International Journal on Data Science and Technology
2022; 8(3): 48-54
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20220803.11
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Multiple Means Based on Multiple Clustering (MMMC) Imputation

Raed Rasheed, Wesam Ashour

Faculty of Engineering, Islamic University of Gaza, Gaza, Palestine

Email address:
rrasheed@iugaza.edu.ps (Raed Rasheed), washour@iugaza.edu.ps (Wesam Ashour)

To cite this article:
Raed Rasheed, Wesam Ashour. Multiple Means Based on Multiple Clustering (MMMC) Imputation. *International Journal on Data Science and Technology*. Vol. 8, No. 3, 2022, pp. 48-54. doi: 10.11648/j.ijdst.20220803.11

Received: August 18, 2022; **Accepted**: September 13, 2022; **Published**: October 11, 2022

Abstract: In recent years, data science has emerged as one of the most significant variables in both the realm of research and the realm of business potential. The existence of missing values is typically observed in real-world datasets, which might present a challenge. There are a variety of methods that can be used to deal with missing values. Imputation methods that are most commonly used to fill in missing data include the mean imputation, the median imputation, and the KNN imputation. The most significant drawback of the mean and mode methods is that, if there are a significant number of missing values, all of those values will be imputed with the same value. This will result in a change to the shape of the distribution, and the variance will be reduced when compared to its value before and after imputation. The more values that are absent, the greater the shrinking that will occur within the variance. In order to address this shortcoming of existing imputations, we have developed a brand-new imputation method. Multiple clustering's serve as the basis for multiple mean calculations (MMMC). When there are missing values in a dataset variable, MMMC imputation will substitute those values with several separate means rather than a single mean. The means obtained from the use of multiple clustering with the other variables contained in the dataset. The findings demonstrate that MMMC is superior to the other imputation strategies in a number of respects.

Keywords: Data Preprocessing, Missing Data, Data Imputation, Clustering

1. Introduction

In recent decades, data science has become increasingly prevalent across a variety of study fields, including but not limited to the fields of medicine, biology, psychology, and climate science [1-5]. They make use of data science in order to further their study or to pave the way for new lines of investigation. The existence of missing values is typically observed in real-world datasets, which might present a challenge. We might say that good outcomes are the result of rich data, whilst poor results are the result of poor data. Nearly all instances of missing can be broken down into one of three main categories [4]. Data that is missing totally at random, data that is missing at random, and/or data that is missing not at random are all referred to as "missing not at random" (MNAR). Data that are MCAR and MAR are referred to as ignorable missing data sometimes, while data that are MNAR are referred to as non-ignorable missing data [6]. There are a variety of methods that can be used to deal with missing values. The majority of these methods are used to impute data and determine the values that are most likely to have been assigned to missing data points in a dataset. These techniques span from more conventional approaches like deletion and single imputation to more contemporary and complex approaches like multiple imputation, model-based procedures, and machine learning techniques. Traditional approaches like these include:

1.1. Missing Data Categories

As mentioned early there are three categories of the missing data. These categories described as:

1) *Missing Completely randomly (MCAR):* An unrelated association between the variable with the missing value and the other variables contained inside the dataset. Examples of typical MCAR include the following: a gender or contact number for a customer's information is missing from the database; a tube containing a blood sample is dropped by accident and breaks; [1] a blood sample tube is broken. Another form of MCAR is the

unintentional loss of surveys. When manual data entry procedures are used in water distribution networks, there is a greater chance for human error, incorrect water reading measurements, errors in instrumentation, changes in experimental design, and other types of errors to occur. These are just some of the reasons why data may be deemed to have MCAR. The immediate consequence of this is that none of the information is there at all; more specifically, the chance that an observation isn't connected to the other variable [1].

2) *Missing randomly (MAR):* A dependent relationship exists between the missing value and other variables within the dataset, but the missing value itself is not dependent on the missing values of the target variable. Rather, the missing value depends on the observed values of other variables. MAR is illustrated by the situation in which a client's income level is unknown but can be guessed based on other characteristics such as the client's profession, experience, and qualification [7]. This is an illustration of the phenomenon.
3) *Missing Not every which way (MNAR):* A dependent relationship that exists between the values that are absent and, as a result, the nature of the variable. MNAR might take place, for instance, when the people living in a rural area decide not to take part in a very extensive survey [7].

