A Novel Algorithm for Imputing the Missing Values in Incomplete Datasets

Hutashan Vishal Bhagat
Sant Longowal Institute of Engineering and Technology
Manminder Singh (manminderfzr@yahoo.com)
Sant Longowal Institute of Engineering and Technology

Research Article

Keywords: Imputation, Imputing values, Missingness Mechanisms, Missing Values, Data Missingness, Data Imputation Model, Incomplete datasets, Root Mean Square Error

Posted Date: June 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1729251/v1

Abstract

In today's world, we completely rely on digital devices to collect data; a failure in such digital devices may result in huge information loss thereby making data mining a more tedious job for a Data Analyst. Missingness to a greater extent in a dataset subsequently comes out with inappropriate results and incomplete data analysis. Therefore, a need to develop an algorithm that can predict the missing values efficiently and accurately. This research paper proposes a novel splitting-based IMV-RE (**I**mputing the **M**issing **V**alues in **R**eal-Time **E**nvironment) algorithm to impute different missing values within a dataset. In the proposed IMV-RE algorithm, an upper limit is set for every class containing missing values that assist the algorithm to predict the missing values more accurately. The experimentation is performed on ten benchmark datasets that include completely numerical values as well as mixed data. Comparative experimental analysis indicates that the proposed IMV-RE algorithm outperforms the existing techniques in sensitivity to Accuracy, Root Mean Square Error (RMSE) and Coefficient of Determination (R^2).

1. Introduction

We are living in a digital world where information can easily be acquired with the help of smart devices, sensors etc. The advent of IoT makes it possible to collect data without the physical intervention of humans. But sometimes, a failure in such devices may result in data loss and hence, affects the subsequent in-depth analysis and data interpretation that provides erroneous results. The rapid increase in the size of datasets has led to emerging of various data mining techniques. To ensure data mining results to be effective and valuable, data scientists must ensure the quality of the collected data. In real-time scenarios, it is usually the case that collected datasets for data analysis may contain some missing values [1]. Therefore, it is not possible for most of the data mining algorithms to directly handle these incomplete datasets. The simplest solution to such problems is case deletion which means removing data having missing values. However, case deletion can be appropriate if the missing rate is small, e.g., 5% and if the missing rate is somehow larger, say 25%, then using 75% of the original dataset might be insufficient to completely reflect the real-world problem, which could also affect the mining results [2]. To ensure the quality of collected datasets, data scientists first pre-process the datasets [3].

In concern to the relationship between the missing data and other data values of the variables in a dataset, missing data mechanism can be assorted as Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) [4, 5]. If the probability of missing data depends upon the observed responses and there exist no relation among the missing values itself, the missing data mechanism is referred to as MAR. Hence, in MAR the reason for missingness in a feature Q of a dataset mainly depends upon the rest of the features within the dataset rather than the Q itself. If there is no relation between the missing values and the set of observed responses, the missingness mechanism is said to be MCAR. In MCAR, the reason for missing in a feature Q depends neither on the other features within the dataset nor on Q itself. If the probability of the missingness in a feature Q depends either on Q itself or on the other features that also contains missing values, the missing mechanism is referred to as MNAR.

Missing value imputation (MVI) techniques provide the best solutions to impute the missing values within the datasets. MVI techniques can be categorized into two categories- MVI techniques based on Statistical Methods (Mean, Mode, Median and Regression) and MVI techniques based on Machine Learning (Neural Networks, Support Vector Machines and Deep Learning Models) [6]. Although, machine learning techniques being complex in nature produce better imputation results, yet they are computationally expensive. For high dimensional datasets, such techniques show high computation time as compared to statistical methods because of which they are not deployed to mission-critical systems [7].

Being inspired from the lower complexity and computation time of statistical MVI techniques, this paper proposed a novel splitting-based imputation approach, namely, IMV-RE. To justify the superiority of the proposed IMV-RE approach, ten benchmark datasets containing both numerical as well as mixed values with varying dimensionality take into consideration and is compared with five existing imputation techniques. Classification accuracy, RMSE and Coefficient of determination (R^2) are considered as the performance metrics with which the proposed IMV-RE approach is evaluated.

2. Literature Review

Missingness within a dataset is a very common problem in statistical analysis. Researchers have come out with numerous MVI techniques depending upon the missing rate, data characteristics and pattern and data correlation. This section gives a brief description of the recent MVI techniques.

