Un nouvel algorithme de forêts aléatoires d'arbres obliques particulièrement adapté à la classification de données en grandes dimensions

Thanh-Nghi Do[*,****], Stéphane Lallich[**]
Nguyen-Khang Pham[***], Philippe Lenca[*,****]

[*]Institut TELECOM, TELECOM Bretagne
UMR CNRS 3192 LabSTICC, Brest, France
tn.do|philippe.lenca@telecom-bretagne.eu
[**]Université Lyon, Laboratoire ERIC, Lyon 2, Lyon, France
stephane.lallich@univ-lyon2.fr
[***]IRISA, Rennes, France
pnguyenk@irisa.fr
[****]Université Européenne de Bretagne, France

Résumé. L'algorithme des forêts aléatoires proposé par Breiman permet d'obtenir de bons résultats en fouille de données comparativement à de nombreuses approches. Cependant, en n'utilisant qu'un seul attribut parmi un sous-ensemble d'attributs tiré aléatoirement pour séparer les individus à chaque niveau de l'arbre, cet algorithme perd de l'information. Ceci est particulièrement pénalisant avec les ensembles de données en grandes dimensions où il peut exister de nombreuses dépendances entre attributs. Nous présentons un nouvel algorithme de forêts aléatoires d'arbres obliques obtenus par des séparateurs à vaste marge (SVM). La comparaison des performances de notre algorithme avec celles de l'algorithme de forêts aléatoires des arbres de décision C4.5 et de l'algorithme SVM montre un avantage significatif de notre proposition.

1 Introduction

Les performances d'un classifieur dépendent de différents facteurs. Parmi ces derniers notons les paramètres nécessaires à son initialisation (par exemple le nombre de classes –ou clusters– pour un algorithme du type k-means), et les données utilisées pour construire le classifieur (par exemple la construction des ensembles d'apprentissage, de test et de validation). Afin d'atténuer l'influence des différents choix possibles et de compenser les limites des différents classifieurs pouvant être utilisés, la combinaison de classifieurs (ou encore méthode ensembliste) a retenu l'attention des chercheurs en apprentissage automatique depuis fort longtemps.

Les méthodes ensemblistes cherchent notamment à réduire la variance (erreur due à la variabilité des résultats en fonction de l'échantillon d'apprentissage) et/ou le biais (erreur de précision non dépendante de l'échantillon d'apprentissage) des algorithmes d'apprentissage (voir