Effets de la qualité des données sur la courbe d'apprentissage des forêts aléatoires

Yves Brostaux*

11 avril 2008

Résumé

Les forêts aléatoires ont été introduites par Leo Breiman (2001) comme un nouvel algorithme d'apprentissage, basé sur l'agrégation d'arbres de décision randomisés. Les effets de l'introduction de bruit et de variables parasites dans l'échantillon d'apprentissage sur la courbe d'apprentissage d'un classificateur de type forêt aléatoire ont été mesurés et comparés aux résultats d'un algorithme classique de génération d'arbre de décision inspiré par la méthode CART de Breiman (1984). Globalement, la vitesse moyenne de l'apprentissage est assez similaire entre les deux algorithmes, mais les forêts aléatoires exploitent mieux les échantillons de petites et de grandes tailles : leur courbe d'apprentissage commence plus bas et n'est pas affectée par la limitation asymptotique présente chez les arbres de décision uniques.

*Premier assistant à l'Unité de Statistique, Informatique et Mathématique appliquées de la FUSAGx