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

Etude des techniques d'imputation de données manquantes dans un contexte de classification.

Baudenne Céline
Supervisé par Prof. Gentiane Haesbroeck
Data Science
Année académique 2020-2021

La qualité des données est une préoccupation centrale de la science des données et les valeurs manquantes en sont une part importante. Les données manquantes peuvent avoir un effet néfaste, en particulier dans le contexte de la classification. Cette étude évalue donc les performances des techniques de traitement de données manquantes en comparant les résultats de classification obtenus après application de techniques de classification sur des ensembles de données imputées au moyen de diverses méthodes d'imputation. Par ailleurs, nous avons également décidé de considérer des données de classification contenant des valeurs atypiques en plus des valeurs manquantes. L'objectif est donc d'étudier si une combinaison de méthodes d'imputation et de classification pourrait atténuer l'effet négatif de la contamination et des valeurs manquantes afin d'obtenir des résultats de classification satisfaisants. À cette fin, nous avons effectué des simulations dans un cadre théorique bien défini afin de maîtriser pleinement les caractéristiques des données. Les ensembles de données ont été construits selon différentes structures de covariance et les valeurs manquantes ont été introduites selon deux mécanismes différents. Des valeurs manquantes ont été introduites dans les données d'entrainement et dans les données de test. Plusieurs méthodes d'imputation et de classification ont été étudiées, y compris, dans les deux cas, des techniques classiques et robustes. L'analyse préalable des différentes situations possibles nous a permis de mieux comprendre le fonctionnement des différentes procédures et de donner un meilleur sens aux résultats. Nos observations nous ont permis de tirer la principale conclusion que, dans un tel contexte, les méthodes robustes devraient être privilégiées tant pour l'imputation que pour la classification. En outre, les caractéristiques des données ainsi que leur contexte devraient toujours être étudiés au préalable afin de choisir les méthodes les plus appropriées.

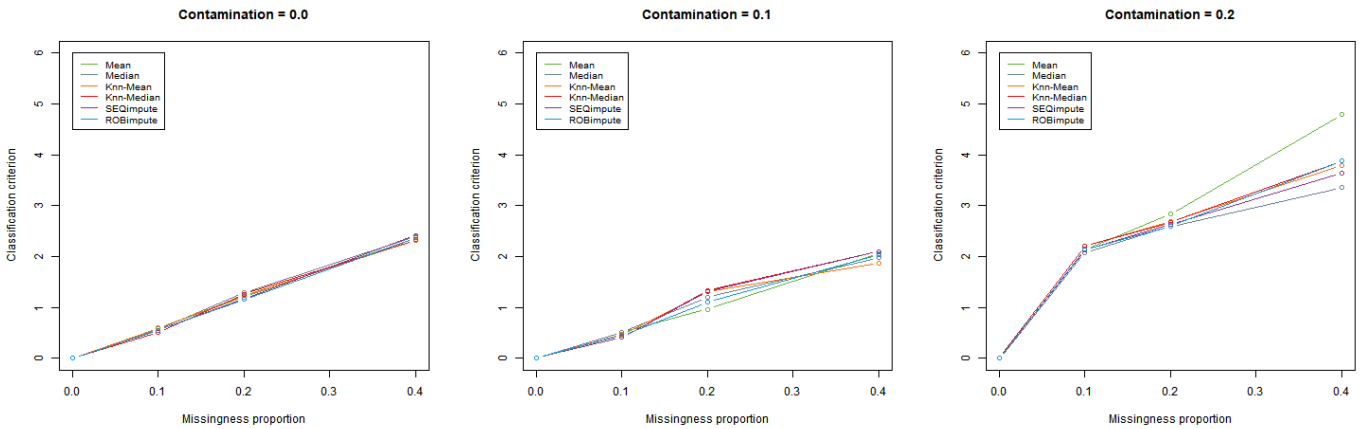

Figure 1: Evaluation des performances de classification pour la méthode de discrimination linéaire robuste sur des données imputées générées dans le cas indépendent, sous MCAR.