

Comparison of performance of data imputation methods in the context of their impact on the prediction efficiency of ranger Random Forest algorithm implementation

Ada Gąssowska, Mateusz Grzyb, Elżbieta Jowik

Abstract

It is very common for datasets to contain incomplete observations. The analysis conducted by ignoring cases with missing data and considering only complete observations is perceived as naive, inefficient and exposed to bias, as missing values may also convey valuable information. To incorporate these sketchy observations, various methods of handling missing data may be considered, including both less and more sophisticated approaches. It turns out that the performance of machine learning classification models moderately depends on the type of approach that is applied to the imputation problem while the time of imputation, when using different techniques, is highly diverse. Presented article focuses on one basic technique (median/mode imputation) and five more sophisticated ones, which origin respectively from: mice,¹ VIM², missRanger³ and softImpute⁴ R packages.

The differences in the results of the classification algorithm for the imputation techniques in respect to included measures are slight, nevertheless, the best outcome was obtained for the hotdeck method from the VIM package. As for the time of imputation, it highly depends on the technique complexity, so the basic method turns out to be the fastest one.

¹Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations in R”. in: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. URL: <https://www.jstatsoft.org/article/view/v045i03>.

²Alexander Kowarik and Matthias Templ. “Imputation with the R Package VIM”. in: *Journal of Statistical Software* 74.7 (2016), pp. 1–16. URL: <https://www.jstatsoft.org/article/view/v045i03>.

³Michael Mayer. *missRanger: Fast Imputation of Missing Values*. R package version 2.1.0. 2019. URL: <https://CRAN.R-project.org/package=missRanger>.

⁴Trevor Hastie and Rahul Mazumder. *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD*. R package version 1.4. 2015. URL: <https://CRAN.R-project.org/package=softImpute>.

1 Introduction

Dealing with missing data is a substantial part of feature engineering as most of the machine learning algorithms do not accept incomplete datasets. However, despite the acknowledged importance of data imputation, in many practical cases it is not handled with proper caution. Basic median/mode imputation is often used, as well as deleting rows with missing data (if the number of missing values is low).

This paper introduces imputation techniques in the context of their impact on the time of filling missing data and the prediction efficiency of ranger⁵ Random Forests algorithm implementation, which is fast and particularly suited for high dimensional data.

2 Methodology

The ranger Random Forests algorithm implementation is particularly suited for high dimensional data. As it does not handle data with missing values, datasets must be imputed before classification.

We used six different imputation functions on each benchmarking dataset:

- **basic method** using `impute()` function from `imputeMissings` package. It imputes missing numeric values with median and categorical variables with mode of other records.
- **mice()** function from `mice` package. The function uses Predictive Mean

⁵Marvin N. Wright and Andreas Ziegler. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R”. in: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: [10.18637/jss.v077.i01](https://doi.org/10.18637/jss.v077.i01).

Matching to impute missing values.

- **missRanger()** function from missRanger package.

The technique uses the ranger package to do fast missing value imputation by chained random forest.

- **hotdeck()** function from VIM package. Each missing value is replaced with an observed response from a “similar” unit.

- **kNN()** function from VIM package. It finds the k closest neighbors to the observation with missing data and then imputes them based on the the non-missing values from its neighbors.

- **softImpute() function combined with mode imputation.**

The first method origins from softImpute package and is applied to impute numeric variables. Mode imputation is used for categorical ones imputation.

Function using imputeFAMD() from missMDA package was also created but it was unable to perform imputation on all of benchmarking datasets due to issues with convergence.

Imputation with use of Amelia package also proved impossible due to occurrence of highly correlated variables in some benchmarking datasets.

To automatize our work we created specialized functions: **split_and_impute()** and **train_and_test()**.

- **Split_and_impute()** divides data into train and test sets and imputes it with specified imputer.

It takes four arguments:

- id - id of benchmarking dataset to impute,
- imputer - one of the described earlier imputation functions,
- train_size - size of train set (values from 0 to 1, default value is 0.8),
- save - whether to save imputed dataset (default value is TRUE)

and returns 3 values:

- train - imputed train dataset,
- test - imputed test dataset,
- time - time of imputation

- **Test_and_train()** performs crossvalidation on train set and makes the predictions for test set made with specified learner, trained on whole train set. Based on mentioned results it calculates AUC, BACC and MCC measures for both crossvalidation and test set testing stages.

