QDC 2008

Actes du 4e Atelier

Qualité des Données et des Connaissances

En conjonction avec EGC 2008

29 Janvier 2008
Nice, France

Organisé par
Stéphane Lallich, Philippe Lenca et Fabrice Guillet

Évaluation de critères asymétriques pour les arbres de décision

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

*Laboratoire ERIC
5, av. pierre Mendès-France 69600 Bron, France
{abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
**Université de Genève
40 bd du Pont-d'Arve CH-1211 Geneva 4, Switzerland
Gilbert.ritschard@unige.ch

Résumé. Nous proposons dans cet article d'évaluer la qualité d'arbres de décision construits sur des jeux de données déséquilibrées avec une mesure d'entropie asymétrique. En effet, différents critères d'éclatement asymétriques ont été proposés pour tenir compte du déséquilibre des classes lors du choix du meilleur éclatement. Après la construction de l'arbre se pose le problème de l'assignation d'une classe à chaque feuille: une règle tenant compte de l'asymétrie doit être adoptée pour déduire des règles de prédiction à partir de l'arbre. Comment évaluer les résultats de ces modèles de prédiction? Nous considérons les courbes ROC et les graphiques rappel / précision pour évaluer les arbres de décisions sur des jeux de données déséquilibrés, en comparant les arbres construits sur un critère asymétrique avec ceux construits sur un critère symétrique.

1 Introduction

L'apprentissage sur données déséquilibrées est un problème important en fouille de données (Provost (2000); Barandela et al. (2003)). Un jeu de données est déséquilibré quand la distribution des modalités de la classe est très éloignée de la distribution uniforme. C'est le cas dans de nombreux exemples réels : dans le domaine médical, pour prédire une maladie rare ; dans l'industrie pour prédire une panne ; ou encore dans le domaine bancaire, pour détecter les clients insolvables ou les transactions frauduleuses. Dans ces exemples, un état rare de la variable classe (malade, panne, non solvable, frauduleux) doit être détecté en priorité. Les méthodes d'arbre standard ne tiennent pas compte de ces spécificités et optimisent simplement un critère global, ce qui implique que tous les individus sont classés dans la classe majoritaire, soit celle qui minimise le taux d'erreur global. Ce type de modèle de prédiction est inutile car il n'apporte pas d'information. Dans les arbres de décision, ce problème intervient lors de deux étapes. Premièrement, pour choisir la meilleure variable et le meilleur point d'éclatement pour la création d'une nouvelle partition, les algorithmes classiques utilisent une mesure d'entropie, comme l'entropie de Shannon ou l'entropie quadratique. Ces mesures considèrent que la distribution uniforme (pour laquelle le nombre d'individus de chaque classe est le même)

est la situation la plus incertaine. Cependant, si par exemple dans le monde réel 1% des personnes sont malades, obtenir la règle pour laquelle 50% des individus sont malades pourrait être intéressant et apporter de l'information à l'utilisateur du modèle. L'utilisation des mesures d'entropie classiques empêche l'obtention de ce type de règle et donc de règles pertinentes pour la prédiction d'une classe rare. Le second aspect important des arbres de décision est la règle d'assignation d'une classe à chaque feuille. Une fois que l'arbre est construit, chaque branche définit la prémisse d'une règle. La conclusion de la règle dépend de la distribution des classes dans la feuille. Les algorithmes classiques concluent à la classe la plus fréquente dans la feuille, mais cette méthode n'est pas pertinente : dans l'exemple précédent où 1% des personnes sont malades, une règle menant à une feuille où la fréquence de la classe 'malade' est de 30% conclura 'non malade'. Si l'on considère l'importance de prédire correctement la classe minoritaire, il pourrait être plus intéressant de conclure 'malade'. Cela provoquerait un nombre d'erreurs global plus important, mais moindre sur la classe rare et donc un meilleur modèle. Des critères asymétriques ont été proposés pour gérer le déséquilibre dans les arbres de décision. Comment ces critères influencent-ils l'apprentissage ? Quelles mesures de performance doit-on utiliser pour évaluer l'intérêt d'utiliser ces critères ? Nous proposons de considérer les courbes ROC pour évaluer la structure des arbres, et les graphes rappel / précision pour mesurer la performance des modèles de prédiction sans avoir à fixer une règle d'assignation. La section suivante présente les critères symétriques et asymétriques. En section 3 nous proposons une méthode d'évaluation pour comparer des arbres de décision basés sur ces différentes mesures. La section 4 présente notre évaluation et nos résultats. Nous finirons par la section 5 qui conclue et propose des travaux futurs.

