Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare

Tan Duy Le
School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
tanld@jaist.ac.jp

Razvan Beuran
School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan

Yasuo Tan
School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan

***Abstract*—In healthcare research, the reliability of input data is essential. However, missing data is a common incident in this field for various reasons. Current research mainly focuses on developing new data imputation methodologies, while there is a need for studying on a global evaluation of existing algorithms. In this research, we compared the performance of four influential missing data imputation algorithms, Regularized Expectation-Maximization (EM), Multiple Imputation (MI), kNN Imputation (kNNI) and Mean Imputation on two real health care datasets: (1) MHEALTH dataset and (2) the University of Queensland Vital Signs dataset. Under the Missing Completely At Random (MCAR) assumption, Root Mean Squared Error (RMSE) and execution time were used as best performing evaluation criteria. The experimental analysis suggests that EM is the imputation algorithm which is expected to be a good choice to deal with the problem of missing data in the healthcare area.**

***Index Terms*—missing data, healthcare, comparison, EM, MI, Mean, kNNI**

I. Introduction

In the healthcare field, especially in healthcare monitoring systems, the reliability of input data is extremely important. Accurate healthcare decisions can only be made with accurate input data. To execute healthcare tasks, the application expects to process sequences of complete instances collected from sensors. However, for various reasons such as equipment errors, incorrect measurements, limitations in the data acquisition process or faulty sampling, missing data is a typical problem. A missing value is defined as an attribute that has not been sampled in the data set, or that was never recorded. The presence of missing value not only makes the conduct of data analysis complicated but also poses severe concerns for scientists. Sophisticated handling methods are required to achieve a better accuracy if there are more than 5% missing samples [1].

Many efforts have been made, and a large body of research regarding technologies for substituting missing data with statistical prediction, which is defined as "missing data imputation", have been proposed. However, the main focus of the current study is on developing new imputation methodologies, while there is a lack of research on a global evaluation of existing methods, especially on healthcare data. Healthcare data is longitudinal, complex and unstructured data. Therefore, researchers can not treat healthcare data as the normal type of data. In addition, details on the performance of each imputation methodology can provide guidelines to obtain the more appropriate methodological decision in practice.

This research compares four influential missing data imputation algorithms, Regularized Expectation-Maximization (EM), Multiple Imputation (MI), kNN Imputation (kNNI) and Mean Imputation on two real healthcare datasets. Based on two evaluation criteria: Root Mean Squared Error (RMSE) and execution time, the result of the comparison is generated.

The remainder of this research is organized as follows. The selection of the most influential missing data imputation algorithms, missing data patterns, datasets, evaluation criteria and data analysis procedure are discussed in Section II. Section III provides the experimental results. Finally, the paper ends with conclusions in Section IV.

II. Methods

Based on various comprehensive research, Regularized EM, MI, kNNI and Mean Imputation are indicated as the most influential missing data imputation algorithms for healthcare. The experiment was conducted by analyzing two well-established datasets called MHEALTH and the University of Queensland Vital Signs. We introduced 5% to 45% of missing values to the datasets under Missing Completely at Random (MCAR) assumption. After 1000 simulations for each percentage of missing value for each dataset, the final result was obtained by averaging Root Mean Squared Error (RMSE) and execution time.

A. Most Influential Missing Data Imputation Algorithms

The imputation algorithms were selected based on their usage, reference, popularity, standardization, smart, variability,

978-1-5386-6113-0/18/$31.00 ©2018 IEEE

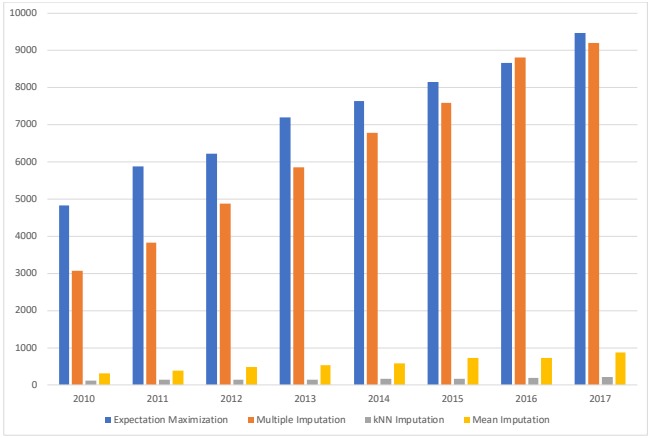

Fig. 1. The evolution of academic publications concerning missing data imputation algorithms from 2010 to 2017 in healthcare area.

and extension, which are proposed in the healthcare data research community. Various comprehensive studies [2], [3] and [4] indicated Regularized EM, MI, kNNI and Mean Imputation as the influential missing data imputation algorithms.

