Journal of Big Data

RESEARCH Open Access

Adaptive multiple imputations of missing values using the class center

Kritbodin Phiwhorm[1], Charnnarong Saikaew[2], Carson K. Leung[3], Pattarawit Polpinit[1] and Kanda Runapongsa Saikaew[1*]

*Correspondence:
krunapon@kku.ac.th
[1] Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand
Full list of author information is available at the end of the article

Abstract

Big data has become a core technology to provide innovative solutions in many fields. However, the collected dataset for data analysis in various domains will contain missing values. Missing value imputation is the primary method for resolving problems involving incomplete datasets. Missing attribute values are replaced with values from a selected set of observed data using statistical or machine learning methods. Although machine learning techniques can generate reasonably accurate imputation results, they typically require longer imputation durations than statistical techniques. This study proposes the adaptive multiple imputations of missing values using the class center (AMICC) approach to produce effective imputation results efficiently. AMICC is based on the class center and defines a threshold from the weighted distances between the center and other observed data for the imputation step. Additionally, the distance can be an adaptive nearest neighborhood or the center to estimate the missing values. The experimental results are based on numerical, categorical, and mixed datasets from the University of California Irvine (UCI) Machine Learning Repository with introduced missing values rate from 10 to 50% in 27 datasets. The proposed AMICC approach outperforms the other missing value imputation methods with higher average accuracy at 81.48% which is higher than those of other methods about 9 – 14%. Furthermore, execution time is different from the Mean/Mode method, about seven seconds; moreover, it requires significantly less time for imputation than some machine learning approaches about 10 – 14 s.

Keywords: Big data, Data mining, Incomplete data, Machine learning, Class center, Missing value imputation

Introduction

Big data has become a critical technology for developing novel solutions in a wide variety of fields. For instance, large and complex amounts of structured and unstructured data are growing at high-speed rates [1]. The development of these big data will enhance the discovery of useful information, such as hidden patterns and unknown correlations, that can be useful in many fields, including healthcare, financial, manufacturing, and social life [2], such as the failure or malfunction of the sensors that provide information or partial observation of an object of interest because of some hidden phenomenon.

Many real-world applications suffer a common drawback, missing or unknown data. For example, some results may be lost in an industrial experiment due to mechanical faults during the data gathering procedure. Likewise, some tests cannot be done in medical diagnosis because some medical tests may not be appropriate for certain patients, or the medical report proforma permits the omission of specific qualities.

The quality of data [3] is a significant concern to them for conducting effective data analytics. Although the outcome of data analysis tasks depends on several factors such as attribute selection, algorithm selection, and sampling techniques, a critical dependency relies upon the efficient handling of missing values [4]. The data is either missing or incorrectly entered by a human, which results in an incorrect prediction [5, 6], as missing values degrade performance. Therefore, missing data is a significant issue in big data analytics, as it can significantly increase the cost of computation and skew the results [7]. As a result, data quality is a fundamental requirement for big data processing, and data quality suffers when missing values are present [8].

A data analysis algorithm cannot handle incomplete datasets directly by itself. The simplest way to deal with this problem is case deletion, which means directly removing all the data of cases with missing values [9, 10]. However, if the missing value rate is high, the deletion approach affects the remainder of the complete data and can reduce the accuracy of the results [11]. As a result, reliable imputation techniques are necessary to consider the matter of missing data. Additionally, imputation of missing data can aid in the maintenance of the completeness of a dataset, which is critical in small-scale data mining projects and big data analytics.

To date, missing value imputation (MVI) has been proposed as a promising solution for incomplete datasets [12–19]. MVI can be broadly classified into statistical and machine learning techniques. The mean and mode are common statistical MVI technique measurements that typically require a short time to compute. However, machine learning MVI techniques, such as support vector machine (SVM), and random forest (RF) methods, require a long computation time to achieve high accuracy [20–23]. On the other hand, the k-nearest neighbor (KNN) technique [24] requires much less imputation time than other machine learning techniques [25–28]. However, the KNN method performs only an online search of the nearest neighbors through the Euclidean distance function [29].

Among the KNN-based methods, Troyanskaya et al. [30] and Daberdaku et al. [31] presented a weighted KNN algorithm for the missing data imputation. Further, Cheng et al. [32] proposed a KNN method that used purity to enhance the performance of K nearest neighbors. Fan et al. [33] proposes the weighted KNN approach, which uses the inverse of the Euclidean distance as the weight for each data point. Of all these KNN-based weighted methods, the set of nearest neighbors is computed by the weight distance between the data of missing values and the complete data.

Sometimes, although more complex algorithms might produce better imputation results, they will generally require a higher computational cost, which is a consideration in machine learning techniques versus statistical techniques [34]. However, most machine learning techniques are usually more computationally expensive than many statistical techniques, due to the model training and construction process.

Recently, class center-based missing value imputation was proposed to produce relatively better imputation results at a lower computational cost [35–37]. The class center is based on the mean of the data samples in a specific class, which is similar to the idea of the cluster center (or centroid) applied in the k-mean algorithm [38]. Thereafter, the Euclidean distances between each data sample and the class center are measured, to define a threshold for the later imputation guideline in "Materials and methods" section.

This study aims to propose an algorithm for missing value imputation such that it achieves high accuracy, yet requires minimal time. This presents a novel imputation method: the adaptive multiple imputations of missing values using the class center approach (AMICC). The key contributions of our work are (1) the class center is based on the mean/mode of the data samples and replaces values appropriately, according to the attribute type of the dataset, (2) the proposed adaptive threshold value follows the standard deviation (STD) values, where the computation can indicate how data are spread out over a range of normal and filter outliers, and (3) for outlier data, the missing values are replaced with more appropriate values by using the median and the average weight distance values of the class.

The remainder of this article is organized as follows: "Materials and methods" section presents the related work and describes the proposed model, the diversity operator, and the AMICC algorithm design. The experimental results are presented in "Experiments and results" section, while "Discussion" section offers a discussion and the conclusions are presented in "Conclusions" section.

Materials and methods

In this section, we first present the missing value imputation in "Missing value imputation" section. "Our adaptive multiple imputations of missing values using class centers" section details our proposed method, the AMICC algorithm.

Missing value imputation

MVI uses a statistical or machine learning method to estimate the observed data chosen to replace the missing values. The simplest statistical methods for continuous and discrete variables are mean and mode imputation [39], respectively. Besides statistical techniques, MVI also uses machine learning methods to estimate the observed data chosen to replace the missing values. For instance, MVI analyzes a pattern classification task where the missing feature is employed as the target output for the classification model. The rest of the complete features are the input attributes used to train and test the model [40].

One of the most widely used machine learning techniques is KNN imputation [41], where missing values are imputed using the values calculated from the nearest neighbor observed data. In finding the nearest neighbors, the preferred choice in nearest neighbor classification is to define the Euclidean distance, which is defined as:

$$dist(x_i, x_j) = \sqrt{\sum_{n=1}^{N} [x_{ni} - x_{nj}]^2} \tag{1}$$

where function $dist(x_i, x_j)$ computes the distance between the instance x_i and x_j, N is the number of attributes or features, and x_{ni} represents the ith instance in the nth attribute.

The baselines compared with the proposed method are usually based on statistical and machine learning techniques. Two other well-known machine learning techniques for MVI are the SVM [42] and RF [43] algorithms. The SVM algorithm uses kernel functions for the nonlinear mapping of an original feature space into a higher dimensional feature space to build a hyperplane. The RF algorithm is an ensemble of decision tree classifiers, which establishes the outcome based on the predictions of the decision trees. RF predicts the outcome by taking the average or mean of the output from various trees.

Our adaptive multiple imputations of missing values using class centers

Several real-world datasets often found with not-a-numbers (NaNs), blank fields, or other placeholders may have missing values. Training a model with a dataset of many missing values can drastically impact the quality of the machine learning or statistical model, resulting in higher computational costs. If the quantum of missing data is large, the efficiency will fluctuate accordingly.

As indicated in "Missing value imputation" section, missing values are commonly replaced by mean/mode. Hence, in the AMICC method, the class center is based on the mean/mode of the data samples in a specific class. In the MVI approach, the AMICC method replaces the missing values with the mean or mode depending on the attribute type. In the outlier data of each class, the AMICC method identifies the threshold values for checking the outlier data; the threshold values are calculated based on the distances between the class centers and their correspondence to the complete data.

In Fig. 1, the AMICC approach comprises three modules: the first focuses on data preprocessing in "Data pre-processing" section, the second calculates the threshold identification in "Threshold identification" section and the third imputes the missing values in "Imputation of missing values" section. These three modules are described in the following sub-sections.

