Traitement de données volumineuses par ensembles d'arbres aléatoires

Pierre Geurts*
Université de Liège
Département d'Électricité, Électronique et Informatique
Service de méthodes stochastiques
Institut Montefiore
Sart Tilman B28
B-4000 Liège Belgique
* Chargé de recherches, FNRS, Belgique
p.geurts@ulg.ac.be
http ://www.montefiore.ulg.ac.be/~geurts

Résumé. Cet article présente une nouvelle méthode d'apprentissage basée sur un ensemble d'arbres de décision. Par opposition à la méthode traditionnelle d'induction, les arbres de l'ensemble sont construits en choisissant les tests durant le développement de manière complètement aléatoire. Cette méthode est comparée aux arbres de décision et au bagging sur plusieurs problèmes de classification. Grâce aux choix aléatoires des tests, les temps de calcul de cet algorithme sont comparables à ceux des arbres traditionnels. Dans le même temps, la méthode se révèle beaucoup plus précise que les arbres et souvent significativement meilleure que le bagging. Ces caractéristiques rendent cette méthode particulièrement adaptée pour le traitement de bases de données volumineuses.

1 Introduction

Actuellement, un des domaines de recherche les plus actifs en apprentissage automatique est l'étude des méthodes basées sur un ensemble de modèles (Dietterich 2000a). La plupart de ces méthodes consistent à se servir d'une méthode d'apprentissage classique pour construire plusieurs modèles et ensuite à agréger les prédictions de ces modèles pour donner une prédiction finale, potentiellement meilleure que les prédictions individuelles. Les méthodes d'ensemble diffèrent essentiellement par la manière dont sont construits les modèles ainsi que par la manière d'agréger leurs prédictions.

Une des catégories de méthodes les plus populaires est constituée par les méthodes de type perturbation et combinaison. Ces méthodes consistent à perturber un algorithme d'apprentissage de manière à ce qu'il fournisse des modèles différents à partir d'un même échantillon d'apprentissage. Les prédictions sont ensuite agrégées par un simple moyennage ou un vote à la majorité dans le cas de la classification. Le représentant le plus populaire de cette famille de méthode est le bagging (Breiman 1996) dans lequel les différents modèles sont obtenus en effectuant un ré-échantillonnage de l'ensemble d'apprentissage avant chaque étape d'induction. Ces méthodes d'ensemble ont beaucoup de succès récemment principalement à cause de l'amélioration de précision qu'elles apportent aux méthodes traditionnelles, telles que les arbres de