ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2021.0347

A COMPARISON OF MISSING DATA HANDLING TECHNIQUES

S. David Samuel Azariya[1], V. Mohanraj[2], J. Jeba Emilyn[3] and G. Jothi[4]
[1,2,3]Department of Information Technology, Sona College of Technology, India
[4]Department of Computer Applications, Sona College of Arts and Science, India

Abstract

Missing data is a regular concern on data that professionals have to deal with. Efficient analysis techniques have to be followed to find interesting patterns. In this study, we are comparing 16 different imputation methods namely Linear, Index, Values, Nearest, Zero, slinear, Quadratic, Cubic, Barycentric, Krogh, Polynomial, Spline, Piecewise Polynomial, From derivatives, Pchip and Akima. These techniques are performed on real time UCI dataset and are under Missing Completely at a Random (MCAR) assumption, our result suggests the nearest, zero, quadratic and polynomial imputation methods which provides above 96% of accuracy when compared to the other techniques.

Keywords:

Missing Data, Imputation Methods, Missing Completely at Random

1. INTRODUCTION

Missing data is a common concern in data analytics. Donald Rubin formalised the analysis of missing data with the idea of the missing process. Three types of missing data generation methods can be implemented, namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [1].

The occurrence of missing values is entirely at random with respect to MCAR, not connected to any variable. MAR suggests that only experiential data is linked to the missing values. MNAR suggests that an experimental and non-experimental component is correlated with the missing values. Missing data is a very difficult task in the preprocessing stage of data that affects the accuracy of decision-making [2].

In the recent years, many researchers have been proposing lot of methods for missing data imputation. The various kinds of missing data imputation Technique, such as Linear, Index, Values, Nearest, Zero, slinear, Quadratic, Cubic, Barycentric, Krogh, Polynomial, Spline, Piecewise Polynomial, from derivatives, Pchip and Akima, are explored in this research. The efficiency of the imputation techniques is evaluated using various kinds of machine learning data sets such as iris, wine, credit card and Boston housing.

These datasets are taken from the UCI machine learning repository. In this research the popular machine learning dataset Irish is used to explore the performance of the imputation methods. It also compared the performance of different imputation algorithms based on benchmark classification algorithms such as Naive Bayes and Decision Table.

The results of the experiments suggest that the imputation approach should be used depending on the conditions of the data. From the experimental results it is also believed that spline technique performs better than the existing approaches.

The remaining paper is structure is as follows: The related literature is reviewed in section 2. The datasets are outlined in section 3. The different missing data imputation algorithms have been explained in section 4. Missing data-generated processes are addressed in section 5. Section 6, the experimental findings are discussed. In section 7, the conclusion and potential directions are provided.

2. RELATED WORKS

Recently, numerous methods have been developed to impute the missing values. Existing imputation methods study done by Mean, fuzzy K-means (FKM), K-Nearest Neighbors (KNN), Singular Value Decomposition (SVD), Bayesian Principal Component Analysis (bPCA) and multiple imputations by chained equations (MICE) [3] and another comparison done by using deletion of missing value, mode imputation, Hot-deck imputation and most similar value filling [4]. In [5] the impact of missing value on the performance of a machine learning model was discussed and illustrated in case of brain disorders dataset. In [6] the performances of several statistical and machine learning imputation methods are evaluated to predict recurrence in patients in an extensive real breast cancer data set.

3. DATASETS

The iris dataset is a basic and beginner-friendly dataset containing details about petal and sepal sizes of the flowers. The dataset has 3 classes in each class with 50 instances, so it contains 150 rows with only 4 columns [7].

The credit card dataset has information about transactions and it classified fraudulent or genuine. It has 284807 instances and 31 attributes. Using these details companies can develop fraudulent detecting systems [8].

The wine dataset helps to predict wine quality and it contains various chemical information. It has 4898 instances, each with 14 variables. It's good for regression and classification work [9].

The Boston housing dataset was used for pattern recognition. It contains details about the Boston houses and crime rate, rent, number of rooms, etc. It has 506 rows and 14 column variables. This dataset can be used to forecast house prices [10]. The summary of the dataset is presented in Table.1.

Table.1. Dataset summary

Datasets	No. of Instances	No. of Attributes
Iris	150	4
credit card	284807	31
wine	4898	14
Boston housing	506	14

4. MISSING DATA HANDLING TECHNIQUES

The ‘linear’ technique skips the index and regards the values as being spaced equally. The only approach allowed on MultiIndexes. The ‘time’ works on regular and higher precision information to interpolate the interval length provided. The ‘index’ and ‘values’ using the index 's real numerical values. The ‘pad’ using current values, fill in NaNs. The ‘nearest’, ‘zero’, ‘slinear’, ‘quadratic’, ‘cubic’, ‘spline’, ‘barycentric’ and ‘polynomial’ these approaches use the index's numerical values. Both 'polynomial' and 'spline' enable you to define an order as well (int). The ‘krogh’, ‘piecewise_polynomial’, ‘spline’, ‘pchip’, ‘akima’ wrappers all over the methods of SciPy [11] interpolation of related names. The ‘from_derivatives’ in the Bernstein foundation, create a piecewise polynomial consistent with stated values and derivatives at breakpoints.