EM [5] is a meta-algorithm applied to optimize the maximum likelihood of data by repeating two steps until coverage: uses other variables to impute a value (Expectation step) and then checks whether that is the most probable value (Maximization step). Since its proposal, the citation accounted in Google Scholar is more than 51359 citations. Therefore, for missing value imputation, EM is one of the first successful solutions which applies maximum likelihood as a guaranteed approach. In 2016, there was 5500 research applied EM in healthcare application indexed by Google Scholar.

MI [6] is a statistical algorithm for handling incomplete data sets. MI creates $M > 1$, however, usually $M \leq 10$ complete datasets from the original data, where each complete dataset is analyzed separately and then combined to produce one set of overall results. There are three required steps for the application of this algorithm: imputation, analysis, and pooling. Since its proposal, the impact of this method is notable in the literature with nearly 15000 citations while over 8800 research projects applied MI in healthcare applications accounted by Google Scholar.

kNNI defines each sample or individual with its closest k neighbors in a multi-dimensional space and then imputes the missing data with a given variable by averaging non-missing values of these k neighbors. In spite of being cited and compared in thousands of research projects, the application of kNNI in the healthcare field is still small compared with EM and MI algorithms. Since its proposal, there were only around 800 projects applied kNNI to solve problems in healthcare.

Mean Imputation [7] is a method where the missing value is imputed by the mean of the available values. In this method,

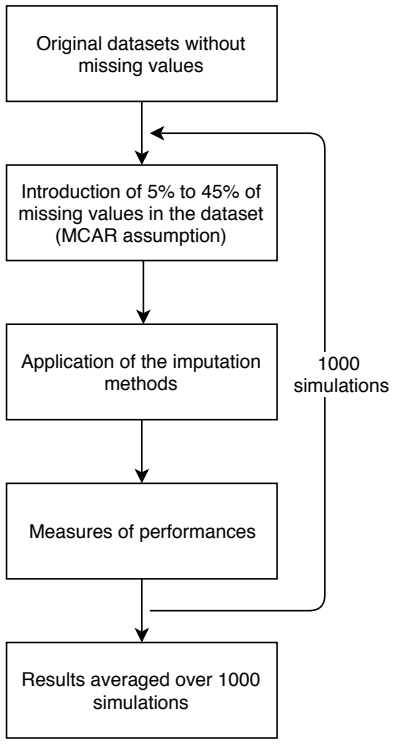

Fig. 2. Data analysis flowchart [11].

the sample size is maintained, however, the variability in the data is reduced. Therefore, the standard deviations and the variance estimates tend to be underestimated. However, because of its simplicity, Mean Imputation is widely used by researchers, particularly in case the rate of missing data is very small. Since its proposal, there were only around 6,490 projects applied Mean Imputation to solve problems in health care.

Figure 1 presents the evolution of academic publications concerning missing data imputation algorithms. Publication statistics were acquired from Google Scholar; the search query is defined as the subfield name of algorithms and at least one of "medical" or "health" appeared, for example, "kNN Imputation" AND "medical" OR "health".

B. Missing Data Patterns

Little & Rubin [8] classified missing data into three types:

- Missing completely at random (MCAR) when the missing values are randomly distributed across all observations,
- Missing at random (MAR) when the missing values are not randomly distributed across observations but are distributed within one or more sub-samples,
- Missing not at random (MNAR) when the missing values are neither randomly distributed across observations nor distributed within one or more sub-samples; the value of the missing variable is related to the reason it is missing.