1.2. Methods of Handling Missing Data

Two strategies were found to handling missing data [8]. The first strategy is ignoring missing values and the second strategy is to imputing missing values.

1) *Ignoring Missing Values:* The method of ignoring missing data consists of skipping through any samples that have information that is missing. This strategy is commonly employed and has a propensity to become the default option for dealing with missing data. The fact that the quantity of the dataset is decreased is the most significant drawback of using this method [9]. There are three primary methods for ignoring missing data: listwise deletion, pairwise deletion, and variable deletion. Listwise deletion removes samples from consideration when calculating a specific variable; pairwise deletion removes samples from consideration when calculating another variable; and variable deletion removes the variable entirely if it contains missing values. When doing an analysis using a complete sample, it is necessary to disregard any observations that lack values for a variable of interest. Because of this, the analysis is restricted to only those instances in which all of the values have been observed, which typically results in a skewed estimate and a loss of precision [10]. To conduct an analysis using pairwise deletion, we use each and every sample from the time period in which the variables of interest were gathered. It does not exclude the complete unit, but rather uses the greatest amount of data that can be obtained from each individual unit. The benefit of using this method is that it maintains the maximum amount of data that can be analyzed, despite the fact that some of its variables do not have any values. This method has the disadvantage of using a separate sample size for each of the variables that are being studied [10]. The sample size for each separate analysis is greater than the sample size for the overall analysis [9].
2) *Imputation of missing values:* A process that takes a missing value and fills it in with some possible other values [11]. It is the goal of the various imputation approaches to provide an accurate assessment of the parameters of the population. This is done to ensure that the power of knowledge mining and data analysis techniques is not diminished. The optimal technique to deal with the missing data is contingent on the quantity of data that is missing. Although there is no hard-and-fast rule to determine what percentage of missing data is unacceptable, it is usually preferable to try and do comparison of findings before and after imputation if quite 25% of the data is absent [9].

The remainder of the paper is structured as follows: Section 2 describes the literature review and previous work in this field, Section 3 presents the missing data imputation, Section 4 describes proposed imputation technique, Section 5 illustrates the experiments and evaluation of experimental results, Section 6 summarizes the overall conclusions of the paper.

2. Literature Review

Nishanth and Ravi [12] proposed a machine learning method known as a probabilistic neural network. This method yielded more effective results when compared to the mean, K-Nearest Neighbour (K-NN), Hot Deck (HD), and a decision tree strategy. In their research on the quality of the air, Gómez-Carracedo et al. [13] found that using multiple imputation approaches led to more variable outcomes than using single imputation methods did. Garca-Laencina et al. [14] investigated and contrasted a variety of pattern categorization strategies for the purpose of dealing with missing data. They presented a top-down pattern classification flowchart, in which they categorized the many different approaches to missing data into four different groups. They emphasized machine-based solutions and highlighted both the positives and negatives associated with using such solutions. In the study by Galán et al. [15], genetic algorithms were utilized to fill in the blanks of missing data in the knowledge and skills domain. In order to conduct research on the surface temperature, Wang and Chaib-draa [16] utilized a web Bayesian framework that included Gaussian Process Regression. The authors came to the conclusion that their method is superior to other Gaussian process methods such as sparse pseudo-input Gaussian process (SPGP) and sparse spectrum Gaussian process (SSGP). In an earlier study, Blend and Marwala [17] conducted an analysis of Human Immunodeficiency Virus (HIV) and acquired immunodeficiency syndrome (AIDS) data. As part of this

study, they contrasted an auto-associative neural network (AANN), a neuro-fuzzy (NF) system, and a hybrid system that combines AANN and NF. It was discovered that the AANN performed better than the NF system by a median of approximately 6%, however the hybrid technique performed approximately 16% better in terms of accuracy than either the solo AANN or NF systems. However, the computational efficiency of the hybrid system was decreased by fifty percent. A new tensor-based imputation method that supported canonical polyadic (CP) decomposition was reported by Dauwels et al. [18], and the authors compared it to mean imputation, regression imputation, and K-Nearest Neighbors. Their proposed method was evaluated using medical questionnaires, and the results demonstrated an improvement in imputation accuracy. It is well documented in the literature [19] that tensor-based imputation methods are frequently utilized approaches in traffic information systems and road sciences. Techniques based on tensor decomposition are also utilized in the fields of psychology, chemometrics, signal processing, bioinformatics, neuroscience, web mining, and computer vision [20].