Data pre-processing

In the data pre-processing section, there are some differences in UCI dataset experiments and missing data types based on the missing completely at random (MCAR)

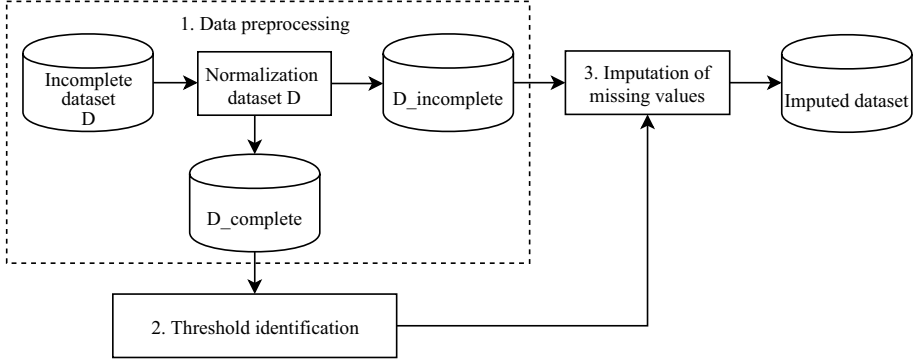


Fig. 1 Procedure of the proposed method

which the presence of missing data does not depend on the input values perse [44]. Therefore, in large datasets plagued by MCAR missing data, samples with missing values can be discarded without biasing the distribution of the remaining data.

This study simulated missing rates of 10%, 20%, 30%, 40%, and 50% [9, 15, 20, 32, 35, 41] to compare the proposed method to the imputation methods listed in the UCI datasets experiment. As shown in Eq. (2), the missing rate is a percentage of the total number of missing values in the dataset. All variables except the class attribute had their missing values simulated.

$$missing\ rate = \frac{number\ of\ missing\ values \times 100}{number\ of\ examples \times number\ of\ features} \tag{2}$$

The missing rate of 50% in this study is the highest when the number of examples and features is considered. For example, consider the Blood dataset, which contains 748 examples and nine features, as illustrated in Table 1. According to equation (2), there were 3,366 missing values when the missing rate was set to 50% (50 = 3366 × 100 / (748 × 9)).

Additionally, normalization is a technique frequently used during the data preparation process. The goal of normalization is to change the values of numeric columns in the dataset to a standard scale, without distorting differences in the ranges of values. The incomplete dataset must be normalized in the domain [0,1], as normalized data on the same scale avoids the effect of different attribute ranges on distance calculation. Thereafter, the incomplete dataset is divided into two subsets: one is the incomplete data containing the missing values for later imputation, and the other is the complete data without missing values for calculating the initial values of the next step.

For example, Fig. 2 shows a three-class incomplete dataset with ten feature dimensions ($F = 10$), in which the question marks represent attributes with missing values. Class i ($i = 1$ to N; $N = 3$) of D, denoted by D_i, is divided into $D_{i\_complete}$ and $D_{i\_incomplete}$.

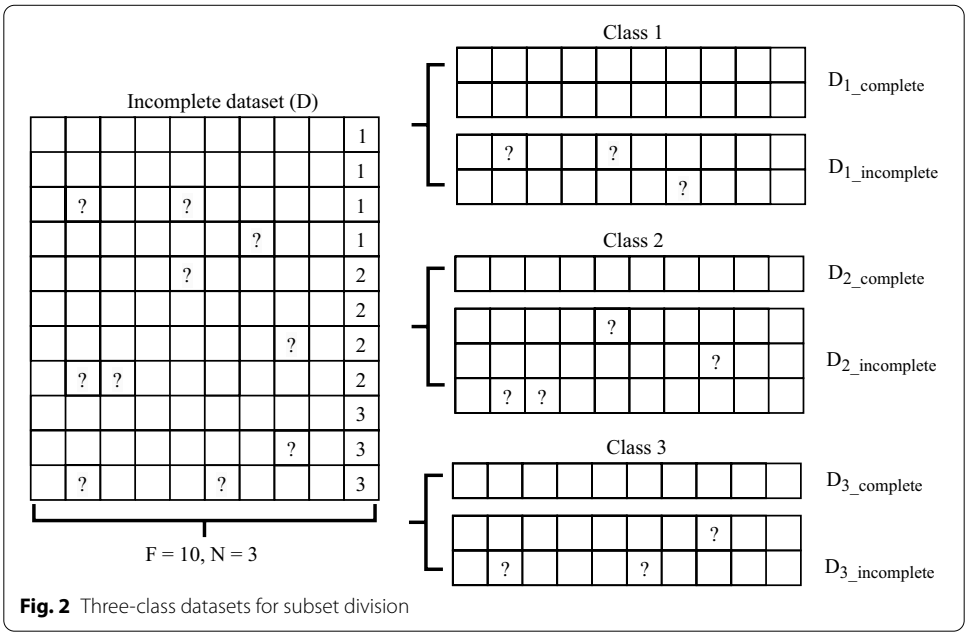


Fig. 2 Three-class datasets for subset division

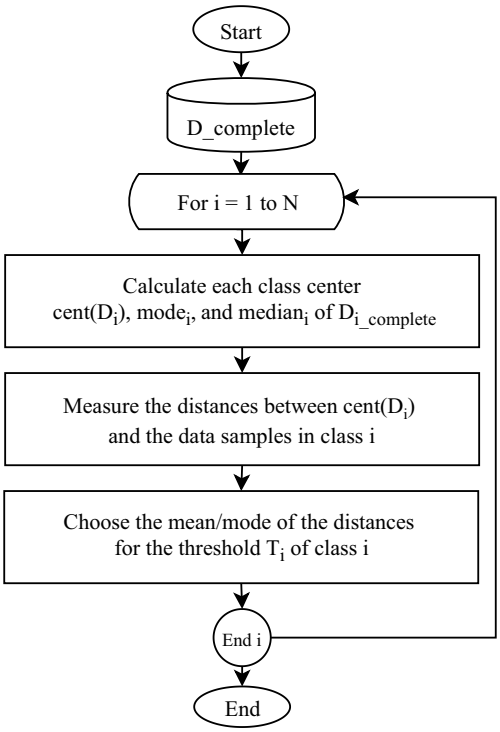


Fig. 3 Threshold identification process

Threshold identification

Figure 3 shows that the process of identifying the threshold based on the distances between the class centers and their correspondences to the complete data described in more detail below.

From the incomplete dataset D containing N classes, dataset D is divided into complete ($D\_complete$) and incomplete ($D\_incomplete$) subsets, where $D\_incomplete$ contains missing

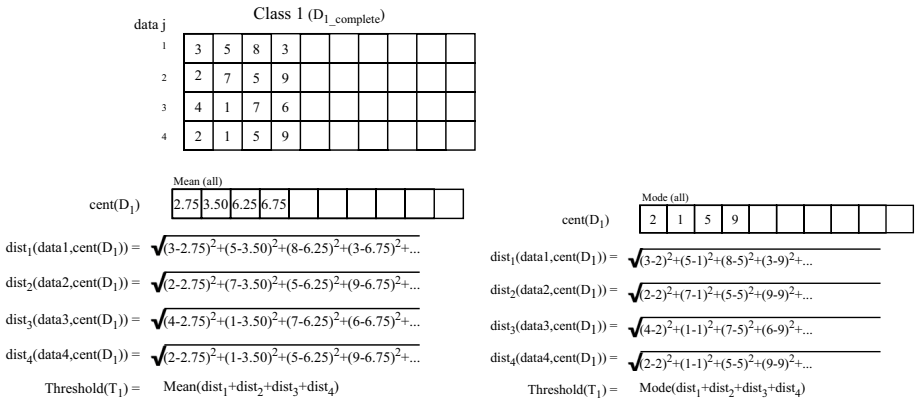


Fig. 4 Examples of the class center and distance calculations for Class 1 in (**a**) a numerical dataset and (**b**) a categorical dataset

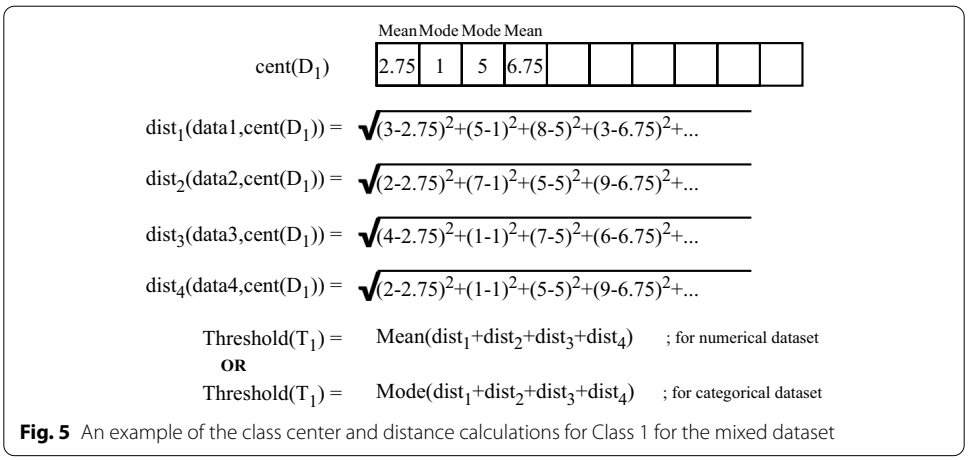


Fig. 5 An example of the class center and distance calculations for Class 1 for the mixed dataset

values. For the i-th class of $D_{i\_complete}$, the class center ($cent(D_i)$), mode, and median are calculated. When computing the class center values for a numerical attribute, the mean is used as the class center. Otherwise, if the attribute is categorical, the mode value is the class center value.

