Les forêts d'arbres extrêmement aléatoires : utilisation dans un cadre non supervisé

Kevin Dalleau*, Miguel Couceiro*
Malika Smail-Tabbone*

*Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Résumé. Dans ce travail, nous présentons une nouvelle méthode permettant le calcul de similarités entre objets basée sur les forêts d'arbres extrêmement aléatoires. L'idée principale de notre méthode est de séparer les données de manière itérative jusqu'à ce qu'une condition d'arrêt soit respectée, et de calculer une similarité basée sur la co-occurrence des instances dans les feuilles de chaque arbre obtenu. Nous évaluons la méthode sur un ensemble de jeux de données synthétiques et réels. Cette évaluation est basée sur la comparaison des similarités moyennes entre instances ayant la même étiquette aux similarités moyennes entre instances d'étiquette différente. Ces mesures sont comparables aux notions de similarités intracluster et intercluster, mais ont pour intérêt d'être agnostiques aux choix d'une méthode de clustering en particulier. L'étude empirique montre que la méthode permet effectivement de distinguer les individus n'appartenant pas aux même clusters. Les forêts d'arbres extrêmement aléatoires non supervisées ont des propriétés intéressantes, telles que : (i) l'invariance aux transformations monotones de variables, (ii) la robustesse aux variables corrélées, et (iii), la robustesse au bruit. Enfin, nous présentons les résulats obtenus par l'applicaton d'un algorithme de clustering hiérarchique agglomératif, en utilisant les matrices de similarité obtenues par notre méthode. Les résultats obtenus sur des jeux de données homogènes et hétérogènes sont prometteurs.

1 Introduction

De nombreux algorithmes d'apprentissage non supervisé se basent sur une mesure de similarité ou distance entre instances. Bien qu'il existe un grand nombre de métriques décrites dans la littérature, en pratique, l'ensemble de métriques disponibles est grandement réduit par les caractéristiques des données et de l'algorithme choisi. Le choix d'une distance peut impacter fortement la qualité d'un clustering.

Shi et Horvath proposent dans (Shi et Horvath (2006)) la méthode des *Unsupervised Random Forest* (URF), dérivant des random forests (RF, Breiman (2001)). Leur méthode permet de calculer des distances entre instances non étiquetées en utilisant les forêts d'abres aléatoires. Une fois la forêt construite, les données d'entraînement sont passées dans chacun des arbres. Chaque feuille contenant un nombre limité d'instances, et toutes les instances terminant dans les mêmes feuilles pouvant être considérées comme similaires, il est possible de définir une