General-Purpose Open-Source Program for Ultra Incomplete Data-Oriented Parallel Fractional Hot Deck Imputation

UP-FHDI

In Ho Cho
Depart. of Civil, Construction and Environment Engineering
Iowa State University
Ames, Iowa, USA
icho@iastate.edu

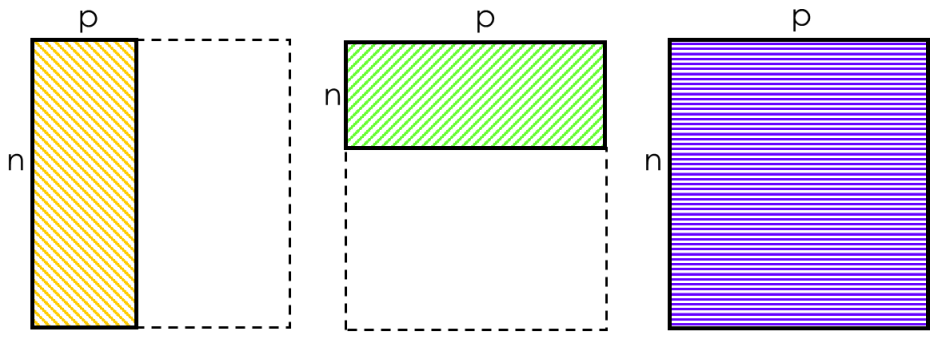
Jae-Kwang Kim
Department of Statistics
Iowa State University
Ames, Iowa, USA
jkim@iastate.edu

Yicheng Yang
Depart. of Civil, Construction and Environment Engineering
Iowa State University
Ames, Iowa, USA
yicheng@iastate.edu

ABSTRACT

There emerges a strong need for a large/big data-oriented imputation method for accelerating data-driven scientific discovery in the new era of big data and powerful computing. Imputation is a statistics-based procedure to fill in missing data, and there exists a wide spectrum of methods. Still, they are often not applicable for large/big incomplete data and require difficult statistical assumptions. With support from NSF (OAC-1931380), we developed the ultra data-oriented parallel fractional hot-deck imputation (UP-FHDI [1,2]) which is a general-purpose, assumption-free software for handling item nonresponse in big incomplete data by leveraging the theory of FHDI and parallel computing. Here, "ultra" data means a data set with high dimensions and many instances (i.e., concurrently big-p and big-n; see Figure). UP-FHDI inherits the strength of FHDI [3] that can cure multivariate missing data by filling each missing unit with multiple observed values without requiring any prior distributional assumptions.

UP-FHDI adopts a parallel file system that supports inter-processor communication and allows simultaneous access from multiple compute servers to the hard drive to optimize memory usage by managing essential data in memory and other data on the hard drive. Meanwhile, we use the Optimal Overload IO Protection System with UP-FHDI to dynamically adjust the intensive and simultaneous IO workload during a job to avoid global file system performance degradation. Exploring the strength of this parallel file system, we provide full details of ultra data-oriented parallelisms on significant steps of UP-



(a) big-n data (b) big-p data (c) ultra data

Figure. Types of incomplete datasets: (a) $n \gg K^p$; (b) $n \leq K^p$; (c) n and p are both very large, where K is the number of categories of imputation cells. Individual continuous variable is categorized into K in FHDI.

FHDI: cell construction, estimation of cell probability using expectation maximization, pa rallel imputation, and parallel variance estimation, respectively. The cell construction step adopts the parallel k-nearest neighbors method for deficient donor selection to break the computational bottleneck of cell-merging scheme of serial FHDI. The sure independence screening is embedded into the UP-FHDI for ultrahigh dimensional variable reduction and thus overcomes the curse of dimensionality. Besides the parallel Jackknife method, UP-FHDI implements another computationally efficient variance estimation using parallel linearization techniques.

We validate the UP-FHDI's accuracy by conducting Monte Carlo simulations. Results confirm that UP-FHDI can handle an ultra dataset with one million instances and 10,000 variables concurrently and do not have a specific limitation on data volume. UP-FHDI exhibits promising scalability with different ultra datasets, and its practical computational performance has a good agreement with the cost models of speedup and execution time. Furthermore, we confirm UP-FHDI's positive impact on the subsequent deep learning performance. Since machine learning models may over-fit with ultrahigh dimensional data, we adopted a two-stage feature selection method that leverages the mutual information and graphical LASSO to reduce ultrahigh dimensionality to a small subset as a pre-processing remedy.

We provide full documentation to illustrate how to deploy, compile, and perform UP-FHDI for ultra incomplete data curing. To maximize the benefit of broad researchers, many synthetic and practical example data sets for UP-FHDI are made publicly available in IEEE Data Port.

Keywords—Parallel Imputation; Incomplete Data Curing; Fractional Hot Deck Imputation

REFERENCES

[1] Y. Yang, Y. Kwon, J. K. Kim, and I. Cho, "Ulta data-oriented parallel fractional hot-deck imputation with efficient linearized variance estimation," *IEEE Transactions on Knowledge and Data Engineering*, 2021. (under review)

[2] Y. Yang, J. Kim, and I. Cho, "Parallel Fractional Hot Deck Imputation and Variance Estimation for Big Incomplete Data Curing," IEEE Transactions on Knowledge and Data Engineering 2020. [DOI: 10.1109/TKDE.2020.3029146].

[3] J. Im, I. Cho, and J. Kim, "FHDI: An R Package for Fractional Hot-Deck Imputation for Multivariate Missing Data," The R Journal, 2018, vol. 10(1), 140-154. [https://journal.r-project.org/archive/2018/RJ-2018-020/index.html].