PhUSE US Connect 2019

Paper ML04

Hit and Miss: An evaluation of imputation techniques from Machine Learning

Karma Tarap, Janssen Research & Development, Basel, Switzerland
Antoine Stos, Actelion Pharmaceuticals Ltd, Basel, Switzerland

ABSTRACT

Missing data in clinical trials can have serious consequences on statistical findings. This is of concern in Real World Data due to the high degree of missingness typically observed.

In this paper, we provide a broad survey of multivariate imputation techniques from Machine Learning and an empirical imputation testing strategy to compare against the current state of the art in Clinical Imputation. The best performing model on our data is then used to partially impute baseline characteristics for an ongoing clinical submission.

We also propose a novel imputation technique, missXGB based on the XGBoost algorithm, that scales better than existing imputation methods.

INTRODUCTION

The volume of Real-World Data (RWD) available continues to grow exponentially as technology and integrated electronic medical records have made this information increasingly accessible and useful for outcomes research and regulatory submissions. While randomized clinical trials remain the gold standard for the assessment of efficacy, there is increasing interest for converting RWD into Real-World Evidence (RWE) that can be used for informed healthcare decision making (IPSOR 2018).

A continued challenge faced with using Real World Data to augment Clinical Trials is in dealing with the prevalence of missing data. Single imputation methods such as Last Observation Carried Forward (LOCF), Worst Observation Carried Forward (WOCF) and mean imputations are known to under-represent the variance of variables as well as potentially change the correlation structure of the data (Little and Rubin, 2014).

The main objective of this paper is to assess and compare different multivariate imputation methods from Machine Learning literature to help solve the missing data problem. A secondary objective is to propose an empirical imputation pipeline that can be used for future testing purposes. Finally, we will introduce a new scalable discriminative method **(missXGB**) that we will compare against the winning model.

This research is done in the context of a real clinical submission where imputation of baseline covariates was necessary for the pooling of multiple clinical trials and registry studies.

Background

Missing data is a pervasive problem in long term observational studies. This problem is further compounded given there can be many differing reasons why variables are not populated: It could be that the certain assessments were not taken either by accident or on purpose; the patient did not attend; or the record was lost.

In simple cases, where only a small portion of values are missing, simple imputation methods may be sufficient. For example, complete-case analysis which omits all observations with missing values to conduct statistical analysis can be used if the data is missing completely at random without biasing results (Little and Rubin, 1987).

Consequently, the impact of missing data depends on the underlying missing data generating mechanism. Missing data is typically categorized as one of the following: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) as summarized in Table 1.

Missing completely at random means that the data is missing with no relation to either observed or unobserved values and data is equally likely to be missing.

Missing not at random is when the missing pattern depends on the observed values. This type of missingness is what we can model as if we can impute the missing value based on other non-missing parameters. An example of this case could be that certain laboratory assays of special interest are only performed when extreme lab tests are observed.

A final, stricter category of missing pattern is when the values are missing not at random (MNAR). This is when we cannot explain the missing mechanism with the data collected, but the underlying reason is due to some extraneous factor. Perhaps the patient did not come to clinic due to religious holidays. In practice, real-world data is likely a mixture of these different mechanisms.

Mechanism of Missing Data	**Assumption**
Missing Completely at Random (MCAR)	$f(\mathcal{M}\|\mathbf{X}^{obs},\mathbf{X}^{miss}) = f(\mathcal{M})$
Missing at Random (MAR)	$f(\mathcal{M}\|\mathbf{X}^{obs},\mathbf{X}^{miss}) = f(\mathcal{M}\|\mathbf{X}^{obs})$
Not Missing at Random (NMAR)	$f(\mathcal{M}\|\mathbf{X}^{obs},\mathbf{X}^{miss})$ is a function of $\mathbf{X}^{miss}$

Table 1: Statistical assumptions of mechanisms used to generate patterns of missing data M for data set X (Bertsimas 2018).