Regularized Expectation Maximization (EM) algorithm for imputing missing value which is based on the iterated analysis of the linear regressions among the missing and non-missing variables is proposed in [8]. The experimental test is carried out using regularized EM technique over the climate data. The regularized EM technique is further enhanced in [9] and two novel methods kEMI and kEMI$^+$ are proposed by the authors. The proposed techniques are based on the information fusion mechanism that uses Dempster-Shafer fusion to fuse the most appropriate estimates. Due to the high computation time of the missForest [10], the authors in [11] proposed a new technique known as mForest that can achieve ten times less computation time than missForest. A Column-wise Guided Data Imputation (cGDI) technique that divides the complete samples from the incomplete samples and selects the most suitable imputation method separately for each feature is proposed in [12].

In [13], the authors proposed a CBC-IM (Class Based Clustering approach for Imputation) technique for missing values imputation especially for medical datasets. CBC-IM technique first partitioned the dataset into complete and incomplete variables and then uses the Euclidean distance and Fuzzy measure to find out the similarity between the two records. Another similar approach known as CCMVI (Class Center based Missing Value Imputation) is proposed in [7]. CCMVI is a two-step process that first defines the threshold value for every class containing missing values and then finds the appropriate estimate from the complete samples considering the threshold values. An approach based on the linear regression technique known as CLR (Cumulative Linear Regression) is proposed in [14]. The incomplete variables are first cumulated and incorporated in the linear regression equation to replace the missing values in the next incomplete variable.

The KNNI (K Nearest Neighbor Imputation) is the most extensively used MVI technique. In [15], the authors proposed a feature weighted grey KNNI technique that uses the combination of relevant feature information and grey rational based k nearest neighbors to impute the missing values. Another approach that makes combine use of Multi-Layer Perceptron and KNN to impute the multiple missing values simultaneously is proposed in [16]. This combination results in an increase in the performance with an increase in the time complexity of the algorithm. In [17], the authors proposed two techniques CBRL (Cumulative Bayesian Ridge with Less NaN) and CBRC (Cumulative Bayesian Ridge with high Correlation). CBRL has used the most appropriate features within the dataset that contain a lesser number of missing values whereas CBRC is used to select the best feature that gives a high correlation with the target feature.

Correlation Maximization-based Imputation Methods (CMIM) are proposed in [18] that first find the highly correlated segments of the data and use linear regression estimator to impute the missing values. The authors use a two-step method to handle large missing gaps in [19] known as Ratio-Based Imputation (RBI). In RBI the MVI is done by using machine learning models whereas the analysis is done by data fusion technique in CPS (Cyber-Physical Systems) datasets. Considering the missing values in a dataset an optimization problem, the authors in [20] proposed an optimal method known as BNII. The BNII technique is a two-stage approach: firstly, using the Bayesian Network relationship among different attributes is calculated and secondly, in an iterative manner imputation is done till local maximum posterior probability is reached. A novel architecture based on Particle Swarm Optimization (PSO) for cleaning the data and then utilizing the K-means to calculate the fitness value as well as to narrow down the search space is proposed in [21]. The ontology makes PSO replace missing values more accurately but in worst scenarios, this approach has a time complexity of $\Theta(n^3)$.

A novel method known as Modulo 9 proposed in [22] impute the missing values within the interval of [0–9] and then use congruency with addition and multiplication to make an appropriate estimation. The authors compared the Modulo 9 approach with the eleven robust MVI techniques and outperform all of them. Considering generalization for datasets that contain mixed-type data, the authors in [23] proposed a tuple-oriented region splitting imputation technique known as RESI (Region-Splitting Imputation). RESI technique first uses the entropy weight method to assign weights to the attributes and split the data into complete and incomplete subsets based on their integrity rate. The model is trained over a complete subset that iteratively imputes the next incomplete subset.

From the last few years, researchers have come out with numerous imputation techniques based on meta-heuristic techniques ([24–27],association mining rules[28, 29], dynamic programming techniques[30] and various such hybrid techniques [31–34] to efficiently impute the missing values. The authors in [35] analyzed that there is no such MVI technique that could be considered as a master technique for distinct problems. The key factors that can influence the performance of MVI techniques are the data distribution within the dataset, the missingness rate and characteristics of a dataset [36].