5. PRODUCE MISSING DATA

The process of creating missing data is divided into three key categories as described by Rubin [12].

Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). The first two are also eligible as ignorable missing values mechanisms, for instance in likelihood-based methods to handle missing values, while the MNAR mechanism produces nonignorable missing values.

Let the complete findings be denoted by the Eq.(1)

$$X \in \chi_1 * \ldots * \chi_p \tag{1}$$

let us assume X is a combination of

$$X_j \in \chi_j, j \in \{1,\ldots,p\} \tag{2}$$

where $\dim(\chi_j)=n$ for all j.

The data can be made of numerical and/or qualitative factors, so for any discrete set χ_j can be R^n, Z^n or more commonly S^n for any discrete set S.

Missing values are recorded as NA (not available) and an indicator matrix is established $R \in \{0,1\} n \times p$ such that it is $R_{ij}=1$ if X_{ij} is $R_{ij}=0$ and rather observed. We consider this matrix R the pattern of response (or absence) of the findings X. We may partition the findings X into observed and missing findings according to the same pattern:

$$X = (X_{obs}, X_{mis}) \tag{3}$$

Both X and R are modelled as random variables with $\mathbb{P}X$ and $\mathbb{P}R$ probability distributions, respectively, to describe the different missing value mechanisms. The missing $\mathbb{P}R$ distribution is parameterized by the ϕ parameter.

5.1 MCAR CONCEPT

If the probability that an observation is missing is independent of the variables and observations, the observations are said to be Missing Completely At Random (MCAR): the probability that an observation is missing does not depend on (X_{obs}, X_{mis}). The Eq.(4) is formally denoted by,

$$\mathbb{P}_R(R|X^{obs}, X^{mis}; \phi)=\mathbb{P}_R(R) \quad \forall\phi \tag{4}$$

5.2 IMPUTATION METHODS

For many reasons, several real-world datasets can contain missing values. Imputation is a technique for the replacement and the analysis of the whole data set as if the values imputed were true observed. They also get encoded as NaNs, blanks or anything else. Training a model with a dataset with many missing values can have a dramatic effect on the output of the machine learning model. In order to handle this issue, the missing datain the observations must be removed. However, there is the possibility that important information will lose data. A better strategy would be to impute the missing values. Recently, there are many imputation methods are developed.

To fill NA values with the interpolate() function in datasets or sequence. This is a very powerful function to fill the missing values. It uses different interpolation methods, rather than hard-coding, to fill the missing values.

In this paper, sixteen different imputation methods are compared namely Linear, Index, Values, Nearest, Zero, slinear, Quadratic, Cubic, Barycentric, Krogh, Polynomial, Spline, Piecewise Polynomial, From derivatives, Pchip and Akima. These techniques are performed on real time UCI dataset and are under Missing Completely at a Random (MCAR) assumption. The numerical example of imputation method is presented in Table.2.

5.3 HANDLING MISSING DATA USING ML

The lack of data is a daily issue a data expert needs to address. Missing data may include missing sequence, incomplete functionality, missing files, incomplete information, data entry error, etc. It is necessary to convert these fields and use them for analysis and modelling before using data with missing data fields. Data manipulation or data loss may be the source of missing values. During data pre-processing the handling of missing data is extremely critical, as many machine learning algorithms do not accept missing values. In this study, various imputation methods are employed to fill the missing values in the data sets. The performance of the imputation methods is evaluated using the machine learning (ML) algorithm such as Naive Bayes (NB) and Decision Tree (DT). Based on the literature reviews these two ML methods have been chosen for the experimental analysis.

5.3.1 Naïve Bayes:

It is based on the theorem of Bayes' assumption that the predictors are independent. Naive Bayes classification assumes that there is no connection between the existences of any particular feature in a class. It is also noted for its simplicity to exceed even highly advanced classification methods. In Bayes theorem the probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$ is determined by means of the Eq.(5),

$$P(c|x) = \frac{P(\chi|C)(c)}{P(x)} \tag{5}$$

5.3.2 Decision Tree:

Decision tree is the most common prediction and classification method. DT is a tree structure flowchart in which every internal node is an attribute test, each branch represents a test outcome and every leaf node (terminal node) is a class mark. This is a visual illustration of a decision and all possible results.

It also helps people define all possible options and weigh every step against the risks and rewards that every choice can produce.

