Imputation Techniques in Medical Studies

Research Problem

A key challenge with constructing machine learning models is missing data especially in the healthcare sector. In our case, we assume that some covariates (i.e., age, BMI, etc.) are missing from the data set, we need to fill-up these values with the best guess, what would be the optimum approach?

Algorithms

A. Traditional statistical inference measurements: Mean, Mode, Median, Correlation, EM, LDA, GMM, MICE, MIST, PCA, LR, SVD, NB and sampling.
B. Shallow machine learning techniques: RF-Proximity Matrix-, SVR, KNN, Clustering, etc.
C. Deep learning: GAN, LSTM and other autoencoders.

Method Design

We can explore two factors through a sensitivity analysis. First, the influence of the magnitude of missing data (how many: 5% 25% 50% 75%?), as in [1], and the influence of the amount of randomness of missing data for model development and performance.

The two recent surveys [1][2] are a good start.

The data set is uploaded to Kaggle by me.

Contact

If interested in this topic, please get in touch: abbas.cheddad[AT]bth.se

References

[1] Lin, WC., Tsai, CF. “Missing value imputation: a review and analysis of the literature (2006–2017),” Artificial Intelligence Review 53, 1487–1509 (2020). https://doi.org/10.1007/s10462-019-09709-4

[2] Thomas, T. and Rajabi, E. (2021), “A systematic review of machine learning-based missing value imputation techniques," Data Technologies and Applications, Vol. 55 No. 4, pp. 558-585. https://doi.org/10.1108/DTA-12-2020-0298.

Suggested Comparison Framework

AUC/F1 score of developed imputation for breast cancer classification models

Methods	5%		25%		50%		75%	
	Random	**N-Random**	**Random**	**N-Random**	**Random**	**N-Random**	**Random**	**N-Random**
Mean								
Median								
Correlation								
Etc.								