Next, the Euclidean distances between $cent(D_i)$ and every data sample in Class i are computed. Figures 4 and 5 show an example of calculating the center of Class 1, $cent(D_1)$, and the distances between $cent(D_1)$ and the other data samples.

Based on the distances, in Fig. 4a, the mean is used for calculating the distances for a numerical dataset; in Fig. 4b, the mode is used for calculating the distances for a categorical dataset; in Fig. 5, the mean or mode is used for calculating distances for a mixed dataset, in which the mean or mode of these distances is used as the threshold (T_1) for Class 1. Thereafter, this step is repeated until the threshold for each class is obtained. The pseudocode for the threshold identification module is shown in Algorithm 1.

Algorithm 1 Threshold identification.

Input: Incomplete dataset D containing F feature dimensions, N classes, and Num data samples
Output: T threshold values for the N_{class} classes

for $j = 1$ to Num **do**
  if $D(j,)$ has missing values **then**
    Obtain the class label and set this class to variable i
    Put $D(j,)$ into $D_{i\_incomplete}$
  else
    Obtain the class label and set this class to variable i
    Put $D(j,)$ into $D_{i\_complete}$
for $i = 1$ to N **do**
  $Avg(i,) =$ Average($D_{i\_complete}$)
  $Mode(i,) =$ Mode($D_{i\_complete}$)
  $Median(i,) =$ Median($D_{i\_complete}$)
  Set the number of rows in $D_{i\_complete}$ to Num_{i_class}
for $j = 1$ to Num_i **do**
  if D is a numerical dataset **then**
    $Distance(j,) =$ dist($D_{i\_complete}(j,), Avg(i,)$)
  else
    $Distance(j,) =$ dist($D_{i\_complete}(j,), Mode(i,)$)
if D is a numerical dataset **then**
  $T_i =$ Average($Distance$)
else
  $T_i =$ Mode($Distance$)

Imputation of missing values

Imputation techniques can be straightforward or quite complicated. These techniques compute the mean/mode of the non-missing values in the complete data and replace the missing values in incomplete data. A single value replaces a missing value for a single imputation, such as the mean of the entire dataset. Multiple imputations are widely accepted as the standard for dealing with missing data in a variety of research fields. Multiple imputations are used to derive unbiased and valid estimates from available data.

In outlier data, the AMICC method checked the normal distribution using the STD value to determine whether the given measurement deviates from the mean. In statistics, STD is a frequently used yardstick of measurement variability. A low STD value indicates that data points are typically very close to the norm, whereas a high STD value indicates that data points span a wide range of values.

Figure 6 shows that the process of the imputation of missing values consists of the following two steps; the first step is to perform a preliminary imputation of the missing value using the mean/mode of each attribute in a class and the second step is to compare the outlier data with STD values. There are two ways to handle outlier data; (1) if STD <= 1, check the outlier data by calculating the distance between the missing value and the class center; if the distance exceeds a threshold, the missing value is considered outlier data and replaced with the median value. Next, (2) if STD > 1, the missing value is considered an outlier. The average weight distance is calculated from the weight distance between the missing value and its nearest neighbors in the complete data. Then, the average weight distance is replaced for the missing value. The proposed method for imputed values is described in detail below.

Step 1: For Class *i*, the incomplete dataset ($D_{i\_incomplete}$) is composed of a missing data sample (*Num*). Figures 7, 8, and 9 illustrate examples of a Class 1 incomplete dataset ($D_{1\_incomplete}$) for numerical, categorical, and mixed datasets, respectively, where the data *j*

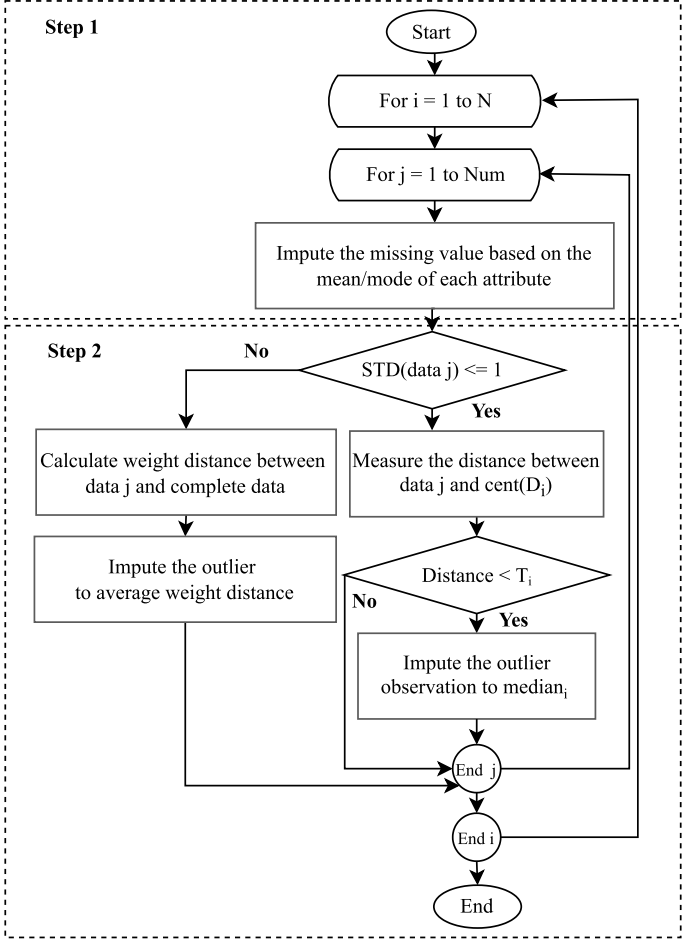
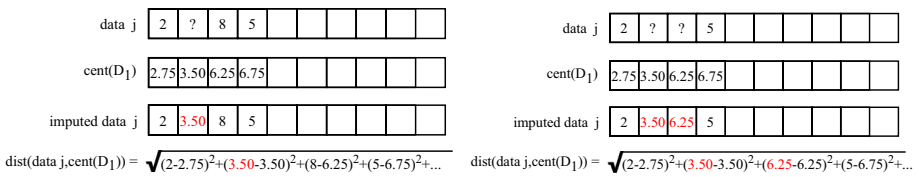
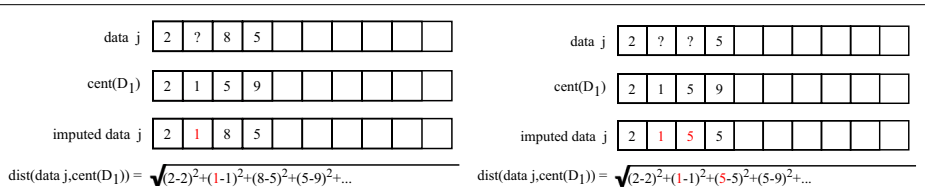


Fig. 6 MVI process



(a) Single missing value **(b)** Multiple missing values

Fig. 7 Example of imputing a numerical dataset for (**a**) a single missing value, and (**b**) multiple missing values



(a) Single missing value **(b)** Multiple missing values

Fig. 8 Example of imputing a categorical dataset for (**a**) a single missing value, and (**b**) multiple missing values

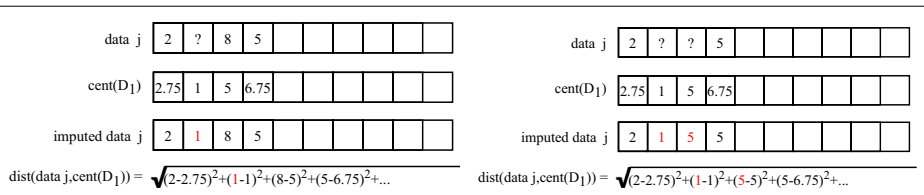


Fig. 9 Example of imputing a mixed dataset for (**a**) a single missing value, and (**b**) multiple missing values

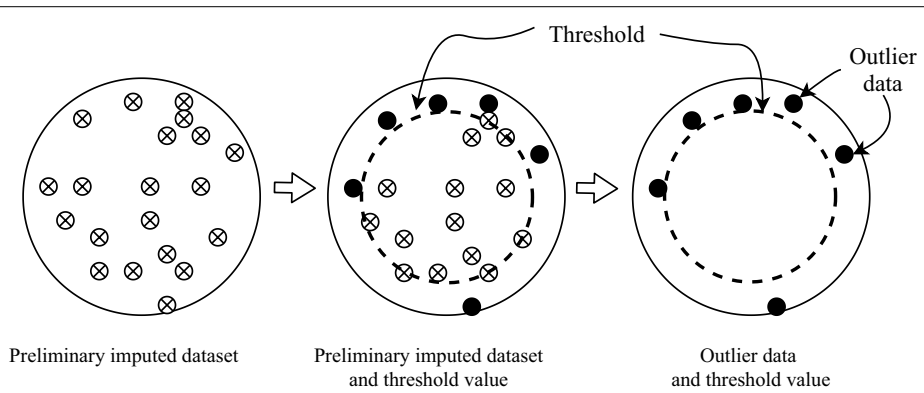


Fig. 10 The imputing outlier data for STD less than one

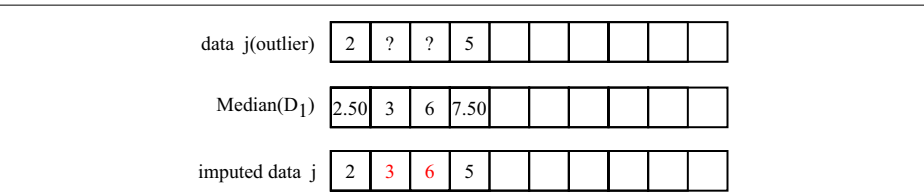


Fig. 11 Example of imputing outlier data with median

($j = 1$ to *Num*) contain one missing value in Figs. 7a, 8a, and 9a and multiple missing values

