Addressing Missing Data in a Healthcare Dataset Using an Improved kNN Algorithm

Tressy Thomas[1 [0000−0002−5270−5497]] and Enayat Rajabi[1 [0000−0002−9557−0043]]

Shannon School of Business, Cape Breton University, Sydney, NS, Canada

Abstract. Missing values are ubiquitous in many real-world datasets. In scenarios where a dataset is not very large, addressing its missing values by utilizing appropriate data imputation methods benefits analysis significantly. In this paper, we leveraged and evaluated a new imputation approach called k-Nearest Neighbour with Most Significant Features and incomplete cases($KNNI_{MSF}$) to impute missing values in a healthcare dataset. This algorithm leverages k-Nearest Neighbour(kNN) and ReliefF feature selection techniques to address incomplete cases in the dataset. The merit of imputation is measured by comparing the classification performance of data models trained with the dataset with imputation and without imputation. We used a real-world dataset, "very low birth weight infants", to predict the survival outcome of infants with low birth weights. Five different classifiers were used in the experiments. The comparison of multiple performance metrics shows that classifiers built on imputed dataset produce much better outcomes. $KNNI_{MSF}$ outperformed in general than the k-Nearest Neighbour Imputation using the Random Forest feature weights($KNNI_{RF}$) algorithm with respect to the balanced accuracy and specificity.

Keywords: missing value· data imputation · knn · healthcare.

1 Introduction

Data has quality if it satisfies the requirements of its intended use[1]. Real-world data are typically incomplete and incompleteness can impair the knowledge discovery process. In order to make data useful for the purpose of analyzing methods, such as data mining and machine learning, a significant amount of time is spent on pre-processing of data. For medical datasets, missing values are unfortunately unavoidable[2]. Incorrect imputation of missing values could lead to inaccurate research as well as wrong predictions[2]. Due to the nature of the domain and significance of applications, it is very important to have highly accurate results.

The missing mechanism is an important concept that defines the connection between the observed and missing variables in a dataset. The missing mechanism gives an account on possible relationships between measured variables and the probability of missing data[3]. According to Little and Rubin's[4] taxonomy, missing mechanisms are missing completely at random (MCAR), missing at

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

random (MAR) and not missing at random (NMAR). In MCAR, the reason of missingness is unrelated to any other observation. MAR arises if the reason for dropout depends on the observed outcomes. The MNAR mechanism depends, in whole or in part, on unobserved measurements itself. Details about the different missing mechanisms can be referenced in many published studies[3] [4].

The complete case analysis is a way of treating missing values in a dataset by ignoring the incomplete cases in the dataset. In many cases, especially in medical domains, this approach can result in loss of information[5]. As the information in the incomplete cases in a dataset is not made useful, the statistical inferences or the model performance may not result in meaningful insights and predictions, particularly when the size of a dataset is not very large. There exists many missing value imputation techniques which can estimate the missing values so that the incomplete cases can be repaired and used for analysis or data modeling purposes without losing information or adding bias to dataset[5]. Replacing the missing data with an appropriate value derived from an observed data is called missing value imputation. Leveraging machine learning algorithms to impute missing values is getting popular due to its applicability. k-Nearest Neighbours(kNN), being one of the simplest and non-parametric instance-based approaches, is widely used in missing value imputation problems[6]. kNN based imputation methods are easy to implement and perform well in a variety of scenarios [6]. The basic kNN based imputation method uses the 'k' nearest neighbor's value to estimate the missing value. In this study, we evaluate the kNN imputation method(KNNI_{MSF}) on a healthcare dataset with a high level of missingness. This new imputation approach utilizes the most significant features with respect to the missing attributes and considers incomplete cases as well to estimate the missing values.

The merit of imputation is evaluated by comparing the performance of classifier algorithms with the dataset without any imputation treatment, with dataset after undergoing KNNI_{MSF} imputation and with dataset imputed using another well-known Weighted kNN imputation based on Random Forest(KNNI_{RF}). The rest of the paper is arranged as follows: section 2 presents the previous evaluation studies on missing value imputation using healthcare datasets. It also presents why this study is different from those previous studies and recommendations. Section 3 provides the details regarding our proposed imputation approach. The next section presents the experimentation settings, the dataset we used in this study along with the evaluation metrics. It is followed by Section 5 where the results and inferences regarding the experiments are discussed. Our paper ends with the conclusion section which presents the significant observations and future scope for this research.