The function takes seven arguments:

- train - imputed train dataset without missing values,
- test - imputed test dataset without missing values,
- learner - a classification learner which returns probabilities. In this case we will be using Ranger Random Forest learner,
- target - name of target variable in the dataset,
- postive - label of positive class (default value is '1'),
- folds - number of folds used for crossvalidation (default value is 5),
- title - title for crossvalidation metrics plot (default value is no title)

and returns various values needed for the comparison of methods of imputation:

- cv_plot - plot of Receiver Operating Characteristics curve (ROC), AUC, BACC and MCC achieved during crossvalidation stage,
- mean_auc - mean AUC achieved during crossvalidation stage,
- mean_bacc - mean BACC achieved during crossvalidation stage,
- mean_mcc - mean MCC achieved during crossvalidation stage,
- test_auc - AUC achieved during test stage,
- test_bacc - BACC achieved during test stage,

- test_mmc - MCC achieved during test stage,
- learner - learner trained on whole train set.

After calling `split_and_impute()` and `test_and_train()` functions on each dataset with six different imputers we were able to compare performance of each tested imputation function in the context of their impact on the prediction efficiency of Ranger Random Forest algorithm.

The prediction effectiveness was assessed in relation to the area under the ROC curve (AUC⁶), balanced accuracy (BACC⁷) and Matthews correlation coefficient (MCC⁸) measures. Time of each imputation was also measured, and taken into consideration.

3 Results

Each imputation function was performed on twelve benchmarking datasets with missing data, divided into test and train sets. We gathered the predictions for test sets from ranger Random Forest model fitted on train sets, and compared them using AUC, BACC and MCC (see Table1, Figure1, Figure2 and Figure3). We also measured the time of each imputation and compared it (Figure 4).

It should be mentioned that any differences in results are only due to a change in the imputation technique. Each of them performed on the same sets with identical divisions into train and test sets.

4 Conclusions

All of the imputers performances were relatively similar. The best mean AUC, mean

BACC and mean MCC were all scored by VIM_hotdeck imputer. It also proved to be fast, with imputation times of a few seconds. Mice and VIM_knn imputer were only slightly worse, however Mice imputation took a lot of time for large datasets (up to 9 hours). SoftImputer acquired the worst score regarding AUC measure and it the second worst regarding BACC and MCC. MissRanger's scores were slightly better, but not as good as any of the imputation from VIM package. Basic median/mode imputer was the quickest one, but its mean scores were the lowest (except of AUC where it scored second lowest one).

method	mean AUC	mean BACC	mean MCC
basic	0.897	0.821	0.657
mice	0.907	0.828	0.672
missRanger	0.898	0.827	0.669
softImpute	0.893	0.822	0.664
VIM_hotdeck	0.908	0.837	0.690
VIM_knn	0.904	0.832	0.678

Table 1: Mean AUC, BACC and MCC for each method

It seems that according to our reaserch the best imputer is hotdeck() function from VIM package. Its scores are the highest and its time performance relatively good. It is also worth to notice that basic imputer was not much worse than other functions, while its imputation time was very low even for large datasets.

⁶Flach Peter, Hernandez-Orallo Jose, Ferri, Cèsar "A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance."Proceedings of the 28th International Conference on Machine Learning, ICML 2011

⁷Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M. and Moore, J.H. (2007), A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol., 31: 306-315. DOI: 10.1002/gepi.20211

⁸Boughorbel, S., Jarray, F., El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, 12(6), e0177678. DOI: <https://doi.org/10.1371/journal.pone.0177678>

