第 49 卷 第 7 期
2022 年 7 月

湖南大学学报（自然科学版）
Journal of Hunan University (Natural Sciences)

Vol. 49 No. 7
July 2022

Open Access Article https://doi.org/10.55463/issn.1674-2974.49.7.4

Comparison of Seventeen Missing Value Imputation Techniques

Wafaa Mustafa Hameed[1,2*], Nzar A. Ali[2,3]

[1]Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani , 46001, Kurdistan Region, Iraq,
[2]Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq,
[3]Department of Statistics and informatics, University of Sulaimani, Sulaimani , 46001,Kurdistan Region, Iraq,

Abstract: Copious data are collected and put away each day. That information can be utilized to extricate curiously designs. However, the information that we collect is ordinarily inadequate. Presently, utilizing that information to extricate any data may allow deceiving comes about. Utilizing that, we pre-process the information to exterminate the variations from the norm. In case of a low rate of lost values, those occurrences can be overlooked, but, in the case of huge sums, overlooking them will not allow wanted results. Many lost spaces in a dataset could be a huge issue confronted by analysts because it can lead to numerous issues in quantitative investigations. So, performing any information mining procedures to extricate a little good data out of a dataset, a few pre-processings of information can be done to dodge such paradoxes and, in this manner, move forward the quality of information. For handling such lost values, numerous methods have been proposed since 1980. The best procedure is to disregard the records containing lost values. Another method is ascription, which includes supplanting those lost spaces with a few gauges by doing certain computations. This would increment the quality of information and would extemporize forecast comes about. This paper gives an audit on methods for handling lost information like median imputation (MDI), hot (cold) deck imputation, regression imputation, expectation maximization (EM), support vector machine imputation (SVMI), multivariate imputation by chained equation (MICE), SICE technique, reinforcement programming, nonparametric iterative imputation algorithms (NIIA), and multilayer perceptrons. This paper also explores some good options of methods to estimate missing values to be used by other researchers in this field of study. Also, it aims to help them to figure out what method is commonly used now. The overview may also provide insight into each method and its advantages and limitations to consider for future research in this field of study. It can be a baseline to answer the questions of which techniques have been used and which is the most popular.

Keywords: imputation, mean, mode, data.

十七种缺失值插补技术的比较

摘要：每天都会收集和存放大量数据。 这些信息可以用来解开奇怪的设计。但是，我们收集的信息通常是不充分的。目前，利用该信息提取任何数据可能会导致欺骗。利用它，我们对信息进行预处理以消除规范的变化。在价值损失率低的情况下，可以忽略这些事件，但在巨额资金的情况下，忽略它们不会得到想要的结果。数据集中许多丢失的空间可能是分析师面临的一个巨大问题，因为它可能导致定量调查中的许多问题。因此，执行任何信息挖掘程序以从数据集中提取一些好的数据，可以对信息进行一些预处理来避免这种悖论，并以这种方式提高信息的质量。为了处理此类丢失值，自 1980 年以来已经提出了许多方法。最好的程序是忽略包含丢失值的记录。另一种方法是归属，其中包括通过进行某些计算用一些量规替换那些丢失的空间。这将提高信息的质量，并会即兴预测的发生。本文对处理丢失信息

Received: April 7, 2022 / Revised: May 5, 2022 / Accepted: June 8, 2022 / Published: July 30, 2022

About the authors: Wafaa Mustafa Hameed, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq; Department of Computer Science, Cihan University Sulaymaniya, Sulaymaniya, Iraq; Nzar A. Ali, Asst. Prof. Dr., Department of Statistics and informatics, University of Sulaimani, Sulaimani, Iraq; Department of Computer Science, Cihan University Sulaymaniya, Sulaymaniya, Iraq

Corresponding author Wafaa Mustafa Hameed, wafaa.mustafa@sulicihan.edu.krd

的方法进行了审核，例如中值插补 (计量吸入器)、热 (冷) 甲板插补、回归插补、期望最大化 (电磁场)、支持向量机插补 (SVMI)、链式方程多元插补 (老鼠) 、SICE 技术、强化编程、非参数迭代插补算法 (NIIA) 和多层感知器。本文还探讨了一些很好的方法来估计缺失值，供该研究领域的其他研究人员使用。此外，它旨在帮助他们弄清楚现在常用的方法。该概述还可以深入了解每种方法及其优点和局限性，以供该研究领域的未来研究考虑。它可以作为基线来回答哪些技术已被使用以及哪些是最流行的问题。

关键词：插补，均值，模式，数据。

1. Introduction

Data mining has made great progress in the last few years, but the major challenge is missing data or values. Data mining is the field where experimental data sets are analyzed to discover interesting and potentially useful relationships. Missing data or values in datasets can affect the classifier's performance, leading to difficulty in extracting useful information from datasets. Much information is collected and stored every day. That data can be used to extract interesting patterns. The data that we collect is normally incomplete. Therefore, anyone wishing to use statistical data analysis or data cleaning of any kind will have problems with missing data. We still land on some missing attribute values in a characteristic dataset. People tend to leave the income field empty in surveys; for instance, participants sometimes have no information available or cannot answer the question. Much information may also be lost in the process of gathering data from multiple sources [1]. Using that data to collect some information can now yield misleading results. So, to eradicate the abnormalities, we need to pre-process the data before using it. These instances may be ignored in the case of a small percentage of missing values, but in the case of large amounts, ignoring them will not yield the desired outcome. A lot of missing spaces in a dataset is a big problem. Thus, data pre-processing can be done before performing any data mining techniques to extract some valuable information from a dataset to avoid such errors and thus improve data quality. Fittingly managing lost values is a vital and challenging assignment since it requires:

i) Careful examination of all occurrences of information to recognize the design of missingness within the data;

ii) Clear understanding of diverse ascription strategies. Several techniques have been proposed to handle such missing values since 1980 [2].

