Arbres de Décision Multi-Modes et Multi-Cibles

Frank Meyer, Fabrice Clerot

France Telecom R&D
Avenue Pierre Marzin
22307 Lannion cédex
franck.meyer@francetelecom.com
fabrice.clerot@francetelecom.com

Résumé. Nous présentons une nouvelle méthode d'induction d'arbre de décision appelée MuMTree (pour Multi Models Tree) utilisable pour les modes d'apprentissage supervisé, non supervisé, supervisé à plusieurs variables cibles. Nous présentons les différents principes nécessaires pour réaliser un tel arbre de décision. Nous illustrons ensuite, sur un cas de modélisation multi-cibles, les avantages de cette méthode par rapport à un arbre de décision classique.

1 Introduction

L'approche classique pour modéliser un problème avec N variables cibles est de décomposer le problème en N sous problèmes indépendants et d'utiliser un modèle par variable à expliquer. L'hypothèse sous jacente à cette décomposition est l'hypothèse d'indépendance des variables cibles entre elles. Dans de nombreux cas (données spatiales, données socio-économiques,...) cette hypothèse est fausse. Une méthode générant un seul modèle capable de prédire plusieurs variables cibles à la fois pourrait-elle être plus performante qu'une méthode avec plusieurs modèles spécialisés par variable cible ? Une telle méthode pourrait-elle être utilisée également pour l'apprentissage non supervisé ? Peut-on utiliser le cadre des arbres de décision, réputés pour leur efficacité et leur lisibilité pour construire une telle méthode ? Cet article apporte des réponses à ces différentes questions.

Il existe de très nombreuses méthodes d'induction d'arbre de décision. Les méthodes les plus connues sont pour le mode supervisé, (Kass, 1980), (Beiman, 1984), (Utgoff, 1991), (Quinlan, 1993), (Murthy, 1994) et pour le mode non supervisé, (Chavent, 1998), (Liu, 2000). A notre connaissance, aucune méthode d'induction d'arbres de décision ne permet de générer des modèles à la fois pour le mode d'apprentissage supervisé, non supervisé et multi-cibles.

L'objectif de MumTree est de fournir un principe général de construction d'arbres de décision pour les modes d'apprentissage supervisé et non supervisé, pour les types d'attributs courants (numériques, symboliques). Il en découlera aussi le mode supervisé multi-cibles.