Modern machine learning imputation methods can be categorized as either discriminative or generative (Yoon et al. 2018). Generative classifiers learn a model of the joint probability, p(x, y), of the inputs x and the label y, and make their predictions by using Bayes rules to calculate p(y l x), and then picking the most likely label y. Discriminative classifiers model the posterior p(y l x) directly, or learn a direct map from inputs x to the class labels (Ng and Jordan 2002).

Imputation methods can further be classified as those that provide a single imputed data set, and those that provide multiple datasets (Multiple Imputation). Under multiple imputation, statistical analyses are performed on each imputed data set and the results are pooled (Little and Rubin, 1987). The underlying idea is that one can capture the variability in the predictions coming from the imputation. Typically, classical Machine Learning imputation methods have focused on single imputation (missForest, CART, KNN), however modern deep-learning methods are increasingly supporting multiple imputation (DAE, GAIN).

In this paper, we will consider both single and multiple imputation strategies as well as methods that are discriminative or generative.

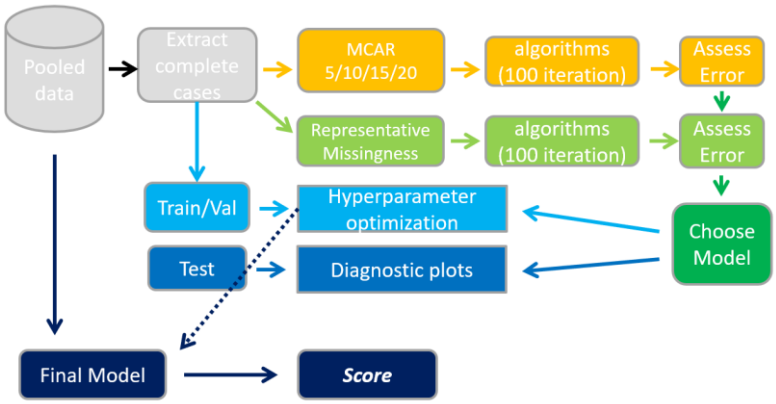

Figure 1 : Imputation testing pipeline consisting of 3 stages: Testing on MCAR, Representative Missing and Final model fitting.

EXPERIMENTAL SETUP

Following from the “No Free Lunch Theorem” which states that “any two optimization algorithms are equivalent when their performance is averaged across all possible problems", in Machine Learning its common practice to assess multiple different models on data as some models will outperform other models, but no model is universally guaranteed to be the best method.

In the context of imputation methods, this means that we cannot can categorically suggest a “best” imputation method for all datasets. Rather, it is the belief of the authors that we should aim to perform imputation using multiple methods on each dataset to identify the best method suited to our data [Figure 1]. Consequently, the methods below were selected to ensure a good coverage of different approaches available from both classical machine learning and deep learning.

Algorithms

1. MICE Predictive-Mean Matching (pmm): An iterative method which imputes missing values from known values in a given dimension using linear regressions. It is commonly used for multiple imputation and can be generalized to multiple missing dimensions using the chained equations process (Buuren and Groothuis-Oudshoorn, 2011). Implemented using the MICE package in R.

2. Classification and Regression Trees as implemented in MICE. Based on Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), Classification and regression trees, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

3. MissForest: Uses random forest trained on the observed values of a data matrix to predict the missing values. It can be used to impute continuous and/or categorical data including complex interactions and non-linear relations. It yields an out-of-bag(OOB) imputation error estimate without the need of a test set or elaborate cross-validation.

4. missXGB: Created using a strategy similar to missForest. The main difference was the use of Gradient Boosted Trees from the XGBoost R package instead of Random Forests. Algorithm is defined in Figure 1.

