
Forêts de Mondrian : vitesses minimax et implémentation en ligne

Jaouad Mourtada*¹

¹Ecole Polytechnique – Ecole Polytechnique Université Paris Saclay – France

Résumé

Les forêts aléatoires, proposées par Breiman (2001), comptent parmi les algorithmes plus couramment utilisés en pratique dans des problèmes de classification et de régression.

En dépit d'une littérature théorique récente sur ces méthodes, celles-ci demeurent encore assez mal comprises en raison de leur aspect "boîte noire".

Nous considérons ici une variante récente de l'algorithme, les forêts de Mondrian introduites par (Lakshminarayanan et al., 2014) pour des raisons computationnelles, dont nous montrons qu'elle se prête à une analyse théorique fine.

En particulier, nous énonçons des résultats d'optimalité au sens minimax en classification et en régression non paramétriques en dimension arbitraire pour les forêts de Mondrian convenablement calibrées.

Nous mettons également en évidence un avantage des forêts de Mondrian sur les arbres du à une réduction du biais pour des fonctions de régression lisses, un effet similaire à celui établi par Arlot et Genuer (2014) pour d'autres modèles de forêts.

Nous discutons enfin une implémentation "en ligne" de l'algorithme, qui repose sur une procédure d'agrégation efficace sur les arbres de décision et possède des propriétés d'adaptativité à la régularité de la fonction de régression.

*Intervenant