Deep Learning Imputation for Unbalanced and Incomplete Likert-type Items

Olushola Soyoye, Kamal Chawla, Zachary Collier, Minji Kong, Yasser Payne, Ann Aviles

Abstract

Asymmetric Likert-type items are skewed-scaled with either no neutral response option or an uneven number of possible favorable and unfavorable responses. Modern missing data methods may be problematic when respondents do not answer asymmetric items because they assume multivariate normality. Alternatively, list-wise deletion and mean imputation assume that data are missing completely at random, which is often unlikely in surveys and rating scales. This article explores the potential of implementing a scalable deep learning-based imputation method. Additionally, we provide access to deep learning-based imputation to a broader group of researchers without requiring advanced machine learning training. We apply the methodology to the Wilmington Street Participatory Action Research (PAR) Health Project.

Objectives

- Multivariate Imputation by Chained Equations (MICE) imputation using deep learning models can (1) avoid making distributional assumptions; (2) handle mixed data types; (3) model nonlinear relationships between variables; and (4) perform well for data with many variables (i.e., high-dimensional settings (Wang et al., 2021). Although there is potential for deep learning to improve MICE imputation, efforts in evaluating deep learning methods in real survey data remain scarce.
- This work proposes imputation of Likert scale data with MICE using deep artificial neural networks (DNN) as an alternative to traditional approaches, because DNN require no imputation model specification or distribution assumptions.

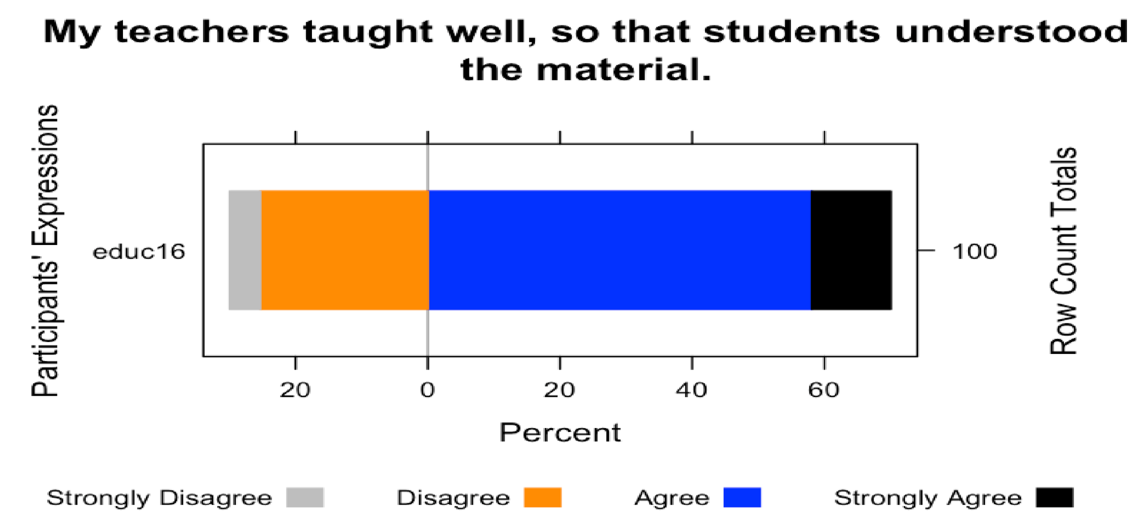
Significance

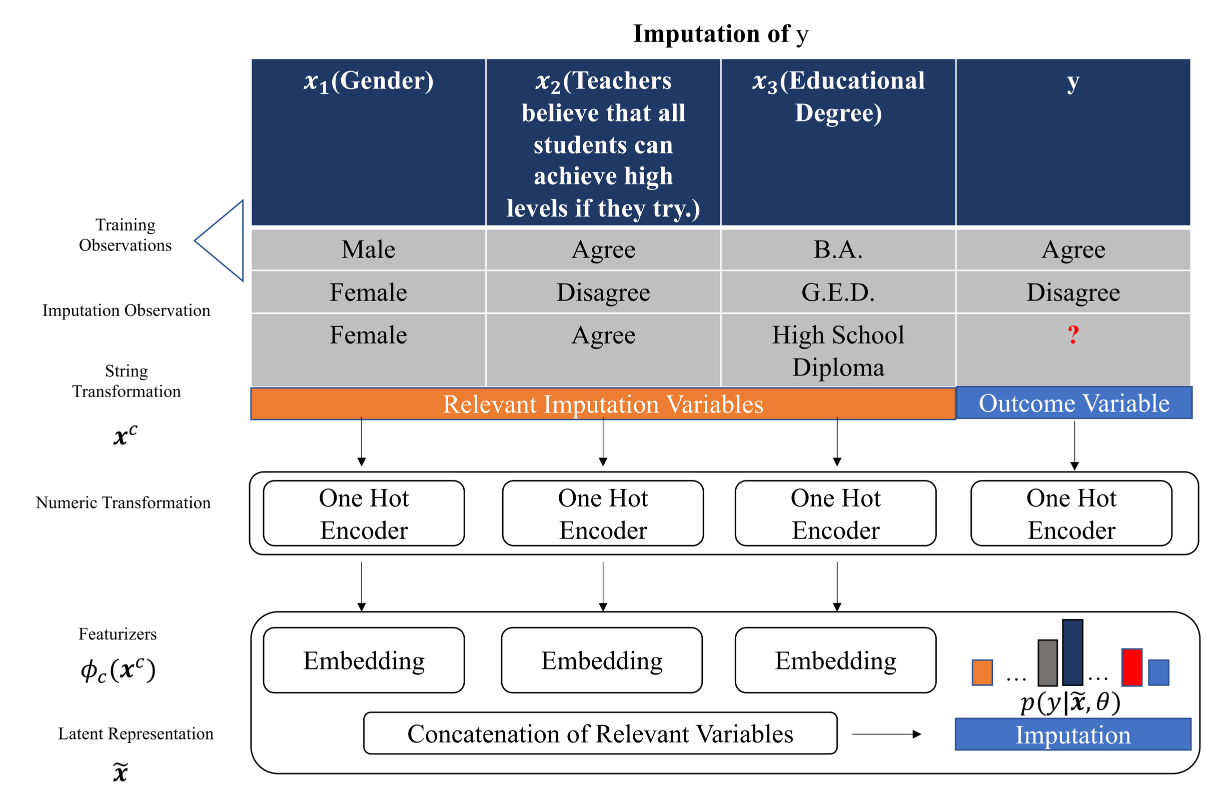
Many existing proposed solutions towards data imputation hold several limitations that have yet to be addressed, all while requiring advanced statistical knowledge and potentially making them inaccessible to many researchers. The illustrated success of implementing a deep learning-based approach in R without requiring advanced statistical background opens up possibilities for a larger group of researchers to utilize imputation methods when handling missing data.