This report illustrates different types of missing values and the techniques used to handle them.

It is imperative to note the contrast between purged and lost values. Purged value implies no value can be doled out; lost value implies a real value for that variable exists but is not accessible or captured in the dataset for a few reasons. The information mineworker ought to separate purged estimation and lost estimation. Sometimes, both values will be treated as lost. Lost information may be due to gear glitch, conflicting with other information hence erased, information not entered due to misconception, and certain information may not be considered critical at the time of information collection. A few information mining calculations do not require substituting lost values as they are planned and created to handle lost values. However, a few information mining calculations cannot bargain with lost values. When utilizing any strategy for managing lost values, it is vital to understand why information is lost [2, 3].

2. Patterns of Missing Values

2.1. Missing Completely at Random (MCAR)

MCAR is the most elevated level of randomness. It suggests that the design of lost value is completely arbitrary and does not depend on any variable which may or may not be included within the examination [3]. It refers to information that does not depend on the interest variable or any other parameter observed in the dataset [4]. When missing values are distributed uniformly across all measurements, then we find the data to be completely randomly missing. For this reason, a quick check is to compare two pieces of data – one with missing observations and the other without missing observations. On a t-test, if there is no mean difference between the two data sets, we can assume that the data is MCAR [5]. Anything that is missing is best ignored, because this type of missing data is rarely found. For instance, if there is water damage to paper forms due to flooding before the data is entered [1, 2], or in a survey, if we get 5% responses missing randomly, it is MCAR [6, 7]. This type is defined by the following equation:

$$P(p_1|X, Y_{0,l}, Y_{m,l)} = f(l, X)$$

where f is a function, i.e., the missing data patterns are determined only by the covariate variables X. Note here that MARX is equivalent to MCAR if there are no covariates in the model [7, 8].

2.2. Missing at Random (MAR)

When a missed value does not depend upon any given or other missed value, it is assumed to be missing at random [8]. Often, the information may not be deliberately missing. If the data meets the requirement that messiness should not rely on *Xl*'s value after accounting for another parameter, we may find an *Xl* entry to be missing at random. Depressed people seem to have less income, for instance, and the reported income now depends on the factor depression. The percentage of missing data among depressed people will be high, as depressed people have lower incomes [1]. In this case, if we get 10% of the data missing for the male responses in a survey and 5% missing for the female responses, then the data is MAR [6]. This type is defined by the following equation:

$$P\left(p_l \middle| X, Y_{0,l}, Y_{m,l}\right) = f(l, X, Y_{0,l})$$

where *f* is a function, i.e., only the covariate variables *X* and the dependent variables have been observed have an effect on the patterns of missing data. Remember that if there is only one dependent variable *Y*, then there is only one missing sequence that does not include any observed dependent variables. For models with one dependent variable, MAR is equivalent to MARX [7].

2.3. Not Missing at Random (NMAR)

If the data are not missing at random or informatively, they are labeled "Not missing at random." Such a situation happens when the messiness process depends on the value of the missing data [4]. This type is defined by the following equation:

$$P\left(p_l \middle| X, y_{0,l}, Y_{m,l}\right) = f(l, X, Y_{0,l}, Y_{m,l})$$

where *f* is a function, i.e., all three types of variables affect the missing data patterns. It is well known how FIML (full information maximum likelihood) estimation performs under all of these conditions [7].

2.4. Missing in Cluster (MIC)

Data are often more missing in some attributes than in others. Also, the missing values in those attributes can be correlated. Using statistical techniques to reveal multi-attribute correlations of missing values is extremely difficult. In this pattern of missing values, data quality is less homogeneous than with MAR. The results of any analytical applications based on the complete data set should be cautious since the sample data are biased in the attributes with many missing values [7, 8].

2.5. Systematic Irregular Missing (SIM)

Data can be missing highly irregularly but systematically. There might be overly missing correlations between the attributes, but these correlations are extremely tiresome to analyze. SIM implies that the data with complete entities are unpredictably under-representative [7]. The data quality with this missing-value pattern is minimal homogeneous compared with that in MAR and less controllable than that with MIC. Applications of any analytical results based on the complete data set are highly questionable [9].

3. Methods of Handling Missing Data

Handling missing data can be done in two different strategies. The first strategy is simply ignoring missing values, and the second is considering the imputation of missing values.

3.1. Ignoring Missing Values

The missing data ignoring technique releases the state that contains missing data. They are mightily used for handling missing data. The earnest problem with this method is that it decreases the dataset size. This is convenient when the dataset has a small number of missing values. There are two common approaches for ignoring missing data:

3.1.1. Listwise Deletion

The complete case analysis approach excludes all observations with missing values for any variable of interest. This approach thus limits the analysis to those observations for which all values are observed. This technique is simple to use but causes a loss of huge data and precision, highly affects variability, and induces bias.

3.1.2. Pairwise Deletion

Pairwise deletion applies to all the cases we analyze, in which the variables of interest are present. It does not exclude the entire unit but uses as much data as possible from every unit. This technique is simple, keeping all available values, i.e., only missing values are deleted. However, it causes the loss of data, which is not a better solution than other methods. The sample size for each analysis is larger than the complete case analysis [2, 10].

