Comparing the effectiveness of six missing value imputation methods for longitudinal classification datasets

Caio Ribeiro, Alex A. Freitas
University of Kent, Canterbury, United Kingdom
cer28@kent.ac.uk, A.A.Freitas@kent.ac.uk

Abstract

Missing values are common in longitudinal datasets because often data is not collected at all waves (time points), and choosing the best method to estimate these values is a challenging task. This work compares the effectiveness of six imputation methods in a classification dataset created from the English Longitudinal Study of Ageing (ELSA), where the classes are age-related diseases and the predictive features are mainly biomedical data. Three of the imputation methods are based on different approaches for computing the mean or mode of a feature (for numeric or nominal features, respectively), two methods use longitudinal (temporal) information, and one method uses the 7-NN (7-Nearest Neighbours) machine learning algorithm. The methods' effectiveness were measured using two criteria: (a) how often each imputation method can be applied (the two longitudinal methods often cannot be applied to some features); (b) how accurately they estimate the missing values. Overall, 7-NN was the most accurate method, but the two longitudinal methods also performed well when they could be applied.

1 Introduction

In machine learning, missing values of variables (features) is a common problem in real-world datasets. Traditionally, missing values can either be removed from the dataset, ignored (i.e., the machine learning algorithm has to handle them during its execution), or replaced by an estimated value. The latter can be done by several methods, usually based on information from the known values in the dataset, to approximate the estimation from the actual value as much as possible.

Longitudinal studies follow a set of subjects over time, and take repeated measures of variables for each subject at different time points. Longitudinal datasets, derived from these studies, are prone to high amounts of missing data, mainly due to attrition (for example, subjects dropping out) [Engels and Diehr, 2003]. Actually, in the longitudinal dataset used in this work, the overall proportion of missing values is 38.5%. This is a strong motivation to investigate the effectiveness of several missing value imputation methods.

We focus on longitudinal datasets to be used as input to classification algorithms. Such datasets are composed of instances (the subjects to be classified) and features, which are variables describing each subject, usually with repeated measures for each time point (called wave) in the dataset.

Classification algorithms aim to predict the value of a nominal class variable for an instance, based on the values of its features. These algorithms use training data (a set of instances with known class values) to create a model for predicting the class of previously unseen instances (test data).

There are many ways to estimate missing values (some particular to longitudinal datasets), and selecting the best imputation method is challenging, since no missing value imputation method is the best choice for all types of features and datasets [Diggle *et al.*, 2002], [Hu *et al.*, 2017], [Mallinckrodt, 2013]. The performance of a method depends on several factors, such as: a) the data distribution [Santos *et al.*, 2017]; b) how the missing data appears in the dataset (missing completely at random, missing at random, or missing not at random) [Diggle *et al.*, 2002], [Mallinckrodt, 2013]; c) the proportion of instances with missing values; d) the availability of information that can be used to make better imputation.

In this study, we performed experiments on a real-world longitudinal dataset created from the English Longitudinal Study of Ageing (ELSA) [Banks *et al.*, 2018], to compare the effectiveness of six missing value imputation methods. The ELSA study interviews its core participants (who are at least 50 years old) repeatedly, over the years prior to their retirement and beyond.

As related work comparing different imputation methods, some studies did not use imputation methods devised specifically for longitudinal data [Belger *et al.*, 2016], [Hu *et al.*, 2017], while others employ the well-known Last Observation Carried Forward method [Gad and Abdelkhalek, 2017], [Minhas *et al.*, 2015], [Zhu, 2014], or several longitudinal methods [Engels and Diehr, 2003]. For our experiments, we selected methods from classical statistics, methods made for longitudinal data, and a machine learning algorithm.