AUC measure

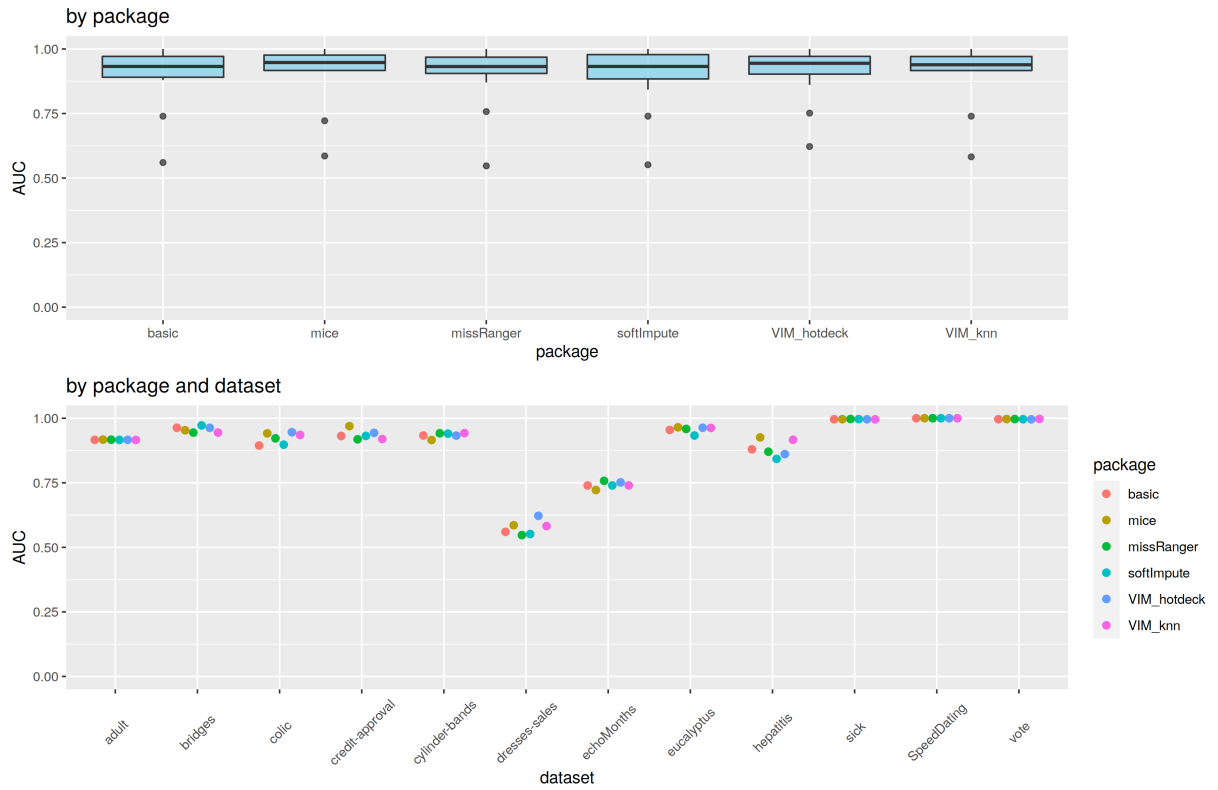


Figure 1: Area Under the Curve for different datasets and different imputations

BACC measure

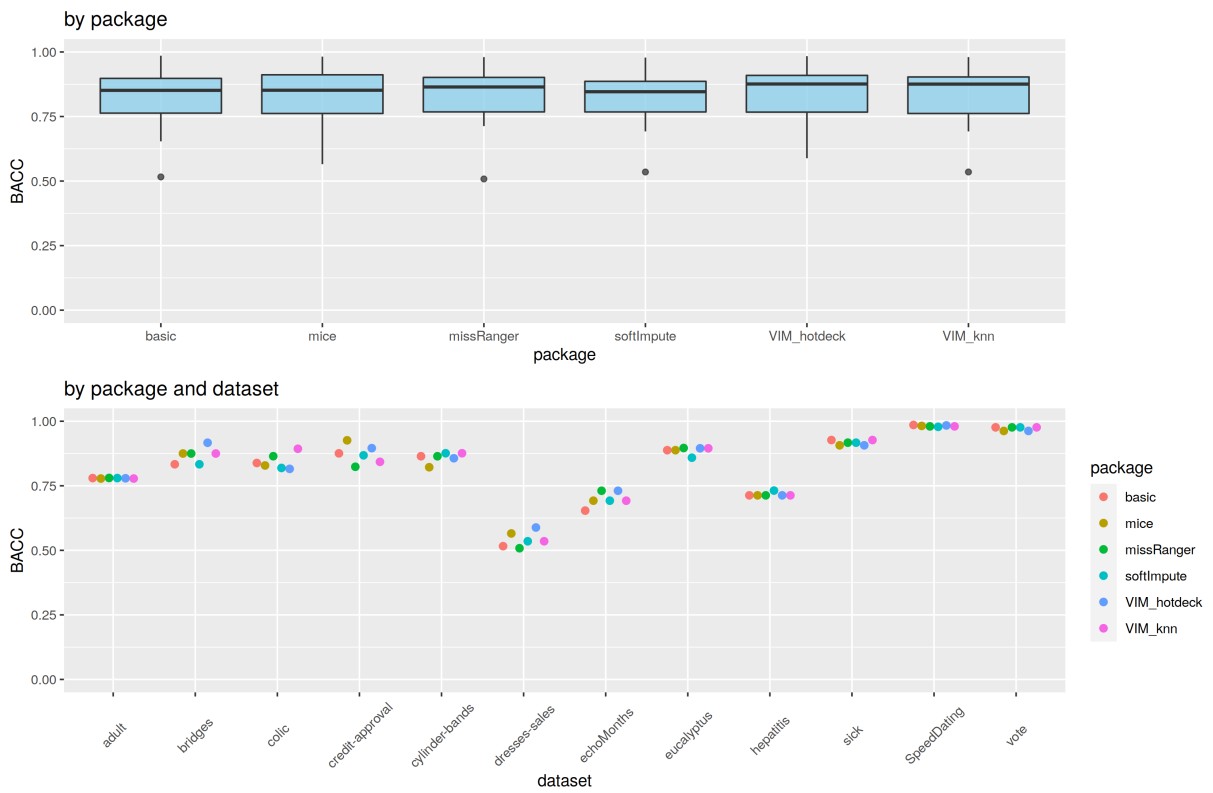