2 Related Works

The missing data problem is crucial in healthcare domain. Hence, several published studies addressed the missing data problem. In one such study[7] the influence of missing value imputation on the classification accuracy was discussed.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

Globally average value, average value within a cluster and average value within the class were the missing value imputation techniques used. The missing values were artificially induced in four healthcare datasets and then imputed before evaluating the impact of missing value imputation experimentally. The comparison of classifier accuracy on different imputed datasets with the complete dataset in this study showed that there can be under– or overestimation of classification accuracy caused by choosing wrong method[7].

Machine learning techniques were found to perform better than the standard mean imputation technique in [13]. Cardiovascular data with missing value frequency up to 30 percentage was used in the experiments[13]. Another study on missing healthcare data imputation is presented in [8]. This research implemented three algorithms in real healthcare dataset and concluded that MICE(Multiple Imputation by Chained Equations) algorithm performs better than Amelia and fuzzy unordered rule induction algorithm imputation(FURIA)[8].

In our study, we have utilized a real-world healthcare dataset with missing values at a higher percentage. The statistical imputation methods recommended in the mentioned studies assume that the missing data are Missing At Random (MAR). We are interested in an imputation technique that can be more generalized but also usable in critical domain applications. It is due to this fact that we are employing one of the easiest non-parametric algorithms (kNN) to implement missing value imputation. We compared our method with KNNI_{RF} algorithm which in general outperforms the other kNN based imputation methods, based on our previous experimentation. KNNI_{RF} is a weighted kNN imputation technique based on Random Forest where the weights for each variable are obtained using Random Forest approach[9] and these weights are used in the distance calculation.

We are presenting a new imputation approach based on kNN algorithm. Our approach considers the incomplete cases also for the estimation but only the relevant features are used for imputing the missing values. This approach is suitable for handling the missing values in small datasets with high missing percentages which we are evaluating in the experiments.

3 Methods

Our approach for the missing value imputation considers the significant features that are relevant for estimating that particular attribute. The steps used in the approach are as follows. The process starts with identifying the attributes with at least one missing value. For missing attribute, feature quality estimation algorithm ReliefF is executed to get the most significant features in the dataset that can predict the missing attribute [10]. ReliefF algorithm accounts the correlation and interaction between the attributes. This is important in estimating the missing values and helps estimate the correct value to replace the missingness. Only complete cases are used for the purpose of selecting the relevant features that can estimate the missing attribute. For each of the missing value of this attribute, Gower distance gd between the instances is calculated based on the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

equation 1[11].

$$gd_{ij} = \frac{\sum_p (\delta_{ijp} d_{ij_p}^f)}{\sum_p (\delta_{ijp})} \tag{1}$$

Where x_i is the missing vector and x_j is observed vector, k is the attribute. For numerical attributes in the instance $(d_{ij_p}^f)$ is calculated by

$$d_{ij_p}^f = \frac{|x_i - x_j|}{|(max_N(x) - min_N(x))|} \tag{2}$$

where N is the total number of instances in the dataset. For other attributes $(d_{ij_p}^f)$ is 1 when the x_i and x_j attribute value differs. Otherwise it is set to 0.

For distance calculation, the features selected from the previous step are used. But for estimation of the missing value, all instances that have selected features and the missing feature present are considered. In the traditional kNN approach only complete cases are used. Utilizing the incomplete cases but with relevant features will provide better estimations, especially when multiple variables are missing in an instance. Similarity between each data point with the missing value instance is calculated as:

$$Sim = \frac{1}{gd + 1} \tag{3}$$

Then the weight for the 'k' neighbour instances are calculated based on the similarity:

$$Wt_k = \frac{Sim_k}{\sum_1^k Sim}. \tag{4}$$

For the estimation of numerical missing value, weighted sum of the nearest neighbour attribute values are used. For nominal values, mode is used to impute the missing value. The steps are iterated for all the missing attribute and its missing values.

