Classification sous contraintes probabilistes par les cartes topologiques

Jihène SNOUSSI et Khalid BENABDESLEM

Université de Lyon, F69622, Lyon, France ;
Université Lyon1, Villeurbanne ;
LIESP, EA4125
{jsnoussi, kbenabde}@bat710.univ-lyon1.fr

Résumé. La classification automatique est un processus non supervisé qui vise à regrouper des données en un ensemble de classes hétérogènes. En outre, Différents travaux ont montré que l'intégration de contraintes peut augmenter le taux de ce processus de classification tout en diminuant le temps d'exécution. Cette nouvelle démarche a connu, ces dernières années, un travail bien étendu. La forme la plus répandue de ces dites contraintes est de type « Must-Link » dont le nom indique l'obligation d'avoir les données dans une même classe, et les contraintes « Cannot-Link » dont le nom indique l'interdiction d'avoir les données dans une même classe. Le travail présenté dans cet article décrit une nouvelle version des cartes topologiques que nous appelons « PrTM » (Probabilistic constrained Topological Map) pour intégrer des contraintes probabilistes. PrTM représente une variante d'un algorithme populaire des cartes topologiques probabilistes GTM (Generative Topographic Mapping). Pour valider notre approche, des comparaisons entre notre proposition « PrTM » et d'autres algorithmes de classification sous contraintes, sont présentées sur différentes bases de données issues de la littérature.

1 Introduction

L'extraction de connaissances par exploration des données fait souvent appel à des processus de classification automatique en mode non supervisé. Effectuer une classification, c'est mettre en évidence d'une part, les relations entre les différentes observations et d'autre part, les relations entre ces observations et leurs caractéristiques (leurs variables). A partir d'une certaine mesure de proximité ou de dissemblance, il s'agit de regrouper un ensemble de données en un ensemble de classes qui soient les plus hétérogènes possible (Saporta, 2006). Cependant, les algorithmes de classification automatique ont accès seulement à l'ensemble des variables; et il n'est fourni aucune information concernant l'affectation d'une observation à une classe. La prise en compte de ces connaissances additionnelles constitue un problème essentiel et un vrai défi pour la recherche actuelle puisqu'il s'agit à la fois de