Figure 2: Balanced Accuracy for different datasets and different imputations

MCC measure

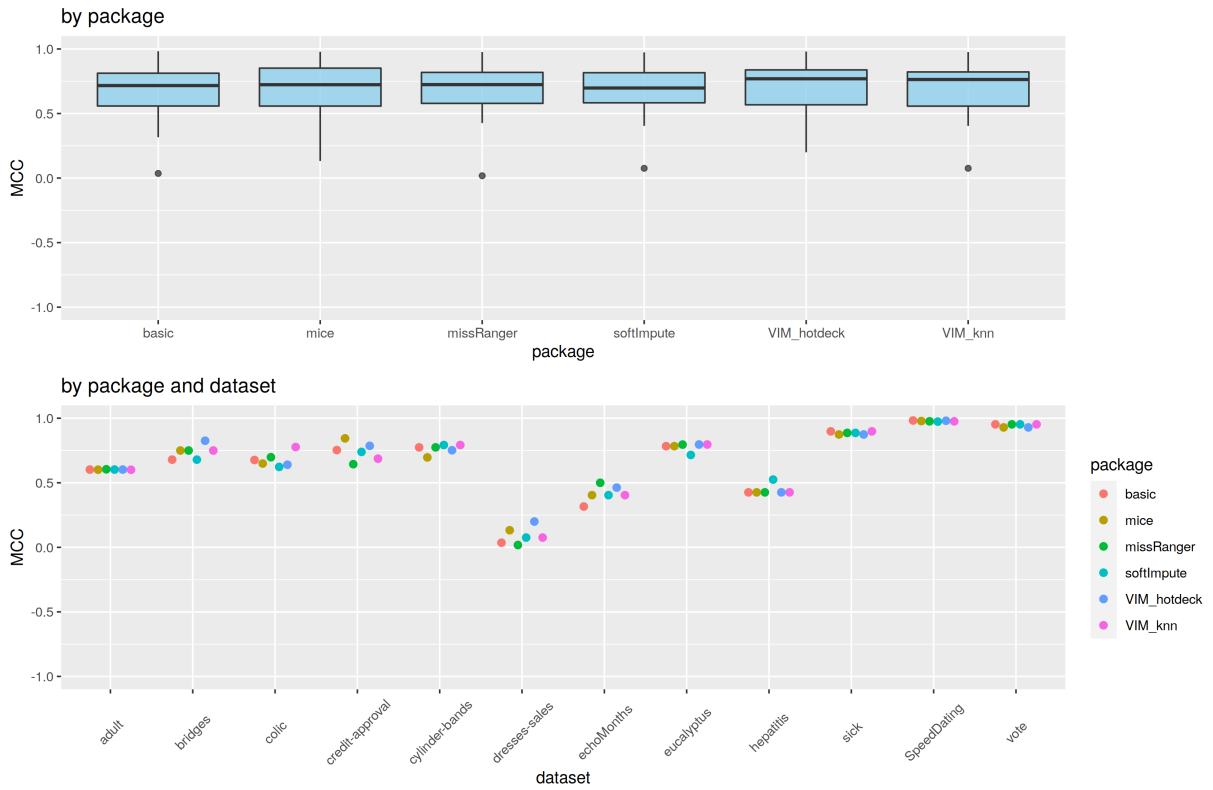


Figure 3: Matthews correlation coefficient for different