Okm : une librairie R pour la classification recouvrante

Guillaume Cleuziou, Léo Rousseau

OKM : une librairie R pour la classification recouvrante

G. Cleuziou et **L. Rousseau**

Laboratoire d'Informatique Fondamentale d'Orléans
Rue Léonard de Vinci - B.P. 6759
F-45067 ORLEANS Cedex 2
guillaume.cleuziou@univ-orleans.fr rousseauleo1@gmail.com

Mots clefs : Classification automatique, classification recouvrante, réallocation dynamique.

1 Introduction

La classification automatique ou clustering consiste à organiser un ensemble d'individus $X = \{x_1, ..., x_n\}$ en classes (ou clusters) de telle sorte que des individus qui se ressemblent soient regroupés au sein d'un même cluster et des individus dissemblables appartiennent à des clusters différents. De nombreuses stratégies de classification ont été envisagées ces soixante dernières années, chacune présentant son lot d'avantages ou d'inconvénients selon la nature ou la quantité de données à traiter, leur dimensionalité, ou la forme des résultats (dendrogrammes, partitions strictes ou floues, concepts, etc.). La présente contribution s'intéresse aux méthodes de partitionnement dites par réallocation dynamique, dont l'algorithme des k-moyennes en est le plus éminent exemple. Les travaux que nous avons menés ont consisté à généraliser le modèle sous-jacent à l'approche k-moyennes afin de générer des classes recouvrantes, c'est-à-dire autorisant chaque individu à appartenir à plusieurs classes. Le modèle OKM (*Overlapping k-means*) [1] est brièvement exposé, s'en suit la présentation d'une première librairie R associée à ce modèle.

2 L'approche OKM

La méthode des k-moyennes est guidée par un critère objectif (moindres carrés) que l'on peut interpréter comme une quantification de l'erreur commise en résumant une classe d'individus à un unique représentant. La minimisation de cette erreur passe ainsi par la recherche des k meilleurs représentants de classes. Nous avons étendu ce principe au cas où chaque individu peut appartenir à plusieurs classes et donc être représenté par plusieurs représentants de classes. Le critère objectif (moindre carré généralisé) sous-jacent à la méthode OKM est donné par

$$J(\Pi, C) = \sum_{i=1}^{n} \|x_i - \phi_{\Pi,C}(x_i)\|^2 \quad avec \quad \phi_{\Pi,C}(x_i) = \frac{\sum_{j=1}^{k} \mathbb{1}_{\{x_i \in \pi_j\}}.c_j}{\sum_{j=1}^{k} \mathbb{1}_{\{x_i \in \pi_j\}}}$$

Dans cette formalisation, Π représente l'ensemble des clusters $\{\pi_j\}_{j=1}^k$, C l'ensemble des représentants $\{c_j\}_{j=1}^k$ et $\phi_{\Pi,C}(x_i)$ une combinaison des représentants des classes de x_i ; la combinaison utilisée dans OKM correspond au centre de gravité de ces représentants. Finalement, l'erreur associée à une classification recouvrante Π est quantifiée par la somme des distances (euclidiennes) entre chaque individu et la combinaison de ses représentants dans la classification.
La minimisation du critère $J()$ est assurée par une approche itérative classique en deux étapes : (1) affectation (ici multiple) de chaque individu aux classes puis (2) mise à jour des représentants des clusters ; chacune des deux étapes assure la décroissance du critère objectif.

3 La librairie R : OKM

Nous avons développé une première librairie[1] R intégrant non seulement l'algorithme OKM mais également une version pondérée (WOKM : *Weighted*-OKM) [2] permettant d'aboutir à des clusters ellipsoïdaux et limitant l'importance des recouvrements entre classes.
Deux fonctions sont proposées (`okm()` et `wokm()`) et correspondent à des approches généralisantes de k-moyennes ; nous avons donc veillé à conserver la forme de la fonction `kmeans()` présente dans la librairie R *stats* installée par défaut :

- `okm(X, centers, iter.max = 10, nstart = 1, visu = FALSE)`
- `wokm(X, centers, iter.max = 10, nstart = 1, B = 2, visu = FALSE)`

Les arguments utilisés sont : `X` : un ensemble d'individus décrits dans une matrice ou un dataframe (e.g. de taille $n \times p$) ; `centers` : un nombre de clusters (entier) ou un ensemble de centres initiaux (matrice ou dataframe) ; `iter.max` : le nombre maximum d'itérations autorisées dans l'algorithme (fixé à 10 par défaut) ; `nstart` : le nombre d'exécutions souhaitées (initialisations différentes) dans le cas où `centers` est un nombre (une seule exécution par défaut) ; `visu` : valeur logique permettant de faire apparaître les détails sur la convergence du critère objectif au cours de l'exécution (FALSE par défaut) ; `B` : paramètre (> 1) permettant de contrôler l'importance de la pondération dans l'approche WOKM.

Chacune des deux fonctions retournera un objet constitué de 4 composantes : `Clusters` : une matrice binaire d'appartenances ($n \times k$) ; `Representatives` : une matrice ($k \times p$) où chaque ligne décrits un représentant de classe ; `Withiness` : la valeur du critère objectif à la dernière itération $J()$; `Overlaps` : le nombre moyen de classes d'appartenance sur l'ensemble des individus.

4 Conclusion

Les premiers développements liés à l'approche OKM étant très récents, de nombreuses variantes sont actuellement à l'étude (nouveaux modèles de pondération, adaptation à la norme L_1, paramétrage des recouvrements, variantes à noyaux, etc.) chacune d'elle permettant de répondre à des besoins applicatifs réels. Une seconde version de la librairie *okm*, plus complète, est actuellement en préparation et sera prochainement déposée sur le site du CRAN.

Références

[1] G. Cleuziou. An extended version of the k-means method for overlapping clustering. In 19th International Conferenceon Pattern Recognition (ICPR'2008), pages 1 4, 2008.

[2] G. Cleuziou. Two variants of the OKM for Overlapping Clustering. Springer, 2010.

[1] Disponible sur : www.univ-orleans.fr/lifo/Members/cleuziou/