3. Missing Data Imputation

Estimating missing data of an observation supported valid values of other variables is termed as Data Imputation [12]. Data imputation techniques are generally had two types Single Imputation and Multiple Imputation.

3.1. Single Imputation

Imputing one plausible value for each missing value of an any variable within the dataset so performing analysis as if all data were originally observed. There are several single data imputation methods:

Imputation with the constant: The constant is substituted for the values when it is lacking. In the event that the variable in question is categorical, it might replace all of the missing values with the value "Missing," "0," or "999" [7].

Mean Imputation: The most frequent approach is filling up data gaps through imputation. It does this by substituting the missing value with the sample mean, median, or mode, depending on how the information is distributed. This strategy is easy to understand and straightforward to put into practice. The most significant drawback of this method is that, if there are a significant number of missing values, all of those values will be imputed to have the same value. This will result in a change to the shape of the distribution, and the variance will decrease when compared to its value before and after imputation. The more values that are absent, the greater the shrinking that will occur within the variance. In many cases, the performance of this strategy can be somewhat enhanced by stratifying the data into subgroups.

Imputation with distributions: In the case of missing values, random values drawn from a known distribution are substituted for them. There is no change in the value that is imputed to the distribution.

Regression Imputation: This method of single imputation can be considered somewhat more complex than others. During this procedure, missing values are filled in with anticipated data based on the non-missing data of other variables that are supported by regression analysis. The linearity of the link between the qualities is assumed to exist inside this methodology. However, the majority of the time, the relationship isn't linear, and as a result, using regression to replace missing values will cause the model to be biased. The distribution shape can be preserved while using this method as opposed to the mean imputation method, which is one of the advantages of using regression imputation. It's possible that this strategy will give biased results, particularly when it comes to MNAR and MAR [10].

KNN Imputation: When there are missing values in a dataset, such values can be "imputed" by copying values from other records within the same dataset that are similar. A distance function is used to determine how similar the two characteristics are to one another. The development of a predictive model for each and every property is not only unnecessary, but it also comes with a number of drawbacks. The analysis of a huge dataset requires a significant amount of time. Additionally crucial is the selection of the k value.

3.2. Multiple Imputation

When using single imputation methods, it is a presumption that the value obtained from a single imputation is the correct one, and the precision is exaggerated. On the other hand, one can never know with complete certainty whether or not imputed values are correct. Because of this, the uncertainty around these imputed values needs to be factored into the procedures for missing data [21]. Therefore, in multiple imputation, rather of replacing a single value for every missing observation, it replaces multiple plausible values to express uncertainty about the correct values to impute. This is done to account for the fact that there may be a range of possible values. As a result, the Multiple Imputation approach produces m distinct complete datasets that contain both observed and imputed values. The same three steps have been used in every multiple imputation approach:

(1) Imputation: The process of imputation of missing data is quite similar to that of single imputation; however, the imputed values are generated "m" times rather than just once. Therefore, there may be m distinct complete datasets when imputation is performed.
(2) Analysis of every dataset: After doing imputation and obtaining "m" distinct datasets, an analysis is performed on each of the "m" datasets.
(3) Pooling: In the end, the findings collected from each of the datasets that were analyzed are compiled.

