Clustering sur données incomplètes avec clusterMI

Vincent Audigier *

26 mars 2024

Résumé

Nous nous intéressons à la classification non-supervisée d'observations incomplètes. Pour cela une nouvelle méthodologie basée sur l'imputation multiple est proposée. Celle-ci consiste en trois grandes étapes : l'imputation selon des modèles dédiés, la classification sur les données imputées, avec l'estimation de l'instabilité associée, et l'agrégation des résultats. Nous revenons sur ces différentes étapes et présentons comment les mettre en œuvre via le package clusterMI disponible sur le CRAN.

Mots-clefs : Clustering – Données manquantes – Imputation multiple

L'imputation multiple fait partie des stratégies classiques pour gérer le problème des données manquantes [Little and Rubin, 2002]. Telle que proposée historiquement, celle-ci consiste à imputer M fois le jeu de données selon un modèle, dit *modèle d'imputation*, à ajuster un modèle dit *d'analyse* sur chacun des tableaux imputés, puis à agréger les résultats selon les règles de Rubin. Ces règles consistent en l'agrégation à la fois en termes d'estimation ponctuelle des paramètres du modèle d'analyse, mais aussi en termes d'estimation des variances des estimateurs associés.

Toutefois, dans le contexte du clustering, une telle méthodologie ne peut pas être appliquée directement. En effet, l'imputation des données nécessite la prise en compte de la structure de groupes sur les individus. Les modèles d'imputation classiquement utilisés, tel que le modèle gaussien multivarié [Schafer, 1997] ne sont alors plus adaptés [Audigier et al., 2021]. Par ailleurs, les règles de Rubin ont été établies pour l'agrégation de coefficients numériques et ne permettent pas directement l'agrégation des partitions issues d'un clustering.

Cette présentation a pour objet de présenter différentes méthodes d'imputation pertinentes pour le clustering, ainsi qu'une façon d'agréger les résultats obtenus suite à l'analyse de chacun des tableaux, tant en termes de partition que d'instabilité associée. Concernant les modèles d'imputation, nous présenterons des approches *par modèle joint*, où une distribution explicite à l'ensemble des variables est effectuée, et des approches *séquentielles*, où seule la distribution conditionnelle de chaque variable est spécifiée. Ces dernières sont connues pour permettre un meilleur ajustement des données. Concernant l'agrégation des résultats, nous présenterons dans un premier temps comment évaluer une instabilité en clustering par bootstrap, comme proposé dans Fang and Wang [2012]. À partir de là, nous proposerons une façon d'agréger les différentes partitions obtenues, par factorisation matricielle non-négative, ainsi que les différentes mesures d'instabilité selon les règles détaillées dans Audigier and Niang [2022].

En résumé, la procédure proposée se présente ainsi :

Imputation Etant donné un jeu de données incomplet, générer M tableaux imputés selon une méthode d'imputation multiple pré-définie (par exemple Kim et al. [2014])

Analyse Pour m dans $\{1 \dots M\}$,

1. construire une partition Ψ_m à partir du $m^{\text{ème}}$ jeu imputé selon la méthode de clustering choisie
2. estimer V_m^{boot} l'instabilité associée par bootstrap

*CEDRIC-MSMDA, CNAM, Paris, vincent.audigier@cnam.fr

Agrégation

1. L'ensemble de partitions $(\Psi_m)_{1\leq m\leq M}$ est agrégé en utilisant une factorisation matricielle non-négative [Li et al., 2007] consistant à rechercher la partition $\bar{\Psi}$ en K groupes telle que

$$\bar{\Psi} = argmin_{\Psi} \sum_{m=1}^{M} \delta(\Psi, \Psi_m)$$

où $\delta(\Psi, \Psi_m)$ indique le nombre de désaccords entre les partitions Ψ and Ψ_m [1].

2. Les mesures d'instabilité $\left(V_m^{boot}\right)_{1\leq m\leq M}$ sont agrégées selon :

$$T = \frac{1}{M}\sum_{m=1}^{M} V_m^{boot} + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} \delta\left(\Psi_m, \Psi_{m'}\right)/n^2$$

Cette méthodologie permet de gérer le problème de données manquantes en classification non-supervisée, que ce soit pour des approches probabilistes ou géométriques. Elle présente l'avantage de ne pas être sensible au problème de *label switching* et permet également d'avoir des nombre de groupes différents lors de la phase d'analyse. Enfin, elle offre la possibilité d'estimer le nombre de groupes dans le cadre d'un jeu de données incomplet, ceci en comparant les mesures d'instabilité des partitions agrégées pour différentes valeurs de K.

Le package associé **clusterMI** offre quatre méthodes d'imputation différentes, dont deux par modèle joint, et deux par approche séquentielle. Différentes méthodes de clustering sont pré-implémentées (k-means, pam, clara, agnes, mélange gaussien, fuzzy c-means) mais il est possible d'envisager d'autres méthodes. Le package offre aussi plusieurs outils de diagnostic afin d'apprécier l'ajustement du modèle d'imputation, la convergence et le choix des modèles d'imputation des approches séquentielle, le nombre de tableaux imputés M, ou encore le nombre de groupes K.

Références

V. Audigier and N. Niang. Clustering with missing data : which equivalent for Rubin's rules ? *Advances in Data Analysis and Classification*, September 2022. doi : 10.1007/s11634-022-00519-1. URL `https://hal.science/hal-03766733`.

V. Audigier, N. Niang, and M. Resche-Rigon. Clustering with missing data : which imputation model for which cluster analysis method ?, 2021.

Y. Fang and J. Wang. Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.*, 56(3) :468–477, 2012. ISSN 0167-9473. doi : 10.1016/j.csda.2011.09.003. URL `https://doi.org/10.1016/j.csda.2011.09.003`.

H. J. Kim, J. P. Reiter, Q. Wang, L. H. Cox, and A.F. Karr. Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3) :375–386, 2014. doi : 10.1080/07350015.2014.885435. URL `https://doi.org/10.1080/07350015.2014.885435`.

T. Li, C. Ding, and M. I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, page 577–582, USA, 2007. IEEE Computer Society. ISBN 0769530184. doi : 10.1109/ICDM.2007.98.

R. Little and D. Rubin. *Statistical Analysis with Missing Data.* Wiley series in probability and statistics, New-York, 2002.

J. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC, London, 1997.

1. $\delta(\Psi, \Psi') = \sum_{(i,i')} \delta_{ii'}$ avec $\delta_{ii'} = 1$ si les individus i et i' appartiennent au même cluster dans une partition et non dans l'autre et $\delta_{ii'} = 0$ sinon