Comparing the Impact of Data Imputation on Time Series Downstream Tasks

Zakhar Tymchenko

Master thesis in Computer Science

Recording time series data is rarely a neat process, as sensor failures frequently occur. Those failures yield data errors and missing values which hinder time series analysis and downstream tasks such as classification or prediction. Several imputation techniques with different effectiveness have been proposed in the literature. Those algorithms are mostly evaluated directly against ground data, isolated from the tasks where the imputed data will be used. A handful of benchmarks have been proposed to evaluate the impact of imputation techniques. Those benchmarks evaluate downstream tasks on point-based tasks and do not consider time series tasks. This work proposes a new comprehensive benchmark designed specifically for downstream evaluation of time series imputation techniques and subsequent analysis of the results. Additionally, the benchmark implements three improvements to the existing state-of-the-art imputation techniques and evaluate them on various datasets. Experimental evaluation shows that each improvement has a niche application, and one in particular is universally usable in most of the cases. The benchmark can be built upon in the future to either deepen the analysis through existing established test pipeline, or to be expanded and include more downstream tasks and imputation algorithms.

Prof. Philippe Cudré-Mauroux