2 Critères asymétriques pour les arbres de décision

Notations et concepts de base On note Ω la population concernée par le problème d'apprentissage. Un individu ω de Ω est décrit par p variables explicatives (ou exogènes)$X_1, \ldots, X_p$. On considère également une variable à prédire C appelée variable endogène, classe ou réponse. Quand il n'y a pas d'ambiguïté, on note la modalité (ou classe) c_i par i. Les algorithmes d'induction d'arbres génèrent un modèle $\phi(X_1, \ldots, X_p)$ pour la prédiction de C représenté par un arbre de décision (Quinlan (1993)). Chaque branche de l'arbre représente une règle. L'ensemble de ces règles forme le modèle de prédiction qui permet de prédire la valeur de la variable endogène pour un nouvel individu dont on ne connaît que les variables exogènes. Les critères d'éclatement qui permettent de choisir la meilleure partition à chaque étape de la génération d'un arbre de décision sont généralement basés sur l'entropie. La notion d'entropie a été définie mathématiquement en dehors du contexte de l'apprentissage (voir Rényi (1960) et Aczel et Daroczy (1975)). L'entropie H d'une partition S à minimiser est généralement une entropie moyenne telle que $H(S) = \sum_{s \in S} p(s) h(p(1|s), \ldots, p(i|s), \ldots, p(n|s))$ où $p(s)$ est la proportion de cas dans le noeud s et $h(p(1|s), \ldots, p(n|s))$ une fonction d'entropie comme l'entropie de Shannon : $H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$. Il existe d'autres mesures d'entropie comme l'entropie quadratique (ou indice de Gini), utilisé dans CART par Breiman et al. (1984) : $H_n(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i(1 - p_i)$.

Critères asymétriques Les propriétés des mesures d'entropie classiques comme celles citées ci-dessus ne sont pas adaptées à l'apprentissage inductif pour les raisons exposés dans

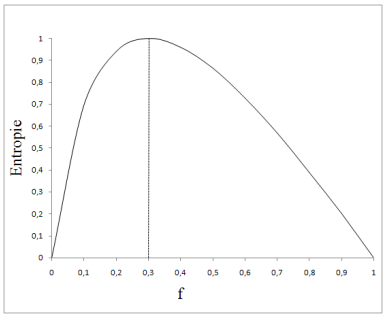

FIG. 1 – *Entropie asymétrique pour* $w = 0.3$

Zighed et al. (2007). En effet, la distribution uniforme n'est pas nécessairement la plus incertaine. C'est pourquoi nous avons proposé une axiomatique permettant de définir une nouvelle famille de mesures plus générales permettant à l'utilisateur de définir la distribution de référence (d'entropie maximale) $W = (w_1, w_2, \ldots, w_n)$. L'entropie asymétrique que nous proposons s'écrit alors : $h_W(f_1, f_2, \ldots, f_n) = \sum_{i=1}^{n} \frac{f_i(1-f_i)}{(-2w_i+1)f_i+w_i 2}$ (figure 1). Une entropie décentrée a également été proposée par Lallich et al. (2007). Cette approche différente propose de transformer les fréquences p_i d'un noeud grâce à une fonction qui change W en distribution uniforme. Pour le cas à deux classes, la fonction de transformation est composée de deux fonctions affines $\pi = \frac{p}{2w}$ si $0 \leq p \leq w$ et $\pi = \frac{p+1-2w}{2(1-w)}$ si $w \leq p \leq 1$. L'entropie décentrée est une entropie classique appliquée sur la distribution transformée. Cette méthode peut être appliquée sur n'importe quelle mesure d'entropie.

