CRAFTML, une forêt aléatoire efficace pour l'apprentissage multi-label extrême

Wissam Siblini*,**, Frank Meyer *
Pascale Kuntz**

*Orange Labs - Lannion, France
prenom.nom@orange.com,
**Laboratoire des Sciences du Numérique de Nantes (LS2N) - Nantes, France
prenom.nom@univ-nantes.fr

Résumé. L'apprentissage multi-label extrême (noté XML pour "eXtreme Multi-label Learning") considère de grands volumes de données où chaque observation est annotée avec quelques labels parmi des centaines de milliers de possibilités. Les méthodes basées sur les arbres, qui divisent hiérarchiquement l'apprentissage en sous-problèmes à petite échelle, sont particulièrement prometteuses dans ce contexte pour réduire les complexités d'apprentissage et de prédiction et pour ouvrir la voie à la parallélisation. Cependant, les meilleures approches actuelles n'exploitent pas la diversification des arbres qui a pourtant montré son efficacité dans les forêts aléatoires et elles ont recours à des stratégies de partitionnement complexes. Pour surmonter ces limites, nous introduisons ici un nouvel algorithme de forêt avec des arbres diversifiés et une stratégie de partitionnement adaptée à l'XML appelé CRAFTML. Des comparaisons expérimentales sur huit jeux de données tirés de la littérature extrême montrent qu'il est plus performant que les autres approches arborescentes de l'état de l'art.

1 Introduction

La classification multi-label a reçu une attention considérable au cours de la dernière décennie et récemment, stimulée par des applications impliquant de grands ensembles de données, elle a été étendue à des problèmes où le nombre de labels peut dépasser le million (Agrawal et al., 2013). Dans ce nouveau contexte, appelé apprentissage multi-label extrême (noté XML en anglais pour eXtreme Multi-Label Learning) où les données sont très creuses et ont un très grand nombre de dimensions, les algorithmes classiques ne parviennent pas à passer à l'échelle ou voient leurs performances prédictives se dégrader. Pour tenter de surmonter ces difficultés, la recherche s'est récemment orientée vers trois directions : (i) utiliser des astuces d'optimisation et la parallélisation sur des "superordinateurs" (Yen et al., 2017), (ii) réduire la dimension des données pour obtenir un problème latent soluble avec des approches multi-label classiques (Bhatia et al., 2015) ou (iii) partitionner hiérarchiquement le problème initial en sous-problèmes à petite échelle (Prabhu et Varma, 2014). La décomposition arborescente de la troisième stratégie présente plusieurs avantages. En découpant l'apprentissage en sous-tâches,