Journal of Big Data

RESEARCH

Open Access

Normalization and outlier removal in class center-based firefly algorithm for missing value imputation

Heru Nugroho*, Nugraha Priya Utama and Kridanto Surendro

*Correspondence:
heru@tass.telkomuniversity.ac.id
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia

Abstract

A missing value is one of the factors that often cause incomplete data in almost all studies, even those that are well-designed and controlled. It can also decrease a study's statistical power or result in inaccurate estimations and conclusions. Hence, data normalization and missing value handling are considered the major problems in the data pre-processing stage, while classification algorithms are adopted to handle numerical features. In cases where the observed data contained outliers, the missing value estimated results are sometimes unreliable or even differ greatly from the true values. Therefore, this study aims to propose the combination of normalization and outlier removals before imputing missing values on the class center-based firefly algorithm method (ON + C3FA). Moreover, some standard imputation techniques like mean, a random value, regression, as well as multiple imputation, KNN imputation, and decision tree (DT)-based missing value imputation were utilized as a comparison of the proposed method. Experimental results on the sonar dataset showed normalization and outlier removals effect in the methods. According to the proposed method (ON + C3FA), AUC, accuracy, F1-Score, Precision, Recall, and AUC-PR had 0.972, 0.906, 0.906, 0.908, 0.906, 0.61 respectively. The result showed combining normalization and outlier removals in C3-FA (ON + C3FA) was an efficient technique for obtaining actual data in handling missing values, and it also outperformed the previous studies methods with *r* and RMSE values of 0.935 and 0.02. Meanwhile, the D_{ks} value obtained from this technique was 0.04, which indicated that it could maintain the values or distribution accuracy.

Keywords: Missing value, Normalization, Outliers, Method, Imputation, Class center, Firefly algorithm

Introduction

In most studies, missing value is a common and serious problem that often leads to biased, inaccurate, or unreasonable conclusions in cases of inappropriate handling [1–10]. When there is no data value for a variable in an observation, it is considered to be missing. The phenomenon of these values is pervasive in clinical studies involving large amounts of data (big data). Meanwhile, big data are extremely large datasets that pose storage and analysis challenges using conventional management techniques [11].

Currently, available analytical methods are only capable of working with complete data [12–15]. Therefore, missing value-related issues present the opportunities of obtaining the right technique as a solution [16].

In classification problems, missing values is a general weakness with the capacity to produce results of an ineffective prediction system [12, 17, 18]. Ignoring these data affects analysis [1, 8, 19–21] or learning outcomes, as well as prediction results on collaborative prediction problems [22]. Furthermore, it has the potential to weaken results and conclusion validities [3, 21]. In the predictive model, incorrect selection of the missing data handling method tends to affect the model's [8, 23] and classifiers' accuracy, as well as performance [24].

According to previous studies, feature normalization has an important effect on classification accuracy [25–28]. Furthermore, in a dataset with numeric feature attributes, the normalization and processing of missing values are regarded as the main problems in the pre-processing stage [29]. Also, a normalized mean interpolation method was developed to solve the missing value in numerical data sets [30]. Several studies have separately analyzed the effects of various normalization techniques and strategies for dealing with missing value on classification performance. However, only a few rated the effects combining the two [29]. Furthermore, applying data normalization has a significant effect on classification and greatly improves the performance of the KNN imputation method [31]. Previous studies have also shown that combining normalization and imputation using the mean is more accurate than traditional mean and median methods [30].

The model-driven imputation algorithm requires that the observable data has no missing values in the dataset, therefore, its characteristics directly affect the results [32]. Training data usually contains noisy data or outliers which affect the final performance of the trained model [33, 34]. From an instance selection perspective, the dataset is bound to contain some noisy data or outliers being observed for missing value imputations. Therefore, the selection is conducted to filter out some of these data, as well as unrepresentative outliers from a given (training) dataset. Also, the selection's performance should be assessed before imputing missing values [32]. Ordinary least squares (OLS) are used to estimate the model's parameters in the linear method. However, the outliers make the estimation of these parameters unreliable [35]. Therefore, the imputation result is not good enough to fulfill the given precision [36], and has a negative effect on the values entered for missing data [37]. To solve this problem, outlier handling should be performed before imputation [36, 37]. The classical method is unable to accurately conduct imputation in the presence of outliers [38], hence, research teams suggested several imputation methods to overcome these problems [37, 39–41].

Previous studies have produced class center-cased firefly algorithm for handling missing values [42], as the development of methods by considering correlation [43]. However, they have failed to consider data normalization and outlier detection before performing the imputation process.

