Un critère d'évaluation Bayésienne pour la construction d'arbres de décision

Nicolas Voisine*, Marc Boullé*, Carine Hue **

* Orange Labs, 2 avenue Pierre Marzin 22300 Lannion
nicolas.voisine@orange-ftgroup.com, marc.boulle@orange-ftgroup.com
** GFI Informatique, 11 rue Louis de Broglie 22300 Lannion, chue@gfi.fr

Résumé. Nous présentons dans cet article un nouvel algorithme automatique pour l'apprentissage d'arbres de décision. Nous abordons le problème selon une approche Bayésienne en proposant, sans aucun paramètre, une expression analytique de la probabilité d'un arbre connaissant les données. Nous transformons le problème de construction de l'arbre en un problème d'optimisation : nous recherchons dans l'espace des arbres de décision, l'arbre optimum au sens du critère Bayésien ainsi défini, c'est à dire l'arbre maximum a posteriori (MAP). L'optimisation est effectuée en exploitant une heuristique de pré-élagage. Des expérimentations comparatives sur trente bases de l'UCI montrent que notre méthode obtient des performances prédictives proches de celles de l'état de l'art tout en étant beaucoup moins complexes.

1 Introduction

La construction d'arbres de décision à partir de données est un problème qui a commencé à être traité en 1963 en construisant le premier arbre de régression pour prédire des variables numériques (Morgan et Sonquist, 1963). Suite à leurs travaux, toute une littérature a vu le jour pour décrire des modèles d'arbre soit pour des variables à prédire numériques, les arbres de régression, soit pour des variables catégorielles, les arbres de décision. On pourra se référer à l'ouvrage *« graphe d'induction »* (Zighed et Rakotomalala, 2000) pour de plus amples détails sur les différentes méthodes d'arbres de décision. Les méthodes CHAID (Kass, 1980) et ID3 (Quinlan, 1986) du début des années 80 sont des méthodes qui restent encore des références à citer. Mais ce sont les méthodes CART (Breiman et al., 1984) et la méthode C4.5 (Quinlan, 1993) dans les années 90 qui sont les références pour évaluer les performances de nouveaux algorithmes. Les premiers algorithmes d'apprentissage automatique d'arbre de décision sont basés sur un pré-élagage. Le principe de construction consiste, à partir de la racine de l'arbre, c'est-à-dire la totalité de l'ensemble d'apprentissage, à choisir parmi toutes les variables explicatives celle qui donne la meilleure partition selon un critère de segmentation. Puis de façon récursive, on applique l'algorithme de segmentation sur les feuilles. Le processus s'arrête quand pour chaque feuille on ne peut plus améliorer le critère de segmentation. Le choix de la variable de coupure et des points de coupure caractérise le processus de segmentation. La plupart des arbres (ID3, CHAID, CART, et C4.5) utilisent la théorie de l'information ou la théorie