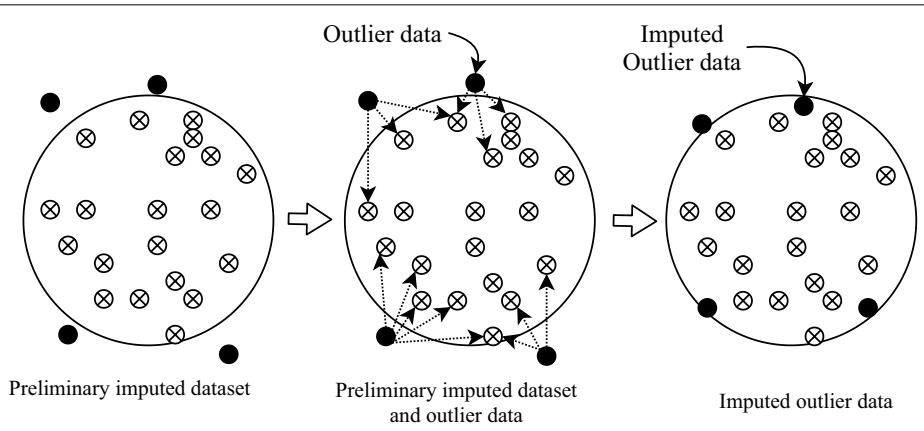


Fig. 12 Imputing outlier data for STD greater than or equal to one

in Figs. 7b, 8b, and 9b. In the examples shown in these figures, the missing feature of data *j*, *cent*(D_1), and imputed values are in the red text. The distance between *cent*(D_1) and the imputed data *j* is calculated and compared with the threshold T_1 in the next step.

Step 2: This step consists of two cases, (1) if STD <= 1, from the preliminary imputed dataset from Step 1, Fig. 10 illustrates how to impute outlier data for STD values less than one, in which the algorithm compares the outlier data to the threshold value of the class. In Fig. 11, for example, if the distance is less than T_1, the imputation process for data *j* is complete; otherwise, each outlier datum is imputed to the median of Class 1.

In the other case (2), if STD > 1, Fig. 12 shows imputed data to the outlier. According to equation (3), the average weight distance [33] is arrived at by calculating the weight distance between the missing value and its nearest neighbors in the complete data.

$$W_i = average\left[\frac{1}{dist(y_i, x_1)} + \frac{1}{dist(y_i, x_2)} + ... + \frac{1}{dist(y_i, x_j)}\right] \tag{3}$$

where W_i is a weight distance of *i*th outlier data, y_i is the *i*th instance of outlier data, and x_1 is the first instance of complete data. From "Missing value imputation" section, function $dist(y_i, x_j)$ computes the distance between the instance y_i and x_j. This step is repeated until the average weight distance for each instance is obtained.

After computing all average weight distances, the outlier datums are imputed to the average weight distance. Algorithm 2 is proposed for missing value imputation.

Algorithm 2 Missing value imputation.

Input: Incomplete data $D_{\_incomplete}$ containing F feature dimensions, N classes, T thresholds; Num number of rows in each class; $NumAttri$ number $Attri$ attribute of missing attributes; W weight distance

Output: Imputed data for $D_{i\_incomplete}$

for $j = 1$ to Num **do**
  W_j = Average($\frac{1}{dist(D_{i\_incomplete}(j,), D_{i\_complete}(1,))} + .. + \frac{1}{dist(D_{i\_incomplete}(j,), D_{i\_complete}(Num_{complete},))}$)
for $i = 1$ to N **do**
  for $j = 1$ to Num **do**
    for $k = 1$ to $NumAttri$ **do**
      if $Attri(,k)$ is a numerical attribute **then**
        $D_{i\_incomplete}(j,k) = Average(D_{i\_complete}, Attri(,k))$
      else
        $D_{i\_incomplete}(j,k) = Mode(D_{i\_complete}, Attri(,k))$
    if STD(j,) <= 1 **then**
      if $D_{\_incomplete}$ is a numerical data **then**
        $Distance$ = dist($D_{i\_incomplete}(j,), Avg(i,)$)
      else
        $Distance$ = dist($D_{i\_incomplete}(j,), Mode(i,)$)
      if $Distance > T_i$ **then**
        $D_{i\_incomplete}(j,) = Median(i,)$
    else
      $D_{i\_incomplete}(j,) = W_j$

Experiments and results

This section presents the performance evaluation and comparison of the proposed AMICC method and statistical and machine learning methods.

Experimental setup

The experimental data included 13 numerical, six categorical, and eight mixed datasets collected from the UCI Machine Learning Repository [45]. These datasets have

been the subject of several studies on machine learning methods and cover examples of datasets of small-, medium-, and large-size [9, 12, 13, 24, 25, 35, 41]. The characteristics of these datasets are shown in Tables 1, 2, and 3. All the datasets show considerable diversity in the number of examples, features, and classes.

In Table 3, the numbers of numerical and categorical attributes are indicated in parentheses for each dataset. The dataset is treated as a numerical dataset if the number of numerical attributes is greater than that of categorical attributes. Otherwise, the dataset is treated as a categorical dataset. For example, as the Abalone dataset consists of seven numerical and one categorical attribute, this mixed dataset is treated as numerical, in which distances are calculations based on the mean. On the other

Table 1 Basic information on the numerical datasets

Datasets	Examples	Features	Classes
Blood	748	9	2
Ecoli	336	7	8
Glass	214	9	6
Ionosphere	351	34	2
Iris	150	4	3
Liver cancer	583	10	2
Optdigits	5,620	64	10
Pima	768	8	2
Sonar	208	60	2
Wine	178	13	3
Yeast	1,484	8	10
Column 2C	310	6	2
Column 3C	310	6	3

Table 2 Basic information on the categorical datasets

Datasets	Examples	Features	Classes
Balance scale	625	4	3
Breast cancer	699	9	2
Lymphography	148	8	4
Promoters	106	57	2
Spect	267	22	2
Tic tac toe	958	9	2

Table 3 Basic information on the mixed datasets

Datasets	Examples	Features	Classes
Abalone (7,1)	4,172	8	29
Acute (1,5)	120	6	2
Card (6,9)	690	15	2
Contraceptive (2,7)	1,473	9	3
German (7,13)	1,000	20	2
Heart (5,8)	303	13	2
Zoo (1,15)	101	16	2
Srinagarind (6,2)	3,467	8	3

Table 4 The total number of missing rate on the numerical datasets

Datasets	Missing rate									
	10%		20%		30%		40%		50%	
	c*	i*	c	i	c	i	c	i	c	i
Blood	6,059	673	5,386	1,346	4,712	2,020	4,039	2,693	3,366	3,366
Ecoli	2,117	235	1,882	470	1,646	706	1,411	941	1,176	1,176
Glass	1,733	193	1,541	385	1,348	578	1,156	770	963	963
Ionosphere	10,741	1,193	9,547	2,387	8,354	3,580	7,160	4,774	5,967	5,967
Iris	540	60	480	120	420	180	360	240	300	300
Liver cancer	5,247	583	4,664	1,166	4,081	1,749	3,498	2,332	2,915	2,915
Optdigits	323,712	35,968	287,744	71,936	251,776	107,904	215,808	143,872	179,840	179,840
Pima	5,530	614	4,915	1,229	4,301	1,843	3,686	2,458	3,072	3,072
Sonar	11,232	1,248	9,984	2,496	8,736	3,744	7,488	4,992	6,240	6,240
Wine	2,083	231	1,851	463	1,620	694	1,388	926	1,157	1,157
Yeast	10,685	1,187	9,498	2,374	8,310	3,562	7,123	4,749	5,936	5,936
Column 2C	1,674	186	1,488	372	1,302	558	1,116	744	930	930
Column 3C	1,674	186	1,488	372	1,302	558	1,116	744	930	930

*c = complete data, i = incomplete data

Table 5 The total number of missing rate on the categorical datasets

Datasets	Missing rate									
	10%		20%		30%		40%		50%	
	c*	i*	c	i	c	i	c	i	c	i
Balance scale	2,250	250	2,000	500	1,750	750	1,500	1,000	1,250	1,250
Breast cancer	5,662	629	5,033	1,258	4,404	1,887	3,775	2,516	3,146	3,146
Lymphography	1,066	118	947	237	829	355	710	474	592	592
Promoters	5,438	604	4,834	1,208	4,229	1,813	3,625	2,417	3,021	3,021
Spect	5,287	587	4,699	1,175	4,112	1,762	3,524	2,350	2,937	2,937
Tic tac toe	7,760	862	6,898	1,724	6,035	2,587	5,173	3,449	4,311	4,311

*c = complete data, i = incomplete data

hand, as the Acute dataset consists of one numerical and five categorical attributes, it is considered as categorical, in which distances are calculations based on the mode.

