SoTree : Auto-organisation topologique et hiérarchique des données

Hanane Azzag, Mustapha Lebbah

Université Paris 13, LIPN-UMR 7030 - CNRS
99, avenue Jean-Baptiste Clément
93430 Villetaneuse, France
{hanane.azzag, mustapha.lebbah}@lipn.univ-paris13.fr

Résumé. Nous proposons dans cet article d'introduire une nouvelle approche pour la classification non supervisée hiérarchique. Notre méthode nommée SoTree consiste à construire, d'une manière autonome et simultanée, une partition topologique et hiérarchique des données. Chaque "cluster" de la partition est associé à une cellule d'une grille 2D et est modélisé par un arbre, dont chaque noeud représente une donnée. Nous évaluerons les capacités et les performances de notre approche sur des données aux difficultés variables. Les résultats préliminaires obtenus sont encourageants et prometteurs pour continuer dans cette direction.

1 Introduction

Le problème de la classification de données est identifié comme une des problématiques majeures en extraction des connaissances à partir de données. Depuis des décennies, de nombreux sous-problèmes ont été identifiés, par exemple la sélection de données ou de variables, la variété des espaces de représentation (numérique, symbolique,...), l'incrémentalité, la nécessité de découvrir des concepts, d'obtenir une hiérarchie, etc. La popularité, la complexité et toutes ces variantes du problème de la classification de données (Jain et al. (1999)) ont donné naissance à une multitude de méthodes de résolution. Ces méthodes peuvent à la fois faire appel à des principes heuristiques ou encore mathématiques.

Dans ce travail, les méthodes qui nous intéressent sont celles qui font de la classification topologique et hiérarchique non supervisée de données (Vesanto et Alhoniemi (2000); Vesanto et Sulkava (2002); Ambroise et al. (1998); Golli et al. (2007)). L'avantage des cartes topologiques est de pouvoir représenter et visualiser un grand ensemble de données, ainsi que les regroupements que l'on peut y effectuer. Elles permettent aussi d'utiliser une représentation cartographique visuelle et familière à l'utilisateur.

Dans ce travail nous cherchons à introduire une nouvelle approche de classification simultanée : hiérarchique et topologique nommée SoTree : Self-organizing Tree. L'idée est de déplacer de manière "autonome" des données sur une grille 2D où chaque cellule représente un arbre de données. Nous disposerons ainsi d'une classification horizontale topologique des données sur la grille et d'une classification verticale hiérarchique au niveau de chaque cellule. La fonction topologique de notre algorithme est inspirée des cartes topologiques de Kohonen