datasets and different imputations

Imputation time

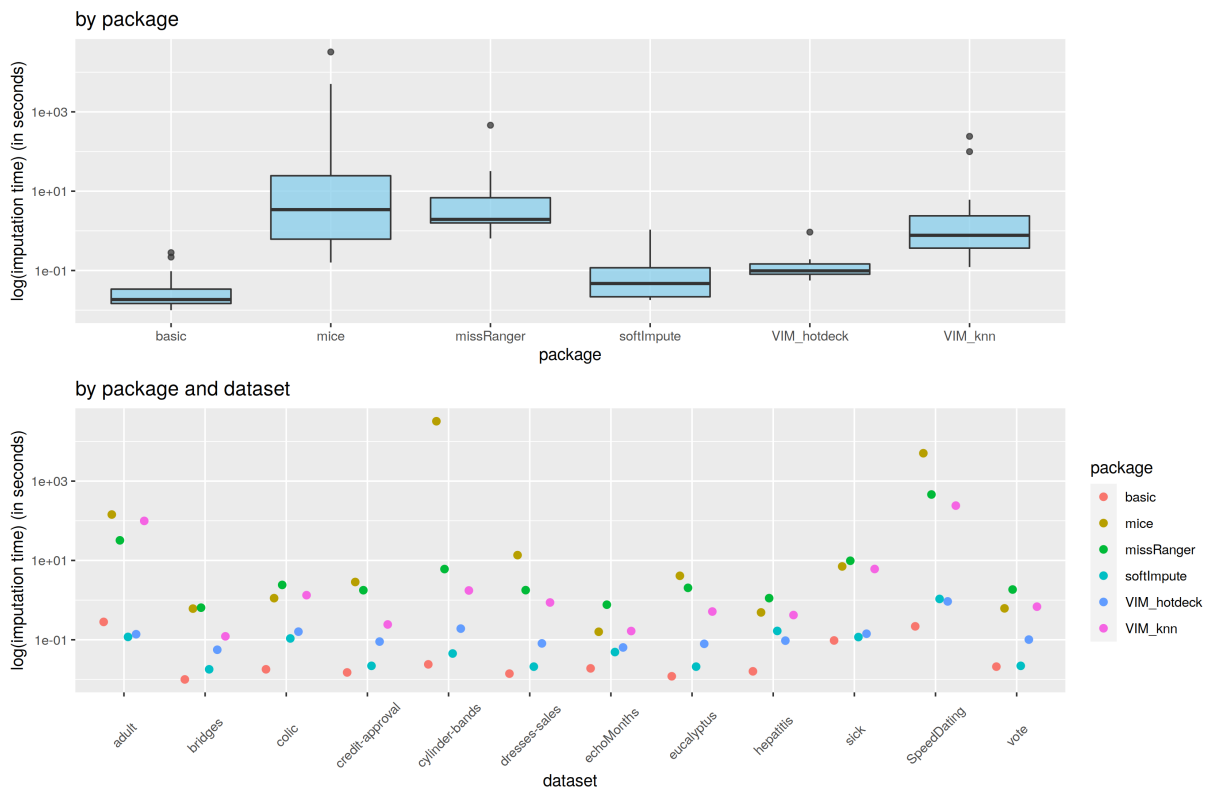


Figure 4: Logarithm of time of imputation for different datasets and different imputations