Choix des conclusions et validation des règles issues d'arbres de classification

Vincent Pisetta*, Gilbert Ritschard**, Djamel A. Zighed*

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

Résumé. Cet article traite de la validation de règles dans un contexte de ciblage où il s'agit de déterminer les profils type des différentes valeurs de la variable à prédire. Les concepts de l'analyse statistique implicative fondée sur la différence entre nombre observé de contre-exemples et nombre moyen que produirait le hasard, s'avèrent particulièrement bien adaptés à ce contexte. Le papier montre comment les notions d'indice et d'intensité d'implication de Gras s'appliquent aux règles produites par les arbres de décision et présente des alternatives inspirées de résidus utilisés en modélisation de tables de contingence. Nous discutons ensuite sur un jeu de données réelles deux usages de ces indicateurs de force d'implication pour les règles issues d'arbres. Il s'agit d'une part de l'évaluation individuelle des règles, et d'autre part de leur utilisation comme critère pour le choix de la conclusion de la règle.

1 Introduction

Les arbres de décision utilisés depuis longtemps en statistique (Morgan et Sonquist, 1963; Kass, 1980) sont devenus, suite aux ouvrages de Breiman et al. (1984) et Quinlan (1993) des outils très populaires pour générer des règles de classification et plus généralement des règles de prédictions. On parle ainsi d'arbre de classification lorsque la variable à prédire est catégorielle et que ses valeurs représentent donc des classes. Cependant, et contrairement à ce que l'expression d'arbre de classification peut laisser entendre, la classification n'est pas le seul intérêt des arbres de décisions. Par exemple, en sciences sociales où il s'agit plus de comprendre comment des prédicteurs peuvent affecter les valeurs prises par la variable à prédire que de classer des individus, ils peuvent avoir un intérêt descriptif, ou encore, comme en marketing notamment, on peut les utiliser dans une optique de ciblage. Dans ce dernier cas, plutôt que de prédire la valeur de la réponse, il s'agit de repérer les profils typiques des individus appartenant à chacune des classes de la variable à prédire. On inverse en quelque sorte le problème en cherchant à caractériser les profils propres à la classe, plutôt que la classe à partir du profil.

L'évaluation de la qualité de l'arbre se fonde le plus souvent sur le taux d'erreur de classification. Ce taux de cas mal classés par les règles, qu'il soit calculé sur les données d'apprentissage, des données test ou encore en validation croisée est évidemment pertinent comme