

# Évaluation des critères asymétriques pour les arbres de décision

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

\*Université Lumière Lyon 2  
{abdelkader.zighed,simon.marcellin}@univ-lyon2.fr  
\*\*Université de Genève  
Gilbert.ritschard@unige.ch

**Résumé.** Pour construire des arbres de décision sur des données déséquilibrées, des auteurs ont proposés des mesures d'entropie asymétriques. Le problème de l'évaluation de ces arbres se pose ensuite. Cet article propose d'évaluer la qualité d'arbres de décision basés sur une mesure d'entropie asymétrique.

## 1 Introduction

L'apprentissage supervisé sur données déséquilibrées fait l'objet de nombreux travaux (Provost (2000)). Pour le cas des arbres de décision, différents auteurs ont proposé d'utiliser des mesures d'entropie prenant en compte l'asymétrie pour la recherche du meilleur éclatement. Nous avons ainsi proposé une axiomatique permettant de définir une famille de mesures asymétriques (Zighed et al. (2007)). Comment évaluer la qualité des arbres construits avec de telles mesures ? En effet, les critères de performances globaux (comme le taux d'erreur) ne prennent pas en compte l'asymétrie des classes. Ceux qui évaluent les performances du modèle sur une seule classe sont tributaires de la règle d'affectation d'une classe dans chaque feuille. Or, dans le cas de données déséquilibrées, la règle majoritaire utilisée habituellement ne convient pas. Nous proposons donc une méthodologie et une évaluation des arbres construits avec une entropie asymétrique.

## 2 Méthodes d'évaluation

Nous avons retenu deux méthodes pour évaluer les arbres de décisions asymétriques : les courbes ROC et les graphes rappel / précision. Les courbes ROC permettent d'évaluer la structure des arbres indépendamment du déséquilibre des classes (Provost et Fawcett (1997)). Les graphes rappel / précision permettent quant à eux d'évaluer les performances du modèle sur une classe, en faisant varier la règle d'affectation. Ces deux méthodes permettent ainsi de tenir compte des deux problèmes cités en introduction.