There are two main approaches to evaluate the performance of missing value imputation methods. The first is to evaluate how missing value imputation changes the accuracy of a classification/regression model [Hu *et al.*, 2017], [Minhas *et al.*, 2015]. The results of such evaluations are dependent on the algorithm used to create the model. The second approach is to replace known values in a synthetic [Gad and Abdelkhalek, 2017], [Zhu, 2014] or real [Belger *et al.*, 2016], [Engels and Diehr, 2003] dataset, and compare the estimated values to the ground-truth, calculating the error rate. For our study, we focused on the second approach, as the results from this classifier-independent approach are easier to generalise.

One important characteristic that sets our study apart from the related works is the number of features in our datasets. The cited studies performed experiments using datasets with under 10 [Engels and Diehr, 2003], [Gad and Abdelkhalek, 2017], [Zhu, 2014], or between 10 and 20 [Hu *et al.*, 2017], [Minhas *et al.*, 2015] longitudinal features (in [Belger *et al.*, 2016] the number of features was not specified). The ELSA dataset used in this work has 45 longitudinal features, with their repeated measures totalling 140 features.

Among the cited works, [Engels and Diehr, 2003] has the most similar evaluation approach. However, in [Engels and Diehr, 2003] the imputation methods were evaluated on just 4 longitudinal features, whereas our dataset has 45 longitudinal features, representing a wider diversity of feature types and distributions. In addition, our work includes continuous, discrete, and unordered nominal features. The latter feature type requires different treatment, and is not present in any of the cited related works.

This article is organised as follows. Section 2 describes the missing value imputation methods used in our experiments. Section 3 presents our dataset creation process. Section 4 has our experimental results and discussion. Section 5 presents our conclusions and future work suggestions.

2 The Missing Value Imputation Methods

Our experiments use six missing value imputation methods [Gad and Abdelkhalek, 2017], [Mallinckrodt, 2013], [Albridge *et al.*, 1988], described below. In the following, $F_{i,t}$ denotes the value of feature F_i at wave t, and I denotes the instance where the missing value is being imputed.

- Global mean/mode: Replace the missing values in feature $F_{i,t}$ by the mean or mode (for numeric or nominal features, respectively) of $F_{i,t}$ over all instances with known values for it in the dataset.
- Class-based mean/mode: Similar to Global mean/mode, but the values used are those in the dataset with the same class value as instance I.
- Age-based mean/mode: Similarly to Global mean/ mode, but the mean/mode is computed over the instances with the same age value as instance I. This is analogous to the class-based imputation method, as age is a key feature to divide the dataset into sections (since in our ELSA dataset all class variables are *age-related* diseases). A similar approach was used by [Zhao *et al.*, 2018], which replaced missing values with the median from individuals with the same age and sex.
- Prev: If a value of feature $F_{i,t}$ is unknown for instance I, input the value of $F_{i,t}$ for I in the previous wave, i.e., input $F_{i,t-1}$. On longitudinal datasets, the Last Observation Carried Forward approach is typically applied, but in our ELSA dataset there is a gap of 4 years between consecutive waves, so we decided to consider only values from the previous wave as viable for imputation.
- PrevNext: If a value of feature $F_{i,t}$ is unknown, and both the values of $F_{i,t+1}$ and $F_{i,t-1}$ are known, replace the missing value by: a) for numeric features, the mean of $F_{i,t+1}$ and $F_{i,t-1}$ for instance I; b) for nominal features, only replace the missing value if both values of $F_{i,t+1}$ and $F_{i,t-1}$ are the same (repeat that value for $F_{i,t}$). Note that, as with the Prev method, only values from the nearest waves are considered viable for imputation, to avoid imputing values too far into the future or too long past.
- K-Nearest Neighbours: Replace the missing value of a feature $F_{i,t}$ by the prediction of a K-Nearest Neighbours algorithm with $k = 7$ (i.e., 7-NN). We evaluated different k values (1,3,5,7,9) in preliminary experiments, and observed little difference in the average error values. In this work we report results for $k = 7$, which produced the best results overall. We leave the possibility of optimising the value of k in more detailed experiments for future work. To avoid the problem of high dimensionality, 7-NN calculates the Euclidean distance between instances using only a feature subset, namely the values of F in other waves, and the subject's age and gender. Note that, in cases where the value of F_i was not known in all other waves for the current instance, we considered this method could not be applied. The missing value is replaced by the mean (for numeric features) or mode (for nominal features) value of $F_{i,t}$ among the 7 nearest neighbours. Note that the 7-NN can be considered a longitudinal imputation method, in the sense that its distance function takes into account the variation of a feature's values across the waves (time points).

