Qualité d'ajustement d'arbres d'induction

Gilbert Ritschard*, Djamel A. Zighed**

*Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch http://mephisto.unige.ch
**Laboratoire ERIC, Université Lumière Lyon 2
zighed@univ-lyon2.fr http://eric.univ-lyon2.fr

Résumé. Cet article discute des possibilités de mesurer la qualité de l'ajustement d'arbres d'induction aux données comme cela se fait classiquement pour les modèles statistiques. Nous montrons comment adapter aux arbres d'induction les statistiques du khi-2, notamment celle du rapport de vraisemblance utilisée dans le cadre de la modélisation de tables de contingence. Cette statistique permet de tester l'ajustement du modèle, mais aussi l'amélioration de l'ajustement qu'apporte la complexification de l'arbre. Nous en déduisons également des formes adaptées des critères d'information AIC et BIC qui permettent de sélectionner le meilleur arbre en termes de compromis entre ajustement et complexité. Nous illustrons la mise en œuvre pratique des statistiques et indicateurs proposés avec un exemple réel.

Mots clés : arbre d'induction, qualité d'ajustement, tests du khi-2, comparaison d'arbres

1 Introduction

Les arbres d'induction (Kass, 1980; Breiman et al., 1984; Quinlan, 1993; Zighed et Rakotomalala, 2000) sont l'un des outils les plus populaires d'apprentissage supervisé. Ils consistent à rechercher par éclatements successifs de sommets, une partition de l'ensemble des combinaisons de valeurs des prédicteurs optimale pour prédire la variable réponse. La prédiction se fait simplement en choisissant, dans chaque classe de la partition obtenue, la modalité la plus fréquente de la variable à prédire. Bien que leur utilisation première soit la génération d'arbres de décisions pour la classification, les arbres d'induction fournissent une description de la façon dont la distribution de la variable à prédire est conditionnée par les valeurs des prédicteurs. Ils nous indiquent par exemple comment la répartition entre clients solvables et insolvables est influencée par les attributs âge, sexe, niveau d'éducation, profession, etc. En ce sens les arbres d'induction sont donc des outils de modélisation de l'influence des prédicteurs sur la variable à prédire au même titre que par exemple la régression linéaire, la régression logistique ou la modélisation log-linéaire de tables de contingence multi-dimensionnelles. C'est essentiellement à cet aspect de modélisation descriptive, et en particulier à l'évaluation de la qualité de la description fournie par un arbre induit que nous nous intéressons dans cet article.

En modélisation statistique, qu'il s'agisse de régression linéaire, d'analyse discri-