MULTIPLE IMPUTATION FOR CLUSTERING ON INCOMPLETE DATA

Vincent Audigier [1], Ndèye Niang[1]

[1] CEDRIC Lab, MSDMA Teamn CNAM, (e-mail: vincent.audigier@cnam.fr, n-deye.niang_keita@cnam.fr)

ABSTRACT: We present how MI can be considered for addressing missing values in the context of clustering. For achieving this goal, we present a novel imputation method entitled FCS-homo, as well as a pooling method for the set of partitions obtained from each imputed data set. The proposed methodology is evaluated using a simulation study in comparison with state of the arts methods. We start by treating the case where the observations are generated from a gaussian mixture model with missing at random values. The study is completed by experiments based on various real data sets where the distribution of the data is unknown. These first results tend to show that multiple imputation is a efficient method for handling missing data in clustering, especially when the data distribution is unknown.

KEYWORDS: clustering, missing data, multiple imputation

1 Introduction

Among methods for addressing missing values, direct methods (DM) and multiple imputation (MI) are probably the most commonly considered. DM can be described as methods consisting in adapting the analysis methodology to be applied on incomplete data. This can be achieved by optimising a criterion based on incomplete data rather than complete data. DM are theoretically appealing, but they require a dedicated methodology for each analysis method. On the contrary, MI consists in separating the missing data issue to the analysis by proceeding in three steps. The first step is the imputation step, which consists in replacing each missing values by several plausible values. At the end, several imputed datasets are available. The second step consists in analysing each imputed dataset according to the analysis method wished. Finally, the third step consists in pooling the several analysis results to obtain a unique one. By separating the imputation step and the analysis step, MI allows applying any statistical analysis when missing values are imputed and consequently is less analysis method dependent that DM. However, it can also introduce bias if the imputation method is not well chosen in regard to the analysis model.

Several DM have been proposed to perform clustering with missing values. For instance, Marbac *et al.* (2019) proposed an EM algorithm to estimate parameters from a gaussian mixture model, Chi *et al.* (2016) proposed to extend kmeans criterion for accounting for missing values, while Hathaway & Bezdek (2001) extended fuzzy c-means algorithm by an *optimal completion strategy.* However, addressing missing values by MI remains challenging in clustering for at least two reasons. Firstly, because the imputation step requires specific models. Indeed, available imputation methods are generally based on the assumption that observations are drawn from a unique distribution, which is obviously inconsistent with the underlying assumptions made in cluster analysis. The second reason is that the way to pool partitions obtained at the second MI step is unclear. Indeed, the pooling rules in MI are theoretically applied on the parameters from a generalized linear model and not on a categorical variable as a partition of observations. Thus, addressing missing values in clustering by MI is not straightforward.

In this work, we propose a novel methodology for addressing missing values in clustering by MI. It consists in a novel imputation method entitled FCS-homo as well as a novel pooling rule.

2 Method

2.1 FCS-homo

Fully conditional specification (FCS) (van Buuren *et al.*, 2006) consists in imputing missing data by assuming a distribution for each variable conditionally to the others and then impute each variable sequentially according to each ones. FCS methods are often used in practice since they allow a better fit of the imputation model. More precisely, let $P(X_j|X_{-j};\zeta_j)$ be the distribution of X_j $(1 \leq j \leq p)$ conditionally to other variables, denoted X_{-j}, and parameterized by ζ_j. For instance, $P(X_j|X_{-j};\zeta_j) = \mathcal{N}(X_{-j}\beta,\sigma^2)$ with $\zeta_j = (\beta,\sigma)$. Then, FCS methods impute the mth data set as follows:

- initialize missing values of $\mathbf{X}$ by random draws from observed values
- for j in 1 ... p
  - a generate ζ_j based on observed individuals on X_j
  - b impute $X_j$according to $P(X_j|X_{-j};\zeta_j)$
- repeat until convergence

In a context of cluster analysis, we propose a FCS method which accounts for the cluster data structure. To achieve this goal, each regression model is

conditional to a supplementary variable W indicating the cluster of each observation. Let $Z = (W,X)$ be the incomplete data set gathering the cluster variable W, which is unknown and considered as fully missing, and X the incomplete data set. Then, the algorithm involves two main steps: imputation of Z given W and vice versa. Generating Z given W is performed using regression models including an intercept specific to each cluster $P(Z_j|Z_{-j},W;\zeta_j) = \mathcal{N}(Z_{-j}\beta+\mu_w,\sigma^2) \quad \zeta_j = (\beta,\sigma,\mu_w)$ while generating W given Z is performed using linear discriminant analysis (see Audigier *et al.* (2021) for more details).

2.2 Pooling

Given M imputed data set, we denote Ψ_m the partition obtained from the data set m. This partition can be obtained from any clustering algorithm (e.g. k-means). The set $(\Psi_m)_{1\leq m\leq M}$ is pooling using Non Negative matrix Factorization which consists in looking at the partition $\bar{\Psi}$ such as

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

with $\delta(\Psi,\Psi_m)$ the number of disagreements * between Ψ and Ψ_m. An associated instability can also be computed as proposed in Audigier & Niang (2022).

3 Results

The proposed methodology is evaluated by comparison with DM approaches under MAR mechanisms. For this purpose, we focus on three clustering techniques: the Gaussian mixture model, the k-means and the fuzzy c-means. The study is first carried out on data simulated according to a Gaussian mixture model in which we vary the separability of the clusters, their number, their size and their correlation structure. Missing data are generated according to different mechanisms varying by their nature (MCAR or MAR) and by the rate of missing values. In a second step, both approaches are compared on different real data sets where the distribution is not known but where a cluster structure is well identified. In both cases, the three clustering techniques are applied using the theoretical number of clusters and the missing data are handled either directly or by multiple imputation. The resulting partitions are

*$\delta(\Psi,\Psi') = \sum_{(i,i')} \delta_{ii'}$ with $\delta_{ii'} = 1$ if individuals i and i' are in the same cluster for a given partition and not for the second, while $\delta_{ii'} = 0$ otherwise

then compared to the expected partition according to the adjusted Rand index (ARI).

4 Discussion

The study illustrates that the use of multiple imputation for handling missing values in clustering generally improves the partition quality for geometric clustering methods, namely k-means and fuzzy c-means, compared to direct k-pod and optimal completion strategy approaches (respectively). As for the results on the parametric Gaussian model approach, similar performances are observed when the data are derived from a Gaussian mixture. Nevertheless, significant differences are observed on real data where the direct methods often lead to lower ARI.

Thus, these first results tend to show that MI is a efficient method for handling missing data in clustering, especially when the data distribution is unknown. Moreover, this technique allows to apply any clustering method on incomplete data, whereas direct methods remain specific to the clustering technique considered.

References

AUDIGIER, V., & NIANG, N. 2022. Clustering with missing data: which equivalent for Rubin's rules? *Advances in Data Analysis and Classification*, Sept.

AUDIGIER, V., NIANG, N., & RESCHE-RIGON, M. 2021. *Clustering with missing data: which imputation model for which cluster analysis method?*

CHI, J. T., CHI, E. C., & BARANIUK, R. G. 2016. k-POD: A Method for k-Means Clustering of Missing Data. *The American Statistician*, **70**(1), 91–99.

HATHAWAY, R. J., & BEZDEK, J. C. 2001. Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **31**(5), 735–744.

MARBAC, M., SEDKI, M., & PATIN, T. 2019. Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification*, 1–19.

VAN BUUREN, S., BRAND, J., GROOTHUIS-OUDSHOORN, C., & RUBIN, D. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064.