Vers une classification non supervisée adaptée pour obtenir des arbres de décision simplifiés

Olivier Parisot, Yoanne Didry, Pierrick Bruneau, Thomas Tamisier

Département Informatique, Systèmes et Collaboration (ISC)
Centre de Recherche Public - Gabriel Lippmann, Belvaux, Luxembourg
parisot@lippmann.lu

Résumé. L'induction d'arbre de décision est une technique puissante et populaire pour extraire de la connaissance. Néanmoins, les arbres de décision obtenus depuis des données issues du monde réel peuvent être très complexes et donc difficiles à exploiter. Dans ce cadre, cet article présente une solution originale pour adapter le résultat d'une classification non supervisée quelconque afin d'obtenir des arbres de décision simplifiés pour chaque cluster.

1 Introduction

Utilisée à l'origine comme un outil d'aide à la décision, les arbres de décision sont très populaires en fouille et en analyse visuelle des données. Ce succès s'explique notamment par le fait qu'ils utilisent un formalisme transparent et simple à comprendre (Murthy, 1998).

En pratique, la génération automatique d'arbres de décision depuis des données est possible grâce à l'induction d'arbre de décision, une technique bien connue (Quinlan, 1986). Malheureusement, les arbres de décision générés depuis des données issues du monde réel peuvent être très grands et difficiles à exploiter. De nombreuses approches de simplification ont donc été proposées. La plus connue, l'élagage (*pruning*) (Breslow et Aha, 1997), consiste à supprimer les parties de l'arbre qui ont un faible pouvoir explicatif. D'autre part, il existe des approches qui travaillent directement sur les données, en utilisant des méthodes de *preprocessing* (Parisot et al., 2013a). Enfin, la classification non supervisée (ou *clustering*) est un outil particulièrement utile pour la fouille de données, et à priori, cette technique peut donc être également exploitée pour obtenir des arbres de décision plus simples : en conséquence, une étude récente a proposé une nouvelle approche de classification non supervisée pour obtenir des arbres de décisions plus simples (Parisot et al., 2013b).

Dans cet article, nous présentons une méthode permettant d'adpater une classification non supervisée existante afin de simplifier l'arbre de décision obtenu à partir de chaque cluster.

2 Contribution

La méthode proposée (Figure 1) a pour but d'adapter une classification non supervisée pouvant être obtenue au moyen de toute technique existante non hiérarchique (k-means, EM,