Etude de la robustesse de RMixmod (package de classification par modèles de mélanges) en cas de chevauchement de classes

F. Langrognet[a]

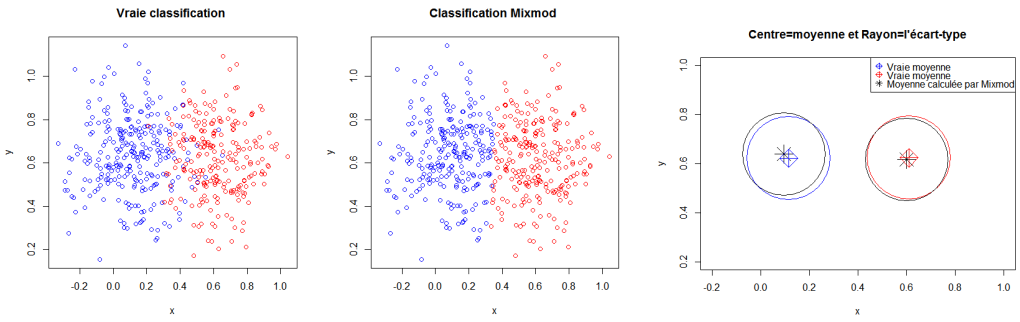
[a]Laboratoire de Mathématiques de Besançon
UMR 6623 - Université de Franche-Comté - CNRS
16, route de Gray - 25030 Besançon
florent.langrognet@univ-fcomte.fr

Mots clefs : classification, modèles de mélanges, critères de sélection, chevauchement

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 près 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 situation complexes.
L'une des difficultés en classification des données réside dans la capacité à donner des bons résultats en cas de chevauchement entre plusieurs classes : trouver le bon nombre de classes, les bons paramètres et les bonnes affectations des individus à ces classes (labels).
L'étude consiste à tester RMixmod sur des jeux de données simulées en contrôlant le degré de chevauchement (ω) entre les classes (grâce au package MixSim) sur des données quantitatives.

Capacité de RMixmod à retrouver les paramètres et la classification

Dans un 1^{er} temps, on supposera connus le nombre de classes (2). Il s'agit alors d'analyser la capacité de RMixmod à retrouver les labels et les paramètres des classes en augmentant le degré de chevauchement.
A titre d'exemple, à partir du jeu de données (Figure 1 (a)) présentant un chevauchement modéré ($\omega = 0.15$), on constate que RMixmod retrouve la classification initiale (Figure 1 (b)) et les paramètres initiaux (Figure 1 (c)) avec une grande précision.



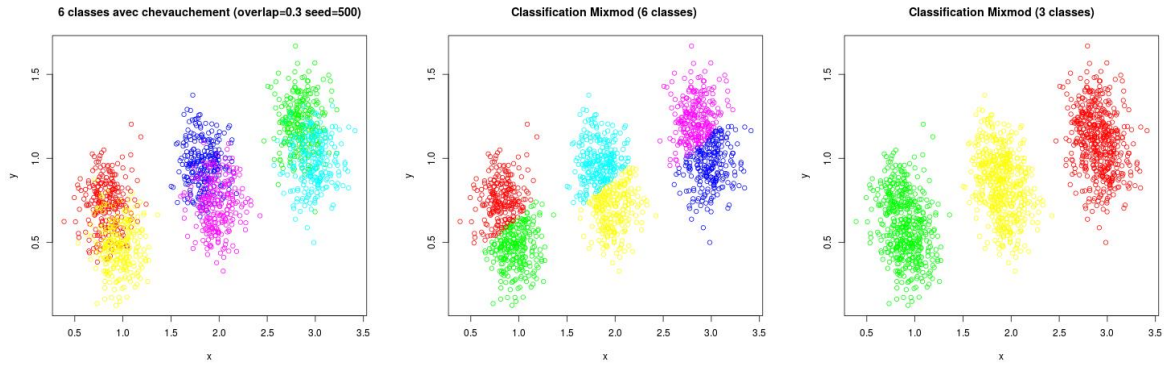
(a) (b) (c)

Figure 1: 2 classes avec chevauchement modéré

Capacité de RMixmod à retrouver le bon nombre de classes (ainsi que les paramètres et la classification)

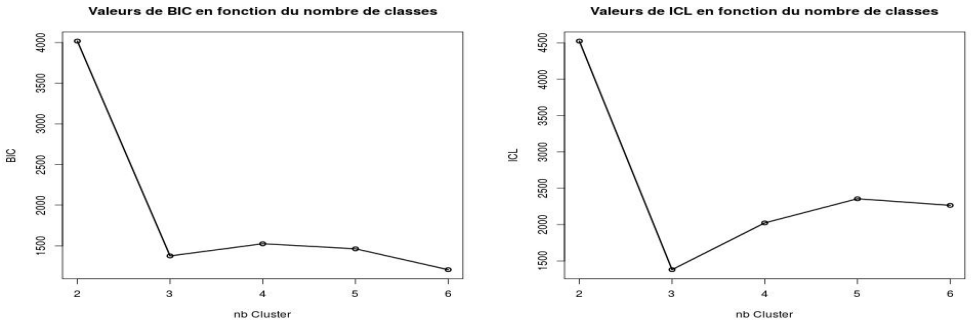
On se place ici dans une situation avec 6 classes présentant des chevauchements de classes 2 à 2. L'objectif est de répondre à la question cruciale du choix du nombre de classes. Les critères de sélection disponibles dans RMixmod permettent d'y répondre efficacement en tenant compte des objectifs de l'utilisateur.

Ainsi, dans le cas d'un chevauchement (par couple de classes) défini par $\omega = 0.3$ (Figure 2 (a)), le critère BIC (Bayesian Information Criterion) permet de retrouver la classification en 6 composants (Figure 2 (b)) alors que le critère ICL (Integrated Completed Likelihood), privilégiant des classes bien séparées permet de choisir une classification en 3 composants (Figure 2 (c)).



(a) (b) (c)

Figure 2: 6 classes avec 3 chevauchements



(a) (b)

Figure 3: Valeurs des critères BIC (a) et ICL (b) en fonction du nombre de classes

Références

[1] Lebret R., Iovleff S., Langrognet F., C. Biernacki, G. Celeux, G. Govaert (2013). Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. A paraître dans *Journal of Statistical Software*.

[2] Biernacki C., Celeux G., Govaert G., Langrognet F., (2006). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis*, vol. 51/2, pp. 587-600

[3] Ranjan Maitra, Wei-Chen Chen, Volodymyr Melnykov (2012). MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms; R. *Journal of Statistical Software*, Vol. 51, Issue 12, Nov 2012