3.2. Single Imputation

Single imputation procedures produce a precise value for a dataset's missing real value. This method necessitates a lower computing cost. Researchers have proposed a variety of single imputation strategies. The typical strategy is to analyze other responses and select the greatest possible response. The value can be calculated using the mean, median, or mode of the variable's available values. Single imputation can also be done using other methods, such as machine learning-based techniques. Imputed values are considered actual values in single imputation. Single imputation ignores the reality that no imputation method can guarantee the true value. Single imputation approaches ignore the imputed values' uncertainty. Instead, in future analysis, they recognize the imputed values as actual values [11, 12].

3.3. Multiple Imputations

Using distinct simulation models, multiple imputation methods yield several values for imputing single missing data. In addition, these strategies use imputed data variability to generate various credible responses. Multiple imputation methods are sophisticated, but unlike single imputation, they do not suffer from bias values. In multiple imputations, each missing data point is replaced with m values acquired through m iterations (where $m > 1$ and m is between 3 and 10) [6]. This technique uses a statistical method to deal with missing values. It performs through three stages:

- *Imputation:* generate *m* imputed data sets from a distribution which results in *m* complete data sets. The distribution can be different for each missing entry.
- *Analysis:* In this stage, *m* imputed data sets are analyzed. It is known as complete data analysis.
- *Pooling:* With simple rules, the output obtained after the data analysis is pooled to obtain the final result. The resulting inferences from this stage are statistically valid if the methods to create imputations are 'decent.'

The multiple imputation method is used to substitute missing values with possible solutions. The missing data set is transformed into complete data set by using suitable imputation methods that can then be analyzed by any standard analysis method.

Therefore multiple imputations have become popular in the handling of missing data. In this method, the process is repeated multiple times for all variables with missing values as the name indicates and then analyzed to combine m number of imputed data sets into one imputed data set [7, 11].

4. Missing Value Imputation Techniques

4.1. Mean Imputation

The mean imputation technique is used to calculate the mean of a missing value by using the corresponding attribute value. This technique is simple to use; it is built into most statistical packages, and it is faster compared with other techniques. It introduces good results with a small amount of data, but it gives poorer results for big data. This model is suitable for only MAR; it is not useful for MCAR [8, 13]:

$$\hat{x}_{ij} \sum_{i:x_{ij} \in c_k} \frac{x_{ij}}{n_k}$$

where n_k represents the number of non-missing values in the j-th feature of the k-th class, C_k, that are missing [7, 8].

4.2. Median Imputation (MDI)

Because of the effect of the presence of outliers on the mean, it seems better to use the median instead of the mean to ensure robustness. In this situation, the missing data are replaced by the median of all known values of that attribute in the class where the instance with the missing feature belongs. This method is also considered a choice when the distribution of the values is skewed. It should be assumed that the value x_{ij} of the k-th class, C_k, is missing. It can be replaced the following [7]:

$$\hat{x}_{ij} = median_{(i:x_{ij} \in c_k)}\{x_{ij}\}$$

4.3. Hot (Cold) Deck Imputation

In this technique, the missing value is replaced by the estimated distribution of the observed data. It can be implemented in the two following stages:

- Partitioning the data into clusters;
- Replacing missing values within a cluster.

The missing values are filled with the variable mean or mode of a cluster. Using a random Hot Deck, to substitute the missing value, an observed value of an attribute is selected randomly. The cold deck imputation method is similar to HD, but it takes data sources other than the current dataset. Using the hot deck, the missing values are imputed by realistically obtained values, which allows avoiding distortion in the distribution; however, little empirical work for accuracy estimation creates problems if there is no other sample with a close relation throughout the dataset [8, 10, 11].

4.4. K-Nearest Neighbor Imputation (KNNI)

KNNI involves specifying the similarity between two values and replacing the missing value with a similar one using Euclidean distance. The advantages of this approach are as follows:

- Suitable for datasets with both qualitative and quantitative attribute values;
- No need to create a predictive model for each attribute of missing data, which is helpful for multiple missing values.

An obstacle to this approach is when KNN looks for the most similar instances, and that the algorithm searches the entire dataset to find these similar instances [12].

4.5. Regression Imputation

This technique can be applied by using known values for the construction of the model and then calculating the regression between variable ends, using the model to calculate the missing values. The results from applying this technique are more accurate than those of mean imputation. The calculated data save deviations from mean and distribution shape, but the degree of freedom is distorted and may increase relationships [10].

4.6. Most Common Method

Instead of ignoring records with missing values, this method replaces missing spaces with certain values. For categorical attributes, it replaces the missing values

with the most common attribute value of the corresponding attribute or with the mode. The numerical attributes' missing values are replaced by the average or mean of the corresponding attribute [14, 15].

4.7. Expectation Maximization Imputation (EMI)

There are three types of clustering algorithms:

- *Hard clustering:* Clusters do not overlap, meaning that each element either belongs or does not belong to a cluster.
- *Soft clustering:* Clusters may overlap; that is, with different degrees of freedom, the elements can belong to multiple clusters at the same time.
- *Mixed models:* A probabilistic approach to doing soft clustering is used. Each cluster corresponds to a generative model that is typically Gaussian or multinomial [2].