4. Proposed Imputation Technique

As was previously stated, the most significant drawback associated with the mean and the other imputation procedures that were presented is that each missing value in the variable

would be replaced with the same value. When this is done, the shape of the distribution shifts, and the variance experiences a reduction when it is compared to both before and after the imputation. Therefore, the primary goal of the method that has been provided is to locate a distinct mean value for each cluster contained within the variable. The formation of these clusters was determined by the imputed variable in addition to each other variable taken individually. The MMMC imputation that was proposed is broken up into five stages. The first step is to arrive at the correlation matrix for the dataset by calculation. The second phase involves utilizing the k-means clustering algorithm to group the missing value variable with each other variable in a separate fashion. Third, the mean of each cluster is determined by calculating the weighted correlation of the variables. The next thing that was determined was the mean of the means for each cluster over all of the iterations. As a last step, calculate the mean of the means of each cluster and impute it into the value that is absent within the same cluster. After the correlation matrix of the dataset has been calculated, the further steps will be described as follows:

Cluster phase: If the dataset with missing values consists of n variables, we choose each variable contains missing values v_i and create sub datasets $\{v_i, v_j\}$ where $i, j \in \{1, 2,... n\}$ and $j \neq i$ means j denoted all other variables separately. Then, performs k-means clustering with k = m to all sub dataset $\{v_i, v_j\}$.

Calculate clusters mean phase: In this phase the means of V_i in the sub dataset $\{v_i, v_j\}$ within the cluster k calculated and denoted by $means_{j,k}$ where j is the variable index and k is the cluster index where $k \in \{1, 2,... m\}$. After this phase we will have $((n-1) \times m)$ means where n is the number of variables in the original dataset and m is the number of clusters used in k-mean.

Calculate mean of means phase: After finding $((n-1) \times m)$ means we calculate the mean of each cluster $mean_k$ where $k \in \{1, 2,... m\}$.

Data imputation phase: In this phase for each missing value in V_i imputed with $mean_k$ within the same cluster k.

Figures 1 and 2 describe the proposed MMMC imputation in details where it received the original dataset with missing values and return the imputed dataset. It starts with correlation matrix calculation then each $variable_i$ in the dataset and let i equal to index of the variable needs to impute. After that, for each $variable_j$ in the dataset if the index of $variable_i$ not equal to index of $variable_j$ perform k-mean clustering and produce set of clusters $clusters_j$ for $variable_i$ associated with $variable_j$. Then, for each cluster $cluster_{j,k}$ in the clusters $clusters_j$ calculate the mean of values in $cluster_{j,k}$ and calculate the weights for the $variable_i$ and $variable_j$ correlation. Repeat the previous steps for all variables having missing values in the dataset. Finally, for each mean in the calculated means $means_{j,k}$ calculate the mean of cluster means multiplied by its weight then replace all missing values in the current variable within the cluster with mean have been calculated. Then do all the steps again for all variables with missing values and return the imputed dataset.

