Impact of Data Imputation Methods in Data Analytics for Healthcare Data

Anu Maria Sebastian
David Peter

Cochin University of Science and Technology, Cochin, India

Abstract. The healthcare industry has a lot of data which could be used effectively to predict or classify diseases with the help of data mining and machine learning techniques. However, the missing data is a very common occurrence in healthcare and can have grave impacts on the conclusions that can be drawn from the data. Developing a generalized imputation strategy that can be used across a variety of datasets is difficult as each dataset has its own attributes, characteristics, and intrinsic structures. The objective of this paper is to classify the popular data imputation methods for healthcare data and analyze and compare their performance.

Key words: data imputation, machine learning, statistical methods, data analytics, data preprocessing, healthcare.

Introduction

In the preprocessing stage of data analytics, many times the data with missing attribute values are either omitted or assigned a NULL value. It is likely that this transformation makes the data less realistic. Data imputation (DI) is the process of replacing missing data with substituted values. The pattern of missingness can be monotonous or arbitrary.

The missing data can be of the following types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). If there is a pattern in the missing data and the rest of the data cannot explain it, then it is MNAR. But if the data can explain the missed data pattern then it is MAR and if there is no pattern to the missing of data, it is MCAR.

The study by Henry et al. (2013: 115-126) observed that reweighted estimating equations produce the least biased and the missing indicator variables produce the most biased outcomes. The complete case analysis, replacement with observed frequencies, and multiple imputations (MI) impart moderate bias in the results. The variable selection plays an important role in data analytics as it identifies the important variables that are associated with the outcomes and also improves the prediction accuracy of the resulting models.

The variable selection methods are basically designed to work better with fully observed data, and thus missing data can be problematic. DI is the most popular method for handling the missing data because of its ease of use, and statistical methods are commonly used for the variable selection to perform the imputation. Machine learning (ML) methods are also widely used for data imputation.

Eirola et al. (2017: 195-206) demonstrated the benefits of applying suitable DI algorithms in achieving increased prediction accuracy with the same prediction models. Finding another dataset of a similar property could be a difficult task (Chowdhury et al., 2017: 13-19). The popular imputation algorithms used in healthcare include Multiple Imputation by Chained Equation (MICE), Fuzzy Unordered Rule-based Induction Algorithm (FURIA) and Amelia.

Classifying The Data Imputation Methods

The techniques for handling the missing data can be classified into the following approaches such as complete case analysis/case deletion, model-based imputation (such as maximum likelihood estimation (MLE), expectation-maximization (EM), Gaussian mixture models, etc.) direct imputation through statistical methods, and using ML methods. This paper focuses on the DI methods under statistical and ML approaches which are more scientific.

The popular statistical methods used for DI include mean imputation, hot-deck imputation, regression imputation, and interpolation and extrapolation. In mean imputation, the mean of the values of that attribute is substituted at the missing instance. Hot-deck imputation puts a random value from the sample values of the attribute and the randomness adds in some variance. Imputation through regression predicts a value for the missing instance. Therefore, it preserves the dependencies among the attributes. In stochastic regression imputation a random residual value is added as the predicted value. The interpolation and extrapolation imputation is used for longitudinal data which is done by creating a function to derive a set of data points within the range of known values. Banjar et al. (Banjar et al., 2017: 11-25) developed a linear interpolation imputation algorithm to handle the missing values in medical data for modeling predictors for patients with leukemia.

The common data imputation methods that utilize ML algorithms include k-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Self-Organizing Maps (SOMs), etc. KNN is an algorithm which is used to find a match closest out of the K neighbors in a multi-dimensional space. Both continuous and discrete imputation can be done with KNN. MLP learns the structure or association between the input and the output. Here learning is done by adjusting the weights of the connection between neurons of the network layers. A SOM neural network is a multivariate method for data analysis that is capable of modeling even with nonlinearities. Fig.1 depicts the DI workflow. As DI handles numerical and categorical data differently, analyzing the type of data is important. This is to be followed by the analysis of the type of missingness (whether it is MCAR, MAR, or MNAR). In most of the cases, every dataset will have its own intrinsic structure which is imposed by a correlation between the attributes present in it. The missing values are then handled by appropriate imputation method which is later evaluated with performance metrics like accuracy, Area Under Curve (AUC), Root Mean Square Error (RMSE), etc.