We conducted the comparison under MCAR assumption. The missing ratio for the dataset is defined as follows:

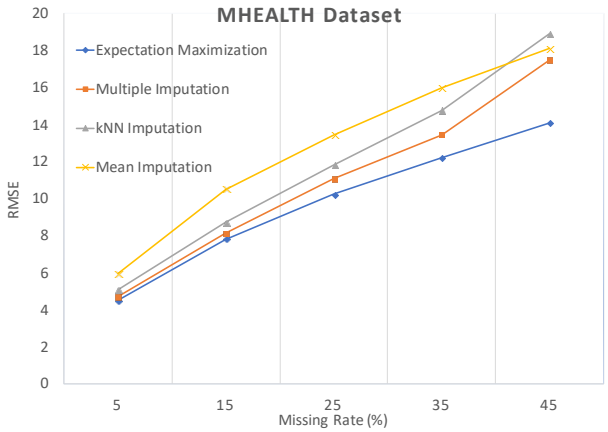

Fig. 3. The evolution of RMSE with a wide range of missing rates in MHEALTH Dataset.

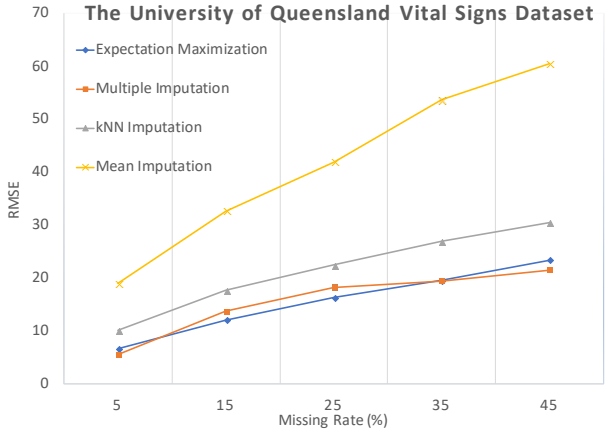

Fig. 4. The evolution of RMSE with a wide range of missing rates in The University of Queensland Vital Signs Dataset.

$$p = \frac{\text{the number of the missing values}}{\text{the number of the total values}} \quad (1)$$

C. Datasets

We analyzed two well-established datasets called MHEALTH [9] and the University of Queensland Vital Signs [10].

MHEALTH dataset is a well-established dataset which consists of 161280 lines of data. These data represent body motion and vital signs records of ten volunteers of diverse profile performing 12 physical activities in total 10 minutes. The sensor positioned on the chest provides 2-lead ECG measurements. The collected information can be potentially used for basic heart monitoring, checking for various arrhythmias or looking for the effects of exercise on the ECG.

The University of Queensland Vital Signs Dataset is a high-quality, high-resolution, and multiple-parameter monitoring vital signs dataset. The dataset represents a wide range of patient monitoring data and vital signs recorded during 32 surgical cases where patients underwent anesthesia at the Royal Adelaide Hospital for the duration ranging from 13 minutes to 5 hours (median 105 minutes), divided into 10 minutes period. The essential data are the electrocardiograph, pulse oximeter, and arterial blood pressure.

D. Evaluation Criteria

The imputing performance is evaluated via the Root Mean Squared Error (RMSE) and execution time.

Root Mean Square Error (RMSE) measures the differences between the predicted values (the imputed values) $X_i^{imputed}$ and the actually observed values (the true values) X_i^{obs}. This metric is the measure of accuracy for continuous variables. Therefore, RMSE is employed by most studies when compare two datasets. The more RMSE is, the less effective method is. The RMSE formula is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2}{n}} \quad (2)$$

Additionally, the execution time was also considered as the measure of performance of each imputation algorithm.

E. Data Analysis Procedure

The data analysis procedure of this research is conducted following the process proposed by Peter et al. [11] illustrates in Figure 2. Since the original datasets do not contain any missing values, a range from 5% to 45% of missing values was artificially created under MCAR assumption. The simulated missing values were imputed by employing the most influential missing data imputation algorithms. The performances including RMSE and execution time (expressed in seconds) were measured. In order to achieve accurate results, each dataset and each percentage of missing value was performed 1000 simulations. The final result was obtained by averaging over 1000 simulations.

