of *LNAI*, pp. 171–179.

Goodman, L. A. et W. H. Kruskal (1954). Measures of association for cross classifications, i. *JASA I*(49), 732–764.

Japkowicz, N. (Ed.) (2000a). *AAAI'Workshop on Learning from Imbalanced Data Sets*, Number WS-00-05 in AAAI Tech Report.

Japkowicz, N. (2000b). The class imbalance problem : Significance and strategies. In *Int. Conf. on Artificial Intelligence*, pp. 111–117.

Japkowicz, N. et S. Stephen (2002). The class imbalance problem : A systematic study. *Intelligent Data Analysis 6*(5), 429–450.

Jinyan, L. et L. Huiqing (2002). Kent ridge bio-medical data set repository. Technical report. http ://sdmc-.lit.org.sg/GEDatasets.

Kubat, M. et S. Matwin (1997). Addressing the curse of imbalanced data sets : One-sided sampling. In *International Conference on Machine Learning*, pp. 179–186.

Kvalseth, T. O. (1987). Entropy and correlation : some comments. *IEEE Trans. on Systems, Man and Cybernetics 17*(3), 517–519.

Lallich, S. (2002). Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches, Université Lyon 2, France.

Lallich, S., P. Lenca, et B. Vaillant (2007c). Construction of an off-centered entropy for supervised learning. In *Int. Symp. on Applied Stochastic Models and Data Analysis*. 8 p.

Lallich, S., B. Vaillant, et P. Lenca (2005). Parametrised measures for the evaluation of association rule interestingness. In *Int. Symp. on Applied Stochastic Models and Data Analysis*, pp. 220–229.

Lallich, S., B. Vaillant, et P. Lenca (2007a). Construction d'une entropie décentrée pour l'apprentissage supervisé. In *QDC/EGC 2007*, pp. 45–54.

Lallich, S., B. Vaillant, et P. Lenca (2007b). A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability 9*, 447–463.

Ling, C. X., Q. Yang, J. Wang, et S. Zhang (2004). Decision trees with minimal costs. In *Int. Conf. on Machine Learning*.

Liu, A., J. Ghosh, et C. Martin (2007). Generative oversampling for mining imbalanced datasets. In *Int. Conf. on Data Mining*, pp. 66–72.

Liu, X.-Y., J. Wu, et Z.-H. Zhou (2006). Exploratory under-sampling for class-imbalance learning. In *IEEE Int. Conf. on Data Mining*, pp. 965–969.

Liu, X.-Y. et Z.-H. Zhou (2006). The influence of class imbalance on cost-sensitive learning : An empirical study. In *IEEE Int. Conf. on Data Mining*, pp. 970–974.

Loh, W.-Y. et Y.-S. Shih (1997). Split selection methods for classification trees. *Statistica Sinica 7*, 815–840.

Marcellin, S., D. Zighed, et G. Ritschard (2006a). An asymmetric entropy measure for decision trees. In *Information Processing and Management of Uncertainty in knowledge-based systems*, pp. 1292–1299.

Marcellin, S., D. Zighed, et G. Ritschard (2006b). An asymmetric entropy measure for decision trees. In *IPMU 2006*, Paris, France, pp. 1292–1299.

Michie, D., D. J. Spiegelhalter, et C. C. Taylor (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

Quinlan, J. (1975). *Machine Learning*, Volume 1.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal* (27), 379–423, 623–656.

Shih, Y.-S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing 9*, 309–315.

Simovici, D. A. et S. Jaroszewicz (2006). Generalized conditional entropy and a metric splitting criterion for decision trees. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Volume 3918 of *LNAI*, pp. 35–44.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology* (76), 103–154.

Visa, S. et A. Ralescu (2005). Issues in mining imbalanced data sets - A review paper. In *Midwest Artificial Intelligence and Cognitive Science Conf.*, pp. 67–73.

Wehenkel, L. (1996). On uncertainty measures used for decision tree induction. In *Information Processing and Management of Uncertainty in Knowledge based Systems*, pp. 413–418.

Weiss, G. M., K. McCarthy, et B. Zabar (2007). Cost-sensitive learning vs. sampling : Which is best for handling unbalanced classes with unequal error costs ? In *Int. Conf. on Data Mining*, pp. 35–41.

Weiss, G. M. et F. Provost (2001). The effect of class distribution on classifier learning. TR ML-TR 43, Department of Computer Science, Rutgers University.

Weiss, G. M. et F. Provost (2003). Learning when training data are costly : The effect of class distribution on tree induction. *J. of Art. Int. Research 19*, 315–354.

Yang, Q. et X. Wu (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making 5*(4), 597–604.

Zhou, Z.-H. et X.-Y. Liu (2006). On multi-class cost-sensitive learning. In *Nat. Conf. on Artificial Intelligence*, pp. 567–572.

Zighed, D., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In *Extraction et Gestion des Connaissances*, pp. 81–86.

Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'Induction – Apprentissage et Data Mining*. Hermes.

Summary

In supervised learning, real data are often highly imbalanced. In the case of decision trees, three types of improvements can be carried out: to the split function, to the rule of decision and to the pruning procedure. Our contribution concerns the split function, for which we have proposed a method to off-center usual entropies. In this paper, we report on experiments carried out on 25 data bases using C4.5, which show the excellent results of off-centered entropy facing the Shannon entropy, including in case of *bagging*.