6. EXPERIMENTAL RESULTS

In this section, an assessment criterion is conducted using the proposed method.

6.1 ASSESSMENT CRITERIA

In this study, the machine learning techniques are employed in two ways (1) to evaluate the efficiency of the imputation methods (2) to identify the best imputation method to fill the missing values.

In this experiment we utilized varies classification evaluation metrics likely Precision, Recall, Root Mean Square Error (RMSE), and Accuracy to evaluate the performance of the various imputation methods. The datasets are split into training and testing datasets (Training 80%, Testing 20)

6.2 ASSESSMENT ANALYSIS

The performance of imputations methods is compared to four evaluation criteria.

Precision isn't limited to problems with binary classification. Precision(P) is calculated as the sum of true positives across all classes in an imbalanced classification problem with more than two classes, divided by across all classes the sum of True Positives (TP) and False Positives (FP)

$$P = TP/TP+FP \tag{6}$$

Recall (R) is determined as the number of true positives divided by the total number of true positive and False Negatives (FN).

$$R=TP/(TP+FN) \tag{7}$$

Root Mean Squared Error (RMSE) is somewhat similar to Mean Absolute Error, the only difference being that the variation between the original values and the expected values is taken by RMSE as the average of the square. The benefit of RMSE is that the gradient is measured more easily, while Mean Absolute Error requires complex linear programming software to measure the gradient. The effect of larger errors becomes more noticeable than smaller errors, so the model can now concentrate more on larger errors.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{8}$$

6.3 RESULTS AND DISCUSSION

In this experimental analysis, sixteen imputation methods are compared and the performances are analysed. The popular machine learning dataset Irish is used to explore the performance of the imputation methods. The Irish is widely used dataset to evaluate the efficiency of the imputation methods. The performances are evaluated using the classification algorithms namely Naive Bayes and Decision Table. The results of each imputation method are presented in Table.2. In Table.3, the precision, recall, root mean square error value, and the accuracy of before and after imputation method is recorded.

Table.2. The Numerical Example of Imputation methods.

Data Series (with missing values)	**Imputation methods**															
	Linear	**Index**	**Values**	**Nearest**	**Zero**	**slinear**	**Quadratic**	**Cubic**	**Bary**	**Krogh**	**Poly-nomial**	**Spline**	**Piecewise polynomial**	**From derivatives**	**Pchip**	**Akima**
12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
NaN	3	3	3	5	5	3	4.5	6.75	6.75	4.57	6.75	6.75	6.75	4.57	3	3
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table.3. The Results of various evaluation metrics

S. No	**Imputation Methods'**	**Algorithm**	**Precision**		**Recall**		**RMSE**		**Overall accuracy**	
			Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
1	Linear	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1483	94.6309	95.333
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.667
2	Index	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1483	94.6309	95.333
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.666
3	Values	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1483	94.6309	95.667
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.667
4	Nearest	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1468	94.6309	95.333
		Decision Table	0.920	0.953	0.0916	0.953	0.2161	0.1682	91.9463	95.333
5	Zero	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1467	94.6309	95.333

		Decision Table	0.920	0.953	0.0916	0.953	0.2161	0.1682	91.9463	95.333
6	Slinear	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1483	94.6309	95.333
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.667
7	Quadratic	Naïve	0.946	0.960	0.946	0.960	0.1586	0.1493	94.6309	96
		Decision Table	0.920	0.940	0.0916	0.940	0.2161	0.1872	91.9463	94
8	Cubic	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1549	94.6309	95.333
		Decision Table	0.920	0.900	0.0916	0.900	0.2161	0.2308	91.9463	90
9	Barycentric	Naïve	0.946	0.960	0.946	0.960	0.1586	0.1555	94.6309	96
		Decision Table	0.920	0.632	0.0916	0.627	0.2161	0.3569	91.9463	62.667
10	Krogh	Naïve	0.946	0.340	0.946	0.340	0.1586	0.6347	94.6309	34
		Decision Table	0.920	0.873	0.0916	0.873	0.2161	0.2508	91.9463	87.333
11	Polynomial	Naïve	0.946	0.960	0.946	0.960	0.1586	0.1493	94.6309	96
		Decision Table	0.920	0.940	0.0916	0.940	0.2161	0.1872	91.9463	94
12	Spline	Naïve	0.946	0.961	0.946	0.960	0.1586	0.1551	94.6309	96
		Decision Table	0.920	0.906	0.0916	0.906	0.2161	0.2348	91.9463	90.667
13	Piecewise Polynomial	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1476	94.6309	95.333
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.667
14	From derivatives	Naïve	0.946	0.953	0.946	0.953	0.1586	0.1476	94.6309	95.333
		Decision Table	0.920	0.927	0.0916	0.927	0.2161	0.2067	91.9463	92.667
15	Pchip	Naïve	0.946	0.960	0.946	0.960	0.1586	0.1423	94.6309	96
		Decision Table	0.920	0.933	0.0916	0.933	0.2161	0.2017	91.9463	93.333
16	Akima	Naïve	0.946	0.960	0.946	0.960	0.1586	0.1433	94.6309	96
		Decision Table	0.920	0.940	0.0916	0.940	0.2161	0.1898	91.9463	94