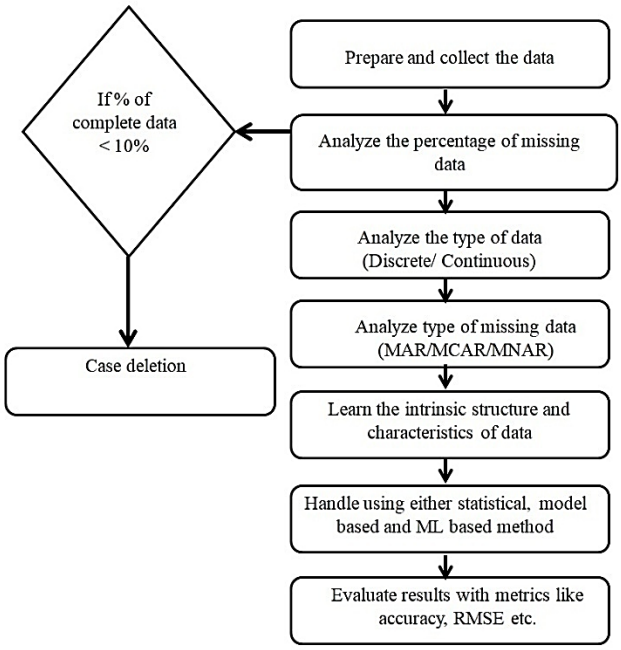

Fig. 1. The workflow of Data Imputation

Zhang et al. (2017: 57-66) conducted a survey to study the methods to handle the missing continuous data of participants in randomized controlled trials. Richter et al. (2018: 1-14) had given a review of the statistical and ML imputation methods for handling the missing values to model cancer risk. Pearson et al. (2018: 34-42) briefed the methods used for handling the missing observations in the meta-analysis of exercise therapy interventions in patients with heart failure. Wilson et al. (2018: 94–106) had given a review on how to deal with the irregular observational data for control applications using statistical methods. In the following sections, we analyze the statistical and ML-based DI methods.

Data Imputation Using Statistical Methods

A class center-based approach for classification with missing value imputation (CCMVI) was proposed by Tsai et al. (2018: 124-135). The imputation was done based on a threshold value and the Euclidian distance was computed. This method was applied to breast cancer and blood datasets from the UCI repository, which included categorical, numerical, and mixed data types. The imputation diagnostics for continuous data was done with classification accuracy and RMSE as the measures. For categorical data, the hit ratio was used as the metric. The imputation accuracy achieved by CCMVI for the categorical variables were slightly lower compared to the other methods such as mode, SVM, Feature weighted grey-KNN (FKNN), and Weighted voting Random Forests (WRF). This approach outperformed the other MVI approaches like KNN, SVM-based radial basis function kernel, weighted voting random forests, and mean and mode substitution for numerical datasets.

Perkins et al. (2017: 568-575) used the Collaborative Perinatal Project data for the imputation experiment. Both the multiple imputation (MI) and the inverse probability weighting (IPW) were experimented in this work. The missing data were grouped under the three cases of MAR, MNAR, and MCAR, and the imputation for each case was performed. This prototypical study tried to estimate the association of smoking during

pregnancy with risk of spontaneous abortion using linear logistic regression. Complete case analysis, MI, and augmented IPW resulted in similar conclusions for MCAR. For MAR, the complete case analysis had given an erroneous result which implied that smoking prevents spontaneous abortion whereas, both MI and augmented IPW had given improved results. For MNAR, the complete case analysis and IPW failed to use the information available in most of the incomplete data cases. Thus, the complete case analysis in some cases can result in spurious outcomes for MAR and so it is suggested to apply MI to handle the missing data in such cases.