3 Experimental Methodology

Consider a set of n missing value imputation methods $S = \{M_1, ..., M_n\}$, and a dataset with a set of d features $\{F_1, ..., F_d\}$. For each feature F_i at wave k ($F_{i,k}$) in a dataset, we create a subset of the original dataset, composed of all the instances with known values for $F_{i,k}$ (removing instances where $F_{i,k}$'s value is missing). This subset is hereafter called the known data subset for $F_{i,k}$. Then, each method from S has its average error rate measured in a 5-fold cross-validation performed in that known data subset. That is, the known data subset for the current feature $F_{i,k}$ is randomly partitioned into 5 folds of about the same size, and each imputation method is executed 5 times, each time using a different fold as a held-out "validation" subset, and the other four folds as the "estimation" subset. This process is summarised in Figure 1.

In the validation subset, the known values of $F_{i,k}$ are temporarily hidden from the imputation method being evaluated, and the method uses all instances in the estimation subset to determine the best value to be imputed for each instance in the validation subset. The estimated values are then compared with the true, known values of $F_{i,k}$ in the validation set, and an error measure is computed.

If $F_{i,k}$ is nominal, the error is 0 or 1, depending on whether or not the estimated value matched the known value, respectively. If $F_{i,k}$ is numeric, the error is the absolute value of the

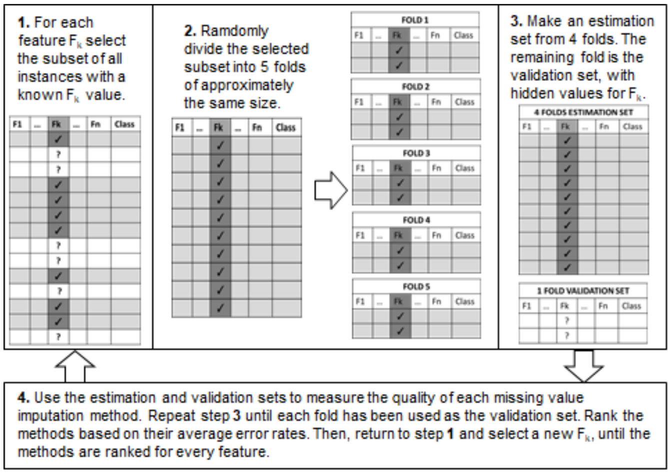
1. For each feature F_k select the subset of all instances with a known F_k value.
F1 ... Fk ... Fn Class
2. Ramdomly divide the selected subset into 5 folds of approximately the same size.
FOLD 1
FOLD 2
FOLD 3
FOLD 4
FOLD 5
3. Make an estimation set from 4 folds. The remaining fold is the validation set, with hidden values for F_k.
4 FOLDS ESTIMATION SET
1 FOLD VALIDATION SET
4. Use the estimation and validation sets to measure the quality of each missing value imputation method. Repeat step 3 until each fold has been used as the validation set. Rank the methods based on their average error rates. Then, return to step 1 and select a new F_k, until the methods are ranked for every feature.

Figure 1: Cross-validation approach to evaluate missing value imputation methods.

difference between the estimated and known values of $F_{i,k}$. The error associated with each method is the average of its errors over all instances in the validation subset, over the 5 iterations of the cross-validation procedure.

The methods are then ranked based on their average error, where the smallest-error method is assigned rank 1 and the largest-error method is assigned the last rank. If two methods have the same average error, they share a rank (e.g.: if two methods tie for the first place, both get a rank of 1.5).