Precision value of each imputation algorithms: Two separate classifiers such as Naïve Bayes and decision tree are employed to examine the efficiency of the imputation methods. The precision accuracy of each method of imputation is compared with that of the before imputation dataset. In this figure, it is noted that the Krogh imputation method gives the lowest precision value for Naïve Bayes and decision tree, i.e. 0.34 and 0.873, respectively. Similarly, as compared to the other imputation methods, nearest, zero, quadratic and polynomial imputation techniques give maximum precision value.

Recall value of each imputation algorithms: Based on the classification algorithm evaluation parameters, the efficiency of the imputation algorithms will be evaluated. The recall value of each imputation process is compared with that of the dataset prior to the imputation. It is noted in Fig.2, that the lowest recall values are recorded by the Krogh method of imputation. After imputing the data values using the nearest, zero, quadratic and polynomial imputation methods, both Naïve Bayes and decision tree classifiers produce the highest recall value.

RMSE of each imputation algorithms: In contrast to the other algorithms, the Krogh imputation method produces the highest error value, which means that it gives the lowest efficiency. It is often assumed that, with all the imputation algorithms, the Naïve Bayes algorithm achieves better results than the decision tree.

Accuracy value of each imputation algorithms: Initially, the accuracy of classification is determined for the before and after imputation dataset. The efficiency of the imputation methods is measured using Naïve Bayes and Decision Table, the classification algorithm. With respect to Naïve Bayes, this produces the lowest classification accuracy for the Krogh imputation method.

7. CONCLUSION

This paper investigates sixteen different imputation methods. Some traditional classification assessment measures, such as precision, recall, RMSE, and overall accuracy, are employed for evaluating the efficiency of the methods. From the experimental results, it is noted that Krogh imputation methods provide lowest performance results. It is also noted that, nearest, zero, quadratic and polynomial imputation methods produce highest classification accuracy when compared to other imputation methods. This implies that, these methods impute best values which are nearest to the original data. Recently, several machine learning approaches and deep learning (DL) methods to determine the missing values were introduced. However, these methods require tremendous computational power and memory. They are extremely suitable for Big data analytics. In future, efficiency of DL strategies for imputing missing values are to be studied and implemented as an extension of this research work.

REFERENCES

[1] R.J. Little and D.B. Rubin, "*Statistical Analysis with Missing Data*", Wiley Press, 2019.

[2] J. Sim, J.S. Lee and O. Kwon, "Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications", *Mathematical problems in Engineering*, Vol. 2015, pp. 1-18, 2015.

[3] Peter Schmitt, Jonas Mandel and Mickael Guedj, "A Comparison of Six Methods for Missing Data Imputation", *Journal of Biometrics and Biostatistics*, Vol. 6, No. 1, pp. 1-6, 2015.

[4] Xueying Xu, Leizhen Xia, Qimeng Zhang, Shaoning Wu, Mingcheng Wu and Hongbo Liu, "The Ability of Different Imputation Methods for Missing Values in Mental Measurement Questionnaires", *BMC Medical Research Methodology*, Vol. 20, No. 42, pp. 1-16, 2020.

[5] R.M. Thomas, W. Bruin and P. Zhutovsky, "*Dealing with Missing Data, Small Sample Sizes, and Heterogeneity in Machine Learning Studies of Brain Disorders*", Academic Press, 2020.

[6] J.M. Jerez, I. Molina and P.J. García-Laencina, "Missing Data Imputation using Statistical and Machine Learning Methods in a Real Breast Cancer Problem", *Artificial Intelligence in Medicine*, Vol. 50, No. 2, pp. 105-115, 2010.

[7] Iris Data Set, Available at https://archive.ics.uci.edu/ml/datasets/Iris, Accessed at 2020.

[8] Credit Card Fraud, Available at https://www.kaggle.com/mlg-ulb/ creditcardfraud, Accessed at 2016.

[9] Wine Data, Available at https://www.kaggle.com/sgus1318/winedata, Accessed at 2020.

[10] The Boston Housing Dataset, Available at https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html, Accessed at 2020.

[11] Scipy, Available at https://www.scipy.org, Accessed at 2020.

[12] D.B. Rubin, "Inference and Missing Data", *Biometrika*, Vol. 63, No. 3, pp. 581-592, 1976.