4.8. Fuzzy K-Means Clustering Imputation (FKMI)

The membership function plays an important role in this technique. This function is assigned to every data object, and it describes the degree to which the data object belongs to the particular cluster. Data objects are not assigned to a concrete cluster, which is mentioned by the centroid of the cluster (i.e., the case of K-means); this is because the data have different degrees of membership with the complete body of K clusters. Unreferenced attributes for every incomplete part of the data are replaced by FKMI based on membership degrees and cluster centroid values. The advantages of this technique are that it gives the best outcome for overlapping data, it has better results than K-means imputation, and data objects may be part of more than one cluster center. However, the disadvantages are the high computation time and noise sensitivity (i.e., low or no membership for noisy objects) [10].

4.9. Support Vector Machine Imputation (SVMI)

This is a regression-based method used to impute missing values. It takes condition attributes (output) and decision attributes. Then, the SVMI is applied to predict values of missing condition attributes. The advantages of this technique are that it is efficient for high-dimensional spaces and shows efficient memory consumption. However, there is also a drawback of using this technique, which is that it shows poor performance if the number of samples is much less than the number of features [10, 16].

4.10. Most Common Imputation (MCI)

In this imputation technique, clusters are first formed by applying K-means clustering. As in KNN, in this technique, the nearest neighbors are found using clusters. All the instances in each cluster are referred to as nearest neighbors of each other. Then, missing values are imputed using the same method as is employed by the KNNI imputation method. This procedure is fast, and therefore, it is good for application to big datasets. This algorithm minimizes the intra-cluster variance. Here, too, the value of the K parameter is an important factor, and it is difficult to predict its value. In addition, this algorithm does not guarantee global minimum variance [2].

4.11. Multivariate Imputation by Chained Equation (MICE)

MICE expects information to be lost arbitrarily (damage), assuming that the likelihood of a lost variable depends on the watched information. In addition, MICE gives numerous values inside the input of one lost estimation by arranging relapse (or other reasonable) models, using its "method" parameter to make calculations. In MICE, each lost variable is treated as a dependent variable, and other information inside the record is treated as an independent variable. To begin with, MICE foresees lost information utilizing the winning information of other factors. At that point, it replaces lost values utilizing the expected values and makes a dataset called ascribed dataset. By cycle, it makes numerous ascribed datasets. Each dataset is, at that point, analyzed utilizing standard measurable investigation procedures, and numerous investigation results are given [17, 18].

4.12. SICE Technique

It pretends the probability of a missing variable depends on the observed data. It provides multiple values in the place of one missing value by creating a series of regression models. Each missing variable is treated as a dependent variable, and other data in the record are treated as an independent variable. It predicts missing data using the existing data of other variables. Then, it replaces missing values using the predicted values and creates an imputed dataset. It achieves a 20% higher F-measure for binary data imputation and 11% less error for numeric data imputations than its competitors with similar execution times. It imputes binary, ordinal, and numeric data. It performed better for the imputation of binary and numeric data. It is an excellent choice for missing data imputation, especially for massive datasets where MICE is impractical because of its complexity. However, it could not show better performance than MICE for the case of ordinal data [6].

4.13. Reinforcement Programming

It imputes missing data by learning a policy to impute data through an action-reward-based experience. It imputes missing values in a column by working only on the same column (similar to univariate single imputation) but imputes the missing values in the column with different values, thus keeping the variance in the imputed values. It is generally used as a dynamic approach for calculating missing values using machine learning approaches. It can converge and

solve imputation problems by using exploration and exploitation [19, 20].

4.14. Nonparametric Iterative Imputation Algorithms (NIIA)

It is an iterative imputing of the missing values in a dataset. It identifies some missing values and then computes the complete values used to estimate these incomplete values. Then, these missed values imputed are used for further analysis of other incomplete instances, and the process repeats until the values of the dataset are filled [21].

4.15. Multilayer Perceptrons

Multilayer perceptrons are the development technique using artificial neural networks. It runs on a multilayer and uses different learning processes to train the network [22].

4.16. Utilizing Uncertainty-Aware Predictors and Adversarial Learning MLP UA-Adv

It imputes the missing values so that the adversarial neural network cannot distinguish real values from imputed ones. Also, to account for the uncertainty of imputed values using confidence scores obtained from our adversarial module. The adversarial module aims to discriminate imputed values from real ones, estimate a missing entry with high accuracy, work well with small datasets, and utilize a novel adversarial strategy to estimate the uncertainty of imputed data [24].

Table 1 Review of different techniques to handle missing values

Techniques of Missing Value	Short Review	Advantage	Disadvantage
Leastwise Deletion	- Deletion of cases containing missing values (entire row is deleted) - High loss of information due to deletion of the entire row - High effect on variability - Loss of precision and inducing bias.	- Simple to use.	-Loss of huge data - Loss of precision - High effect on variability - Inducing bias
Pairwise Deletion	- Deletion of records only from a column containing missing values - Less loss of information by keeping all available values - Less effect on variability - Less loss of precision and inducing bias	- Simple to use. - Keeping all available values, i.e., only missing values are deleted	- Loss of data, - Not a better solution than other methods
Mean Imputation	- Replacing MVs with the arithmetic mean of data - Resultant mean and SD after the imputation may be much higher than those of the original - Not a good substitution method	- Simple to use - It is built into most of the statistical packages	- Resultant Mean and SD after the imputation may be much higher than those of the original - Affected by outliers
Median imputation (MDI)	- Missing data are replaced by the median of all known values of that attribute in the class where the features belong	- Good choice when the distribution of the values is skewed	- Not affected by outliers
Hot (cold) deck imputation	- Missing value is replaced by the estimated distribution of the observed data	- Avoiding distortion in the distribution	- Empirical for accuracy estimation - Creating a problem if any other sample has no close relation to the entire manner of the dataset.
Regression Imputation	- Replacing MVs with the values predicted from observed values Regression equation: $Y = \alpha 0 + \alpha 1\ X$	- A very easy and simple technique - Calculated data saves deviations from mean and distribution shape	- Only applicable if data is linearly separable, i.e., there is a linear relationship between the attributes - Degree of freedom is distorted and may raise the relationship
Expectation Maximization (EM)	- This iterative method finds maximum likelihood in two steps: expectation (E step) and maximization (M step). Iteration continues until the algorithm converges	- MVs are imputed by realistically obtained values, which allows avoiding distortion in the distribution	- A little empirical work for accuracy estimation - Creating a problem if any other sample has no close relation to the entire manner of the dataset
Fuzzy K-means clustering imputation (FKMI)	- FKMI substitutes unreferenced attributes for every uncompleted data based on membership degrees and cluster centroid values	- The best outcome for overlapping data - Better than k-means imputation - Data objects may be part of	- High computation time. - Noise sensitive, i.e., low or no membership

