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.

3 Expérimentations et résultats

Nous avons mené des tests sur 11 jeux de données, dont 2 sont des données réelles issues du dépistage du cancer du sein. La proportion de la classe minoritaire varie de 4% à 35%. Nous avons construit deux modèles en 10-validation croisée : un arbre de décision utilisant l'entropie quadratique, et un arbre utilisant l'entropie asymétrique. Le critère d'arrêt a été fixé à 3% de gain d'information, et la distribution de référence de l'entropie asymétrique au déséquilibre du jeu de départ. Nous observons les résultats sur la classe minoritaire. Nous pouvons résumer les résultats obtenus en trois points. Premiérement, le critère AUC calculé à partir des courbes ROC est systématiquement supérieur en utilisant l'entropie asymétrique. Deuxièmement, la comparaison des courbes ROC entre les deux types d'arbres montre que la courbe ROC de l'entropie asymétrique est dominée sur la partie gauche du graphique (seuil d'acceptation élevé) mais domine sur la partie droite. Enfin et de la même manière, les graphes rappel / précision montrent qu'à rappel égal, l'entropie asymétrique est moins précise pour les forts seuils d'acceptation mais permet une meilleure précision sur les faibles seuils.

4 Conclusion et perspectives

Ainsi si on utilise un seuil d'acceptation bas pour classer les feuilles d'un arbre de décision, les courbes ROC comme le critère rappel / précision encouragent l'utilisation d'une entropie asymétrique lorsque les jeux de données sont déséquilibrés. Nous considérons quatre pistes pour étendre notre travail. D'une part, l'utilisation d'un critère adaptatif cherchant à s'écarter à chaque éclatement de la distribution du noeud parent. Deuxièmement, l'élaboration d'un critère d'arrêt adapté aux données déséquilibrées. Le présent travail nous permettra de proposer une méthode pour le choix de la règle d'affectation. Enfin, les outils que nous proposons pour adapter les arbres de décision aux données déséquilibrées seront étendus pour les cas à plus de deux modalités.

Références

Provost, F. (2000). Learning with imbalanced data sets. *Invited paper for the AAAIŠ2000 Workshop on Imbalanced Data Sets*.

Provost, F. J. et T. Fawcett (1997). Analysis and visualization of classifier performance : Comparison under imprecise class and cost distributions. *Knowledge Discovery and Data Mining*, 43–48.

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

Summary

To build decision trees on imbalanced datasets, authors proposed asymmetric entropies. Then the problem of evaluating those trees has to be solved. This paper proposes to evaluate the quality of decision trees based on asymmetric entropy measure.