Chapitre 7 : Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée

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

*Département d'économétrie, Université de Genève
**Laboratoire ERIC, Université de Lyon 2
gilbert.ritschard@unige.ch, {abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
http://mephisto.unige.ch, 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. Une première expérimentation sommaire est présentée.

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'écarter 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