Abstract

The treatment of missing data has become a mandatory step for performing valid data analysis in most scientific research fields. In fact, researchers have found that dealing with missing data avoids misleading data analysis and improves the quality and power of the research results [1]. According to the authors in [2,3], the missing values in a data set could be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR), a categorization that should be taken into consideration to deal with the problem of missing data. The number of observations, the types of variables, and the percentage of missing values in a data set are also important characteristics that should be contemplated before dealing with missing values. Understanding the missing data case helps the researchers to identify the imputation techniques that best handles the missing data problem. However, the development of procedures to impute categorical data is not significantly available as the procedures focused on continuous data imputation [1]. This study compares six different imputation methods to find the one that performs the most appropriate treatment for categorical data, type ordinal, in a breast cancer dataset.
References

1. Finch, W. 2010. “Imputation methods for missing categorical questionnaire data: A comparison of approaches”. Journal of Data Science, vol. 8(8), pp. 361-378.
2. Rubin, D. 1976. “Inference and missing data”. Biometrika, vol. 63(3), pp. 581-592.
3. Little, R. and Rubin, D. 2002. “Introduction” in Statistical Analysis with Missing Data, 2nd ed., John Wiley & Sons, Inc., pp. 3-23.
4. de Leeuw, D. and Huisman, M. 2003. “Prevention and treatment of item nonresponse”. Journal of Official Statistics, vol. 19(2), pp. 153-176.
5. Schmitt, P., Mandel, J., and Guedj, M. 2015. “A comparison of six methods for missing data imputation”. Journal of Biometrics & Biostatistics, vol. 6(1), pp. 1-6.
6. Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., and Schafer, J. L. 1997. “Analysis with missing data in prevention research” in The science of prevention: methodological advances from alcohol and substance abuse research, vol. 1, pp. 325-366.
7. van der Heijden, G. J., Donders, A. R., Stijnen, T., and Moons, K. G. 2006. “Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example”. Journal of Clinical Epidemiology, vol. 59(10), pp. 1102-1109.
8. Schafer, J. L. and Graham, J. W. 2002. “Missing data: our view of the state of the art”. Psychological Methods, vol. 7(2), pp. 147-177.
9. Myers, T. A. 2011. “Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data”. Communication Methods and Measures, vol. 5(4), pp. 297-310.
10. Bhattacharyya, G. and Johnson, R. 2014. Statistics: Principles and Methods. 7th edition. John Wiley & Sons, Inc. [E-book] Available: Safari e-book.
11. Bruce, P. and Bruce, A. 2017. Practical Statistics for Data Scientists. 1st edition. O’Reilly Media, Inc. [E-book] Available: Safari e-book.
12. Wolberg, W. 1992. “Breast Cancer Wisconsin (Original) Data Set”. Internet: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
13. Olinsky, A., Chen, S., and Harlow, L. 2003. The comparative efficacity of imputation methods for missing data in structural equation modeling. European Journal of Operational Research, vol.151(1), pp. 53-79.
14. Shrive, F. M., Stuart, H., Quan, H., and Ghali, W. A. 2006. “Dealing with missing data in a multi-question depression scale: a comparison of imputation methods”. BMC medical research methodology, vol. 6(1), pp. 57.
15. Hu, L., Huang, M., Ke, S., and Tsai, C. 2016. “The distance function effect on k-nearest neighbor classification for medical datasets”. SpringerPlus, vol. 5, pp.1-9.
16. García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., and Verleysen, M. 2009. “K nearest neighbors with mutual information for simultaneous classification and missing data imputation”. Neurocomputing, vol. 72(7-9), pp. 1483-1493.
17. Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. 2011. “Multiple imputation by chained equations: what is it and how does it work?”. International journal of methods in psychiatric research, vol. 20(1), pp. 40-49.
18. Mazumder, R., Hastie, T., and Tibshirani, R. 2010. “Spectral regularization algorithms for learning large incomplete matrices”. Journal of Machine Learning Research, vol. 11, pp. 2287-2322.
Index Terms

Computer Science

Information Sciences

Keywords

MCAR; categorical data; ordinal data.