Classification en présence d'outliers (données aberrantes) avec RMixmod (package de classification par modèles de mélanges)

Florent Langrognet*[1]

[1]Laboratoire de Mathématiques de Besançon (LM-Besançon) – CNRS : UMR6623, Université de Franche-Comté – UFR Sciences et techniques 16 route de Gray 25 030 Besançon cedex, France

Résumé

Les modèles de mélanges offrent un cadre probabiliste flexible et efficace pour traiter des problématiques de classification supervisée ou non supervisée. L'objectif du projet MIXMOD est de diffuser un ensemble logiciel de classification des données par modèles de mélanges à un large spectre d'utilisateurs via plusieurs composants logiciels. La bibliothèque de calcul mixmodLib (C++) en est la pierre angulaire, résultat d'un travail de 15 ans sur la robustesse et la rapidité de calcul. Le package RMixmod, ensemble de fonctions pour R, interfacé avec mixmodLib (grâce à RCPP) est devenu un outil de référence pour la classification des données. Intégrant de nombreuses fonctionnalités (algorithmes de type EM, critères de sélection, modèles parcimonieux, stratégies d'initialisation, ...), cet ensemble logiciel permet de traiter des données quantitatives, qualitatives et mixtes, y compris dans des situations complexes.
Lorsque le jeu de données contient des individus parasites (c'est-à-dire ayant des valeurs aberrantes, encore appelés outliers) la classification devient alors particulièrement difficile (trouver le bon nombre de classes, affecter le bon label aux vrais individus, ...).
En présence d'outliers, il peut être tentant d'appliquer un pré-traitement pour nettoyer le jeu de données avant de le soumettre à un logiciel de classification. Mais ces méthodes sont généralement peu efficaces.
A l'opposé, on peut considérer que la classification doit s'effectuer sur l'ensemble des individus avec une classe supplémentaire (celle des outliers).
L'étude consiste à mettre à l'épreuve RMixmod sur ce type de problématique. Nous nous intéressons plus particulièrement au cas où des outliers, répartis selon une loi uniforme, viennent s'ajouter à des individus issus de 2 lois gaussiennes.
La flexibilité des modèes de mélanges (ici gaussiens) permet non seulement de retrouver les classes d'origine mais également de faire apparaître une classe contenant les outliers.
RMixmod peut donc être utilisé efficacement sur des jeux de données contenant des outliers.

*Intervenant