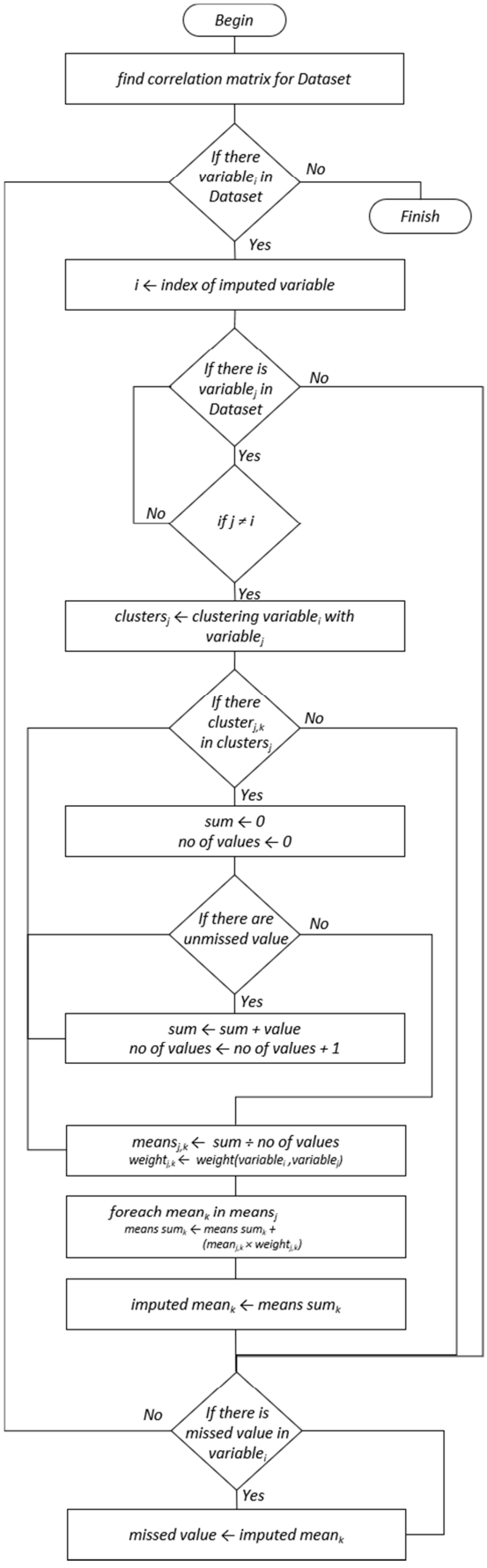

Figure 1. *MMMC imputation flowchart.*

```
Algorithm 1:  MMMC
Dataset ← { The dataset with missing values with n variables }
find correlation matrix for Dataset
foreach variable_i in Dataset
    i ← index of imputed variable
    foreach variable_j in Dataset
        if j ≠ i then
            clusters_i ← clustering variable_i with variable_i
            foreach cluster_j,k in clusters_j
                sum ← 0
                no of values ← 0
                foreach unmissed value in variable_i in cluster_j,k
                    sum ← sum + value
                    no of values ← no of values + 1
                end foreach
                means_j,k ←  sum ÷ no of values
                weight_j,k ←  weight(variable_i, variable_j)
            end foreach
        end if
    end foreach
    foreach means_j in means_j,k
        means sum_k ← 0
        foreach mean_k in means_j
            means sum_k ← means sum_k + (mean_j,k × weight_j,k)
        end foreach
        imputed mean_k ← means sum_k
    end foreach
    foreach missed value in variable_i in cluster_j,k
        missed value ← imputed mean_k
    end foreach
end foreach
return Imputed Dataset
```

Figure 2. *Multiple Means Based on Multiple Clustering Algorithm.*

5. Experiments and Evaluation

5.1. Methodology

The performance of the suggested MMMC imputation method is compared with the performance of other methods within this subsection. In this study, we analyze the performance of several imputation methods, including single imputation methods, the Mean Imputation, the Median Imputation, the KNN Imputation, and the suggested MMMC Imputation. These methods of imputation are exclusively utilized when working with numeric datasets. The datasets that were used for this paper might be found in the UCI Machine Learning Repository [22].

The outline of every dataset is given in Table 1.

The five different datasets described in Table 1 obtained from UCI machine learning repository. Then we injected varying randomly percentage (10%, 20%, 30%, 40%, and 50%) of missing values in each dataset. The missing values are then imputed using imputation methods namely mean imputation, median imputation, KNN imputation, and proposed MMMC imputation. Mean and Median imputation is done by using mean and median of pandas package in Python. KNN imputation and MMMC imputation are done by using sklearn, pandas, and numpy packages in Python.

Table 1. *Description of the dataset used [22].*

#	Dataset	Dataset Description	Ins.	Att.
1	Glass Identification	Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis, in determining whether the glass was a type of "float" glass or not	214	11
2	Indian Liver Patient Dataset	This dataset contains 416 liver patient records and 167 nonliver patient records. The dataset was collected from north east of Andhra Pradesh, India.	583	10
3	Seeds Dataset	Measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment.	210	7
4	Breast Cancer Coimbra	Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.	643	11
5	Breast Cancer Wisconsin (Prognostic)	Prognostic Wisconsin Breast Cancer Database	116	10

There are several different ways to quantify the performance of an imputation method, including accuracy, relative accuracy, MAE (mean absolute error), and root mean squared error (RMSE). These can be used to evaluate the effectiveness of an imputation approach (root mean square error). RMSE is one of the performance indicators that is considered to be the most representative and is utilized extensively in imputation research [23]. The performance was evaluated based on the average of the Normalized RMSE (NRMSE) values. Because scales are unique to each feature of the dataset, the normalized root-mean-square error (RMSE) is the statistic that should be used [9]. The following is an explanation of the formula that is used to calculate NRMSE and Mean NRMSE.

$$\text{NRMSE} = \sqrt{\frac{mean((original\ value - imputed\ value)^2)}{\max(\text{original value}) - \min(\text{original value})}}$$

$$Mean\ \text{NRMSE} = \frac{\sum_{i=1}^{n} NRMSE}{n}$$

Number of variables in the dataset denoted by n.

