World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:12, No:07, 2018

Effect of Genuine Missing Data Imputation on Prediction of Urinary Incontinence

Authors : Suzan Arslanturk, Mohammad-Reza Siadat, Theophilus Ogunyemi, Ananias Diokno

Abstract : Missing data is a common challenge in statistical analyses of most clinical survey datasets. A variety of methods have been developed to enable analysis of survey data to deal with missing values. Imputation is the most commonly used among the above methods. However, in order to minimize the bias introduced due to imputation, one must choose the right imputation technique and apply it to the correct type of missing data. In this paper, we have identified different types of missing values: missing data due to skip pattern (SPMD), undetermined missing data (UMD), and genuine missing data (GMD) and applied rough set imputation on only the GMD portion of the missing data. We have used rough set imputation to evaluate the effect of such imputation on prediction by generating several simulation datasets based on an existing epidemiological dataset (MESA). To measure how well each dataset lends itself to the prediction model (logistic regression), we have used p-values from the Wald test. To evaluate the accuracy of the prediction, we have considered the width of 95% confidence interval for the probability of incontinence. Both imputed and non-imputed simulation datasets were fit to the prediction model, and they both turned out to be significant (p-value < 0.05). However, the Wald score shows a better fit for the imputed compared to non-imputed datasets (28.7 vs. 23.4). The average confidence interval width was decreased by 10.4% when the imputed dataset was used, meaning higher precision. The results show that using the rough set method for missing data imputation on GMD data improve the predictive capability of the logistic regression. Further studies are required to generalize this conclusion to other clinical survey datasets.

Keywords : rough set, imputation, clinical survey data simulation, genuine missing data, predictive index
Conference Title : ICDM 2018 : International Conference on Data Mining
Conference Location : Dublin, Ireland
Conference Dates : July 23-24, 2018