Arbres de Décision

Ricco RAKOTOMALALA
Laboratoire ERIC
Université Lumière Lyon 2
5, av. Mendés France
69676 BRON cedex
e-mail : rakotoma@univ-lyon2.fr

Résumé

Après avoir détaillé les points clés de la construction d'un arbre de décision à partir d'un petit exemple, nous présentons la méthode CHAID qui permet de répondre de manière cohérente à ces spécifications. Nous la mettons alors en œuvre en utilisant un logiciel gratuit téléchargeable sur Internet. Les opérations sont décrites à l'aide de plusieurs copies d'écrans. L'accent est mis sur la lecture et l'interprétation des résultats. Nous mettons en avant également l'aspect interactif, très séduisant, de la construction des arbres. De manière plus générale, nous essayons de mettre en perspective les nombreuses techniques d'induction d'arbres en faisant le bilan de l'état actuel de la recherche dans le domaine.

Mots-clés : Arbres de décision, segmentation, discrimination, apprentissage automatique

Abstract

In this paper, we show the key points of the induction of decision trees from a small dataset and we present the CHAID algorithm. Using a free software, the induction algorithm is detailed with several screenshots. We put emphasis on the interpretation of results and the interaction potentiality of the method. In a more general way, we try to give a comprehensive survey of the numerous variants which have been developed these last years.

Keywords: Decision Tree, Induction Tree, Supervised machine learning, Data mining

1 Introduction

La construction des arbres de décision à partir de données est une discipline déjà ancienne. Les statisticiens en attribuent la paternité à Morgan et Sonquist (1963) qui, les premiers, ont utilisé les arbres de régression dans un processus de prédiction et d'explication (AID – Automatic Interaction Detection). Il s'en est suivi toute une famille de méthodes, étendues jusqu'aux problèmes de discrimination et classement, qui s'appuyaient sur le même paradigme de la représentation par arbres (THAID -- Morgan et Messenger, 1973 ; CHAID -- Kass, 1980). On considère généralement que cette approche a connu son apogée avec la méthode CART (Classification and Regression Tree) de Breiman *et al*. (1984) décrite en détail dans une monographie qui fait encore référence aujourd'hui.

En apprentissage automatique, la plupart des travaux s'appuient sur la théorie de l'information. Il est d'usage de citer la méthode ID3 de Quinlan (Induction of Decision Tree – Quinlan 1979) qui, lui même, rattache ses travaux à ceux de Hunt (1962). Quinlan a été un acteur très actif dans la deuxième moitié des années 80 avec un grand nombre de publications où il propose un ensemble d'heuristiques pour améliorer le comportement de son système. Son approche a pris un tournant important dans les années 90 lorsqu'il présenta la méthode C4.5 qui est l'autre référence incontournable dès lors que l'on veut citer les arbres de décision (1993). Il existe bien une