Table 1. Performance improvement using statistical di methods for healthcare data

Reference	Imputation Method	Performance Metric(s)
Tsai et al., 2018: 124-135	CCMVI	Average accuracy 61.5% (for numerical data) and 78.1% (for categorical data); RMSE is minimized to 12.24
Tan, 2017: 43-49	MI	Average odds ratio 1.26
Hu, 2017: 112-120	0-value imputation	Average AUC 0.0893 and average bias - 0.0036
Gomes, 2015 515-528	MI	6% increase in positive alert.
Ayilara et al., 2019	MI with auxiliary variable	Average RMSE 1.1 (for MCAR) and 1.09 (for MAR)

Tan et al. (2017: 43-49) illustrated the benefits of DI to estimate the prevalence of dementia and mild cognitive impairment (MCI). The data from the respondents of Chinese Veteran Clinical Research were used for this study. Stratified weighting (SW), IPW, hot-deck imputation, ordinal logistic regression, and MI were experimented for this. A sensitivity test was performed by comparing MAR, MCAR, and MNAR. It was observed that imputing the missing values had a larger influence on the estimation of dementia prevalence. Discarding the missing values underestimated the prevalence of dementia. The SW method was recommended when the information available was limited and MI could give the lowest rate of misclassification and computed higher estimates of dementia and MCI prevalence. Full information maximum likelihood (FIML) had given a comparable performance as MI. Table 1 compares the performance of the different statistical methods for DI. The metrics used for the comparison include accuracy, Area Under the Curve (AUC), Root Mean Square Error (RMSE), F1 score (the harmonic mean of precision and recall), sensitivity, specificity, etc. It is observed that CCMVI yields more accuracy with categorical data than numerical data. The CBCC-IM-EUC method gives better accuracy than other methods. Table 2 compares the percentage of missing data handled by the different statistical DI methods for healthcare data. Except for MI, the listed methods can improve the performance with imputation near to 50% of missing data.

Table 2. Percentage of missing data handled by statistical di methods

Reference	Imputation Method	Percentage of missing data handled
Tsai et al., 2018: 124-135	CCMVI	50%
Tan, 2017: 43-49	MI	23.06%

Hu, 2017: 112-120	0- imputation	50%
Gomes, 2015: 515-528	MI	8%
Ayilara et al., 2019	MI with auxiliary variable	50%

Hu et al. (2017: 112-120) experimented different statistical DI methods such as mean, 0-imputation, imputing normal values, and MICE on the data taken from the University of Minnesota clinical data repository. In general, imputed data is giving better results than non-imputed data in the detection of Surgical Site Infection (SSI). Among the methods considered for comparison, 0-imputation achieved the highest average AUC value of 0.893. Though it appears to be a bit counter-intuitive, the relationship of two significant features with the parameter temperature is causing this observation. The signs of the maximum and minimum post-operative temperatures are automatically correcting for the bias making 0-imputation as the best imputation method among them.

Gomes et al. (2015 515-528) proposed a new strategy to address the missing data in Patient-reported outcome measures (PROMs) through MI. They observed that the performance was varying with the assumptions they made for the non-responsiveness. This approach could improve the positive alarm rate by 6%. Ayilara et al. (2019) used the registry data available at regional joint replacement registry for Manitoba, Canada to study the impact of missing data in Patient-Reported Outcomes (PRO). With this data, they have experimented three imputation techniques including complete case analysis, MLE, and MI with and without auxiliary variables. Ten imputations using MI with auxiliary variable had given the best performance among them by reducing the bias and RMSE by 50% and 45%, respectively. The DI caused increase in the precision of computing the PRO scores.

Data Imputation Using Ml Based Methods

Yelipe et al. (2018: 487-504) imputed the missing values in medical data using class-based clustering for classification. The distance or similarity (for fuzzy data) for each record to the cluster centers are computed here. The datasets used for the experimentation were Iris, Hepatitis, and Wisconsin datasets. All the records were expressed as vectors with dimensionality equal to the number of class labels obtained through k-means clustering. For numerical missing values, the mean value of the attribute was substituted and for nominal missing values, the corresponding attribute value of a similar record was chosen. The classification performance was evaluated with Class-based clustering with imputation (CBCC-IM-EUC), SVM, KNN, and C4.5 algorithms. It was observed that the imputation could improve the classification performance.

