Construction interactive d'arbres de décisions avec des variables mixtes

François Poulet

ESIEA – Pôle ECD
38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé
53000 Laval
poulet@esiea-ouest.fr
http://visu.ecd.free.fr

Résumé. Nous présentons une extension d'algorithmes de construction interactive d'arbres de décisions aux cas des variables intervalle et taxonomiques. Les algorithmes présentés sont ainsi capables de traiter indifféremment des variables continues, intervalles et taxonomiques (ou un quelconque mélange de ces trois différents types). Ce type d'approche, centrée utilisateur, est ce qui caractérise la fouille visuelle de données. Ces algorithmes peuvent fonctionner en mode 100% manuel (c'est l'utilisateur seul qui créé l'arbre de décision) ou en mode mixte (coopération entre l'utilisateur et une méthode automatique pour trouver la meilleure coupe pour le noeud courant de l'arbre, ici basée sur des SVM : Séparateurs à Vaste Marge). Après avoir décrit l'adaptation de ces algorithmes aux cas des variables intervalles et taxonomiques, nous présentons les résultats que nous avons obtenus sur différents ensembles de données artificiels.

1. Introduction

L'extraction de Connaissances dans les Données (ECD) peut être définie [Fayyad et al., 1996] comme le processus non trivial de découverte de connaissances valides, nouvelles, potentiellement utilisables et compréhensibles dans les données. Dans la plupart des outils existant, la visualisation n'intervient en général que lors de deux étapes du processus de fouille : dans l'une des toutes premières étapes pour voir les données ou leur distribution, et dans l'une des toutes dernières étapes pour voir le résultat obtenu. Entre ces deux étapes, il y a exécution d'un algorithme automatique. Le rôle de l'utilisateur est donc "simplement" de régler les paramètres de l'algorithme, puis de lancer son exécution et d'attendre les résultats.

De nouvelles méthodes sont récemment apparues [Ankerst et al., 2001], [Poulet, 1999], [Wong, 2001] essayant d'augmenter l'implication de l'utilisateur dans le processus de fouille notamment par le biais d'un rôle plus important de la visualisation [Aggarwal, 2001], [Shneiderman, 2002]. Cette nouvelle approche s'appelle la fouille visuelle de données.

Nous présentons deux algorithmes de classification supervisée et plus précisément de construction interactive d'arbres de décision. Ces algorithmes de classification supervisée utilisent à la fois les capacités humaines en reconnaissance de formes et la puissance de calcul des algorithmes automatiques dans une approche centrée utilisateur.