15 Jun 2020

Handling missing data for unsupervised learning with an application on a FITBIR Traumatic Brain Injury (TBI) Dataset

Louis Steinmeister

Dacosta Yeboah

Gayla Olbricht
Missouri University of Science and Technology, olbrichtg@mst.edu

Tayo Obafemi-Ajayi

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/math_stat_facwork/1010

Handling missing data for unsupervised learning with an application on a FITBIR Traumatic Brain Injury (TBI) Dataset

Louis Steinmeister, Dacosta Yeboah, Gayla Olbricht, Tayo Obafemi-Ajayi, Bassam Hadi, Daniel Hier, Donald C. Wunsch II

Background The problem of missing data and imputation have been widely discussed amongst specialists. However, many data scientists and applied statisticians fail to appropriately consider this issue. Often, it seems intuitive to discard observations containing missing data or simply to substitute means. This can lead to disastrous consequences, particularly in an era of exponentially increasing data volumes. In the following, we show how inappropriate handling of missing data and an insufficient analysis of the censoring mechanism can lead to a bias, overconfidence in the estimation of parameters, could challenge the reproducibility of obtained results, and may distort the structure of the dataset.

Methods The latter is demonstrated to be particularly impactful in unsupervised learning approaches, such as clustering or PCA, which aim to describe different aspects of the data distribution. Our goal is to provide guidance to practitioners in dealing with missing data by 1) summarizing the different mechanisms of missingness and outlining their implications, 2) by identifying the role of domain experts and helping statisticians and data scientists ask the right questions, 3) by outlining appropriate imputation techniques for the remaining cases, which reduce potential biases and tend to impose less artificial structure on the data. Lastly, we suggest ways in which to assess and incorporate imputation uncertainty.

Results As data volumes grow, dealing with missing data has become increasingly important. We present a pipeline for evaluating and dealing with missingness, incorporating valuable inputs from domain experts, and assessing imputation uncertainty.

Conclusions and Significance to the Warfighter It is demonstrated how this pipeline is employed to impute missing data in a TBI dataset to reduce artificial structure which may lead to misidentification and misclassification of TBI-subphenotypes (TBI is a common condition for deployed soldiers in war zones). Furthermore, the guidance provided in this paper is expected to assist practitioners in more accurately assessing uncertainty arising from missing data. Thereby, we hope to reduce unidentified biases and analytical errors, while improving the reproducibility of obtained results in military and biomedical research leading to more accurate and objective models.

Learning Objectives

1. Understanding the mechanisms of missingness and their implications for subsequent analyses.
2. Incorporating knowledge from domain experts to obtain information about the data generating process and logical rules applying to the data.
3. Gaining familiarity with appropriate statistical imputation techniques to reduce biases and artificial structure leading to more objective outcomes.

Disclaimer & Acknowledgement This research was sponsored by the Missouri University of Science and Technology Mary K. Finley Endowment and Intelligent Systems Center; the Army Research Laboratory (ARL) and the Lifelong Learning Machines program from DARPA/MTO, and it was accomplished under Cooperative Agreement Number W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Format We would like to submit this abstract for all formats.

This abstract was accepted for a poster presentation at the Military Health System Research Symposium 08/2020. This can be verified and the abstract will be accessible on the official website (mhsrs.amedd.army.mil/SitePages/Home.aspx) after 06/23/2020 by creating an account.

Suggested Citation:
Steinmeister, L., Yeboah, D., Olbricht, G., Obafemi-Ajayi, T., Hadi, B., Hier, D., & Wunsch II, D. (2020). *Handling missing data for unsupervised learning with an application on a FITBIR Traumatic Brain Injury (TBI) Dataset*. Military Health System Research Symposium.