Feature Projection KNN (FP-KNN) classifier model for imbalanced and incomplete medical data was proposed by Porwik et al. (2016: 644-656). The Fibrosis dataset was used for this study. The classifier contained an ensemble of homogenous KNN classifiers. With FP-KNN, there was no need for feature normalization and the neighbor area scaling method could be used in case of distance measure ambiguity. A weight factor was used for handling imbalance in data and FP-KNN might not work well for low dimensional datasets. The imputation diagnostics were carried out using Receiver Operating Characteristics (ROC), G-Measure (GM), and Cohen's Kappa. The subset of features chosen by GM had given better overall accuracy. Zhong et al. (Zhong et al., 2016: 307-316) introduced a two-stage granular DI approach for handling the missing data and did the experiment on blood dataset. A fuzzy clustering known as Fuzzy C-Means (FCM) was developed for the imputation. The structure of the data was important in this approach.

The number of clusters and the fuzzification coefficient were the prime factors which decided the performance of the imputation.

D. Ferreira-Santos and P.P. Rodrigues (2017) studied the impact of imputation in Bayesian network structure learning for diagnosis of sleep disorder. The Obstructive Sleep Apnea Dataset of 318 patients was used for this study. The percentage of missing data ranged from 0% to 97%. Using statistical significance, the variables were ranked for data imputation. The 10-nearest neighbors' imputation was done for each new variable included and the odds ratio was calculated to evaluate the imputation. Naive Bayes (NB) and Tree Augmented Naive Bayes (TAN) were the classifiers used. NB with 10 variables achieved an accuracy of 70.85%, and sensitivity of 95.07%, whereas TAN with 10 variables achieved an accuracy of 68.71%, and sensitivity of 89.05%. NB with 16 variables achieved an accuracy of 70.79%, and sensitivity of 94.36%, whereas TAN with 16 variables achieved an accuracy of 69.78%, and sensitivity of 90.33%.

García-Laencina et al. (2015: 125-133) developed an imputation method for the prediction of 5-year survival of breast cancer patients with missing discrete values. The breast cancer dataset from the Institute Portuguese of Oncology of Porto was used with prediction models such as KNN, classification trees, logistic regression, and SVM. More than 40% of the data had a minimum of three values missing. Imputation techniques used were Mode imputation (Mimp), Expectation-Maximization imputation (EMimp) and KNN imputation (KNNimp). The best results were with KNN and KNNimp combination, which had more than 81% of accuracy and more than 0.78 of area under the ROC.

Shukla et al. (2018: 199-208) proposed a new strategy for handling the missing data with self-organizing map density-based spatial clustering of applications with noise (SOM-DBSCAN) approach. The study used breast cancer incidence dataset from the SEER program. The missing values were imputed with the values of the patient data belonging to the same cluster rather than the entire dataset. This method was very useful as a preparatory step for the classification of data with missing data to improve accuracy. Table 3 compares the performance of the different ML methods for DI. It is observed that the KNN method gives better accuracy than other methods. Table 4 presents the percentage of missing data handled by the different ML-based DI methods. The autoencoders improves the performance with imputation even with 50% of missing data.

Table 3. Performance improvement using ml based di methods for healthcare data

Reference	Imputation Method	Performance Metric(s)
Yelipe et al., 2018: 487-504	CBCC-IM-EUC	Average accuracy 95.685%
Porwik, 2016: 644-656	FP-KNN with GM	Specificity 0.83 and sensitivity 0.67
Zhong et al., 2016: 307-316	FCM	Average AUC of 0.3855
Ferreira-Santos and Rodrigues, 2017	NB-10	Accuracy 70.85%
	NB-16	Accuracy 70.79%
García-Laencina et al., 2015: 125-133	KNN	Accuracy 81%
Beaulieu-Jones and J. Moore, 2017: 207-218	Autoencoder	MAR: Average RMSE 0.10
		MNAR: Average RMSE 0.19
Kim et al., 2018: e1006106	RIDDLE	Accuracy 66.8%

B. Beaulieu-Jones and J. Moore (2017 207-218) investigated the impact of autoencoders for DI. Data used for this work was taken from Pooled Resource Open-Access Amyotrophic Lateral Sclerosis (ALS) Clinical Trials Database (PRO-ACT). The missing values were simulated in MAR and MNAR cases at varying missing percentages, limited to a maximum of 50%. It was observed that MAR achieved better imputation accuracy with 0.10 RMSE and MNAR achieved better results with RMSE of 0.19. Also, these encoders were able to give best ALS prediction accuracy with an average RMSE of 0.32.