4 Dataset Creation

For our experiments, we used a real-world longitudinal dataset, created from data extracted and processed from the English Longitudinal Study of Ageing (ELSA) [Banks *et al.*, 2018]. The dataset spans 4 waves (time points), where each pair of consecutive waves is separated by a 4-year period.

The dataset involves the classification task of machine learning. It has 140 features (40 unordered nominal and 100 numeric features), of which only age and gender do not have missing values. Most features refer to biomedical data, mainly results from tests performed during nurse visits to each subject in waves 2, 4, 6 and 8 of the study. Most features were measured in all those 4 waves, but many measurements have missing values due, e.g., to subjects dropping out of the study or entering the study at a later wave. In total, 38.5% of all feature values (across all 4 waves) are missing. The dataset also has the diagnoses (positive or negative value) of 10 different age-related diseases as class variables, whose values have to be predicted by a classifier.

We preprocessed the data to ensure all features had the same measurement unit and were represented by a single value at each wave (e.g., features with multiple values per wave, from repeated tests, were averaged to produce a single value per wave). As there were no definitive variables indicating whether a subject was diagnosed with each of the 10 age-related diseases at each wave, we inferred the class values from questions in the ELSA questionnaire such as: whether the subject had been recently diagnosed with the disease, and whether the patient confirms the disease diagnosis from past waves. For more details on the dataset's creation, please see [Pomsuwan and Freitas, 2017].

5 Computational Results and Discussion

For each imputation method we computed: *a)* its applicability, i.e., for how many missing values the method could be applied; *b)* its normalised average error rate, for nominal and numeric features separately; and *c)* its average error-based rank.

Regarding the applicability of each method, 7-NN could not be applied to features that didn't have repeated measurements (2 out of the 138 features with missing values). All the mean/mode-based methods (Age-based, Class-based, and Global mean/mode) can be applied to every feature, except that the Age-based method was not applicable in rare cases where there was no known feature values for subjects of the same age of the current instance's subject. The PrevNext and Prev methods could not be applied in many cases, by definition; as the Prev method requires the feature to have a known value in the previous wave, and the PrevNext method requires a known value in both the previous and the next wave in the dataset, which is even less common.

In the last column of Tables 1 and 2, we show the error rates for an "Oracle" method, which is not a valid imputation method but indicates an over-optimistic error measure, as follows. The Oracle knows the error rates of each of the 6 imputation methods on the validation set (for each cross-validation fold), for each feature, even though it does not know the true value of the feature for each instance in the validation set. Hence, to estimate missing values in the validation set, for each feature, the Oracle first uses the method with the smallest error on the validation set to estimate all missing values for which that method can be applied. Next, if there are still missing values for that feature in the validation set, the Oracle uses the second smallest-error imputation method in all applicable cases, and so on, until all missing values have been estimated. Hence, the Oracle's error rate is a kind of unrealistic lower bound for the imputation error rate.

Tables 1 and 2 show the mean error rate over the nominal features and the mean normalised absolute error over the numeric features, for each method. Table 1 considers the entire dataset, whilst Table 2 considers only the feature values where all methods could be applied. In each Table, the smallest error in each row, excluding the over-optimistic Oracle's error, is shown in bold face.

The normalisation method used was min-max, where the absolute error for each estimated feature value is divided by the difference between the minimal and maximum values observed for that feature, producing errors in the [0..1] range.

Considering all features (Table 1), 7-NN was the most accurate method for both nominal and numeric features. As the PrevNext and Prev methods could not be applied in many cases, they were assigned the maximum error value (1) in those cases, which greatly hindered their results. The difference between the error rates of the best imputation method and the Oracle is smaller for nominal features (1.1%) than for numeric features (1.8%).

Table 1: Error rates (in [0..1]) of the imputation methods, computed by 5-fold cross-validation. For nominal features each value represents the mean error rate (over 39 features) and for numeric features each value is the mean normalised absolute error (over 99 features).