For the purpose of this study, we have used Python and PyCharm 2021.2.2 (Community Edition) as a tool for data manipulation, data imputation, and analyzing performance of different imputation methods.

5.2. Results

This section describes the result of the evaluation of the imputation methods namely mean, median, KNN, and MVOC. We switch missing value by one value without taking into consideration uncertainty of the imputation.

Calculated results for the Mean NRMSE are presented in the accompanying Tables 2-6 for each dataset, with the results broken down according to the percentage of imputed data and the type of imputation technique used. In this table, the values in each column represent the proportion of missing data, and the values in each row represent the imputation technique that was applied to the missing data. The imputation strategies NRMSE for the Glass Identification dataset are displayed in Table 2, together with all possible personates of randomly injected missing value pairs. Which suggests that the suggested imputation method is about equivalent to the average of mean and median imputation, but it performs marginally better than KNN imputation in forty percent of cases where data are absent. The results presented in Table 3 show that the proposed imputation technique is the most successful solution for addressing all of the missing value percentages. According to Table 4, the suggested imputation method is superior to the KNN imputation method in both 40% and 50% of cases where values are absent. The proposed MMMC imputation method outperforms previous imputation techniques, as shown by Tables 5 and 6, respectively.

Table 2. *Mean NRMSE for Glass Identification dataset.*

Method	10%	20%	30%	40%	50%	Avg.
Mean	0.124653	0.147063	0.195916	0.215429	0.27113	0.19083
Median	0.130787	0.154562	0.206467	0.226851	0.285639	0.20086
KNN	0.072811	0.106694	0.152762	0.186952	0.25498	0.15483
MMMC	0.097336	0.123153	0.172546	0.19713	0.254267	0.16888

Table 3. *Mean NRMSE for Indian Liver Patient dataset.*

Method	10%	20%	30%	40%	50%	Avg.
Mean	0.513109	0.849228	0.894408	0.978381	1.189	0.88482
Median	0.539853	0.886486	0.929974	1.02071	1.23229	0.92186
KNN	0.499625	0.855539	0.947905	0.97274	1.26605	0.90837
MMMC	0.429588	0.763264	0.784845	0.889478	1.11171	0.79577

Table 4. *Mean NRMSE for Seeds dataset.*

Method	10%	20%	30%	40%	50%	Avg.
Mean	0.125373	0.196374	0.24171	0.275423	0.320415	0.23185
Median	0.123409	0.20123	0.248602	0.282996	0.328716	0.23699
KNN	0.046867	0.073124	0.130682	0.188239	0.248212	0.13742
MMMC	0.059424	0.104526	0.141623	0.180614	0.240271	0.14529

Table 5. *Mean NRMSE for Breast Cancer Coimbra dataset.*

Method	10%	20%	30%	40%	50%	Avg.
Mean	0.631758	0.856069	1.13785	1.3676	1.53841	1.10633
Median	0.641561	0.934067	1.14517	1.42666	1.57031	1.14355
KNN	0.592852	0.850684	1.0959	1.32645	1.55444	1.08406
MMMC	0.584378	0.79687	1.11827	1.29956	1.48164	1.05614

Table 6. *Mean NRMSE for Seeds dataset.*

Method	10%	20%	30%	40%	50%	Avg.
Mean	12.3393	3.73631	4.41683	13.1527	15.1027	9.74956
Median	12.3258	4.10635	4.80912	13.3004	15.1719	9.94271
KNN	12.3643	3.98121	5.46716	13.2615	15.1682	10.04847
MMMC	12.2589	3.62034	4.29035	13.1344	15.1642	9.69363

6. Conclusion

In this paper, we propose a novel method of imputation called Multiple Means Based on Multiple Clustering (MMMC). When there are missing values in a dataset variable, MMMC imputation will substitute those values with several separate means rather than a single mean. The means that were created by utilizing multiple clustering with the other variables in the dataset. In order to evaluate the effectiveness of the proposed imputation method, five distinct datasets were taken from the machine learning repository at UCI. As a measure of overall performance, we have relied on the Mean of Normalized RMSE, abbreviated as NRMSE. The bulk of the datasets that were evaluated show significant improvement when using the proposed MMMC imputation technique to impute missing values. The results show that MMMC imputation method is more efficient than other methods in most experiments.