		more than one cluster center	degree for noisy objects
Support Vector Machine Imputation (SVMI)	- Takes condition attributes (here, decision attribute, i.e., output) and decision attributes (here, conditional attributes). SVMI then would be applied for predicting values of the missed condition attribute	- Efficient in large dimensional spaces - Efficient memory consumption	- Poor performance if the number of samples is much less than that of features
K-nearest neighbor imputation (KNN)	- Determining the similarity between two values and replacing the missing data with similar ones using Euclidean distance	- Avoiding distortion in the distribution as missing values are imputed by realistically obtained values - No need to create a predictive model - Helpful for multiple missing value	- Obstacle approach since the algorithm searches all of the data set - Prediction of the value of k is quite a difficult task.
Most Common Imputation (MCI)	- It replaces the missing value with the most common attribute or mode. - The numerical attribute's missing value is replaced by the average of the mean corresponding attribute	- Fast and good for applying to big datasets - Reducing the intra-cluster variance to a minimum	- Difficulty predicting the value if the number of the elements is too big - Does not guarantee global minimum variance
Multivariate Imputation by Chained Equation (MICE)	- It pretends the probability of a missing variable depends on the observed data - It provides multiple values in the place of one missing value by creating a series of regression models - Each missing variable is treated as a dependent variable, and other data in the record are treated as an independent variable - Predicting missing data using the existing data of other variables. Then, it replaces missing values using the predicted values and creates an imputed dataset	- *Flexibility:* Each variable can be modeled using a model tailored to its distribution - Managing imputation of variables defined only on a subset of the data - Incorporating variables that are functions of other variables - Does not require monotone missing-data patterns	- Lacking a theoretical rationale - Difficulties when specifying the different imputation models
SICE technique	It is an extension of the popular MICE algorithm. Two variants of SICE are presented: SICE-categorical and SICE-numeric to impute binary, ordinal, and numeric data. In addition, twelve existing performance algorithms are implemented to predict house price imputation methods and compare their performance with SICE.	- Achieves 20% higher F-measure for binary data imputation and 11% less error for numeric data imputations than its competitors with similar execution times. - Imputing binary, ordinal, and numeric data - Performs better for the imputation of binary and numeric data - An excellent choice for missing data imputation, especially for massive datasets, where MICE is impractical because of its complexity	- It could not show better performance than MICE for the case of ordinal data
Reinforcement Programming	- Imputing data through an action-reward-based experience - Imputing missing values in a column by working only on the same column but imputing the missing values in the column with different values, thus keeping the variance in the imputed values. It is generally used as a dynamic approach for calculating missing values using machine learning approaches.	- Performing well compared to other univariate single imputation and ML-based imputation approaches. Our approach produces an MAE of 0.0183 compared to 0.0271, 0.0201, and 0.0208 for mean, median, and the most frequent value-based univariate single imputations, respectively.	- Using numeric data variables only
Nonparametric Iterative Imputation algorithms (NIIA)	The NIIA method imputes each missing value several times until the algorithm converges. In the first iteration, all complete instances are used to estimate missing values. The information within incomplete instances is utilized since the second iteration. It selects some missing values (which can be all missing values in an attribute or a certain missing value). All complete instances are used to estimate the selected missing values. The information within incomplete instances is used from the second iteration onwards. The instances imputed in this imputation iteration are	- Easily captures the distribution of a dataset - High accuracy when datasets have a high missing ratio.	- Some datasets that the NIIA approach cannot perform that well

	treated as observed data (or complete instances) for imputing the remained missing attributes. This process repeats until all missing attributes are imputed.		
Multilayer Perceptrons	- Data sets with categorical variables - Based on neural networks - It runs on a multilayer and uses different learning processes to train the network	- For data sets with only quantitative variables - The analyzed models provide good and similar results.	- Requiring a high computational cost - Requiring training time
Utilizing uncertainty-aware predictors and adversarial learning MLP UA-Adv Imputer	- Training well with small datasets and utilizing a novel adversarial strategy to estimate the uncertainty of imputed data - Proposing a novel hybrid loss function that enforces the imputers to generate values for missing data that, on the one hand, obey the underlying data distribution so that it can confuse the well-trained adversarial module and, on the other hand, predict existing non-missing values accurately - The run time of the methods shows that they are efficient and have less execution time than peer imputer models - A very simple structure - Working with any feature type and small data	- An important role in the overall performance - Less runtime compared to OT imputers	- Training well with small datasets

Table 2 Comparing different techniques according to the dataset used in the application