Kim et al. (2018: e1006106) demonstrated that Race and ethnicity Imputation from Disease history with Deep Learning (RIDDLE) was giving the best results in estimating the missing racial and ethnic data when compared to competing methods like SVM, Random Forest (RF), logistic regression, and gradient boosted trees. There were four types of races considered with varying percentage of the population. The RIDDLE approach could give better results with an accuracy of 0.668, AUC of 0.833, and F1-score of 0.652.

Table 4. Percentage of missing data handled by ml based di methods

Reference	Imputation Method	Percentage of missing data handled
Yelipe et al., 2018: 487-504	CBCC-IM-EUC	48%
Porwik, 2016: 644-656	FP-KNN with GM	36%
Zhong et al., 2016: 307-316	FCM	45%
García-Laencina et al., 2015: 125-133	KNN	17.99%
Beaulieu-Jones and J. Moore, 2017: 207-218	Autoencoder	50%
Kim et al., 2018: e1006106	RIDDLE	30%

Conclusion

This paper classifies and analyses the performance of different data imputation methods for healthcare data. Developing a generalized imputation method for a variety of datasets is difficult as each dataset has its own attributes, characteristics, and intrinsic structures. Comparing the performance of DI algorithms also is difficult as the performance is very much dependent on the attributes of the dataset. In general, MI has a better range of results when compared to single imputation. MI can also accommodate the uncertainty factors and facilitate precise classification or prediction with missing observations. Complete case analysis is the least recommended imputation technique for MCAR data because it discards information possessed by incomplete cases. Imputing MNAR needs to consider the parameters of interest which are then to be modeled carefully.

In general, the imputation accuracy increases with increased referential population size. One challenge in imputation is to fill in the missing values by creating minimal bias. The imputation algorithms are developed with an assumption that the data follows a normal distribution that may not be true always. Use of deep networks is a good option to impute data as it can represent inherent features and correlations in it. Including inference for imputation may be useful in cases where the specific data are not directly available in the database. When the data missing percentage is considerably high using

variants of Generative Adversarial Networks for imputation is a good option. ML-based DI methods generally take more imputation times than statistical methods except for a few algorithms like KNN.

Inappropriate methods to handle the missing data can lead to misleading results and can adversely affect the classification and prediction results. This can be fatal in the case of healthcare applications. Avoiding incomplete data records also may cause imprecise results as it discards the information possessed by those data, and not advisable in the case of healthcare applications. Suitably combining the statistical methods and ML methods for the imputation to achieve better results is suggested as a topic for further research.

References

Ayilara, O. F., Zhang, L., Sajobi, T., Sawatzky, R., Bohm, E., Lix, L. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and Quality of Life Outcomes, 17(106). Available at: https://doi.org/10.1186/s12955-019-1181-2

Banjar, H., Ranasinghe, D., Brown, F., Adelson, D., Kroger, T., Leclercq, T., White, D., Hughes, T., Chaudhri, N. (2017). Modelling Predictors of Molecular Response to Frontline Imatinib for Patients with Chronic Myeloid Leukaemia. Plos One, 12(1). Available at: https://doi.org/10.1371/journal.pone.0168947

Beaulieu-Jones, B., Moore, J. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. Biocomputing, 22, 207-218

Chowdhury, M.H., Islam, M.K., Khan, S.I. (2017). Imputation of missing healthcare data. 2017 20th International Conference of Computer and Information Technology (ICCIT), 13-19.