3 Évaluation des arbres dans le cas déséquilibré

Mesures de performance Il existe différentes mesures pour évaluer un modèle de prédiction. La plupart sont basées sur la matrice de confusion qui croise la classe réelle des individus du jeu d'apprentissage avec la classe prédite par le modèle, et permet de calculer les taux de vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN). Certaines mesures évaluent les performances d'un modèle sur une modalité spécifique comme le taux de rappel ($\frac{VP}{VP+FN}$) et le taux de précision ($\frac{VP}{VP+FP}$). La F-mesure est la moyenne harmonique du rappel et de la précision. D'autres mesures ne distinguent pas les classes : on peut citer le taux d'erreur global, la sensibilité et la spécificité. Ces dernières sont moins intéressantes pour nous, car de par leur construction elles favorisent la classe majoritaire (on peut néanmoins citer la mesure Pragma Thomas et al. (2007) qui permet à l'utilisateur de spécifier l'importance accordée à chaque classe ainsi que ses préférences en termes de rappel et de précision). Le rappel et la précision sont donc les mesures les plus adaptées concernant la prédiction d'une classe rare spécifique. Cette classe rare sera également appelée classe d'intérêt, et les individus appartenant à cette classe les individus positifs.

La matrice de confusion est obtenue en appliquant une règle d'affectation à chaque feuille de l'arbre. Ceci n'est pas problématique quand la règle d'affectation est la règle majoritaire. Mais avec un critère asymétrique cette règle n'est plus adaptée (Marcellin et al. (2006)) : si

l'on considère que la pire situation est la distribution W, et donc que la probabilité de la classe i est w_i dans le cas le plus incertain, alors aucune décision ne peut être prise pour les feuilles présentant cette distribution. Ainsi les feuilles où la classe d'intérêt est mieux représentée que dans le cas de référence ($f_i > w_i$) devraient être affectées à la classe i. Cette règle simple et intuitive pourrait être remplacée par un test statistique, comme nous l'avons proposé avec l'intensité d'implication Ritschard et al. (2007) par exemple. Pour éviter les limitations d'une règle, nous ferons varier le seuil de décision entre 0 et 1 pour obtenir un graphique rappel / précision sur la classe d'intérêt. Ceci nous permet de voir si une méthode domine l'autre pour les différents seuils de décision possibles.

Courbes ROC Les courbes ROC (*Receiver operating characteristics*) constituent un outil adapté à la visualisation des performances d'un classifieur pour une classe spécifique. De nombreux travaux en exposent les principes (Egan (1975); Fawcett (2006)). Premièrement, un score (probabilité d'appartenance à la classe d'intérêt) doit être calculé pour chaque individu. Pour les arbres de décision, ce score est la proportion d'individus positifs dans la feuille où il a été classé. Puis tous les individus sont représentés dans un espace taux de faux positifs / taux de vrais positifs, de manière cumulative, du mieux noté au moins bien noté. Une courbe ROC proche de la diagonale principale indique que le modèle n'apporte aucune information utilise au sujet de la classe. *A contrario* une courbe ROC présentant un point en $[0, 1]$ signifie que le modèle sépare parfaitement les individus positifs des individus négatifs. L'aire située sous la courbe (*Area Under Curve*, AUC) synthétise l'information de la courbe ROC.

