Study of missing data imputation techniques for the purpose of classification

Auteur : Baudenne, Céline
Promoteur(s) : Haesbroeck, Gentiane
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/11228

Study of missing data imputation techniques for the purpose of classification.

Baudenne Céline
Supervised by Prof. Gentiane Haesbroeck
Data Science
Academic year 2020-2021

Data quality is a central concern in data science and missing values are an important part of it. Missing data can have a detrimental effect in particular in the context of classification. This is the reason why missing values should be handled with the main objective of improving the classification results. This study thus evaluates the performances of missing data treatments by comparing the classification results obtained after applying classifiers on data sets imputed by means of various imputation methods. In addition, we decided to take a further step in addressing data quality issues by considering classification data containing atypical values in addition to missing values. The objective is thus to investigate whether a combination of imputation and classification methods could mitigate the potential adverse effect of contamination and missingness in order to achieve satisfactory classification results. For this purpose, we performed simulations within a well-defined theoretical framework in order to have full control over the data characteristics. Data sets have been built according to different covariance settings and missing values have been introduced following two missingness mechanisms. Both training and testing sets were affected by missing values. Several imputation methods and classification methods have been investigated, including, in both cases, classic and robust techniques. The prior analysis of the different possible situations enabled us to better understand how the different procedures operate and better make sense of the results. Our observations enabled us to draw the main conclusion that, in such context, robust methods should be favoured both for imputation and classification methods. Moreover, the characteristics of the data as well as its context should always be studied beforehand in order to choose the most appropriate methods.

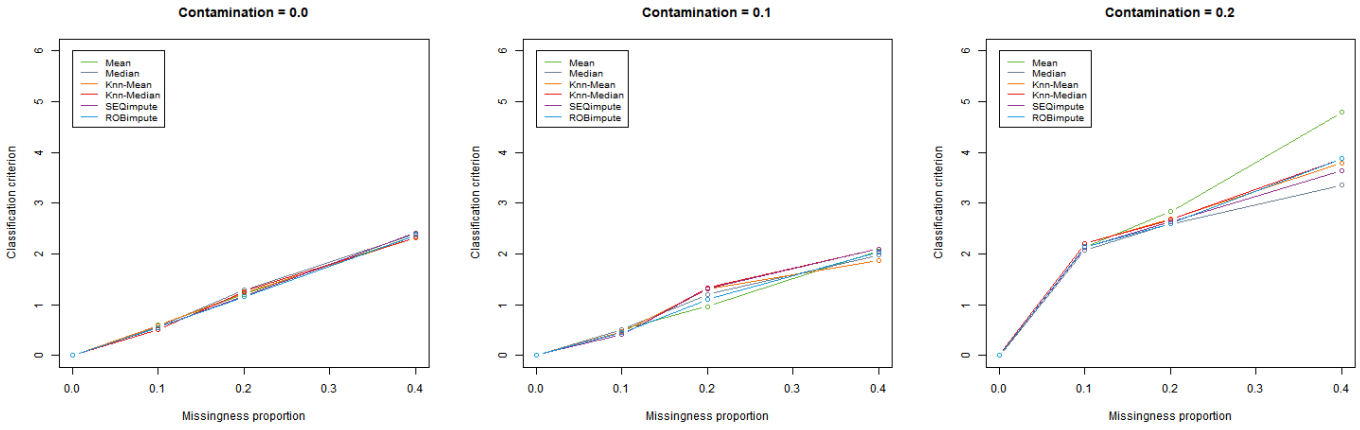

Figure 1: Evaluation of the classification performances, as a function of the missingness proportion, of the Robust Linear Discrimination applied on imputed data sets generated in the independent setting, under MCAR, for different proportions of contamination.