Eirola, E., Akusok, A., Björk, K.-M., Johnson, H., Lendasse, A. (2017). Predicting Huntington's Disease: Extreme Learning Machine with Missing Values. Proceedings in Adaptation, Learning and Optimization Proceedings of ELM-2016, 195-206.

Ferreira-Santos, D., Rodrigues. P. P. (2017). Improving Diagnosis in Obstructive Sleep Apnea with Clinical Data: A Bayesian Network Approach. 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS).

García-Laencina, P.J., Abreu, P.H., Abreu, M.H., Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in Biology and Medicine, 59, 125-133.

Gomes, M., Gutacker, N., Bojke, C., Street, A. (2015). Addressing Missing Data in Patient-Reported Outcome Measures (PROMS): Implications for the Use of PROMS for Comparing Provider Performance. Health Economics, 25(5), 515-528.

Henry, A.J., Hevelone, N.D., Lipsitz, S., Nguyen, L.L. (2013). Comparative methods for handling missing data in large databases. Journal of Vascular Surgery, 58 (5), 115-126.

Hu, Z., Melton, G., Arsoniadis, E., Wang, Y., Kwaan, M., Simon, G. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. Journal of Biomedical Informatics, 68, 112-120.

Kim, J.S., Gao, X., Rzhetsky, A. (2018). RIDDLE: Race and ethnicity Imputation from Disease history with Deep Learning. PLOS Computational Biology, 14(4), e1006106.

Pearson, M.J., Smart, N. A. (2018). Reported methods for handling missing change standard deviations in meta-analyses of exercise therapy interventions in patients with heart failure: A systematic review. Plos One, 13 (10), 34-42.

Perkins, N.J., Cole, S.R., Harel, O., Tchetgen, E.J.T., Sun, B., Mitchell, E.M., Schisterman, E.F. (2017). Principled Approaches to Missing Data in Epidemiologic Studies. American Journal of Epidemiology, 187(3), 568-575.

Porwik, P., Orczyk, T., Lewandowski, M., Cholewa, M. (2016). Feature projection k-NN classifier model for imbalanced and incomplete medical data. Biocybernetics and Biomedical Engineering, 36(4), 644–656.

Richter, A.N., Khoshgoftaar, T.M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial Intelligence in Medicine, 90, 1-14.

Shukla, N., Hagenbuchner, M., Win, K. T., Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. Computer Methods and Programs in Biomedicine, 155, 199-208.

Tan, J.-P., Li, N., Lan, X.-Y., Zhang, S.-M., Cui, B., Liu, L.-X., He, X., Zeng, L., Tau, L.-Y., Zhang, H., Wang, X.-X., Wang,' L.-N., Zhao, Y.-M. (2017). The impact of methods to handle missing data on the estimated prevalence of dementia and mild cognitive impairment in a cross-sectional study including non-responders. Archives of Gerontology and Geriatrics, 73, 43-49.

Tsai, C.-F., Li, M.-L., Lin, W.-C. (2018). A class center based approach for missing value imputation. Knowledge-Based Systems, 151, 124-135.

Wilson, E.D., Clairon, Q., Henderson, R., Taylor, C.J. (2018). Dealing with observational data in control. Annual Reviews in Control, 46, 94-106.

Yelipe, U., Porika, S., Golla, M. (2018). An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. Computers & Electrical Engineering, 66, 487-504.

Zhang, Y., Flórez, I.D., Lozano, L.E.C., Aloweni, F.A.B., Kennedy, S.A., Li, A., Craigie, S., Zhang, S., Agarwal, A., Lopes, L.C., Devji, T., Wiercioch, W., Riva, J.J., Wang M., Jin, X., Fei, Y., Alexander, P., Morgano, G.P., Zhang, Y., Carrasco-Labra, A., Kahale, L.A., Akl, E.A., Schünemann, H.J., Thabane, L., Guyatt, G. H. (2017). A systematic survey on reporting and methods for handling missing participant data for continuous outcomes in randomized controlled trials. Journal of Clinical Epidemiology, 88, 57-66.

Zhong, C., Pedrycz, W., Wang, D., Li, L., Li, Z. Granular data imputation: A framework of Granular Computing. Applied Soft Computing, 46, 307-316.