

Interactive Clustering Tree: Une méthode de classification descendante adaptée aux grands ensembles de données

Ricco Rakotomalala*, Tanguy Le Nouvel**

* Laboratoire ERIC - Université Lyon 2
5 av. Mendès France - 69500 Bron.
ricco.rakotomalala@univ-lyon2.fr.

** SPAD, 63 Boulevard de Ménilmontant - 75011 Paris.
tlenouvel@spad.eu

Résumé. Nous présentons une nouvelle méthode de segmentation non supervisée, particulièrement bien adaptée aux gros volumes de données et aux besoins opérationnels du secteur Banque Assurance. Cette méthode de classification descendante hiérarchique présente la segmentation finale sous la forme d'un arbre de décision dont l'appartenance aux classes (ou segments) dépend de règles logiques faisant intervenir les variables de l'analyse. De ce fait, la méthode hérite des propriétés inhérentes aux arbres de décision (interactivité, choix des variables de coupure, élagage/développement de l'arbre). Il en résulte une segmentation très simple d'interprétation et directement opérationnelle pour l'affectation d'un nouvel individu à l'une des classes. Enfin, elle intègre la possibilité de construire l'arbre avec d'autres variables que celles dont on mesure l'inertie. En ce sens, nous pouvons la considérer comme une généralisation au cas multicibles, plusieurs variables à prédire simultanément, des méthodes supervisées. Cet article présente les fondements théoriques de la méthode et s'appuie sur un exemple pratique pour illustrer les résultats obtenus et les comparer aux méthodes usuelles.

1 Introduction

La typologie, appelée également classification ou apprentissage non-supervisé, vise à segmenter une population en groupes homogènes d'individus du point de vue d'un ensemble de variables, de telle sorte que : deux individus d'un même groupe se ressemblent le plus possible ; deux individus de groupes distincts diffèrent le plus possible. La typologie synthétise et résume de manière concise une réalité complexe matérialisée par un ensemble de variables. Cette méthode exploratoire est un formidable outil pour exploiter les masses d'informations gigantesques stockées par les compagnies d'assurance et les organismes bancaires ainsi que les sociétés de vente par correspondance. La multiplicité des thématiques abordées : Risque, Tarification, Marketing opérationnel, Fidélisation, étude des points de vente, sont autant de terrains d'application de ce type de méthode.