In Tables 4, 5, and 6 summarize the total number of missing rates of numerical, categorical, and mixed dataset types, respectively, where c is the number of complete data, and i is the number of incomplete data. In Section "Data pre-processing", as shown in equation (2), the missing rate is a percentage of the total number of missing values in the dataset. All variables except the class attribute had their missing values simulated. The simulated missing rates of 10%, 20%, 30%, 40%, and 50%.

The missing rate of 10% in this study is the lowest when the number of examples and features is considered. For example, consider the Spect dataset, which contains 267 examples and 22 features, as illustrated in Table 2. According to equation (2), there were 587 missing values (incomplete data) when the missing rate was set to 10% ($10 = 574 \times 100 / (267 \times 22)$).

Table 6 The total number of missing rate on the mixed datasets

Datasets	Missing rate									
	10%		20%		30%		40%		50%	
	c*	i*	c	i	c	i	c	i	c	i
Abalone (7,1)	30,038	3,338	26,701	6,675	23,363	10,013	20,026	13,350	16,688	16,688
Acute (1,5)	648	72	576	144	504	216	432	288	360	360
Card (6,9)	9,315	1,035	8,280	2,070	7,245	3,105	6,210	4,140	5,175	5,175
Contraceptive (2,7)	11,931	1,326	10,606	2,651	9,280	3,977	7,954	5,303	6,629	6,629
German (7,13)	18,000	2,000	16,000	4,000	14,000	6,000	12,000	8,000	10,000	10,000
Heart (5,8)	3,545	394	3,151	788	2,757	1,182	2,363	1,576	1,970	1,970
Zoo (1,15)	1,454	162	1,293	323	1,131	485	970	646	808	808
Srinagarind (6,2)	24,962	2,774	22,189	5,547	19,415	8,321	16,642	11,094	13,868	13,868

*c = complete data, i = incomplete data

Next, K-fold cross-validation was used to decrease the bias of the test results [46]. This is an effective method of improving the evaluation and comparison of learning algorithms by dividing the data into K segments. In each iteration, one of the K segments is used to examine the model, and the other K-1 segments are combined to form a training set. This study used a tenfold cross-validation intelligent classifier system to reduce the bias associated with random sampling [47, 48].

In the classification phase, after different techniques individually imputed the missing values of the incomplete training subset, each training subset was used to train an SVM classifier. The testing subset was used to examine the classification accuracy of the SVM classifier. The MCAR mechanism for the incomplete dataset was implemented ten times for each missing rate, to avoid biased results, as indicated in "Missing value imputation" section.

During the MVI process, the proposed AMICC approach was compared to baseline approaches consisting of statistical methods (Mean/Mode imputation), machine learning methods (SVM [42], KNN [27], and RF [43]), and a class center-based MVI approach (CCMVI [35]). In statistical MVI methods, the mean/mode are common statistical measurements used to replace all missing values with the mean/mode value. In machine learning MVI techniques, SVMs are effective in various pattern recognition problems and provide superior classification performance due to their modeling flexibility. KNN is the most widely used data mining technique and was developed based on missing value imputation. Additionally, the RF significantly improves correlation; it is reasonable for missing data ranging from moderate to high. On the other hand, CCMVI is based on determining the class center and using the distances between the class center and other observed data to define a threshold for later imputation.

In the evaluation phase, the accuracy of the results obtained from the model is defined, as described by Eq. (4).

$$\begin{aligned} Accuracy &= \frac{\sum_{i=1}^{N} f(n_i)}{N}, n_i \in N \\ f(n) &= \begin{cases} 1 & \text{if } classify(n) = nc \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where N is the number of data points in the dataset to be classified (the test set), $n \in N$, and nc are the original class of item n. Function f is equal to 1 if $classify(n) = nc$; otherwise it is 0.

The root-mean-square error (RMSE) is a commonly used metric for comparing the actual values to the values imputed by various MVI techniques [12, 22, 41]. This measure is solely appropriate for numerical data values [35]. The RMSE of a model prediction for an estimated variable of X_{model} is given below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(X_{obs} - X_{model})^2}{N}} \tag{5}$$

where X_{obs} is the observed value and X_{model} is the modeled value. This study used the RMSE to measure the error of the imputation method because a relatively high RMSE is undesirable. The smaller the error is, the more accurate the model.

In addition to classification accuracy, the hit rate is the number of hits divided by the size of the test dataset. The predicted rating is called a hit if its rounded value is equal to the actual rating identified in the test dataset. The hit rate can be used to evaluate the performance of a model for categorical data [35], as it represents the percentage of instances where the model correctly predicts the actual class of an observation, as described by Equation (6).

$$Hit\ rate = \frac{n_{hits}}{n_{total}} \times 100\% \tag{6}$$

where n_{hits} is the number of hits associated with the actual rating and n_{total} is the number of test samples.

The performance of the proposed predictive model was measured using the accuracy, RMSE, and hit rate to determine the efficiency of the proposed model compared to that of other existing methods.

Accuracy analysis

In "Experimental setup" section, Tables 7, 8, and 9 summarize the average classification performance of numerical, categorical, and mixed dataset types, respectively. The results in Table 7 show the average classification of the Mean method for numerical datasets, while the results in Table 8 show the average classification of the Mode method for categorical datasets. The average results indicated that the AMICC method achieved the highest accuracy across all dataset types, at least 79.349%, 87.865%, and 77.721%, respectively, and outperformed the other methods significantly ($p < 0.001$). Similarly, Table 9 shows the average results for real data from Srinagarind hospital, revealing that the AMICC approach attained the maximum accuracy of at least 81.094%.

Subsequently, the CCMVI method outperformed the SVM, KNN, and RF methods, all of which produced comparable average results, whereas the Mean/Mode method performed poorly.

Table 7 Average classification accuracies of the MVI methods for the numerical datasets

Dataset	Accuracy (%)					
	Mean	SVM	KNN	RF	CCMVI	AMICC
Blood	73.409	73.832	75.508	74.808	76.535	76.194
Ecoli	70.708	70.833	75.310	75.697	77.107	77.480
Glass	60.056	60.766	63.804	64.869	63.664	69.876
Ionosphere	90.729	90.712	92.308	91.351	93.168	91.377
Iris	88.333	88.627	93.787	94.000	95.080	90.756
Liver cancer	57.977	58.255	58.725	58.771	58.777	57.976
Optdigits	53.820	55.247	54.234	67.799	66.024	97.897
Pima	64.886	64.895	65.353	65.069	65.385	79.366
Sonar	53.923	53.990	56.567	55.904	58.346	87.532
Wine	85.022	83.933	86.764	86.404	89.112	98.127
Yeast	36.956	37.208	39.071	39.365	39.058	47.906
Column 2C	67.729	67.677	67.858	67.607	67.839	84.452
Column 3C	48.381	48.219	48.535	48.555	48.626	72.602
Average	65.533 (6)	65.707 (5)	67.525 (4)	68.477 (3)	69.132 (2)	79.349 (1)

Table 8 Average classification accuracies of the MVI methods for the categorical datasets

Dataset	Accuracy (%)					
	Mode	SVM	KNN	RF	CCMVI	AMICC
Balance scale	60.310	60.963	60.451	63.034	67.222	79.285
Breast cancer	95.285	95.282	95.951	96.026	96.321	98.713
Lymphography	74.662	75.703	76.959	75.594	82.892	87.072
Promoters	78.943	79.698	86.792	84.868	94.189	95.472
Spect	79.371	79.251	79.026	79.401	80.389	86.941
Tic tac toe	74.342	76.929	80.965	80.403	78.854	79.708
Average	77.152 (6)	77.971 (5)	80.024 (3)	79.888 (4)	83.311 (2)	87.865 (1)

Table 9 Average classification accuracies of the MVI methods for the mixed datasets

Dataset	Accuracy (%)					
	Mean/Mode	SVM	KNN	RF	CCMVI	AMICC
Abalone	17.950	17.968	18.957	19.060	20.074	30.384
Acute	93.900	96.800	98.867	96.933	98.850	99.389
Card	56.814	56.832	56.919	56.748	57.675	91.681
Contraceptive	52.986	53.101	55.458	56.451	59.589	59.249
German	69.958	70.994	70.176	69.988	71.774	83.707
Heart	54.851	54.970	55.492	54.990	56.198	86.359
Zoo	84.475	85.545	86.614	88.139	89.980	89.901
Srinagarind	73.416	73.174	74.002	73.735	74.420	81.094
Average	63.044 (6)	63.673 (5)	64.561 (3)	64.506 (4)	66.070 (2)	77.721 (1)

Figures 13, 14, and 15 show the average classification performance of numerical, categorical, and mixed datasets, respectively. Each bar represents the variation in accuracy findings over five distinct missing rate ranges (10–50%). The high accuracy