Feature type (number)	**7-NN**	**Age-based**	**Class-based**	**Global mean**	**Prev**	**PrevNext**	**Oracle**
Nominal (39)	**0.083**	0.09	0.112	0.113	0.441	0.749	0.072
Numeric (99)	**0.090**	0.102	0.103	0.104	0.402	0.656	0.072
All features (138)	**0.088**	0.099	0.106	0.106	0.413	0.684	0.072
Applicability: % of cases replaced	81.79%	97.08%	100.00%	100.00%	35.57%	2.95%	100.00%

Table 2: Error rates (in [0..1]) of the imputation methods, computed by 5-fold cross-validation, considering only instances where the PrevNext and Prev methods were applicable. For nominal features each value represents the mean error rate (over 14 features), and for numeric features each value is the mean normalised absolute error (over 40 features).

Feature type (number)	**7-NN**	**Age-based**	**Class-based**	**Global mean**	**Prev**	**PrevNext**	**Oracle**
Nominal (14)	0.075	0.099	0.108	0.108	**0.067**	0.086	0.062
Numeric (40)	0.096	0.086	0.101	0.101	0.075	**0.061**	0.059
All features (54)	0.081	0.096	0.102	0.103	0.073	**0.067**	0.06

Regarding the percentage of missing values in the dataset that could be replaced by each method, 7-NN was not applicable in almost a fifth of the cases, involving instances where there were no known values of the feature being replaced in other waves. Importantly, for most missing values in our dataset, the longitudinal methods Prev and PrevNext could not be applied, either because the method was not applicable for that feature or because the previous and/or next value was unknown in the current instance.

On the other hand, when we only considered the error rates for features where all methods could be applied (Table 2), both longitudinal methods outperformed all others, with PrevNext obtaining the smallest average error rate over the 54 features (roughly 39% of the features in the dataset). Hence, when applicable, the longitudinal methods provided more accurate estimations of missing values than the other methods, including the more complex machine learning method 7-NN. The difference between the error rates of the best method and the Oracle are 0.5% and 0.2%, for nominal and numeric features respectively.

We ranked the imputation methods based on their average error rate for each feature, where a method's error rate for a feature is the average of its error over all estimated values of that feature across all waves (time points). The average ranks (over the 43 features) were as follows: 2.76 for the 7-NN, 2.69 for Age-based mean/mode, 3.22 for Class-based mean/mode, 3.58 for Prev, 3.84 for Global mean/mode, and 4.67 for PrevNext. Note that, as we assigned the maximum error value to the PrevNext and Prev when they could not be applied, their average ranks have been negatively impacted.

Then, we performed the Wilcoxon signed rank test, applying the Holm-Bonferroni Sequential Correction [Demšar, 2006], to compare the feature ranks of the method with the smallest average error rate, 7-NN, to all others, pairwise, with a significance level of $\alpha = 0.05$. In every test result, there was statistically significant evidence against the null hypothesis that the methods' performances were equivalent, meaning 7-NN's performance was indeed superior to each of the other methods. The *p-values* obtained in the tests were: $6.0\mathrm{e}{-9}$, $5.59\mathrm{e}{-9}$, $5.58\mathrm{e}{-9}$, $6.01\mathrm{e}{-9}$ and $5.58\mathrm{e}{-9}$, when 7-NN was compared against Age-based mean/mode, Class-based mean/mode, Prev, Global mean/mode and PrevNext, respectively.

6 Conclusions

We compared the applicability and estimation accuracy of six missing value imputation methods. The experiments were performed on a dataset created from the English Longitudinal Study of Ageing (ELSA), with 138 features of biomedical data, across four waves of the study.

We analysed how well each method estimated the missing values in a classifier-independent scenario. For the experiments, we hid some known feature values from the imputation methods, and used the other known values to estimate them. Then, the estimated values were compared to the ground truth (the previously hidden known values), to obtain an average error rate for each imputation method.