References

[1] A. V. D. H. G. S. T. a. M. Donders, "A gentle introduction to imputation of missing values," Journal of clinical epidemiology, vol. 59, pp. 1087-1091, 2006.

[2] O. C. M. S. G. B. P. H. T. T. R. B. D. a. A. R. Troyanskaya, "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, pp. 520-525, 2001.

[3] P. a. H. J. Flyer, "Missing data in confirmatory clinical trials," Journal of biopharmaceutical statistics, vol. 19, pp. 969-979, 2009.

[4] A. a. E. C. Baraldi, "An introduction to modern missing data analyses," Journal of school psychology, vol. 48, pp. 5-37, 2010.

[5] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," Journal of climate, vol. 14, pp. 853-871, 2001.

[6] R. J. A. L. a. D. B. Rubin, "Statistical Analysis with Missing Data".

[7] M. A.-M. A. a. P. P. Osman, "A survey on data imputation techniques: Water distribution system as a use case," IEEE Access, vol. 6, pp. 63279-63291, 2018.

[8] J. P. J. a. K. M. Han, Data mining: concepts and techniques, Elsevier, 2011.

[9] A. P. D. a. R. K. Jadhav, "Comparison of performance of data imputation methods for numeric dataset," Applied Artificial Intelligence, vol. 33, pp. 913-933, 2019.

[10] J. a. G. J. Schafer, "Missing data: our view of the state of the art," Psychological methods, vol. 7, p. 147, 2002.

[11] D. Rubin, "Inference and missing data," Biometrika, vol. 63, pp. 581-592, 1976.

[12] K. a. R. V. Nishanth, "Probabilistic neural network based categorical data imputation," Neurocomputing, vol. 218, pp. 17-25, 2016.

[13] M. A. J. L.-M. P. M. S. a. P. D. Gómez-Carracedo, "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets," Chemometrics and Intelligent Laboratory Systems, vol. 134, pp. 23-33, 2014.

[14] P. S.-G. J. a. F.-V. A. García-Laencina, "Pattern classification with missing data: a review," Neural Computing and Applications, vol. 19, pp. 263-282, 2010.

[15] C. L. F. d. C. J. F. a. S. A. Galán, "Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions," Journal of Computational and Applied Mathematics, vol. 311, pp. 704-717, 2017.

[16] Y. a. C.-d. B. Wang, "An online Bayesian filtering framework for Gaussian process regression: Application to global surface temperature analysis," Expert Systems with Applications, vol. 67, pp. 285-295, 2017.

[17] D. a. M. T. Blend, "Comparison of data imputation techniques and their impact," arXiv preprint arXiv: 0812. 1539, 2008.

[18] J. G. L. E. A. a. P. L. Dauwels, "Tensor factorization for missing data imputation in medical questionnaires," in IEEE, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[19] H. F. G. F. J. W. W. Z. Y. a. L. F. Tan, "A tensor-based method for missing traffic data completion," Transportation Research Part C: Emerging Technologies, vol. 28, pp. 15-27, 2013.

[20] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, pp. 24-40, 2011.

[21] R. a. R. D. Little, "The analysis of social science data with missing values," Sociological Methods & Research, vol. 18, pp. 292-326, 1989.

[22] M. Lichman, "UCI Machine Learning Repository," University of California, School of Information and Computer Science, 2013. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed 24 1 2022].

[23] P. M. J. a. G. M. Schmitt, "A comparison of six methods for missing data imputation," Journal of Biometrics & Biostatistics, vol. 6, p. 1, 2015.