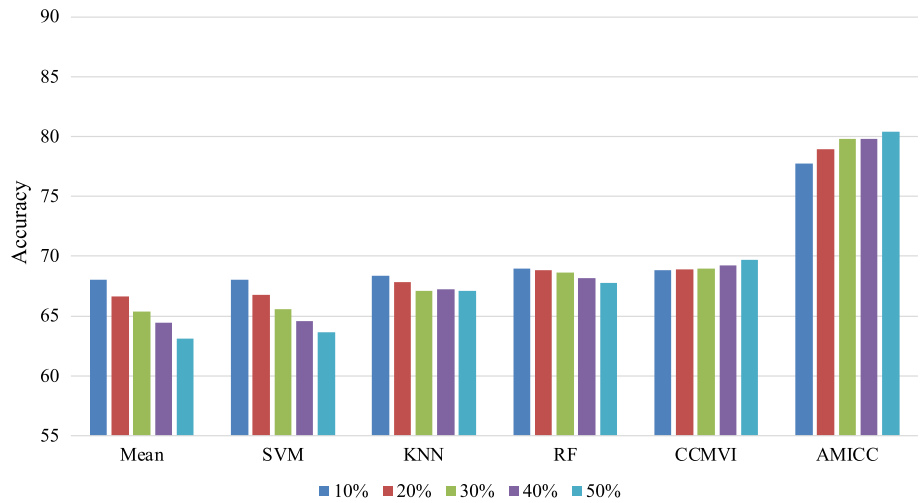


Fig. 13 Classification accuracies for the different MVI methods and the different missing rates in the numerical datasets

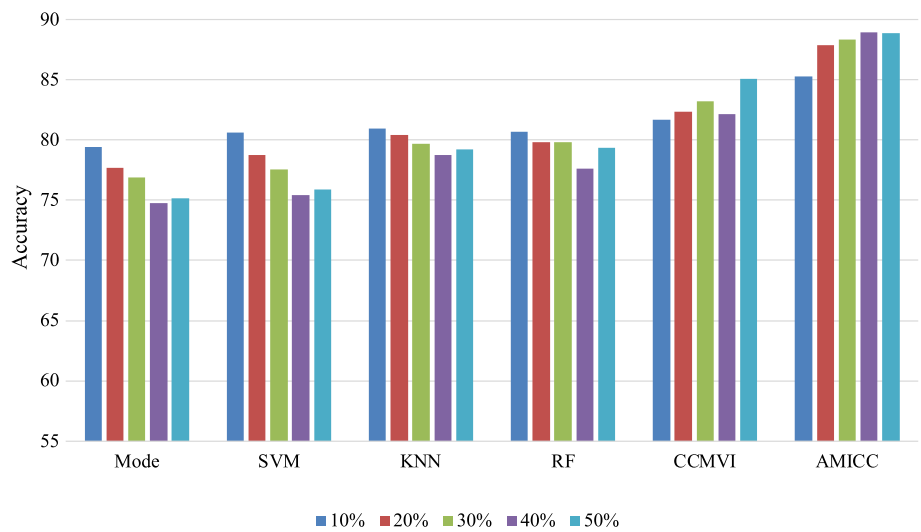


Fig. 14 Classification accuracies for the different MVI methods and the different missing rates in the categorical datasets

results are consistent with a low missing value rate for the Mean/Mode, SVM, KNN, and RF methods [9, 18, 35]. On the other hand, the CCMVI and AMICC methods produce highly accuracy results consistent with a high missing rate and achieve higher accuracy than the other methods.

For AMICC and CCMVI, the results performed well despite high missing values because they replaced missing values with class statistics values. Thus, if the missing rate was high and the number of mean/mode values used to replace missing values increased, the results were highly accurate.

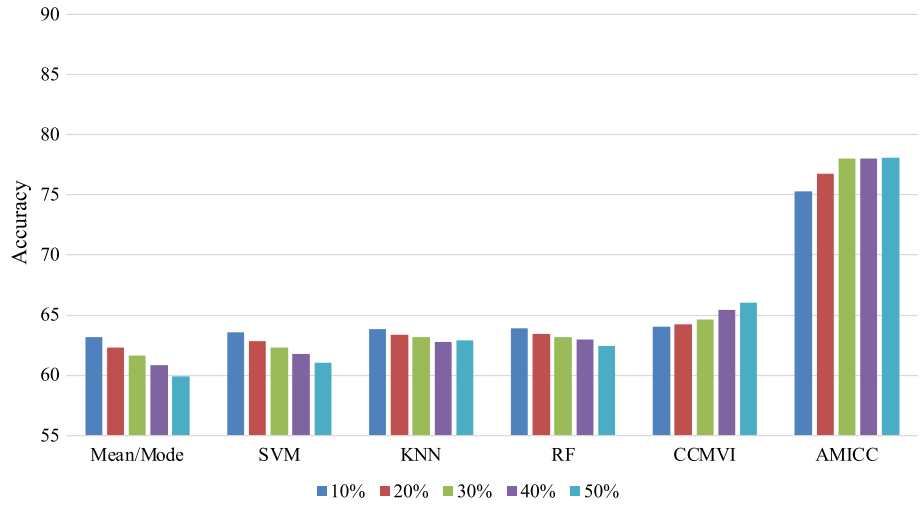


Fig. 15 Classification accuracies for the different MVI methods and the different missing rates in the mixed datasets

Table 10 Average RMSEs of the MVI methods for the numerical datasets

Dataset	RMSE					
	Mean	SVM	KNN	RF	CCMVI	AMICC
Blood	0.515	0.511	0.495	0.502	0.484	0.488
Ecoli	2.231	2.220	1.925	1.855	1.809	1.740
Glass	1.623	1.655	1.524	1.426	1.334	1.126
Ionosphere	0.304	0.304	0.277	0.294	0.261	0.294
Iris	0.338	0.336	0.246	0.243	0.221	0.303
Liver cancer	0.648	0.646	0.642	0.642	0.642	0.648
Optdigits	2.453	2.524	2.591	2.673	2.891	0.705
Pima	0.593	0.592	0.589	0.591	0.588	0.454
Sonar	0.678	0.678	0.659	0.664	0.645	0.353
Wine	0.451	0.464	0.400	0.404	0.364	0.135
Yeast	2.280	2.283	2.212	2.217	2.185	1.961
Column 2C	0.568	0.569	0.567	0.569	0.567	0.386
Column 3C	1.218	1.220	1.216	1.215	1.215	0.716
Average	1.069 (5)	1.077 (6)	1.026 (4)	1.023 (3)	1.016 (2)	

RMSE and hit rate analysis

Tables 10, 11, 12, and 13 illustrate the distribution of the RMSE/hit rate values over all the experiments performed on the numerical, categorical, and mixed datasets, respectively. Each result contains the RMSE/hit rate values attained by each imputation method.

Tables 10 and 11 show the average RMSEs of all missing rates of the MVI methods for the numerical and mixed datasets, respectively. The AMICC method outperformed the other methods, with the RMSEs for numerical and mixed datasets under 0.716 and 0.785, respectively. Following the best approach, the CCMVI method for the numerical under 1.016 and for mixed datasets under 0.993. The other MVI methods demonstrated similar average RMSE results.

Table 11 Average RMSEs of the MVI methods for the mixed datasets

Dataset	RMSE					
	Mean/Mode	SVM	KNN	RF	CCMVI	AMICC
Abalone	3.251	3.251	3.223	3.210	3.217	2.898
Acute	0.229	0.144	0.066	0.158	0.081	0.049
Card	0.657	0.657	0.656	0.657	0.651	0.287
Contraceptive	1.084	1.082	1.044	1.029	0.996	1.026
German	0.548	0.539	0.546	0.548	0.531	0.403
Heart	0.672	0.671	0.667	0.671	0.662	0.369
Zoo	1.157	1.065	0.952	0.903	0.815	0.464
Srinagarind	0.518	0.521	0.511	0.513	0.507	0.502
Average	1.015 (5)	0.991 (4)	0.958 (3)	0.961 (3)	0.933 (2)	0.750 (1)

Table 12 Average hit rates of the MVI methods for the categorical datasets

Dataset	Hit rate (%)					
	Mode	SVM	KNN	RF	CCMVI	AMICC
Balance scale	30.406	30.749	30.179	31.514	33.587	39.360
Breast cancer	33.671	33.645	34.109	34.163	34.272	33.705
Lymphography	47.568	48.095	46.432	46.203	49.608	53.333
Promoters	41.321	41.358	42.660	41.604	46.528	48.365
Spect	79.341	79.094	78.142	79.401	76.360	72.709
Tic tac toe	57.296	58.050	58.177	57.850	57.382	56.451
Average	48.267 (6)	48.499 (3)	48.283 (5)	48.456 (4)	49.623 (2)	50.654 (1)

Table 13 Average hit rates of the MVI methods for the mixed datasets

Dataset	Hit rate (%)					
	Mean/Mode	SVM	KNN	RF	CCMVI	AMICC
Abalone	0.199	0.192	0.300	0.296	0.387	0.100
Acute	46.250	47.067	48.433	48.317	48.667	48.556
Card	2.339	2.362	1.835	1.733	2.467	41.768
Contraceptive	6.955	6.974	8.045	8.087	9.946	13.732
German	0.180	1.026	0.224	0.000	1.774	22.320
Heart	53.195	53.274	54.125	54.185	54.257	50.253
Zoo	18.871	18.832	18.911	19.683	19.802	19.802
Srinagarind	73.187	72.918	73.612	73.606	73.604	73.797
Average	25.147 (6)	25.331 (5)	25.686 (4)	26.738 (3)	26.363 (2)	33.791 (1)

On the other hand, Table 11 illustrates the average result for real data from Srinagarind hospital, demonstrating that the AMICC technique outperformed the other methods, with RMSEs for mixed datasets under 0.502.