The results showed that each of the six imputation methods was the most accurate one for some features in the datasets, which corroborates the notion that no single imputation method is the best for every feature. However, as expected, the most sophisticated method, 7-NN, which is a machine learning algorithm, was the best-performing method for both the nominal and the numeric features of the dataset. In addition, the 7-NN can be considered a longitudinal imputation method, in the sense that it uses all the values of a feature across the different waves (time points) for estimating a missing value. The 7-NN method also had high applicability (81.8%), and the statistical analysis confirmed a significant difference between the estimation accuracy of 7-NN and the other methods.

The two methods devised specifically for longitudinal data (Prev and PrevNext) had very low applicability (35.57% and 2.95%, respectively). However, they had the smallest average error rates over the cases where they could be applied. That shows the value of considering the often ignored temporal aspect of the data when handling missing data in a longitudinal dataset.

Future work could involve experiments with other imputation methods and other longitudinal datasets of ageing.

References

[Albridge *et al.*, 1988] Kim M Albridge, Jim Standish, and James F Fries. Hierarchical time-oriented approaches to missing data inference. *Computers and Biomedical Research*, 21(4):349–366, 1988.

[Banks *et al.*, 2018] J. Banks, M. Blake, S. Clemens, M. Marmot, J. Nazroo, Z. Oldfield, A. Oskala, A. Phelps, N. Rogers, and A Steptoe. English longitudinal study of ageing: Waves 0-8, 1998-2017. [data collection]. *English Longitudinal Study of Ageing: waves 0–8 - 29th Ed.*, 29, 2018.

[Belger *et al.*, 2016] M. Belger, J. Haro, C. Reed, M. Happich, K. Kahle-Wrobleski, J Argimon, G. Bruno, R. Dodel, R. Jones, B. Vellas, et al. How to deal with missing longitudinal data in cost of illness analysis in alzheimer's disease—suggestions from the geras observational study. *BMC Medical Research Methodology*, 16(1):83, 2016.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[Diggle *et al.*, 2002] Peter Diggle, Peter J Diggle, Patrick Heagerty, Patrick J Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002.

[Engels and Diehr, 2003] Jean Mundahl Engels and Paula Diehr. Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, 56(10):968–976, 2003.

[Gad and Abdelkhalek, 2017] Ahmed Mahmoud Gad and Rania Hassan Mohamed Abdelkhalek. Imputation methods for longitudinal data: A comparative study. *International Journal of Statistical Distributions and Applications*, 3(4):72, 2017.

[Hu *et al.*, 2017] Zhen Hu, Genevieve B Melton, Elliot G Arsoniadis, Yan Wang, Mary R Kwaan, and Gyorgy J Simon. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics*, 68:112–120, 2017.

[Mallinckrodt, 2013] Craig H Mallinckrodt. *Preventing and treating missing data in longitudinal clinical trials: a practical guide*. Cambridge University Press, 2013.

[Minhas *et al.*, 2015] Sidra Minhas, Aasia Khanum, Farhan Riaz, Atif Alvi, Shoab A Khan, Alzheimer's Disease Neuroimaging Initiative, et al. Early alzheimer's disease prediction in machine learning setup: Empirical analysis with missing value computation. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 424–432. Springer, 2015.

[Pomsuwan and Freitas, 2017] Tossapol Pomsuwan and Alex A Freitas. Feature selection for the classification of longitudinal human ageing data. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 739–746. IEEE, 2017.

[Santos *et al.*, 2017] Miriam Seoane Santos, Jastin Pompeu Soares, Pedro Henriques Abreu, Hélder Araújo, and João Santos. Influence of data distribution in missing data imputation. In Annette ten Teije, Christian Popow, John H. Holmes, and Lucia Sacchi, editors, *Artificial Intelligence in Medicine*, pages 285–294, Cham, 2017. Springer International Publishing.

[Zhao *et al.*, 2018] Juan Zhao, QiPing Feng, Patrick Wu, Roxana Lupu, Russel A Wilke, Quinn S Wells, Joshua Denny, and Wei-Qi Wei. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *bioRxiv*, 2018.

[Zhu, 2014] Xiaoping Zhu. Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics*, 4(11):933, 2014.