MISSING DATA IMPUTATION FOR SUPERVISED LEARNING

Jason Poulos and Rafael Valle

University of California, Berkeley

Abstract: This paper compares methods for imputing missing categorical data for supervised learning tasks. The ability of researchers to accurately fit a model and yield unbiased estimates may be compromised by missing data, which are prevalent in survey-based political science research. We experiment on two machine learning benchmark datasets with missing categorical data, comparing classifiers trained on non-imputed (i.e., one-hot encoded) or imputed data with different degrees of missing-data perturbation. The results show imputation methods can increase predictive accuracy in the presence of missing-data perturbation. Additionally, we find that for imputed models, missing-data perturbation can improve prediction accuracy by regularizing the classifier.

Key words and phrases: artificial neural networks (ANNs), decision trees, imputation methods, missing data, perturbation, random forests