4 Experimentation Set-up

For the experiments, we used the 'Very Low Birth Weight Infants' dataset. Data on 671 infants with very low (less than 1600 grams) birth weight from 1981-87 were collected at Duke University Medical Center by Dr. Michael O'Shea[12]. There are 671 observations and 26 variables in the dataset. The details of the number of missing attributes by their missing value percentages are given in the Table 1. 78.54% of instances are labelled with the survival outcome as alive and 21.46% are with survival outcome as dead. There are only 174 cases which have all the 25 attributes present. Since most machine learning algorithms typically cannot utilize incomplete cases, in this dataset, about 75 percentage of data will be lost if only complete cases are used for classification. This loss of data can result in a very significant loss of information. This is an example case of how we can utilize imputation to get more data and thus more information to achieve better classification or prediction. To evaluate the merit of imputation,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

Table 1. Missing value percentage in the dataset

Missing Percentage	No of Attributes
1-3%	4
3-9%	11
6-15%	3
15-30%	5
30-60%	2

we have conducted classification using five different classifiers. First, data model was trained and evaluated using only the complete cases from the dataset. Then the dataset was imputed using two different missing value imputation methods KNNI_{RF} and KNNI_{MSF}. The entire dataset was split into training and test in 70:30 ratio. Five fold cross validation was repeated five times for training the model. Model evaluation and comparison is done with the test data . The classifiers used here for the prediction of survival outcome of the infant are Logistic Regression(LR), Support Vector Machine using Radial basis function kernel(SVM), k-Nearest Neighbour Classifier (kNN), Gradient Boosting Machine(GBM) and Decision Tree Classifier (CT). The positive class in the classification model is the survival outcome 'live' and negative class is 'dead'.

Evaluation Metrics Since the cost of miss-classification is very determinant factor in the evaluation of the model due to the nature of medical domain, we have used multiple metrics such as Accuracy, Balanced Accuracy, Sensitivity and Specificity for the model evaluation. To compare the performance of imputation method, Wilcoxon signed rank test is performed at α =0.1 with null hypothesis: H_0 = The performance of classifiers using KNNI_{RF} imputed dataset is equal to that of KNNI_{MSF} imputed dataset. Alternative hypothesis: H_1 = the performance of classifiers using KNNI_{MSF} imputed data and that using KNNI_{RF} imputed data are not equal.

5 Results

The performance of five classifiers were measured and the evaluation metrics are presented in Tables 2 for the three cases. First is 'Complete cases' where only complete cases from the original dataset was used in modelling and evaluation. The other two, KNNI_{RF} and KNNI_{MSF}, represents the dataset imputed using the KNNI_{RF} and KNNI_{MSF} imputation method respectively. The survival outcome(alive or dead) prediction accuracy for each of the model is given in Table 2. It can be seen from the results that the train models with imputed datasets perform better than that used without any imputation. Also, KNNI_{MSF} resulted in better accurate prediction than KNNI_{RF} in most classifiers. It is evident from the results that the balanced accuracy is very poor for the model trained with the complete cases. The classifiers trained with imputed datasets performed relatively much better with respect to the balanced accuracy. Sensitivity measure,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

which is the measure of how well the classifier predict the positive cases(alive), also suggest a better performance of model trained with imputed data.

Table 2. Comparison of Evaluation metrics of the Classifiers

Missing Value Handling	**LR**	**SVM**	**kNN**	**GBM**	**CT**
Accuracy Score of the Classifiers					
Complete Cases	82.35	82.35	92.16	90.20	86.27
$KNNI_{RF}$	88.56	88.56	87.06	91.54	89.05
$KNNI_{MSF}$	87.56	89.55	88.06	94.53	89.05
Balanced Accuracy Score of the Classifiers					
Complete Cases	43.75	43.75	48.96	47.92	45.83
$KNNI_{RF}$	80.87	73.15	81.72	85.31	77.80
$KNNI_{MSF}$	81.09	74.63	84.04	88.90	77.80
Sensitivity of the Classifiers					
Complete Cases	87.50	87.50	97.92	95.83	91.67
$KNNI_{RF}$	94.30	93.67	97.47	96.20	97.47
$KNNI_{MSF}$	92.41	93.67	98.10	98.73	97.47
Specificity of the Classifiers					
Complete Cases	0	0	0	0	0
$KNNI_{RF}$	67.44	69.77	48.84	74.42	58.14
$KNNI_{MSF}$	69.77	74.42	51.16	79.07	58.14

Specificity metric, in this case, shows that the model with complete cases is not useful at all in predicting the minority class (survival outcome=dead). The classifier models with $KNNI_{MSF}$ imputed data performed better in predicting the minority class related to that of $KNNI_{RF}$ imputed data. The Wilcoxon signed rank test shows fair evidence against null hypothesis which confirms the performance metrics(Balanced Accuracy and Specificity) of classifiers using $KNNI_{MSF}$ imputed data is greater that that using $KNNI_{RF}$ imputed data. Overall, the performance of $KNNI_{MSF}$ is either comparable or superior to $KNNI_{RF}$ with respect to the evaluated metrics and is a good approach to be used for small datasets with high missingness.