Algorithm 1 Impute missing values with missXGB

Require: X an $n \times p$ matrix, stopping criterion γ

Impute missing to respective column means ;
$k \leftarrow$ vector of sorted indices of column X
w.r.t increasing amount of missing values ;
while not γ **do**
    $X_{old} \leftarrow$ store previously imputed matrix ;
    for s in k **do**
        Fit XGBoost: $y_{obs} \sim x_{obs}$;
        Predict y_{miss} using x_{miss} ;
        $X \leftarrow$ update imputed matrix from y_{miss} ;
    end for
    update γ ;
end while
return Matrix X ;

Figure 2: Pseudocode describing the missXGB algorithm. missXGB is a modified version of the missForest algorithm where initial missing values are either imputed to the column mean or the median depending on the prevalence of outliers. Instead of RandomForest, the XGBoost algorithm is used to gain better scalability and interpretability.

5. Bayesian PCA (bpca): A missing data estimation method based on Bayesian principal component analysis (Oba et al., 2003). BPCA combines an EM approach for PCA with a Bayesian model. In standard PCA data far from the training set but close to the principal subspace may have the same reconstruction error. BPCA defines a

likelihood function such that the likelihood for data far from the training set is much lower, even if they are close to the principal subspace. Implemented using the pcaMethods package in R.

6. MIDA (Denoising AutoEncoders (DAE)): Implementation of MIDA (Multiple Imputation Denoising Autoencoder) as proposed by Gondara & Wang (2017).

Methods

Starting with the pooled clinical data or two Randomized Clinical Trials and two registry trials, we extracted records with complete cases of the 12 baseline characteristics of interest which consisted of ~30% of the original records. Kolmogorov-Smirnov tests were performed to ensure that our data subset was representative of the original pooled dataset.

We initially tested our models under a Missing Completely at Random (MCAR) assumption by randomly inducing missingness into the data assuming each value is equally likely to be missing. This was done to both test the effect of missingness across the models and as potential validation of the models fitted on representative missing values].

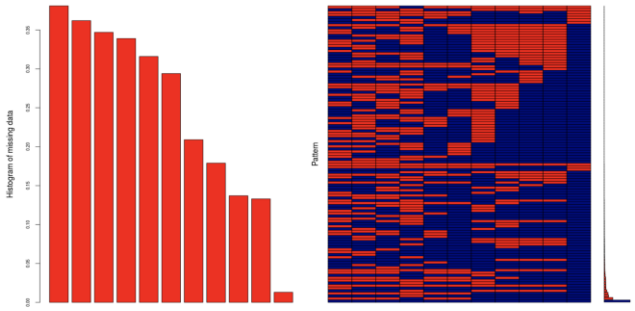

Figure 3: Pattern and frequencies of missing baseline covariates. This information is used to recreate missingness patterns for representative-missingness testing.

As we knew the ground truth of the data, we were able to assess imputation performance by comparing the Normalized Root Mean Squared Error (NRMSE) of the models.

$$\mathrm{NRMSE} = \sqrt{\frac{\mathrm{mean}\left((\mathbf{X}^{\mathrm{true}} - \mathbf{X}^{\mathrm{imp}})^2\right)}{\mathrm{var}\left(\mathbf{X}^{\mathrm{true}}\right)}},$$

Figure 4: NRMSE is a single error metric that quantifies the distance between the True value and imputed values across a data set.

As our data was unlikely to be MCAR we created missing values in our data from the missing value pattern of the pooled data. This was achieved by bootstrap sampling the missing pattern [Figure 3] of the pooled data and applying it to our complete data for 100 iterations. Testing was again performed by assessing the NRMSE of each method. Under this setup, we did not need to assume any special missingness mechanism for each variable as we were sampling the mechanism from the actual data.

Hyperparameter optimization (tuning of the model) was performed on missForest via Bayesian Optimization using the rBayesianOptimization package prior to running on our original data.

Results

We found that under both the MCAR and representative-missing experiments, the best performing method was the Random Forest based missForest algorithm with a NRMSE of 0.27 under the representative-missing experiment [Figure 5]. missForest also provided the lowest NRMSE for each of the 11 covariates.