The main contribution of this paper is to overcome the limitations of state-of-the-art imputation techniques based on the statistical methods. The authors proposed a CBC-IM [13] (Class Based Clustering approach for Imputation) technique for missing values imputation especially for medical datasets. CBC-IM technique first partitioned the dataset into complete and incomplete variables and then uses the Euclidean distance and Fuzzy measure to find out the similarity between the two records. Another similar approach known as CCMVI is proposed [7]. CCMVI is a two-step process that first defines the threshold value for every class containing missing values and then finds the appropriate estimate from the complete samples considering the threshold values. Although, the CCMVI and IM-CBC techniques have less time complexity, yet their dependency on complete samples make them unsuitable for the imputation especially when a dataset has high missingness rate or when there is an unavailability of at least one data sample without missing values within a class. The major goal of any MVI technique is to replace the missing values in such a manner that the overall data integrity, structure, and trends of the data are maintained. The proposed IMV-RE MVI technique shows better performance than the commonly used existing single as well as multiple imputation techniques. Unlike the CCMVI technique, the IMV-RE technique successfully imputes the missing values even the missing rate is high. Unlike the missForest [10], MICE (Multiple Imputation using Chained Equations) and other meta-heuristic techniques [24–27], the proposed IMV-RE has low time complexity. Because of the simplicity of the proposed IMV-RE algorithm, it is significant for high dimensional datasets also.

3. Imputing The Missing Values In Real-time Environment

The proposed IMV-RE is a splitting-based data imputation algorithm that is the robust statistical MVI technique. The proposed algorithm first decides the upper limit for each feature within a cluster which is further used for imputing the missing values so that the imputed values lie much nearer to the exact values. The following sub-section discussed the step-by-step process of the proposed imputation algorithm.

3.1 Imputation Algorithm (IMV-RE)

Figure 1 shows the whole process from start to the end of IMV-RE algorithm and each step is described as below:

Step 1

Initially a dataset X_D is given that has D dimensions and C_N number of classes.

Step 2

Cluster the data samples into their respective classes along with the missing data samples using the labels given in case of labeled datasets and for unlabeled datasets, K-Means clustering is used to cluster the similar data samples such that C_1, C_2, C_3,…… C_N numbers of classes are there.

Step 3

For each class C_i (where $i = 1$ *to* N), replace all the missing values with zero. Calculate the centered value ($Cent_i$) and standard deviation ($SDev_i$) for class C_i.

Step 4

Calculate the L_2 norm between the $Cent_i$ and each data sample in class C_i. Select the mid-point value as an upper limit ($UpLimC_i$) for class C_i.

Step 5

For each class C_i obtained from the **Step 2**, separate all the data samples containing missing values from the complete data samples into a cluster such that cluster C_i contains L_{miss_sam} numbers of missing samples.

Step 6

For each missing sample L_{miss_sam} containing N_miss number of missing values, impute each missing value with $Cent_i$ calculated for the class $C_i$in **Step 3**. Now, calculate the L_2 norm between the $Cent_i$ and the imputed L_{miss_sam} sample. If the value is smaller than the $UpLimC_i$ (calculated in **Step 4**), then imputed value is finalized else +/- standard deviation ($SDev_i$) is applied to new imputed value in a sequential order and corresponding the L_2 norm between the $Cent_i$ and the imputed value is calculated.

Step 7

The final value to be imputed is the value in correspond to smaller value calculated for the L_2 norm between the $Cent_i$ and the new imputed L_{miss_sam} sample.

4. Experimental Implementation

This section is divided into three subsections. The first subsection represents the benchmark datasets, the second subsection represents the experimental setup and the third subsection represents the metrics used to evaluate the proposed IMV-RE algorithm.

4.1 Datasets

Table 1 depicts the brief description of the datasets used for the experimentation. A total of ten datasets, that includes 8 numerical and 2 of mixed data type, from the UCI machine learning repository are collected with the number of samples ranging from 70 to 5000, the number of features ranging from 4 to 206 and the number of classes ranging from 2 to 10.

4.2 Experimental Setup

The MCAR (Missing Completely at Random) missing value mechanism is used to intentionally add missing values in the above datasets with missingness rates of 10%, 20%, 30%, 40% and 50% of the total data within datasets. A fixed random seed is used for generating the missing values in every dataset to avoid biased results.

The proposed IMV-RE approach is used to impute the missing values and is compared with five baseline approaches: Class Center based Missing Value Imputation (CCMVI), K-Nearest Neighbor Imputation (KNNI), MiceForest [37], IterativeImputer [38]and SimpleImputer [39]. The experimental work is performed on python IDE Spyder version 5.2.1 in Anaconda Navigator usinglaptop PC DELL G5, Intel Core i7 processor, RAM 12GB, 512GB SSD.