Tables 12 and 13 show the average hit rates, also known as recall or sensitivity, of all missing rates of the MVI methods for the categorical and mixed datasets, respectively. The AMICC method outperformed the other methods with the hit rate for categorical and mixed datasets at 50.654% and 33.791%, respectively. The CCMVI

method was the next best method for categorical and mixed datasets at 49.623% and 19.614%, respectively. The Mean/Mode, SVM, KNN, and RF methods demonstrated similar average hit rate results.

Additionally, Table 13 shows the average hit rate for real data from Srinagarind hospital, indicating that the AMICC technique outperformed the other methods, with a hit rate at 73.797% for mixed datasets.

Figure 16 shows the boxplots of the average RMSEs of the missing values for the MVI methods on the numerical and mixed datasets. Each boxplot displays the average RMSE result for all the missing rates. The red pluses indicate the means, and the blue horizontal line within each box shows the median. The AMICC approach had an RMSE value less than those of all the other MVI methods on the numerical and mixed datasets, with values under 1.

Figure 17 shows the boxplots of the average hit rate of the missing values of the MVI methods on the categorical and mixed datasets. Each boxplot displays the average hit rate result for all missing rates. The AMICC approach had a hit rate higher than those of all the other MVI methods. In addition, the AMICC approach had a median value that marked the midpoint of the data in the interquartile range, showing that a hit rate dataset was normally distributed.

Execution time analysis

When choosing a suitable algorithm for missing value imputation, it is necessary to consider not only algorithm accuracy but also algorithm execution time.

Table 14 shows the overall average execution time of the MVI methods for the datasets. The average execution time results show that the Mean/Mode method required the least execution time at 9.612 s. The KNN method had the second-fastest execution time of 10.905 s, increasing approximately 1.293 s over the Mean/Mode method. The KNN method's execution time was much faster than that of the other machine learning techniques because the KNN algorithm is a lazy learning method that does not require a model learning process.

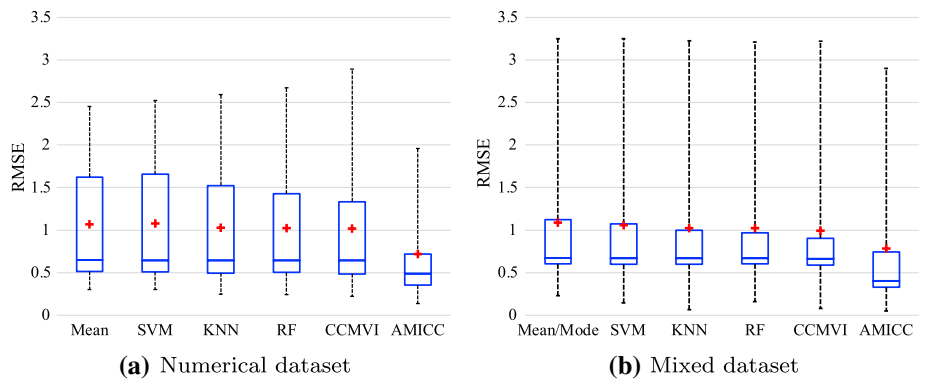
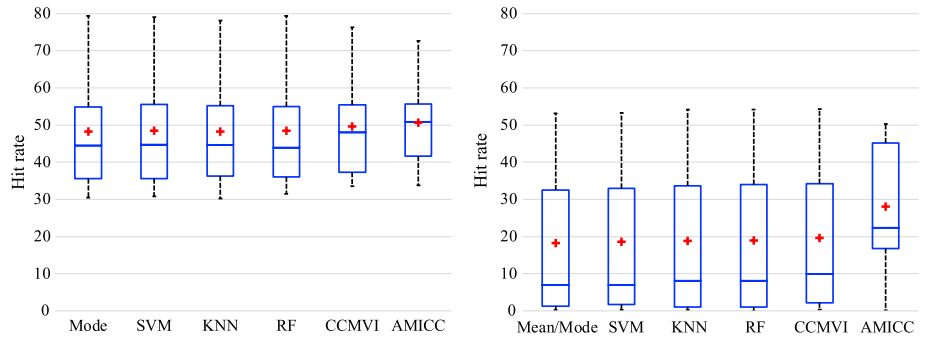


Fig. 16 Boxplots of the average RMSE values for the MVI methods on (**a**) the numerical, and (**b**) the mixed datasets



(a) Categorical dataset **(b)** Mixed dataset

Fig. 17 Boxplots of the average hit rate values for the MVI methods on (**a**) the categorical, and (**b**) the mixed datasets

Table 14 Average execution time of the MVI methods

Dataset type	execution time (second)					
	Mean/Mode	SVM	KNN	RF	CCMVI	AMICC
Numerical dataset	3.520	33.271	4.123	32.296	18.299	8.319
Categorical dataset	23.212	45.004	26.394	35.875	37.627	34.944
Mixed dataset	2.105	10.144	2.198	11.262	10.012	5.912
Average	9.612 (1)	29.473 (6)	10.905 (2)	26.478 (5)	21.979 (4)	16.392 (3)

Table 15 Average classification accuracies for comparison algorithms of the proposed method

Algorithms	Dataset types		
	Numerical	Categorical	Mixed
Baseline	64.821	67.441	61.562
+ Checking attribute type (1)	65.660	75.436	61.664
+ Defining class center (2)	69.220	80.701	64.634
(1) + (2)	69.246	81.597	64.901
(1) + (2) + Replacing outlier with STD	70.857	82.161	72.041
(1) + (2) + Replacing outlier with weight distance and median values (3)	79.349	87.865	77.721

The AMICC method was the third-fastest at 16.392 s and increased from Mean/Mode about 6.780 s; it took more execution time because it needed to calculate the class center distance of the imputed missing data and replace the missing values [27, 35].

Discussion

The experimental results show that the proposed method with multiple imputations using the class center of the average outperforms the other MVI methods. In Table 15, the AMICC approach comprises three important algorithms: the first algorithm (1), which focuses on check attribute type; the second algorithm (2), which defines the class

center; and finally, the third algorithm (3), which replaces outlier with weight distance and median values. These three algorithms are described in the following sections.

Table 15 summarizes the average classification accuracies for the comparison algorithms of the proposed method. AMICC verified the type of attribute before replacing the missing values. From Section "Imputation of missing values", lines 3 to 7 of the pseudo-code in Algorithm 2 for the imputation of missing values checks the attribute type and replaces the missing values with the mean/mode value. For example, when checking attribute type was added to the baseline, the accuracy improved to 65.660%, 75.436%, and 61.664% for the numerical, categorical, and mixed dataset types, respectively. The performance was enhanced because if missing data was substituted with an inappropriate value for the attribute type, the imputed value became noise. In other words, if the pseudo-code for checking the attribute type was removed, the result will be as described in row 1 of Table 15 for the algorithm Baseline.

In addition, the AMICC method outperformed the others because it defined a class center algorithm. The class center was calculated using the mean of the data samples within a particular class, which is similar to the cluster center or centroid concept, which can represent the content of a class. Section "Threshold identification" defines a class center algorithm for lines 8 to 12 of the pseudo-code in Algorithm 1. For example, when a defined class center algorithm was included, the accuracy rose to 69.220%, 80.701%, and 64.634% for the numerical, categorical, and mixed datasets, respectively. Additionally, the efficiency was boosted because if missing data were replaced with an outer class mean value that is not suitable for the class center, the imputed value became inaccurate.

Furthermore, the AMICC method specified a threshold value for outlier detection and replaced it with weight distance and median value. Section "Threshold identification", lines 18 to 21 of the pseudo-code in Algorithm 1, illustrates the threshold identification for a defined threshold value. The imputation of missing value occurred in Algorithm 2, lines 15 to 18 of the pseudo-code. For example, when imputed missing values were combined with weight distance and median values, the accuracy enhances to 79.349%, 87.865%, and 77.721% for the numerical, categorical, and mixed datasets respectively. By comparison, the proposed method outperformed the CCMVI method, which relied on threshold values and replaced outliers by subtracting and adding STD values. Other MVI techniques, such as the SVM, KNN, RF, and Mean/Mode, did not provide a threshold or check for outlier data; consequently, if missing data were replaced and then became outlier data, the imputed value became noise.

