Machine Learning on Biomedical Datasets: Solving the Missing Values Problem without Imputation Methods

Marius FERSIGAN[a,*] and Marius MĂRUŞTERI[b]

[a] Iuliu Haţieganu University of Medicine and Pharmacy, Louis Pasteur Str., no. 6, 400349 Cluj-Napoca, Romania
[b] University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Gheorghe Marinescu Str., no. 38, 540139 Târgu Mureş, Romania
E-mails: marius.fersigan@algunion.com; marius.marusteri@umftgm.ro

* Author to whom correspondence should be addressed

Abstract
Background: Not all datasets are created equal. There are some happy scenarios when the researcher has the luxury of curating the dataset and ensuring all the desired fields are filled. However, especially when retrieving data from large EMR databases out in the wild, missing values are the norm (e.g., only some patients would have blood sugar readings, only a portion of the patients had transaminases determined, etc.). But the vast majority of machine learning models are not supporting missing values - hence, traditionally, we call the imputation methods to the rescue. But when a significant portion of values is missing, the imputation methods might insert incorrect data. Also, dropping the cases with missing values is not feasible when working with EMR data (too many instances are removed). *Aim*: To implement a missing-values-proof ensemble modeling method without sacrificing the predictive power. *Materials and Methods*: Using a realistic synthetic patient generator (Synthea), we generated large FHIR datasets under various settings. We implemented a cartesian genetic programming model to develop an automated acceptance test that, in turn, extracted missing-values-free subsets from the training partition of the original dataset. One of the core functions of the acceptance test is to ensure the potential missing values patterns are not going to insert bias in the model training step. We trained one or more models on each missing-values-free training subset. The resulted models are integrated into a global ensemble model. *Results*: Our approach resulted in superior model performance metrics without the need for missing values imputation methods. *Conclusions*: Machine learning on datasets containing missing values is feasible when employing specialized training dataset generation pipelines. Removing the missing values imputation methods from the workflow eliminates potentially incorrect data insertion, resulting in more robust models.

Keywords: Ensemble methods; Machine Learning; Missing data; Imputation methods