Table 1
Datasets Description

Datasets	Datasets type	No. of instances	No. of features	No. of classes
Waveform	Numerical	5000	21	3
Glass	Numerical	214	9	7
Wheat-Seed	Numerical	210	7	3
Digits	Numerical	1797	64	10
Wine	Numerical	178	13	3
Iris	Numerical	150	4	3
Seeds	Numerical	210	7	3
Ionosphere	Numerical	351	34	2
SCADI	Mixed	70	206	7
Ecoli	Mixed	336	7	8

4.3 Evaluation Metrics

A stratified 5-fold cross-validation using RandomForest Classifier is performed to evaluate the classification accuracy on the imputed datasets. For every dataset, there are five different missing rates that correspond to five different classification accuracies and average of five accuracies are taken into consideration for comparison among different imputation techniques.

In addition to the classification accuracy, two more performance metrics are used to evaluate the performance of the proposed IMV-RE approach.

Root Mean Square Error (RMSE): If x_i and $\widehat{x}$ are the original value and the imputed value of the i^{th} observation respectively, n is total number of samples, then, RMSE is given by the equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})^2}{n}}$$

1

Coefficient of Determination (R^2): Calculated using the following equation:

$$R^2\left(x, \widehat{x}\right) = 1 - \frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

2

where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Moreover, the mean and standard deviation of the imputed datasets are compared with the mean and standard deviation of the actual datasets and percentage error is calculated for all datasets.

5. Results And Discussions

Three metrics are used to evaluate the performance of the proposed IMV-RE algorithm. The IMV-RE algorithm is applied to ten benchmark datasets and the classification results obtained using RandomForest Classifier, RMSE values and Coefficient of determination (R^2) values are compared to CCMVI, KNNI, MiceForest, IterativeImputer, SimpleImputer MVI techniques.

Table 2 depicts the average classification accuracies obtained on the ten benchmark datasets Waveform, Glass, Wheat-Seed, Digits, Wine, Iris, Seeds, Ionosphere, SCADI and Ecoli having 10%, 20%, 30%, 40% and 50% missing rates. The average accuracy of the proposed IMV-RE algorithm achieved for all the ten datasets is 91.58% which is better than the other five MVI approaches MiceForest (81.91%), KNNI (80.95%), SimpleImputer (81.7%) and IterativeImputer (82.81%)

For Waveform, Glass, Wheat-Seed, Digits and Wine datasets the CCMVI technique is not applicable beyond 10% missing rate because of its dependency on the samples without missing values to calculate the threshold value that is utilized to estimate the missing values in a particular class.

Similarly, Table 3 compares the average RMSE values obtained for ten datasets. The results show that the proposed IMV-RE algorithm achieves the lowest RMSE value 0.73 in comparison to the other five MVI techniques. RMSE for CCMVI technique cannot be calculated because of the same above-mentioned reason.

Table 4 compares the average value of coefficient of determination obtained for all the ten datasets. The proposed IMV-RE algorithm achieves average Coefficient of Determination value 0.886 which is higher than all other MVI techniques.

Table 2
Average classification accuracies obtained from distinct MVI techniques

Datasets	IMV-RE	MiceForest	KNNI	CCMVI	SimpleImputer	IterativeImputer
Waveform	0.96443	0.78015	0.76447	-	0.77695	0.79092
Glass	0.80498	0.54197	0.52140	-	0.55801	0.57763
Wheat-Seed	0.95944	0.78583	0.78957	-	0.81168	0.82681
Digits	0.98664	0.77682	0.73919	-	0.65356	0.71446
Wine	0.95883	0.95197	0.93429	0.96990	0.96860	0.96537
Iris	0.97333	0.92267	0.90667	0.97267	0.92933	0.94800
Seeds	0.89143	0.87143	0.88095	0.90286	0.87714	0.87810
Ionosphere	0.91851	0.91059	0.90365	0.92029	0.91285	0.90766
SCADI	0.85275	0.84440	0.83846	0.84725	0.85626	0.84703
Ecoli	0.84753	0.80530	0.81605	0.85595	0.82558	0.82551
Average	**0.91579**	**0.81911**	**0.80947**	**-**	**0.81700**	**0.82815**