From diagnostic plots performed on the imputed MCAR experiment [Figure 6] we observe that our model often manages a prediction very close to the truth - indicated by points falling near the diagonal line. We also observe that when there is insufficient information to perform a good prediction, imputed values tend to stay near the variable mean - the initial value assigned by the algorithm. This pattern is true of other methods that require an initial mean imputation: missXGB, MICE, bPCA and DAE.

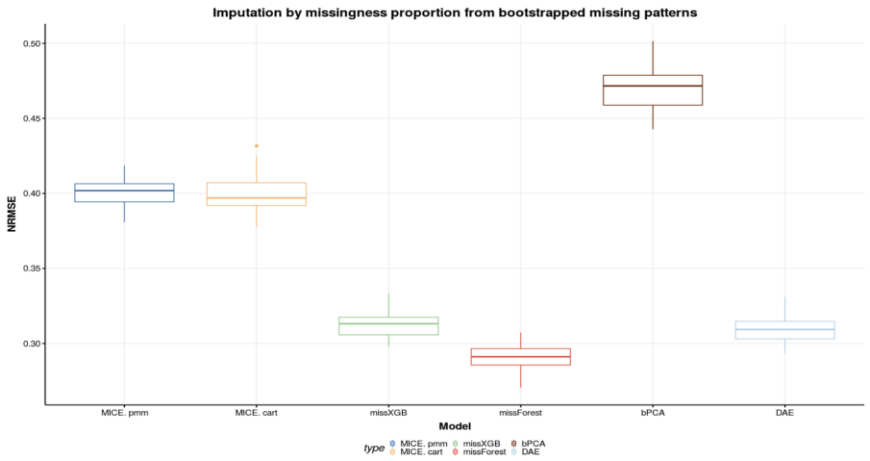

Figure 5: missForest has the lowest NRMSE on our complete-case data with induced representative missing.

Across the board, we observed that methods able to model non-linearities outperformed those that could not (e.g. MIDA vs bPCA respectively).

A further challenge when performing imputation on multiple covariates is in assessing the accuracy of the imputation when we do not have access to the ground truth. This is where the RandomForest algorithms Out-Of-Bag error (a built-in cross-validation) can be useful as a proxy error metric. We observed on our final run that the OOB score was similar to the test NRMSE, indicating that we are likely not over or under-fitting with our choice of hyperparameters.

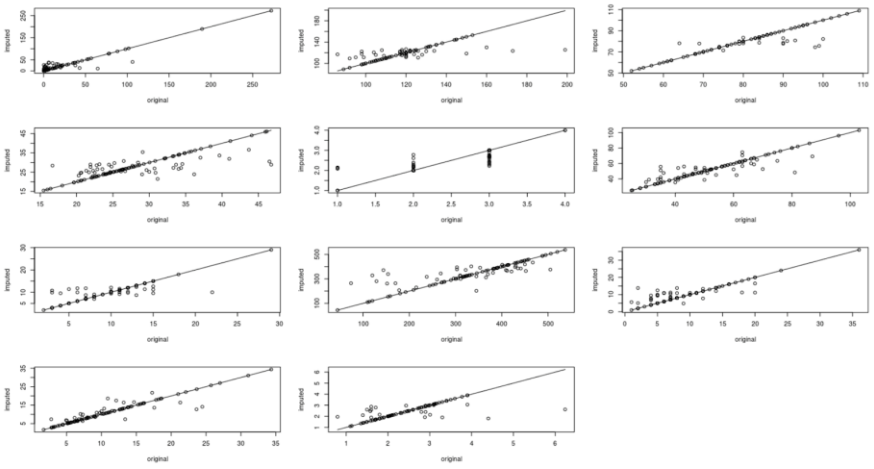

Figure 6: Predicted vs Truth for Baseline covariates with induced MCAR at 30%. Perfect predictions fall on the diagonal line.

Finally, our new method (missXGB) is comparable in performance to state of the art methods (test NRMSE ~ 0.32 on our data) but has a significant speed advantage as demonstrated in Figure 7. As it is based on the XGBoost algorithm, it has the advantage of being more interpretable as we can extract the component decision trees.