Datasets	**Techniques**	**Commentary**	**Resource**
Iris	1. Predictive Mean Matching; 2. Multiple Random Forest Regression Imputation; 3. Multiple Bayesian Regression Imputation; 4. Multiple Classification and Regression Tree (CART); 5. Multiple Linear Regressions using Non-Bayesian Imputation; 6. Multiple Linear Regression with Bootstrap Imputation.	A comparison of different approaches of MICE methods on iris datasets. Efficiency gain with multiple imputations combined with Bayesian regression is that it can better use the available information by accommodating non-linearities.	[18]
1. Iris 2. Credits 3. Adults	1. Mean /Mode; 2. Hot Deck; 3. Expectation Maximization; 4. k-nearest neighbor.	In this paper, the authors compare C5.0 with this newly developed technique known as IITMV and show its performance on different data sets.	[23]
1. Cleveland 2. Heart 3. Zoo 4. Buhl1-300 5. Glass 6. Ionosphere 7. Iris 8. Pima 9. Sonar 10. WaveForm2 11. Wine 12. Hayes-Roth 13. Led7 14. Solar 15. Soybean	1. Mean/mode; 2. Regression; 3. Hot deck; 4. ANN.	The result shows that multilayer perceptrons (MLP) with different learning rules show better results with quantitative datasets than classical imputation methods. In this paper, the type of missing value is missing completely at random (MCAR).	[22]
1. Iris 2. E. coli 3. Breast cancer 1 4. Breast cancer 2	1. Mean 2. K-nearest neighbors (KNN) 3. Fuzzy K-means (FKM) 4. Singular value decomposition (SVD) 5. Bayesian principal component analysis (BPCA) 6. Multiple imputations by chained equations (MICE)	The results show that different techniques are best for different datasets and sizes. MICE is useful for small datasets, but, for big ones, BPCA and FKM are better.	[25]
1. Stock 2. UK statistic 3. Sales 4. Weather	Multiple Linear Regressions	The most basic and simple estimation method based on the autoregressive model. The method is claimed to solve missing values at several time points or columns of the dataset.	[26]

5. Conclusion

The no free lunch theorem, or conservation law, is the theoretical finding that a single learning algorithm is no better than the others. Every algorithm will perform equally well if the same performance measurements are used. However, in reality, this assumption is clearly wrong. Not all algorithms perform the same in a particular field. The best performing algorithms have been found to depend on the type of problem considered, the performance matrix used, and the characteristics of the dataset. For example, MICE and SICE techniques tend to perform well in the medical field, while decision tree algorithms tend to perform well in the sequential field. Therefore, algorithm performance varies from field to field. This study includes 17 techniques and outlines each method's advantages and limitations.

Overall, results were consistent across the different situations and performance measures are summarized in Table 1. The results first suggest that the most popular methods (i.e., mean, KNN, SVD, and MICE) are not necessarily the most efficient. Due to the simplicity of its methodology, the results for Mean are not surprising; this method does not consider the data's underlying correlation structure and thus performs poorly. KNN is a natural improvement over Mean in its exploitation of the observed data structure. MICE is a complex algorithm and its performance appears to be related to the size of the dataset: for small datasets, its performance is fast and efficient, but for large datasets, its performance decreases and becomes time-intensive. Multiple imputation, combined with Bayesian regression, gives better efficiency than other techniques, including Mean, KNN, SVD, and BPCA. However, this combined technique only considers the quality of imputation, which is based on classification methods that do not factor in execution times as exclusion criteria. Consequently, FKM may seem like the method of choice, but its execution time can slow down its performance or efficiency. This study thus considers BPCA as a more suitable solution to high-dimensional data. A summary of the different techniques and their applications to datasets can be found in Table 2. This paper's strength is that it covers most of the techniques that researchers will want to consider as a reference in selecting and combining the most suitable ones for imputing missing values. This study's limitation is that it provides only general conclusions without reference to the specific type of dataset, such as if it is a categorical or a numerical one. Future studies will cover the different types of datasets in more detail.

References

[1] DOSHI B. *Handling Missing Values in Data Mining*, 2010. https://pdfs.semanticscholar.org/3817/b208fe1f40891cc661ea0db80c8fccc56b70.pdf.

[2] GUPTA S., & GUPTA M. K. A Survey on Different Techniques for Handling Missing Values in Dataset. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2018, 4(1): 2456-3307. https://ijsrcseit.com/CSEIT411849

[3] JADHAV A., PRAMOD D., and RAMANATHAN K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 2019, 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

[4] SCHEFFER J. Dealing with Missing Data. *Research Letters in the Information and Mathematical Sciences*, 2002, 3(1): 153-160.

[5] PATIL D. V. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques. *IJCA Special Issue on "Evolutionary Computation for Optimization Techniques"*, 2010: 74-78. https://www.ijcaonline.org/ecot/number2/SPE140T.pdf

[6] KHAN S. I., & HOQUE A. S. M. L. SICE: an improved missing data imputation technique. *Journal of Big Data*, 2020, 7: 37. https://doi.org/10.1186/s40537-020-00313-w

[7] SINGH S. Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods. *Mathematical Journal of Interdisciplinary Sciences*, 2013, 1(2): 75–90. https://doi.org/10.15415/mjis.2013.12015

[8] PRATAMA I., PERMANASARI A. E., ARDIYANTO I., and INDRAYANI R. A review of missing values handling methods on time-series data. Proceedings of the International Conference on Information Technology Systems and Innovation, Bandung, 2016, pp. 1-6. https://doi.org/10.1109/ICITSI.2016.7858189

[9] WANG S., & WANG, H. *Mining Data Quality in Completeness*, 2007. http://mitiq.mit.edu/iciq/PDF/MINING%20DATA%20QUALITY%20IN%20COMPLETENESS.pdf

[10] VAISHNAV R. L., & PATEL K. M. Analysis of Various Techniques to Handling Missing Value in Dataset. *International Journal of Innovative and Emerging Research in Engineering*, 2015, 2(2).