III. Results

After introducing a wide range of missing rates, MHEALTH and the University of Queensland Vital Signs datasets were used with the four imputation algorithms respectively. When the missing data rate is around 5%, there is not a big difference between RMSE curves and execution time among the algorithms.

A. RMSE Analysis

The average performance of each algorithm at each missing rate after 1000 simulations is illustrated in Figures 3 and 4. As expected, the RMSE and execution time curves increased with the increasing of missing rates in all datasets.

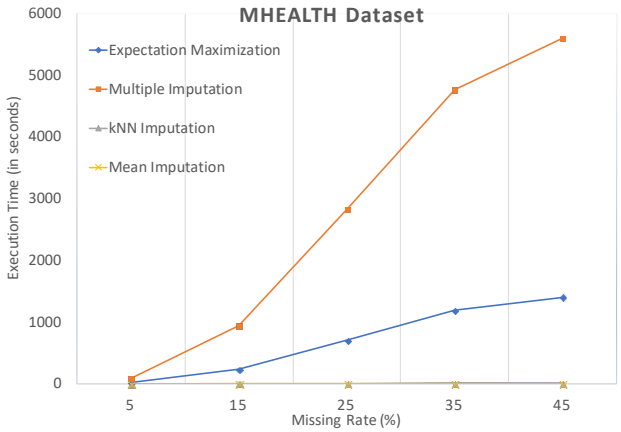

Fig. 5. The evolution of execution time (in seconds) with a wide range of missing rates in MHEALTH Dataset.

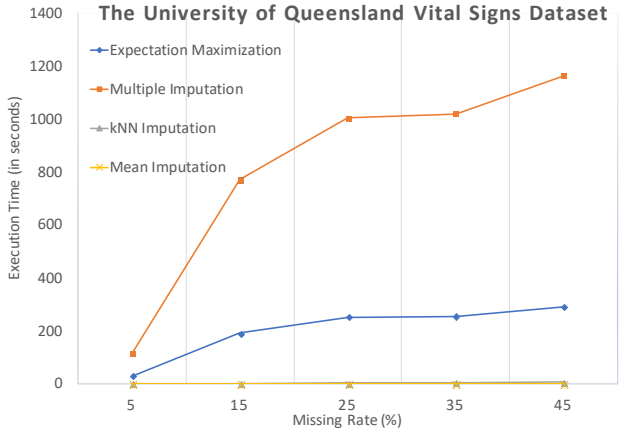

Fig. 6. The evolution of execution time (in seconds) with a wide range of missing rates in The University of Queensland Vital Signs Dataset

According to RMSE, Mean Imputation appeared as the least efficient algorithm. The performances of EM and MI were not consistent between the datasets. In fact, EM performs well with the MHEALTH Dataset while MI achieves better performance with the University of Queensland Vital Signs dataset. However, the RMSE distances between EM and MI in the University of Queensland Vital Signs dataset are not large. kNNI consistently falls between the best and the worst algorithms. Surprisingly, when the missing rate reached 45% in MHEALTH Dataset, the RMSE of kNNI was higher than Mean Imputation.

B. Execution Time Analysis

Figures 5 and 6 show the execution time for each algorithm. Both Mean Imputation and kNNI were all particularly fast with less than 10 seconds duration following the missing data rate. EM was slower, however, the execution time is still reasonable with less than 25 minutes. The execution time of MI was related to the missing data rates, fast on a small missing rate (5%), it reaches 1 hour on the University of Queensland Vital Signs datasets at the highest rate of missing values (45%).

C. Disscussion

Handling missing data is a part of research in the healthcare area. Although there are various alternative techniques to deal with the drawbacks of missing data, there is a need for neutral and well-designed comparison studies in computational sciences. In addition, while attention has been paid to the comparison missing data imputation algorithms for several kinds of data, only few studies have applied real healthcare datasets in the experiments.

In this study, we carried out a neutral comparison of four influential imputation algorithms based on two real healthcare datasets under MCAR assumption. For the validation of the imputation results, RMSE and execution time were analyzed as evaluation criteria.

Table 1 presents the results based on RMSE and execution time. The scores from 1 to 3 indicate the performance, 1 means weak to 3 means excellent. Accordingly, EM is the method of interest with the highest score.