Table 3
Average RMSE obtained from distinct MVI techniques

Datasets	IMV-RE	MiceForest	KNNI	CCMVI	SimpleImputer	IterativeImputer
Waveform	0.68948	0.70871	0.73801	-	0.80543	0.71955
Glass	0.32751	0.37197	0.38146	-	0.38244	0.39661
Wheat-Seed	0.32499	0.46663	0.48078	-	0.63877	0.42437
Digits	2.02541	2.07970	2.25400	-	2.60753	2.53113
Wine	3.54842	4.65323	4.98252	3.79756	7.21529	6.15099
Iris	0.15431	0.21552	0.21250	0.17177	0.40314	0.21158
Seeds	0.14982	0.11402	0.15104	0.15284	0.30043	0.13956
Ionosphere	0.07780	0.05712	0.04592	0.07728	0.06871	0.06354
SCADI	0.00102	0.00082	0.00041	0.00100	0.00182	0.00159
Ecoli	0.02252	0.03085	0.02623	0.02361	0.02984	0.02654
Average	**0.73213**	**0.86986**	**0.92729**	**-**	**1.24534**	**1.06655**

Table 4
Average Coefficient of Determination (R^2) obtained from distinct MVI techniques

Datasets	IMV-RE	MiceForest	KNNI	CCMVI	SimpleImputer	IterativeImputer
Waveform	0.78181	0.75131	0.73399	-	0.71399	0.76373
Glass	0.71782	0.63663	0.60446	-	0.65406	0.65720
Wheat-Seed	0.85907	0.68851	0.71047	-	0.57340	0.77365
Digits	0.67990	0.67332	0.56739	-	0.53716	0.50253
Wine	0.96225	0.96254	0.93562	0.95724	0.93715	0.95333
Iris	0.95632	0.93200	0.92954	0.94407	0.82545	0.93378
Seeds	0.97092	0.97381	0.96951	0.96438	0.90623	0.95981
Ionosphere	0.97309	0.98352	0.98874	0.97279	0.97986	0.98238
SCADI	0.99958	0.99940	0.99981	0.99929	0.99934	0.99948
Ecoli	0.96658	0.94247	0.94557	0.96319	0.95191	0.95639
Average	**0.88673**	**0.85435**	**0.83851**	**-**	**0.80785**	**0.84823**

Figure 2 graphically represents the performance comparison of proposed IMV-RE algorithm over different evaluation metrics.

Table 5
Percentage Error between mean values of actual datasets and imputed datasets

Datasets	Ecoli	Glass	Wheat-Seed	Digit	Wine	Iris	Seed	Ionosphere	SCADI	Waveform
Actual Mean	0.4996	0.0968	6.8967	4.8843	0.0437	3.4645	0.0820	0.2477	0.2033	1.7123
Imputed Mean	0.4988	0.0790	6.9166	4.8887	0.0371	3.4657	0.0749	0.2438	0.2035	1.7153
Percentage Error	**0.1520**	**18.3642**	**0.2890**	**0.0890**	**15.0280**	**0.0332**	**8.5944**	**1.5801**	**0.0914**	**0.1745**

Table 6
Percentage Error between calculated standard deviation of actual datasets and imputed datasets

Datasets	Ecoli	Glass	Wheat-Seed	Digit	Wine	Iris	Seed	Ionosphere	SCADI	Waveform
Actual Stdev	0.144	1.048	1.010	3.684	0.702	0.948	0.632	0.510	0.263	1.520
Imputed Stdev	0.156	0.961	5.311	5.549	0.704	1.968	0.634	0.573	0.951	1.758
Percentage Error	**8.470**	**8.343**	**425.986**	**50.616**	**0.366**	**107.660**	**0.375**	**12.176**	**261.444**	**15.672**

Table 5 represents the percentage error calculated between the mean values of actual datasets and imputed datasets by considering the average of total mean value calculated at different missing rates 10%, 20%, 30%, 40% and 50% using proposed IMV-RE algorithm.

Similarly, the Table 6 represents the percentage error calculated between the standard deviation of actual datasets and imputed datasets by considering the average of total standard deviation value calculated at different missing rates 10%, 20%, 30%, 40% and 50% using proposed IMV-RE algorithm.

6. Conclusion

In this paper, a novel splitting-based IMV-RE algorithm is proposed. The proposed IMV-RE algorithm is a two-step process. In the first step, an upper limit is calculated for every class containing missing values and the second step utilizes the upper limit to impute missing values efficiently. The IMV-RE algorithm only searches within the class to calculate the centered value for the final imputation, unlike the other techniques that need to go through the whole data within the dataset. Hence, the proposed algorithm can be utilized for real-time problems. The proposed IMV-RE algorithm prosperously abstracts the dependency of the CCMVI technique on the complete samples within a class to estimate the missing values. As a result, the proposed IMV-RE algorithm successfully imputes the missing values for higher data missing.