4 Évaluations

Jeux de données et modèles comparés Notre étude est basée sur des arbres de décision évalués en 10-validation croisée pour éviter les problèmes de sur-apprentissage sur la classe majoritaire. Pour chaque jeu de données on considère l'entropie quadratique et l'entropie asymétrique. Le critère d'arrêt choisi pour limiter les problèmes de sur-apprentissage est un gain d'information minimal de 3%, les autres critères d'arrêt classiques comme le support des feuilles ou la profondeur maximale de l'arbre favorisant la classe majoritaire (aucun élagage *a posteriori* n'est appliqué). Nous avons sélectionné 11 jeux de données présentés tableau 1. La classe a toujours deux modalités, et on se concentre sur la prédiction de la moins fréquente. Un premier groupe de jeux de données provient de l'UCI repository (Hettich et Bay (1999)). Pour le jeu de données *letter* (reconnaissance de lettres manuscrites) on considère la reconnaissance de la lettre 'a' face à toutes les autres (*letter_a*), puis les voyelles contre les consonnes (*letter_vowels*). Les classes du jeu de données *Satimage* ont été fusionnées tel que proposé par Chen et al. (2004). Les jeux de données *Mammo1* et *Mammo2* sont des données réelles issues du dépistage du cancer du sein obtenues dans le cadre d'un partenariat industriel. L'objectif est de prédire si des zones situées sur des mammographies numériques sont des cancers ou pas. Ce dernier exemple fournit une bonne illustration de l'apprentissage sur données déséquilibrées : l'oubli d'un cancer peut conduire à la mort de la patiente, ce qui rend la prédiction de cette classe très importante. Une bonne précision est cependant nécessaire, le coût psychologique et monétaire d'une fausse alarme restant très élevé.

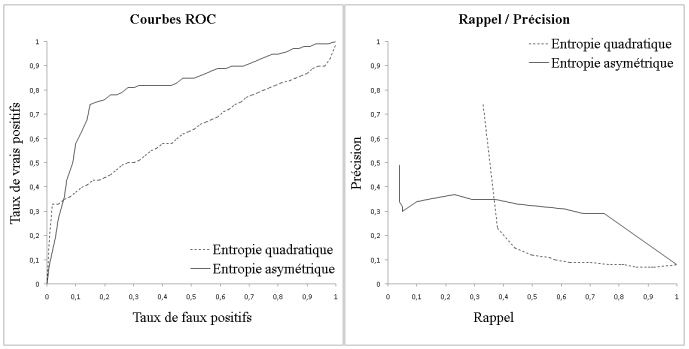

FIG. 2 – *Resultats pour Mammo1*

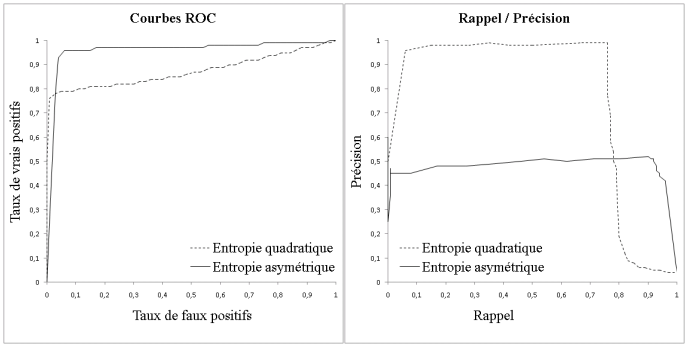

FIG. 3 – *Resultats pour Letter_a*

Résultat et interprétation Le tableau 1 présente les valeurs du critère AUC pour chaque jeu de données. Les figures 2,3,4 et 5 montrent les courbes ROC et les graphiques rappel / précision pour les jeux de données *Mammo1*, *Letter_a*, *Waveform_merged* et *Satimage*.

Les graphiques rappel / précision montrent que pour les forts taux de rappel, la précision est plus élevée avec le critère asymétrique. Ceci signifie que les règles de décision obtenues à partir d'un arbre basé sur l'entropie asymétrique sont plus performantes pour la prédiction de la classe minoritaire. Sur les deux jeux de données réelles (Figures 2) on voit que si on cherche à maximiser le rappel (soit minimiser le nombre de cancers manqués, ou faux négatifs), on obtient moins de faux positifs avec l'entropie asymétrique ; ce qui est l'effet recherché.

