ABSTRACT

Sergius Tadao Hayashi (00000013162)

IMPUTATION USING STATISTICAL AND MACHINE LEARNING METHODS IN FORECASTING LIFE EXPECTANCY

Thesis, Faculty of Science and Technology (2019).

(xv + 90 pages, 18 tables, 19 figures, 2 appendix)

Processing, collecting and reporting data is essential in making decisions. Yet even in a well-designed and controlled study, the occurrence of missing data is not improbable. The occurrence of missing data decreases the statistical power of the dataset and training power for machine learning purposes. This thesis aims to compare six imputation method, three of which are statistical imputation methods and three are machine learning methods for life expectancy data to determine an optimal method for cases with its type of missingness pattern. The life expectancy data consist of 22 variables in relation to social, economic and health of 194 countries collected from World Health Organization's and Wold Bank's database. An artificial dataset was build for simulating the missingness of the original dataset to measure the performance of each method by error metrics. The artificial dataset mimics the original dataset's missingness patterns and the nullity correlation between variables. Imputed artificial dataset were evaluated through its mean squared error, mean absolute error, and mean absolute percentage error while the original dataset were evaluated through its mean and variance changes. Surprisingly, given that the multi layer perceptron had 10^6 iterations, the Hot-Deck and KNN method showed the best results for statistical and machine learning, respectively, with Hot-Deck slighly outperforming KNN.

Keywords: Missing Data, Imputation, Statistical Imputation, Machine Learning Imputation, Mean Imputation, Hot-Deck Imputation, Multiple Imputation, Multi Layer Perceptron, Self Organizing Map, K-Nearest Neighbor.

References: 56 (1990-2018)

ABSTRAK

Sergius Tadao Hayashi (00000013162)

IMPUTATION USING STATISTICAL AND MACHINE LEARNING METHODS IN FORECASTING LIFE EXPECTANCY

Skripsi, Fakultas Sains dan Teknologi (2019).

(xv + 90 halaman, 18 tabel, 19 figur, 2 lampiran)

Mengumpulkan, mengolah dan melaporkan data adalah bagian penting dari pembuatan keputusan. Namun dalam kondisi penelitian yang ditetapkan dan terkontrol, peluang hilangnya sebuah data itu masih mungkin terjadi. Hilangnya data dari sebuah dataset mengurangi kekuatan statistik dari dataset tersbut dan kekuatan data menjadi bahan latihan untuk program *machine learning*. Penelitian ini bertujuan untuk membandingkan enam metode berbeda, tiga darinya adalah metode statistik dan tiga lainnya adalah metode *machine learning* untuk dataset harapan hidup untuk menentukan metode optimal untuk mengimputasi data yang hilang berdasarkan pola kehilangan datanya. Data harapan hidup terdiri dari 22 variabel yang mengandung konteeks sosial, ekonomi dan kesehatan dari 194 negara, diambil dari pusat data *World Health's Organization* dan *World Bank*. Sebuah dataset simulasi dibuat dari dataset asli untuk mengukur performa metode-metode berdasarkan errornya. Data simulasi mencerminkan karakteristik pola kehilangan dari data aslinya. Hasil imputasi dataset simulasi dievaluasi dengan *mean squared error, mean absolute error, and mean absolute percentage error*, di lain sisi hasil imputasi dataset asli dievalusi berdaasarkan perubahan rata-rata dan variansinya. Menariknya, dengan adanya metode *multi layer perceptron* dengan iterasi 10^6, metode *Hot-Deck*, untuk kategori statistik dan metode *KNN*, untuk kategori machine learning menunjukan performa yang lebih baik untuk dimana Hot-Deck memiliki mengungguli KNN.

Kata Kunci: *Missing Data, Imputation, Statistical Imputation, Machine Learning Imputation, Mean Imputation, Hot-Deck Imputation, Multiple Imputation, Multi Layer Perceptron, Self Organizing Map, K-Nearest Neighbor.*

Referensi: 56 (1990-2018)