Conclusions

Big data has been applied to provide effective solutions in several fields. However, much of the collected big data in various domains contain missing values. In this study, we proposed an adaptive multiple imputations of missing values using the class center (AMICC) approach to produce reasonably promising imputation results. The AMICC method is composed of three modules. The first module focuses on data preprocessing; the incomplete dataset must be normalized on the same scale, then split the incomplete and complete data. The second module determines the threshold by calculating the distance between data samples and their associated class centers. Finally, the third module discusses missing value imputation.

The experiments were conducted using numerical, categorical, and mixed datasets. The AMICC method was compared with two statistical techniques (i.e., the CCMVI and Mean/Mode imputation methods) and three well-known machine learning methods (i.e., the SVM, RF, and KNN algorithms). The results showed that the proposed AMICC method outperformed other techniques on all the datasets by achieving the highest accuracy, the lowest RMSE, and the highest hit rate among all six experimented methods. The performance of the AMICC method was superior because it checked the type of attributes in the dataset and replaced values according to the attribute type. For numerical and categorical variables, the AMICC method replaced missing values with class's mean and mode values, respectively. Additionally, it replaced outlier data with weight distance and median values.

As part of our future work, we intend to investigate the following. We wonder whether different distance functions can be used for defining the threshold values and compared to determine the optimal function. Further, in the case of outlier threshold values, an investigation into the method selection, including the median, STD, and mean methods, may result in increased accuracy and faster computation.

Abbreviations

AMICC: Adaptive multiple imputations of missing values using the class center; CCMVI: Class center-based missing value imputation approach; KNN: K-nearest neighbor; MCAR: Missing completely at random; MVI: Missing value imputation; NaNs: Not a Numbers; RMSE: Root mean square error; RF: Random forest; STD: Standard deviation; SVM: Support vector machine; UCI: University of California Irvine Machine Learning Repository.

Acknowledgements

The authors acknowledge the financial support from both funding parties.

Author contributions

The author confirms the sole responsibility for this manuscript fully as a sole author for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. All authors read and approved the final manuscript.

Funding

This research was funded by the Thailand Research Fund under grant No. RDG6050040 and the Faculty of Engineering, Khon Kaen University, Khon Kaen province under grant No. Ph.D-001/2561 as well as NSERC (Canada) and University of Manitoba.

Availability of data and materials

The datasets used in this study appear in http://archive.ics.uci.edu/ml (accessed on 1 May 2021).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interestss

The authors declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Author details

[1]Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand. [2]Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand. [3]Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada.

Received: 8 November 2021 Accepted: 6 April 2022
Published online: 28 April 2022

References

1. Gao Z, Yang Y, Khosravi MR, Wan S. Class consistent and joint group sparse representation model for image classification in internet of medical things. Computer Commun. 2021;166:57–65.
2. Liu Z-G, Pan Q, Dezert J, Martin A. Adaptive imputation of missing values for incomplete pattern classification. Pattern Recogn. 2016;52:85–95.
3. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Practice. 2017;36(1):3–11.
4. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293–314.
5. Rahm E, Do HH. Data cleaning: problems and current approaches. IEEE Data Eng Bull. 2000;23(4):3–13.
6. Chen C, Liu L, Wan S, Hui X, Pei Q. Data dissemination for industry 4.0 applications in internet of vehicles based on short-term traffic prediction. ACM Trans Internet Technol (TOIT). 2021;22(1):1–18.
7. Schinka JA, Velicer WF, Weiner IB. Handbook of Psychology: Research Methods in Psychology, vol. 2. New Jersey: Wiley; 2013.
8. Khan SI, Hoque ASML. Sice: an improved missing data imputation technique. J Big Data. 2020;7(1):1–21.
9. Xia J, Zhang S, Cai G, Li L, Pan Q, Yan J, Ning G. Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recogn. 2017;69:52–60.
10. Ramezani R, Maadi M, Khatami SM. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Eng J. 2018;57(3):1883–91.
11. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing Appl. 2010;19(2):263–82.
12. Sim J, Kwon O, Lee KC. Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. Expert Syst Appl. 2016;46:485–93.
13. Seijo-Pardo B, Alonso-Betanzos A, Bennett KP, Bolón-Canedo V, Josse J, Saeed M, Guyon I. Biases in feature selection with missing data. Neurocomputing. 2019;342:97–112.
14. Doquire G, Verleysen M. Feature selection with missing data using mutual information estimators. Neurocomputing. 2012;90:3–11.
15. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015;90:84–99.
16. Hron K, Templ M, Filzmoser P. Imputation of missing values for compositional data using classical and robust methods. Comput Stat Data Anal. 2010;54(12):3095–107.
17. Lee MC, Mitra R. Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. Comput Stat Data Anal. 2016;95:24–38.
18. Hamidzadeh J, Moradi M. Enhancing data analysis: uncertainty-resistance method for handling incomplete data. Appl Intell. 2020;50(1):74–86.
19. Ispirova G, Eftimov T, Korošec P, Koroušić Seljak B. Might: statistical methodology for missing-data imputation in food composition databases. Appl Sci. 2019;9(19):4111.
20. Folino G, Pisani FS. Evolving meta-ensemble of classifiers for handling incomplete and unbalanced datasets in the cyber security domain. Appl Soft Computing. 2016;47:179–90.
21. Baraldi AN, Enders CK. An introduction to modern missing data analyses. J School Psychol. 2010;48(1):5–37.
22. Amiri M, Jensen R. Missing data imputation using fuzzy-rough methods. Neurocomputing. 2016;205:152–64.
23. Sanit-in Y, Saikaew KR. Prediction of waiting time in one stop service. Int J Mach Learning Computing. 2019;9:3.
24. Zhang S. Cost-sensitive knn classification. Neurocomputing. 2020;391:234–42.
25. Garciarena U, Santana R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst Appl. 2017;89:52–65.
26. Razavi-Far R, Cheng B, Saif M, Ahmadi M. Similarity-learning information-fusion schemes for missing data imputation. Knowledge-Based Syst. 2020;187:104805.
27. Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci. 2013;233:25–35.
28. Yelipe U, Porika S, Golla M. An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. Computers Elect Eng. 2018;66:487–504.
29. Mesquita DP, Gomes JP, Junior AHS, Nobre JS. Euclidean distance estimation in incomplete datasets. Neurocomputing. 2017;248:11–8.
30. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for dna microarrays. Bioinformatics. 2001;17(6):520–5.
31. Daberdaku S, Tavazzi E, Di Camillo B. A combined interpolation and weighted k-nearest neighbours approach for the imputation of longitudinal icu laboratory data. J Healthcare Inform Res. 2020;4(2):174–88.
32. Cheng C-H, Chan C-P, Sheu Y-J. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Eng Appl Artif Intell. 2019;81:283–99.
33. Fan G-F, Guo Y-H, Zheng J-M, Hong W-C. Application of the weighted k-nearest neighbor algorithm for short-term load forecasting. Energies. 2019;12(5):916.
34. Kiasari MA, Jang G-J, Lee M. Novel iterative approach using generative and discriminative models for classification with missing features. Neurocomputing. 2017;225:23–30.
35. Tsai C-F, Li M-L, Lin W-C. A class center based approach for missing value imputation. Knowledge-Based Systems. 2018;151:124–35.
36. Nugroho H, Utama NP, Surendro K. Class center-based firefly algorithm for handling missing data. J Big Data. 2021;8(1):1–14.
37. Nugroho H, Utama NP, Surendro K. Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. J Big Data. 2021;8:129.
38. Sajidha S, Desikan K, Chodnekar SP. Initial seed selection for mixed data using modified k-means clustering algorithm. Arab J Sci Eng. 2020;45(4):2685–703.

39. Silva-Ramírez E-L, Pino-Mejías R, López-Coello M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. Appl Soft Computing. 2015;29:65–74.
40. Kim T, Ko W, Kim J. Analysis and impact evaluation of missing data imputation in day-ahead pv generation forecasting. Appl Sci. 2019;9(1):204.
41. Pan R, Yang T, Cao J, Lu K, Zhang Z. Missing data imputation by k nearest neighbours based on grey relational structure and mutual information. Appl Intell. 2015;43(3):614–32.
42. Pelckmans K, De Brabanter J, Suykens JA, De Moor B. Handling missing values in support vector machine classifiers. Neural Netw. 2005;18(5–6):684–92.
43. Liu C-H, Tsai C-F, Sue K-L, Huang M-W. The feature selection effect on missing value imputation of medical datasets. Appl Sci. 2020;10(7):2344.
44. Little RJ, Rubin DB. Statistical Analysis with Missing Data. New Jersey: Wiley; 2002.
45. Dua D, Graff C. UCI Machine Learning Repository. 2017. http://archive.ics.uci.edu/ml Accessed 1 May 2021
46. François D, Rossi F, Wertz V, Verleysen M. Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing. 2007;70(7–9):1276–88.
47. Ling H, Qian C, Kang W, Liang C, Chen H. Combination of support vector machine and k-fold cross validation to predict compressive strength of concrete in marine environment. Construction and Building Materials. 2019;206:355–63.
48. Jiang G, Wang W. Error estimation based on variance analysis of k-fold cross-validation. Pattern Recogn. 2017;69:94–106.

Publisher's Note