L'analyse des courbes ROC montre que l'utilisation de l'entropie asymétrique améliore le critère AUC (tableau 1). Mais le plus important est la forme des courbes. Les courbes ROC de l'entropie quadratique sont globalement meilleures sur la partie gauche du graphique, c'est-à-dire pour les scores élevés. Puis les deux courbes se croisent, et sur la partie droite le critère asymétrique est toujours dominant. Ainsi plus le score est faible, plus l'entropie asymétrique est adaptée. Nous avons vu en section 2 que pour la prédiction d'événements rares, il est préférable d'utiliser des seuils d'acceptation bas (on accepte une feuille comme appartenant à

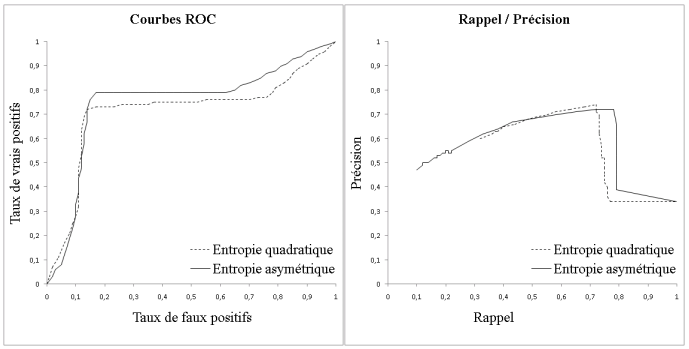

FIG. 4 – *Resultats pour Waveform_merged*

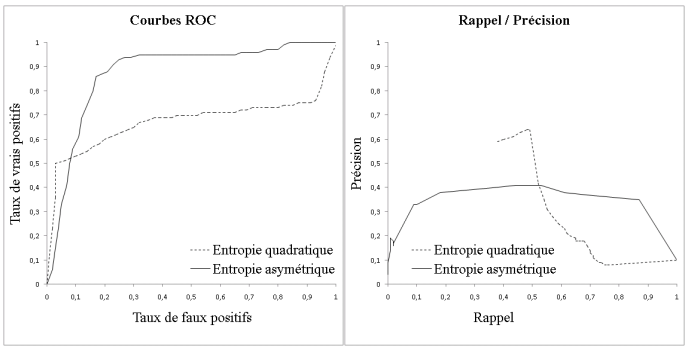

FIG. 5 – *Resultats pour Satimage*

la classe minoritaire si la fréquence observée de la classe minoritaire excède la probabilité de cette classe dans la distribution de référence). Ainsi les courbes ROC indiquent l'utilité d'une entropie asymétrique pour la prédiction d'une classe minoritaire.

Les deux remarques précédentes signifient que pour la recherche de 'pépites' de classe minoritaire, on aura de meilleurs rappel et précision en utilisant un critère asymétrique. En d'autres termes si on accepte de prédire la classe minoritaire pour un seuil inférieur à 50%, alors plus le score est faible, meilleur sera le gain en rappel et précision dus à l'utilisation d'un critère asymétrique.

5 Conclusion

Nous avons donc évalué la manière dont l'utilisation d'un critère d'éclatement basé sur une entropie asymétrique pour construire des arbres de décision sur des jeux de données déséquilibrés influence la qualité de la prédiction de la classe rare. Si les modèles proposés sont moins performants sur des mesures globales comme le taux d'erreur, les courbes ROC comme

Dataset	Nb ind	Nb var	Prop classe min	AUC entropie quadra	AUC entropie asy
Breast	699	9	34%	0.9288	0.9359
Letter_a	2000	16	4%	0.8744	0.9576
Letter_vowels	2000	16	23%	0.8709	0.8818
Pima	768	8	35%	0.6315	0.6376
Satimage	6435	36	10%	0.6715	0.8746
Segment_path	2310	19	14%	0.9969	0.9985
Waveform_merged	5000	40	34%	0.713	0.749
Sick	3772	29	6%	0.8965	0.9572
Hepatisis	155	19	21%	0.5554	0.6338
Mammo1	6329	1038	8%	0.6312	0.8103
Mammo2	3297	1038	15%	0.6927	0.8126