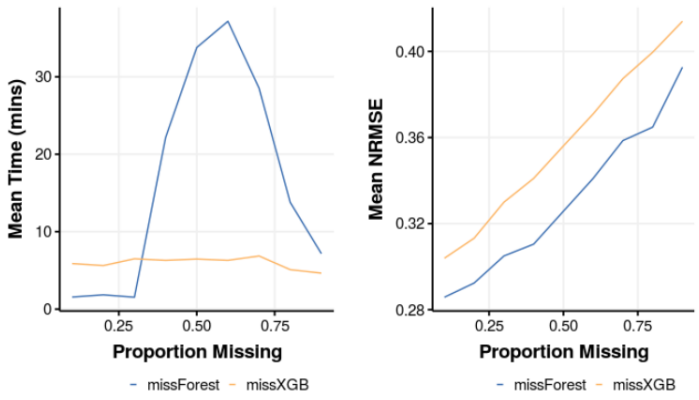

Figure 7: We observe linear scaling of the missXGB compared to missForest (run on 100 replicates of our data under MCAR and 750 trees). The potential degrees of freedom with close to 50% missing causes missForest to run for longer. We also observe that the NRMSE closely trials missForest regardless of proportion missing.

Conclusion

Revisiting our initial objectives, we have succeeded in: Outlining an imputation testing pipeline; Identifying multiple imputation techniques from machine learning to impute our Clinical Data and identified a new scalable imputation model that generalizes to other settings.

If we are to make full use of the exponentially-increasing Real World Data we can collect, we must address the issue of data sparsity. Modern multiple imputation methods (regardless of whether from the field of Statistics or Machine Learning), robust testing strategies and sensitivity analysis will all need to be deployed to ensure that we can unlock the full potential of Real World Evidence.

Keywords: Machine Learning, Deep Learning, Imputation, Real World Evidence

References

Tianqi Chen , Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, California, USA [doi>10.1145/2939672.2939785]

Leo Breiman. Random Forests, Machine Learning, v.45 n.1, p.5-32, October 1 2001 [doi>10.1023/A:1010933404324].

Bishop CM. Pattern recognition and machine learning, information science and statistics. Springer Science+Business Media, LLC; 2007.

Dimitris Bertsimas, Colin Pawlowski, Ying Daisy Zhuo. From Predictive Methods to Missing Data Imputation: An Optimization Approach. Journal of Machine Learning Research 18 (2018) 1-39.

Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. Journal of statistical software, 45(3), 2011.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. Classification and regression trees. CRC press, 1984.

Daniel J Stekhoven and Peter B¨uhlmann. Missforest: non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

Gondara, L. and Wang, K. Multiple imputation using deep denoising autoencoders. arXiv preprint arXiv:1705.02737, 2017.

Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19(16):2088–2096, 2003.

Yachen Yan (2016). rBayesianOptimization: Bayesian Optimization of Hyperparameters. R package version 1.1.0.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14 (NIPS'01), pages 841–848, 2002.

Roderick JA Little and Donald B Rubin. Statistical analysis with missing data. John Wiley & Sons, 2014.

Jinsung Yoon, James Jordon, Mihaela van der Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets. arXiv:1806.02920 [cs.LG]

T O'Neill, R & Temple, R. (2012). The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. Clinical pharmacology and therapeutics. 91. 550-4. 10.1038/clpt.2011.340.

Acknowledgements

We thank our colleagues Brian Hennessy and Matthieu Villeneuve for their continued support and guidance, without which, this paper would not be possible.

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Karma Tarap
Janssen Research & Development
Gewerbestrasse 16, 4123 Allschwil, Switzerland
Email: ktarap@its.jnj.com

Antoine Stos
Actelion Pharmaceuticals Ltd
Gewerbestrasse 16, 4123 Allschwil, Switzerland
Email: astos@its.jnj.com