Arbre BIC optimal et taux d'erreur

Gilbert Ritschard

Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Nous reconsidérons dans cet article le critère BIC pour arbres d'induction proposé dans Ritschard et Zighed (2003, 2004) et discutons deux aspects liés à sa portée. Le premier concerne les possibilités de le calculer. Nous montrons comment il s'obtient à partir des statistiques du rapport vraisemblance utilisées pour tester l'indépendance ligne-colonne de tables de contingence. Le second point porte sur son intérêt dans une optique de classification. Nous illustrons sur l'exemple du Titanic la relation entre le BIC et le taux d'erreur en généralisation lorsqu'on regarde leur évolution selon la complexité de l'arbre. Nous esquissons un plan d'expérimentation en vue de vérifier la conjecture selon laquelle le BIC minimum assurerait en moyenne le meilleur taux d'erreur en généralisation.

1 Introduction

La qualité des arbres de classification, comme pour d'autres classifieurs, est le plus souvent établie sur la base du taux d'erreur de classement en généralisation. Si l'on examine l'évolution de ce taux en fonction de la complexité du classifieur, il est connu qu'il passe par un minimum au delà duquel on parle de sur-apprentissage (*overfitting*). Intuitivement, l'explication de ce phénomène tient au fait qu'au delà d'un certain seuil, plus on augmente la complexité, plus l'arbre devient dépendant de l'échantillon d'apprentissage utilisé, au sens où il devient de plus en plus probable que de petites perturbations de l'échantillon entraîneront des modifications des règles de classification. Lorsqu'il s'agit d'utiliser l'arbre pour la classification, il semble dès lors naturel de retenir celui qui minimise le taux d'erreur en généralisation.

Mais comment s'assurer a priori que l'arbre induit sera celui qui minimisera le taux en généralisation ? Il s'agit de disposer d'un critère qui, tout en se calculant sur l'échantillon d'apprentissage, nous assure que le taux d'erreur sera en moyenne minimum pour tout ensemble de données supplémentaires. A défaut de pouvoir mesurer a priori le taux d'erreur en généralisation, on s'intéresse à la complexité qu'il s'agit de minimiser et l'on tentera de retenir le meilleur compromis entre qualité d'information sur données d'apprentissage et complexité. Le critère BIC (Bayesian Information Criteria) pour arbre que nous avons introduit dans Ritschard et Zighed (2003, 2004) pour comparer la qualité de la description des données fournies par différents arbres nous semble pouvoir être une solution de ce point de vue puisqu'il combine un critère d'ajustement (la déviance) avec une pénalisation pour la complexité (le nombre de paramètres). D'autres critères, dont la description minimale de données (Rissanen, 1983) et le message de longueur minimal, MML, (Wallace et Freeman, 1987) qui combinent également une qualité d'information et une pénalisation pour la complexité pourraient également s'avérer