Methodology

Table 1: Demographic characteristics of survey sample.

Demographic characteristics	Quantitative survey sample (N = 771) n	%
Age		
Males 16-24	192	24.9
Males 25-34	148	19.2
Males 35-44	155	20.1
Females 16-24	86	11.1
Females 25-34	74	9.6
Females 35-44	62	8.0
Females 45-54	54	7.0
Missing	5	0.1
	Mean = 29.86	
Education Level		
High School Diploma	351	45.5
GED	185	24.0
Bachelor's Degree	8	1.0
Other	49	6.3
Missing	173	23.2
Marital Status		
Single without a Significant Partner	421	54.5
Single with a Significant Partner	169	21.9
Legally Married	50	6.5
Living Together (cohabitation)	70	9.1
Common Law Marriage	20	2.6
Married but Separated	14	1.8
Widowed	6	0.8
Missing	22	2.8
Employment		
Full Time	189	24.5
Part Time	135	17.5
Unemployed and Looking for Work	367	47.5
Unemployed and Not Looking for Work	68	8.8
Missing	13	1.7





Data:

Data for this paper were collected from a larger study that chiefly examined attitude towards and experiences with violence in the Northside and Westside section of Wilmington, Delaware. The analytic sample included 771 street-identified Black Americans: 443 men and 328 women, between the ages of 16 and 54.

Training:

- Determine the "meaningful data" to be used in the creation of our imputation model.
- Load the datafile into a pandas Data Frame, with 80% of it split into a training subset and the remaining 20% split into a test subset.
- Datawig, an imputation package, trains DNNs using only non-missing values of the imputed variable, along with other variables in the dataset, for both training and testing subsets.

Results:

Metrics to Determine the Quality of the Imputed Models' Predictions

	Precision	Recall	f1-Score	Support
Agree	**.80** (0.55)	**.81** (0.53)	**.80** (0.54)	101
Disagree	**.51** (0.21)	**.57** (0.25)	**.54** (0.23)	35
Strongly agree	**.82** (0.12)	**.56** (0.12)	**.67** (0.12)	16
Strongly disagree	**.67** (0.00)	**.67** (0.00)	**.67** (0.00)	3
Macro average	**.70** (0.22)	**.65** (0.22)	**.67** (0.22)	155

Note. The evaluation metrics in bold were obtained after performing imputation using MICE with a DNN. The values in parentheses, on the other hand, represent the results obtained when MICE was used in conjunction with multinomial logistic regression.

Conclusion

We introduced DNN to an unfamiliar audience by explaining its advantages over the linear model that may have difficulty predicting Likert-type responses.

We encourage **Datawig** users to explore random search to help them optimize the hyperparameters for more efficient training compared with hill-climbing (Bergstra & Bengio, 2012). Users should also consider regularization to improve the generalization ability and to reduce the complexity of the DNN.

Our study performed single imputation. Multiple imputation differs from single 30 imputation in that it generates a set of possible values for each missing data point, instead of treating the imputed value as an actual observation (Xia & Yang, 2016). Multiple imputation with DNNs may more so mirror the uncertainty of the imputed responses compared with single imputation (W. Leite & Beretvas, 2010).

DNN have shown success in imputation tasks with large datasets, but can struggle with high-dimensional, low-sample-size data, leading to overfitting and unstable gradients (B. Liu et al., 2017). A solution is to choose a DNN architecture that has enough capacity to fit the training data, then use regularization to reduce overfitting (Olson et al., 2018). This approach is similar to training a random forest with many decision trees, then relying on randomization and averaging to reduce variance.

References & Acknowledgement

Take a picture to download the full paper