The Mean Imputation algorithm does not make use of the underlying correlation structure of the data. Therefore, it is not unusual that this algorithm performed poorly in the experiment. kNNI, which utilizes the observed data structure, represented an actual improvement of Mean. However, the RMSE curves of kNNI are not much higher than Mean Imputation in this experiment.

Figure 1 shows the recent interest of researchers on MI and EM algorithms for healthcare data. The number of research applied MI and EM in healthcare are much larger than kNNI and Mean Imputation for seven years.

MI is based on a much more complicated algorithm. Reasonably, MI is the efficient method of missing data imputation. The imputed values are drawn m times from a distribution rather than just once. Therefore, it is also the most time-intensive comparing with other algorithms represented in this research.

TABLE I
THE RESULTS BASED ON RMSE AND EXECUTION TIME

Algorithm	RMSE	Execution Time	Total
EM	***3***	***2***	***5***
MI	3	1	4
KNNI	1	3	4
Mean Imputation	1	3	4

In the experiment, EM appeared to be the most robust imputation algorithm for healthcare data. EM is an interactive procedure in which it uses other variables to impute a value (Expectation), then checks whether that is the value most likely (Maximization). EM re-imputes a more likely value until reaching the most likely value. There are just two steps in EM algorithm, Expectation step (E-step) and Maximization step (M-step). Therefore, the execution time of EM is faster than MI.

Besides, EM preserves the relationship with other variables. Hence, the RMSE curves of EM are lower than KNNI and Mean Imputation. The well RMSE performance of EM under MCAR assumption was also supported by the research of Graham et al. [12].

IV. Conclusion

This research carried out a neutral comparison of four influential missing data imputation algorithms Regularized Expectation-Maximization (EM), Multiple Imputation (MI), kNN Imputation (kNNI) and Mean Imputation based on two well-established healthcare datasets under MCAR assumption. Root Mean Squared Error (RMSE) and execution time were used as best performing evaluation criteria. Experimental results suggest that EM is the best missing data imputation algorithm, as it has both a good RMSE performance and a low execution time.

There are several directions for future research. The appropriateness of a missing data imputation algorithm is contextual and depends on the missing data assumption. The MCAR assumption had been applied in this research. Hence, the MAR and NMAR assumption should be carefully considered in future research. In addition, there is no universal imputation algorithm performs best in every situation. Therefore, further study should implement healthcare datasets with various data types and evaluation criteria.

References

[1] Acuna, Edgar, and Caroline Rodriguez. "The treatment of missing values and its effect on classifier accuracy. Classification, clustering, and data mining applications (2004): 639-647

[2] Garcia, Salvador, Julian Luengo, and Francisco Herrera. "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining." Knowledge-Based Systems 98 (2016): 1-29.

[3] Allison, Paul D. "Missing data". Vol. 136. Sage publications, 2001.

[4] Cheema, Jehanzeb R. "A review of missing data handling methods in education research." Review of Educational Research 84.4 (2014): 487-508.

[5] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological) (1977): 1-38.

[6] Rubin, Donald B., "Multiple imputation for nonresponse in surveys". Vol. 81. John Wiley & Sons, 2004

[7] Donders, A. Rogier T., et al. "A gentle introduction to imputation of missing values." Journal of clinical epidemiology 59.10 (2006): 1087-1091.

[8] Little, Roderick JA, and Donald B. Rubin. Statistical analysis with missing data. Vol. 333. John Wiley & Sons, 2014.

[9] Banos, Oresti, et al. "mHealthDroid: a novel framework for agile development of mobile health applications." International Workshop on Ambient Assisted Living. Springer, Cham, 2014.

[10] Liu, David, Matthias Gorges, and Simon A. Jenkins. "University of Queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research." Anesthesia & Analgesia 114.3 (2012): 584-589.

[11] Schmitt, Peter, Jonas Mandel, and Mickael Guedj. "A comparison of six methods for missing data imputation." Journal of Biometrics & Biostatistics 6.1 (2015): 1.

[12] Graham, John W., Scott M. Hofer, and David P. MacKinnon. "Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures." Multivariate Behavioral Research 31.2 (1996): 197-218.