Une approche algébrique au problème du consensus de partitions

Frédéric Dumonceaux, Guillaume Raschia, Marc Gelgon

Laboratoire d'Informatique Nantes Atlantique
(LINA, UMR CNRS 6241)
Université de Nantes
Rue Christian Pauc, La Chantrerie
44306 Nantes cedex 3, France
prenom.nom@univ-nantes.fr

Résumé. En classification non-supervisée, le consensus de partitions a pour objectif de produire une partition unique, représentant le consensus, à partir d'un ensemble de partitions où chacune est engendrée indépendamment des autres, voire avec des méthodologies différentes. En complément des techniques ayant leur qualité propre en terme de robustesse ou de passage à l'échelle, nous apportons un point de vue original sur le consensus de partitions, c'est-à-dire, par le biais de définitions algébriques qui permettent d'établir la nature des déductions pouvant être réalisées dans une approche systématique (p.ex. un système à base de connaissances). Nous fondons notre approche sur le treillis des partitions pour lequel nous montrons comment peuvent être adjoint des opérateurs dans le but de formuler une expression caractérisant le consensus à partir d'un ensemble de partitions.

1 Introduction

La classification non supervisée (clustering) de données constitue une tâche fondamentale et classique de structuration, pour l'analyse exploratoire de jeux de données. Elle a été l'objet d'un nombre considérable de travaux depuis des décennies, à la fois comme objet général d'analyse de données ou dans des contextes plus appliqués (bioinformatique Hu et Yoo (2004), segmentation d'images Hong et al. (2008), etc ...). Toutefois, la nature même du problème ne fournit généralement pas de vérité-terrain simple, tant sur la composition des classes que leur nombre. Les algorithmes proposés dans la littérature optimisent des critères très divers et font des hypothèses elle aussi diverses sur les propriétés de cohérence intra-classe.

Depuis une dizaine d'années, un axe de recherche, le *clustering d'ensemble*, s'est développé, élaborant des critères et des méthodes pour aggréger différentes partitions d'un même jeu de données, construites par des critères antagonistes.

Plusieurs méthodes sont alors envisageables pour pallier l'incertitude quant à la plausibilité du résultat :