Classification accuracy, RMSE and coefficient of determination (R^2) are the evaluation metrics used to evaluate the performance of the proposed IMV-RE algorithm. The experimental results depict better performance of the proposed IMV-RE algorithm for imputing the missing values under distinct missing rates in comparison to the other state-of-the-art techniques.

Declarations

Conflict of Interest

The authors of this publication declare there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Code/Data Availability

N/A.

Authors' Contributions

Hutashan Vishal Bhagat: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Validation, Data Curation and Visualization.

Manminder Singh: Writing Review, Editing and Supervision

References

1. Kang, H. (2013). The prevention and handling of the missing data.Korean journal of anesthesiology, 64(5),402. https://doi.org/10.4097/kjae.2013.64.5.402
2. Kalkan, Ö. K., Yusuf, K. A. R. A., & Kelecioğlu, H. (2018). Evaluating performance of missing data imputation methods in IRT analyses.International Journal of Assessment Tools in Education, 5(3),403–416. https://doi.org/10.21449/ijate.430720
3. García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (Vol. 72, pp.59–139). Cham, Switzerland:Springer International Publishing
4. Kelkar, B. A. (2022). Missing Data Imputation: A Survey. *International Journal of Decision Support System Technology (IJDSST)*, *14*(1), 1–20. DOI: 10.4018/IJDSST.292446
5. Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793).John Wiley & Sons
6. Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, *48*(1), 5–37
7. Tsai, C. F., Li, M. L., & Lin, W. C. (2018). A class center based approach for missing value imputation. *Knowledge-Based Systems*, *151*, 124–135
8. Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, *14*(5), 853–871
9. Razavi-Far, R., Cheng, B., Saif, M., & Ahmadi, M. (2020). Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, *187*, 104805
10. Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301
11. Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377
12. Petrozziello, A., & Jordanov, I. (2017). Column-wise guided data imputation. *Procedia Computer Science*, *108*, 2282–2286
13. Sammulal, P., Usha Rani, Y., & Yepuri, A. (2017). A class based clustering approach for imputation and mining of medical records (CBC-IM). *IADIS International Journal on Computer Science & Information Systems*, *12*(1), 61–74
14. Mostafa, S. M. (2019). Imputing missing values using cumulative linear regression.CAAI Transactions on Intelligence Technology, 4(3),182–200. https://doi.org/10.1049/trit.2019.0032
15. Pan, R., Yang, T., Cao, J., Lu, K., & Zhang, Z. (2015). Missing data imputation by K nearest neighbours based on grey relational structure and mutual information.Applied Intelligence, 43(3),614–632. https://doi.org/10.1007/s10489-015-0666-x
16. Silva-Ramírez, E. L., Pino-Mejías, R., & López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, *29*, 65–74
17. Mostafa, M., Eladimy, S. S., Hamad, A., S., & Amano, H. (2020). CBRL and CBRC: Novel algorithms for improving missing value imputation accuracy based on Bayesian ridge regression.Symmetry, 12(10),1594. https://doi.org/10.3390/sym12101594

