Un algorithme de classification topographique non supervisée à deux niveaux simultanés

Guénaël Cabanes, Younès Bennani

LIPN - UMR 7030
Université Paris 13 - CNRS
99, av. J-B Clément - F-93430 Villetaneuse
{cabanes, younes}@lipn.univ-paris13.fr

Résumé. Une des questions les plus importantes pour la plupart des applications réelles de la classification est de déterminer un nombre approprié de groupes (*clusters*). Déterminer le nombre optimal de groupes est un problème difficile, puisqu'il n'y a pas de moyen simple pour connaître ce nombre sans connaissance a priori. Dans cet article, nous proposons un nouvel algorithme de classification non supervisée à deux niveaux, appelé S2L-SOM (Simultaneous Two-level Clustering - Self Organizing Map), qui permet de déterminer automatiquement le nombre optimal de groupes, pendant l'apprentissage d'une carte auto-organisatrice. L'estimation du nombre correct de groupes est en relation avec la stabilité de la segmentation et la validité des groupes générés. Pour mesurer cette stabilité nous utilisons une méthode de sous-échantillonnage. Le principal avantage de l'algorithme proposé, comparé aux méthodes classiques de classification, est qu'il n'est pas limité à la détection de groupes convexes, mais est capable de détecter des groupes de formes arbitraires. La validation expérimentale de cet algorithme sur un ensemble de problèmes fondamentaux pour la classification montre sa supériorité sur les méthodes standards de classification à deux niveaux comme SOM+K-Moyennes et SOM+Hierarchical-Agglomerative-Clustering.

1 Introduction

La classification non supervisée, ou clustering, est un outil très performant pour la détection automatique de sous-groupes pertinents (ou *clusters*) dans un jeu de données, lorsqu'on n'a pas de connaissances a priori sur la structure interne de ces données. Les membres d'un même cluster doivent êtres similaires entre eux, contrairement aux membres de groupes différents (homogénéité interne et séparation externe). La classification non supervisée joue un rôle indispensable pour la compréhension de phénomènes variés décrits par des bases de données. Un problème de regroupement peut être défini comme une tâche de partitionnement d'un ensemble d'items en un ensemble de sous-ensembles mutuellement disjoints. La classification est un problème de regroupement qui peux être considéré comme un des plus compétitifs en apprentissage non-supervisé. De nombreuses approches ont été proposées (Jain et Dubes, 1988). Les approches les plus classiques sont les méthodes hiérarchiques et les méthodes partitives.