TAB. 1 – *Jeux de données.*

le comportement du rappel et de la précision en fonction du seuil d'acceptation révèlent que les modèles basés sur l'entropie asymétrique donnent de meilleurs résultats que ceux construits avec une entropie standard, pour les seuils de décisions bas. De nombreux problèmes connexes n'ont pas été abordés dans cet article et feront l'objet de travaux futurs : le premier est le choix de la distribution de référence W, qui pourrait s'adapter pour chaque noeud à la distribution du noeud parent, se rapprochant ainsi des arbres bayésiens (Chai et al. (2004)). Le critère d'arrêt sur les arbres asymétriques devra également être adapté aux jeux déséquilibrés. Le troisième point est la question de la règle d'assignation d'une classe à chaque feuille. On pourra pour cela considérer des règles statistiques, ou des mesures de qualité des règles ; ou utiliser les graphiques que nous avons proposés dans cet article, en cherchant des points optimaux sur le graphe rappel / précision ou la courbe ROC avec le point BEP (*break-even Point*, Sebastiani (2002)), pour trouver le meilleur rapport, ou encore le critère Pragma de Thomas et al. (2007). Enfin, les concepts exposés dans cet article devront être adaptés aux cas à plus de deux modalités. Différents problèmes se posent alors, pour l'extension du critère, les règles d'affectation et l'évaluation des modèles quand deux classes ou plus sont considérées comme pertinentes.

Références

Aczel, J. et Z. Daroczy (1975). *On Measures of Information and Their Characterizations*, Volume 114. NY, S. Francisco, London : 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.

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

Chai, X., L. Deng, Q. Yang, et Ling (2004). Test-cost sensitive naive bayes classification. *Proceedings of the Fourth IEEE International Conference on Data Mining ICDM'04*, 51–58.

Chen, C., A. Liaw, et L. Breiman (2004). Using random forest to learn imbalanced data.

Egan, J. (1975). Signal detection theory and roc analysis. *Series in Cognition and Perception*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letter 27*(8), 861–874.

Hettich, S. et S. D. Bay (1999). The uci kdd archive.

Lallich, S., P. Lenca, et B. Vaillant (2007). Construction d'une entropie décentrée pour l'apprentissage supervisé. *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*, 45–54.

Marcellin, S., D. Zighed, et G. Ritschard (2006). An asymmetric entropy measure for decision trees. *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France*, 1292–1299.

Provost, F. (2000). Learning with imbalanced data sets. *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.

Ritschard, G., D. Zighed, et S. Marcellin (2007). Données déséquilibrées, entropie décentrée et indice d'implication. *4èmes Rencontres Internationales Analyse Statistique Implicative (ASI 4)*.

Rényi, A. (1960). On measures of entropy and information. *4th Berkely Symp. Math. Statist. Probability 1*, 547–561.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv. 34(1)*, 1–47.

Thomas, J., P.-E. Jouve, et N. Nicoloyannis (2007). Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés. *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*.

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

Summary

We propose to evaluate the quality of decision trees grown on imbalanced datasets with a splitting criterion based on an asymmetric entropy measure. To deal with the class imbalance problem in machine learning, especially with decision trees, different authors proposed such asymmetric splitting criteria. After the tree is grown a decision rule has to be assigned to each leaf. The classical bayesian rule that selects the more frequent class is irrelevant when the dataset is strongly imbalanced. A best suited assignment rule taking asymmetry into account must be adopted. But how can we then evaluate the resulting prediction model? Indeed the usual error rate is irrelevant when the classes are strongly imbalanced. Appropriate evaluation measures are required in such cases. We consider ROC curves and recall/precision graphs for evaluating the performance of decision trees grown from imbalanced datasets with an asymmetric splitting criterion, and with a symmetric one.