18. Sefidian, A. M., & Daneshpour, N. (2020). Estimating missing data using novel correlation maximization based methods. *Applied Soft Computing*, *91*, 106249
19. Adhikari, D., Jiang, W., & Zhan, J. (2021). Imputation using information fusion technique for sensor generated incomplete data with high missing gap. *Microprocessors and Microsystems*. <background-color:#cfbfb1;uvertical-align:super;>https://doi.org/10.1016/j.micpro.2020.103636</background-color:#cfbfb1;uvertical-align:super;>103636
20. Lan, Q., Xu, X., Ma, H., & Li, G. (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems with Applications*, *141*, 112926
21. Kamkhad, N., Jampachaisri, K., Siriyasatien, P., & Kesorn, K. (2020). Toward semantic data imputation for a dengue dataset.Knowledge-Based Systems, 196,105803. https://doi.org/10.1016/j.knosys.2020.105803
22. Ngueilbaye, A., Wang, H., Mahamat, D. A., & Junaidu, S. B. (2021). Modulo 9 model-based learning for missing data imputation.Applied Soft Computing, 103,107167. https://doi.org/10.1016/j.asoc.2021.107167
23. Peng, D., Zou, M., Liu, C., & Lu, J. (2021). RESI: a region-splitting imputation method for different types of missing data.Expert Systems with Applications, 168,114425. https://doi.org/10.1016/j.eswa.2020.114425
24. Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, *37*(9), 1322–1331
25. Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. Neurocomputing, 156, 134–142. https://doi.org/10.1016/j.neucom.2014.12.073
26. Priya, R. D., Sivaraj, R., & &Priyaa, N. S. (2017). Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases. *Knowledge-Based Systems*, *133*, 107–121
27. Lobato, F., Sales, C., Araujo, I., Tadaiesky, V., Dias, L., Ramos, L., & Santana, A. (2015). Multi-objective genetic algorithm for missing data imputation.Pattern Recognition Letters, 68,126–131. https://doi.org/10.1016/j.patrec.2015.08.023
28. Wu, C. H., Wun, C. H., & Chou, H. J. (2004, December). Using association rules for completing missing data. In *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)* (pp. 236–241). IEEE. https://doi.org/10.1109/ICHIS.2004.91
29. Wu, J., Song, Q., & Shen, J. (2007, July). An novel association rule mining based missing nominal data imputation method. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)* (Vol. 3, pp. 244–249). IEEE. https://doi.org/10.1109/SNPD.2007.93
30. Nelwamondo, F. V., Golding, D., & Marwala, T. (2013). A dynamic programming approach to missing data estimation using neural networks.Information Sciences, 237,49–58. https://doi.org/10.1016/j.ins.2009.10.008
31. Tang, J., Zhang, G., Wang, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transportation Research Part C: Emerging Technologies*, *51*, 29–40
32. Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25–35
33. Vazifehdan, M., Moattar, M. H., & Jalali, M. (2019). A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. *Journal of King Saud University-Computer and Information Sciences*, *31*(2), 175–184
34. Choudhary, A., Kumar, S., Sharma, M., & Sharma, K. P. (2022). A Framework for Data Prediction and Forecasting in WSN with Auto ARIMA. *Wireless Personal Communications*, *123*(3), 2245–2259. https://doi.org/10.1007/s11277-021-09237-x
35. Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857
36. Sim, J., Kwon, O., & Lee, K. C. (2016). Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. *Expert Systems with Applications*, *46*, 485–493
37. Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, *179*(6), 764–774

38. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1–67

39. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python.The Journal of machine Learning research, 12,2825–2830