[11] RAGHUNATH A. *Survey Sampling Theory and Applications.* Academic Press, Cambridge, 2017.

[12] REBECCA H., & GLAS C. A. W. Modelling Non-Ignorable Missing-Data Mechanisms with Item Response Theory Models. *The British Journal of Mathematical and Statistical Psychology*, 2005, 58(1): 1–17. https://doi.org/10.1348/000711005x47168

[13] PURI A., & GUPTA M. Review on Missing Value Imputation Techniques in Data Mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2017, 2(7): 35-40. https://doi.org/10.32628/CSEIT174405

[14] SASI KUMAR A., & SRIRAM AKRISHNA G. V. Internet of Things Based Clinical Decision Support System Using Data Mining Techniques. *Journal of Advanced Research in Dynamical & Control Systems*, 2018, 10(4): 132-139.

[15] GRZYMALA-BUSSE J. W., GOODWIN L. K., GRZYMALA-BUSSE W. J., and ZHENG X. Handling Missing Attribute Values in Preterm Birth Data Sets. In: ŚLĘZAK D., YAO J., PETERS J. F., ZIARKO W., and HU X. (eds.) *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. RSFDGrC 2005. Lecture Notes in Computer Science*, Vol. 3642. Springer, Berlin, Heidelberg, 2005: 342-351. https://doi.org/10.1007/11548706_36

[16] VAN BUUREN S., & GROOTHUIS-OUDSHOORN K. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 2010, 45(3): 1–67. https://doi.org/10.18637/jss.v045.i03
[17] HAN J., KAMBER M., and PEI J. *Data mining: Concepts and techniques*. 3rd ed. Morgan Kaufmann Publishers, San Francisco, California, 2012. https://sku.ac.ir/Datafiles/BookLibrary/43/Data-Mining-Concepts-and-Techniques-Han.pdf
[18] CHHABRA G., VASHISHT V., and RANJAN J. A Comparison of Multiple Imputation Methods for Data with Missing Values. *Indian Journal of Science and Technology*, 2017, 10(19): 1-7. https://dx.doi.org/10.17485/ijst/2017/v10i19/110646
[19] AWAN S. E., BENNAMOUN M., SOHEL F., SANFILIPPO F., and DWIVEDI G. A reinforcement learning-based approach for imputing missing data. *Neural Computing and Applications*, 2022, 34: 9701–9716. https://doi.org/10.1007/s00521-022-06958-3
[20] RACHMAWAN I. E. W., & BARAKBAH A. R. Optimization of Missing Value Imputation Using Reinforcement Programming. Proceedings of the International Electronics Symposium, Surabaya, 2015, pp. 128-133. https://doi.org/10.1109/ELECSYM.2015.7380828
[21] ZHANG S., JIN Z., and ZHU X. Missing data imputation by utilizing information within incomplete instances. *The Journal of Systems and Software*, 2011, 84(3): 452–459. https://doi.org/10.1016/j.jss.2010.11.887
[22] SILVA-RAMÍREZ E. L., PINO-MEJÍAS R., LÓPEZ-COELLO M., and CUBILES-DE-LA-VEGA M. D. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 2011, 24(1): 121-129. https://doi.org/10.1016/j.neunet.2010.09.008
[23] ALJUAID T., & SASI S. Intelligent Imputation Technique for Missing Values. Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Jaipur, 2016, pp. 2441-2445. https://doi.org/10.1109/ICACCI.2016.7732423
[24] HAMEED W. M., & ALI N. A. Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning. *Periodicals of Engineering and Natural Sciences*, 2022, 10(3): 350-367. http://dx.doi.org/10.21533/pen.v10i3.3110
[25] SCHMITT P., MANDEL J., and GUEDJ M. A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics and Biostatistics*, 2015, 6(1): 1000224. http://dx.doi.org/10.472/2155-6180.1000223
[26] SRIDEVI S., RAJARAM S., PARTHIBAN C., SIBIARASAN S., and SWADHIKAR C. Imputation for the analysis of missing values and prediction of time series data. Proceedings of the International Conference on Recent Trends in Information Technology, Chennai, 2011, pp. 1158–1163. https://doi.org/10.1109/ICRTIT.2011.5972466

参考文:

[1] DOSHI B. 处理数据挖掘中的缺失值，2010年。https://pdfs.semanticscholar.org/3817/b208fe1f40891cc661ea0db80c8fccc56b70.pdf。
[2] GUPTA S. 和 GUPTA M. K. 关于处理数据集中缺失值的不同技术的调查。 国际计算机科学、工程和信息技术科学研究杂志，2018, 4(1): 2456-3307. https://ijsrcseit.com/CSEIT411849
[3] JADHAV A.、PRAMOD D. 和 RAMANATHAN K. 数值数据集数据插补方法的性能比较。应用人工智能, 2019, 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138
[4] SCHEFFER J. 处理缺失数据。信息与数学科学研究快报，2002，3（1）：153-160。
[5] PATIL D. V. 使用基于遗传算法的技术对缺失数据进行多重插补。IJCA特刊 "优化技术的进化计算"，2010：74-78 。https://www.ijcaonline.org/ecot/number2/SPE140T.pdf
[6] KHAN S. I., & HOQUE A. S. M. L. SICE：一种改进的缺失数据插补技术。大数据杂志，2020，7: 37。https://doi.org/10.1186/s40537-020-00313-w
[7] SINGH S. 数据挖掘中缺失值的估计和插补方法的比较。跨学科科学数学杂志，2013，1（2）：75-90。https://doi.org/10.15415/mjis.2013.12015
[8] PRATAMA I.、PERMANASARI A. E.、ARDIYANTO I. 和 INDRAYANI R. 时间序列数据缺失值处理方法综述。信息技术系统与创新国际会议论文集，万隆，2016年，第 1-6 页。https://doi.org/10.1109/ICITSI.2016.7858189
[9] WANG S., & WANG, H. 在完整性方面挖掘数据质量, 2007. http://mitiq.mit.edu/iciq/PDF/MINING%20DATA%20QUALITY%20IN%20COMPLETENESS.pdf
[10] VAISHNAV R. L. 和 PATEL K. M. 分析处理数据集中缺失值的各种技术。国际工程创新与新兴研究杂志，2015，2(2)。
[11] RAGHUNATH A. 调查抽样理论和应用。学术出版社，剑桥，2017年。
[12] REBECCA H., & GLAS C. A. W. 使用项目响应理论模型建模不可忽略的缺失数据机制。 英国数学与统计心理学杂志，2005，58（1）：1-17。https://doi.org/10.1348/000711005x47168
[13] PURI A. 和 GUPTA M. 回顾数据挖掘中的缺失值插补技术。 国际计算机科学、工程与信息技术科学研究杂志，2017, 2(7): 35-40. https://doi.org/10.32628/CSEIT174405
[14] SASI KUMAR A., & SRIRAM AKRISHNA G. V. 使用数据挖掘技术的基于物联网的临床决策支持系统。动力与控制系统高级研究杂志, 2018, 10(4): 132-139.
[15] GRZYMALA-BUSSE J. W.、GOODWIN L. K.、GRZYMALA-BUSSE W. J. 和 ZHENG X. 处理早产数据集中缺失的属性值。在：ŚLĘZAK D.、YAO J.、PETERS J.F.、ZIARKO W. 和 HU X.（编辑）粗糙集、模糊集、数据挖掘和粒度计算。RSFDGrC 2005。计算机科学讲义，卷。 3642 施普林格，柏林，海德堡，2005：342-351。https://doi.org/10.1007/11548706_36
[16] VAN BUUREN S., & GROOTHUIS-OUDSHOORN K. 老鼠：R中链式方程的多元插补。统计软件杂志, 2010, 45(3): 1-67. https://doi.org/10.18637/jss.v045.i03
[17] HAN J.、KAMBER M. 和 PEI J. 数据挖掘：概念和技术。 第三版。 摩根考夫曼出版社，旧金山，加利福尼亚，2012 年。https://sku.ac.ir/Datafiles/BookLibrary/43/Data-Mining-Concepts-and-Techniques-Han.pdf
[18] CHHABRA G.、VASHISHT V. 和 RANJAN J. 缺失值数据的多重插补方法的比较。印度科技杂志，2017,

10(19): 1-7. https://dx.doi.org/10.17485/ijst/2017/v10i19/110646
[19] AWAN S. E.、BENNAMOUN M.、SOHEL F.、SANFILIPPO F. 和 DWIVEDI G. 一种基于强化学习的缺失数据估算方法。神经计算与应用，2022，34：9701-9716。https://doi.org/10.1007/s00521-022-06958-3
[20] RACHMAWAN I. E. W. 和 BARAKBAH A. R. 使用强化编程优化缺失值插补。国际电子研讨会论文集，泗水，2015 年，第 128-133 页。https://doi.org/10.1109/ELECSYM.2015.7380828
[21] ZHANG S., JIN Z., 和 ZHU X. 缺失数据插补利用不完整实例中的信息。系统与软件杂志，2011，84(3)：452–459。https://doi.org/10.1016/j.jss.2010.11.887
[22] SILVA-RAMÍREZ E. L.、PINO-MEJÍAS R.、LÓPEZ-COELLO M. 和 CUBILES-DE-LA-VEGA M. D. 使用多层感知器对完全缺失的随机数据进行缺失值插补。神经网络，2011，24（1）：121-129。https://doi.org/10.1016/j.neunet.2010.09.008
[23] ALJUAID T. 和 SASI S. 缺失值的智能插补技术。计算、通信和信息学进展国际会议论文集，斋浦尔，2016 年，第 2441-2445 页。https://doi.org/10.1109/ICACCI.2016.7732423
[24] HAMEED W. M., & ALI N. A. 利用不确定性感知预测器和对抗性学习提高插补技术性能。 工程与自然科学期刊，2022, 10(3): 350-367. http://dx.doi.org/10.21533/pen.v10i3.3110
[25] SCHMITT P.、MANDEL J. 和 GUEDJ M. 缺失数据插补的六种方法的比较。生物统计学与生物统计学杂志, 2015, 6(1): 1000224. http://dx.doi.org/10.472/2155-6180.1000223
[26] SRIDEVI S.、RAJARAM S.、PARTHIBAN C.、SIBIARASAN S. 和 SWADHIKAR C. 用于分析缺失值和预测时间序列数据的插补。信息技术最新趋势国际会议论文集，钦奈，2011 年，第 1158-1163 页。https://doi.org/10.1109/ICRTIT.2011.5972466