Cao YH, Wu JX. Random subspace sampling for classification with missing data. JOURNAL OF COMPUTERSCIENCE AND TECHNOLOGY 39(2): 472−486 Mar. 2024. DOI: 10.1007/s11390-023-1611-9

Random Subspace Sampling for Classification with Missing Data

Yun-Hao Cao (曹云浩), Jian-Xin Wu (吴建鑫)

Research Objectives

- Scope: Classification with missing data
- Challenge: How to combine classification and imputation *in an efficient way*? How to develop an ensemble method that is effective even when the data *contains many missing values*?
- Contribution: propose an efficient and effective method at different levels of missing data.

Research Method

- Our method mainly contains three parts:
  - Estimating the histogram distributions of each feature.
  - Constructing neural random subspaces.
  - Sampling for missing features in each subspace.

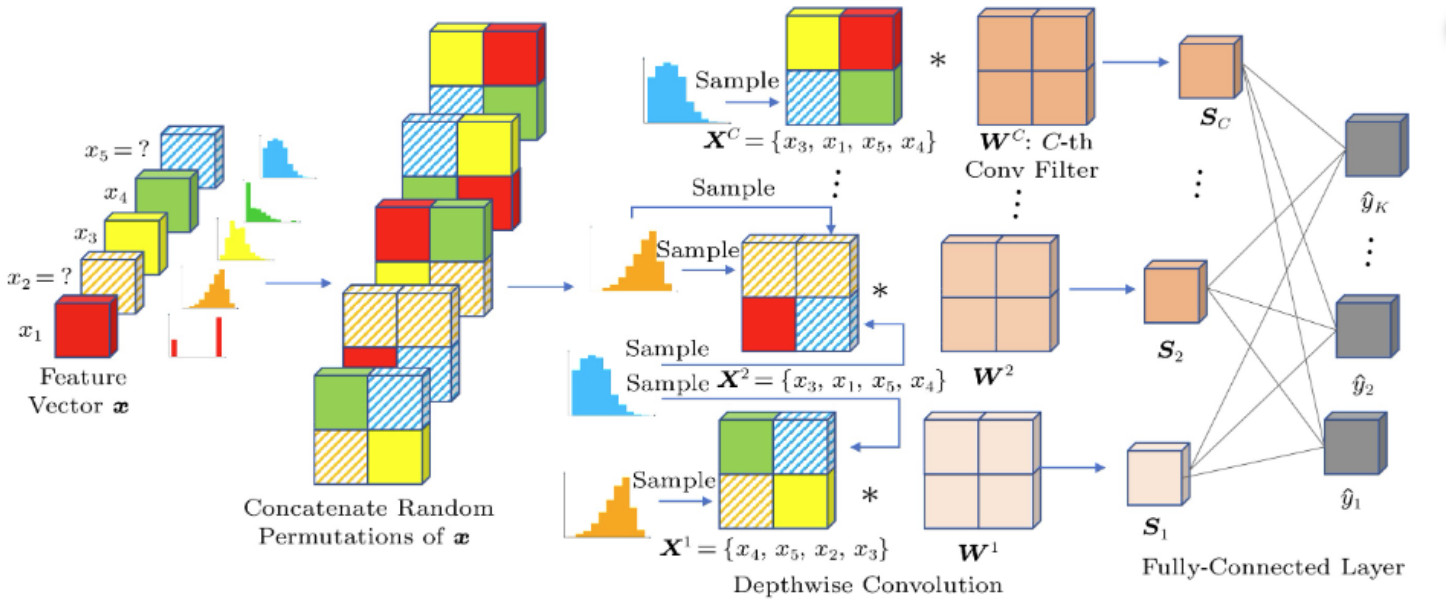

Fig.1. RSS architecture. The input feature vector is $\boldsymbol{x} = (x_1, \ldots, x_5)$, where x_2 and x_5 are missing and marked with diagonal lines. For better illustration, we set $nMul$ to 1 and hence $C = 5$ and $\boldsymbol{X} = \{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^5\}$. "$*$" denotes the convolution operator.

Research Results

- We achieve superior performance on 6 incomplete datasets with inherent missing values and 7 complete datasets at 4 levels of artificially introduced missing values.
- Our method has a larger edge over other methods along with the increase of the portion of missing values (i.e., especially effective for large missing portions).

Research Conclusions

- We proposed a random subspace sampling method RSS for classification with missing data.
- Unlike most established approaches, RSS *does not train on fixed* imputed datasets
- Without the need to train multiple models for ensemble, we use *one single model* for ensemble during inference.
- We will further investigate our method from a theoretical perspective.