Figures

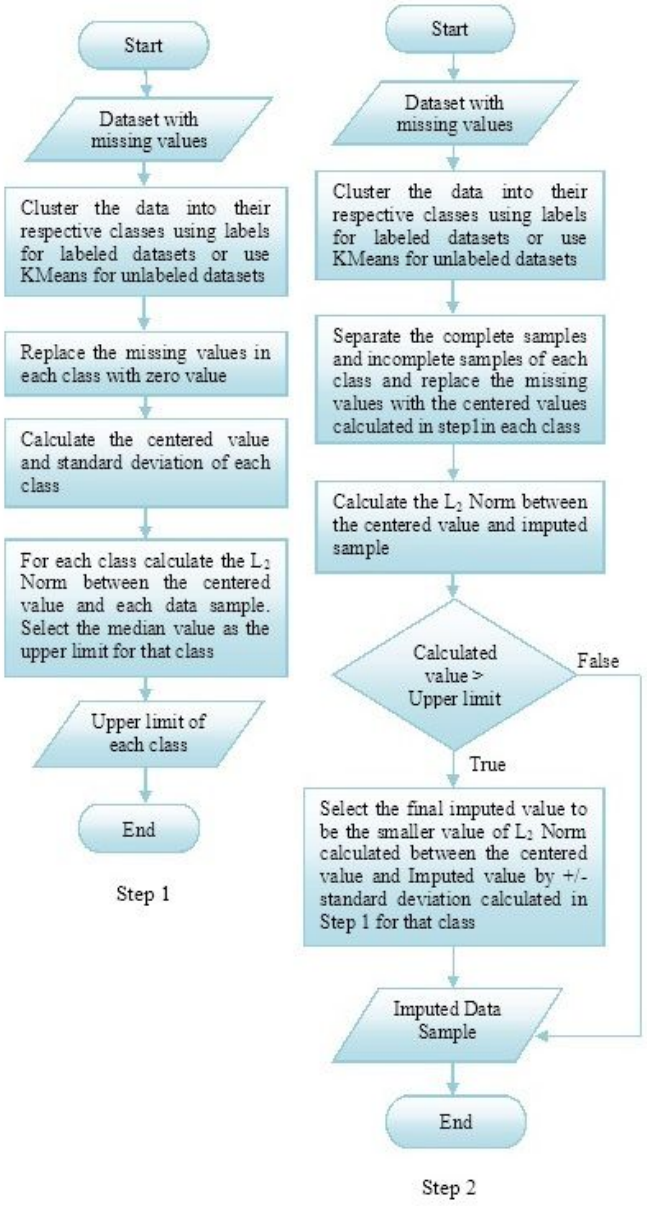

Figure 1

The two step IMV-RE process for Imputation

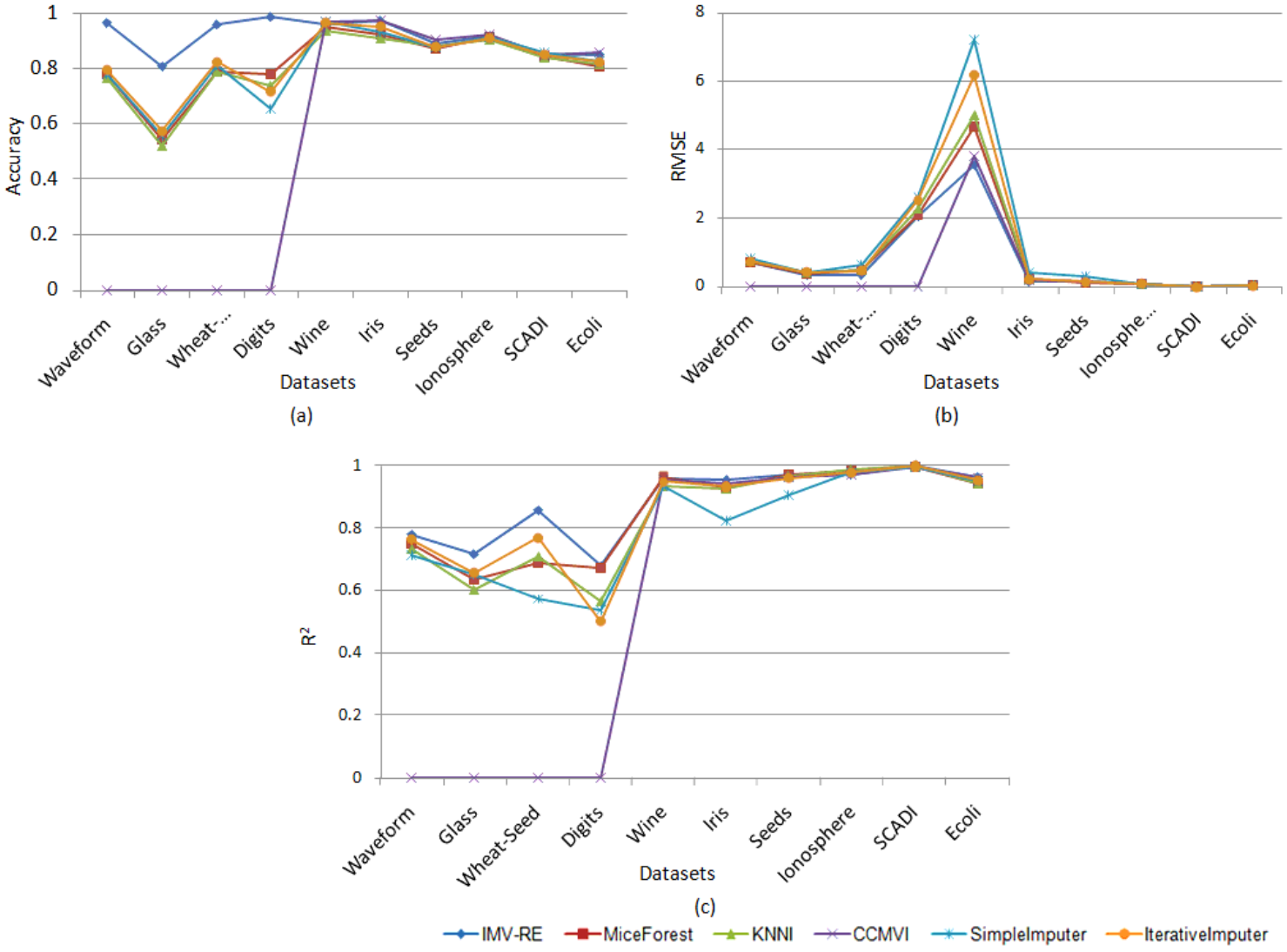

Figure 2

The performance comparison of the proposed IMV-RE algorithm