6 Conclusion and Future Works

Missing value datasets treated using imputation methods can result in better utilization of all available information for data modeling and statistical inferences. Especially in medical domain this can add much value and benefit. Our proposed missing value imputation method was tested on a healthcare dataset with high missingness percentage. The evaluation showed the merit of imputation with improved classifier performance. The comparison of classifiers trained with both complete cases and imputed datasets indicated that the proposed model performance is much better for the classifiers trained with imputed dataset. Also, the $KNNI_{MSF}$ imputation method performed better in general from the accuracy,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17

balanced accuracy and specificity perspectives than the $KNNI_{RF}$ method. It can be concluded that $KNNI_{MSF}$ missing value imputation can treat missing values appropriately and the use of imputed datasets result in better data model training and model performance. In future, this new approach can be tested on more healthcare datasets with missing values to validate its performance.

References

1. Han, J., Kamber, M., Pei, J. 3 - Data Preprocessing. In J. Han, M. Kamber, J. Pei (Eds.), Data Mining (Third Edition) (Third Edit, pp. 83–124). Morgan Kaufmann.(2012).https://doi.org/10.1016/B978-0-12-381479-1.00003-4
2. Schmidt, D., Niemann, M., Lindemann-Von Trzebiatowski, G. (2015). The handling of missing values in medical domains with respect to pattern mining algorithms. In CEUR Workshop Proceedings (Vol. 1492).
3. C. K. Enders Craig K , Applied Missing Data Analysis. The Guilford Press. New York, London (2010)
4. Rubin, Donald B. Inference and missing data. Biometrika 63(3): 581-592. (1976). https://doi.org/10.1093/biomet/63.3.581
5. Bartlett, J. W., Harel, O., Carpenter, J. R. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. American Journal of Epidemiology, 182(8), 730–736. (2014). https://doi.org/10.1093/aje/kwv114
6. A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," Applied Artificial Intelligence, vol. 33, no. 10, pp. 913–933, (2019) https://doi.org/ 10.1080/08839514.2019.1637138
7. Orczyk, T., Porwik, P. Influence of missing data imputation method on the classification accuracy of the medical data. Journal of Medical Informatics Technologies, 22, 111–116.(2013)
8. Chowdhury, Mohaimanul Hoque; Islam, Muhammad Kamrul; Khan, Shahidul Islam . [IEEE 2017 20th International Conference of Computer and Information Technology (ICCIT) - Dhaka, Bangladesh (2017.12.22-2017.12.24)] 2017 20th International Conference of Computer and Information Technology (ICCIT) - Imputation of missing healthcare data. , 1–6.(2017).https://doi.org/10.1109/ICCITECHN.2017.8281805
9. A. Kowarik and M. Templ, Imputation with the R package VIM, Journal of Statistical Software, vol. 74, (2016). https://doi.org/ 10.18637/jss.v074.i07
10. I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 784 LNCS, pp. 171–182,(1994). https://doi.org/10.1007/3-540-57868-4
11. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. In Biometrics (Vol. 27, Issue 4).(1971). https://doi.org/10.2307/2528823
12. O'Shea M, Savitz DA, Hage ML, Feinstein KA. Prenatal events and the risk of subependymal/intraventricular haemorrhage in very low birthweight neonates. Paediatr Perinat Epidemiol. 6(3):352-62. (1992) https://doi.org/ 10.1111/j.1365-3016.1992.tb00775.x
13. Rahman, Mostafizur Davis, Darryl N.. Machine Learning Based Missing Value Imputation Method for Clinical Datasets. IAENG Transactions on Engineering Technologies. 229. (2012).https://doi.org/10.1007/978-94-007-6190-2$_1$9

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_17