This study proposed a combination of normalization and outlier removals as well as normalization on class center-based Firefly Algorithm. It was also developed from preliminary studies conducted by the author [42]. This contributed an evaluation of the normalization and outlier detection combination's impact on several missing data imputation methods, mean imputation (Mean), random value imputation (Rand), linear regression imputation (Reg), multiple imputation (MI), KNN imputation, decision tree (DT) imputation. The combination of outlier removals and normalization on class center-based Firefly Algorithm has never been used in previous studies. Although, it is an efficient technique for determining the actual value in handling the missing and tends to maintain the true distribution of data values.

Related work

The three mandatory technical issues that should be considered in the process of inputting missing data, selecting experimental data sets, methods, and evaluating imputation results are shown in Fig. 1 [44].

The choice of experimental data set is related to the problem area, the filling and the type of test data, missing data mechanism (MCAR, MAR, MNAR) as well as a percentage (missing rate). According to Lin and Tsai (2020), the normalization and outlier detection's consideration was not discussed in the review paper "Missing value imputation: a review and analysis of the literature (2006–2017)". The effect of normalization and various techniques for handling the missing value strategy on classification performance separately was extensively conducted in previous studies. However, only a few assessed the simultaneous combination effect of standardization and missing data handling methods [29]. Some also showed that combining normalization and imputation techniques produced better accuracy values [30, 31, 45].

In addition to normalization of pre-processing, outliers significantly influence the statistical estimation process (for instance, the sample mean and standard deviation), resulting in either excessively high or low values [46]. Several missing data imputation methods including mean, linear regression, multiple, and class center-based, utilize the mean value. Generally, the training data contains noisy data or outliers with the ability to reduce the learning model's final performance [33, 34]. Therefore, it is necessary to select instances in the observed data set for missing values imputation and to determine the selection performance of instances from the observed data set before the imputation [32]. According to other studies, outliers play an important role in the imputation method's performance. In cases where a dataset contains these data points, mixed models with high flexibility can produce deviations from the true data pattern [47].

It was also reported that imputation results were strongly influenced by the presence of outliers [35–37]. Therefore, outlier handling should be conducted before imputation [36, 37]. Currently, the classical method is unable to perform imputation accurately in

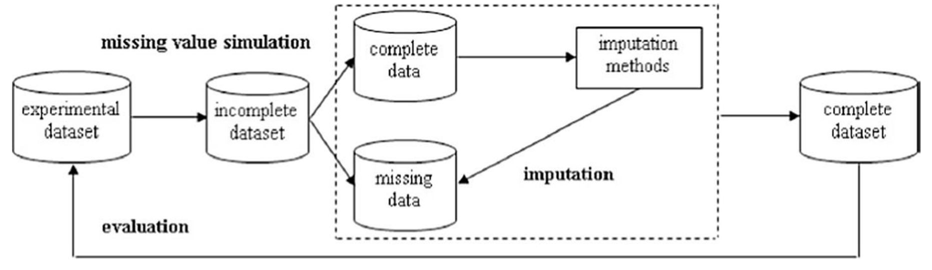


Fig. 1 The experimental design procedure for missing data imputation experimental [44]

the presence of outliers [38], however, various techniques have been proposed as a solution to this problem [37, 39–41].

Research methodology

Methods of handling the missing data

Missing value imputation (MVI) which is a basic method for resolving incomplete dataset problems has been studied for several years [44]. The method of dealing with missing data largely depends on the type and requirements. There are two imputation methods, namely statistics and machine learning [48, 49], which create approximations from chosen observable data to replace the missing value. For each of these real values in the dataset, a single imputation technique was used to generate a specific number because it is less computationally intensive and more cost-effective. Research teams proposed a variety of single-imputation methods. Furthermore, other responses analysis was used to select the highest-scoring response as a general rule of thumb. Alternatively, the value can be obtained by calculating the mean of the available variable values [50].

Regression imputation was utilized when the data available estimate the predictive values in place of missing. Meanwhile, imputation techniques for regression were highly dependent on missing variables [51]. The average vector estimation and covariance matrix of data were calculated using sub-matrices that contained no missing values [52]. Other statistical techniques used for data imputation were multiple and random imputation. Several of these values can be imputed using the multiple imputation method. The three stages of this method are as follows, imputation, analysis, and pooling [53]. Imputation variance was eliminated, while item distribution was preserved, and estimates of the distribution function through imputed data were shown to be consistent and asymptotically normal when using the random imputation method. Imputation variance in the estimated distribution function cannot be eliminated but can be significantly reduced by this method [54]. The statistical methods widely used in previous studies are expectation maximization (EM), linear regression (LR), least squares (LS), and mean/mode are shown in Fig. 2.

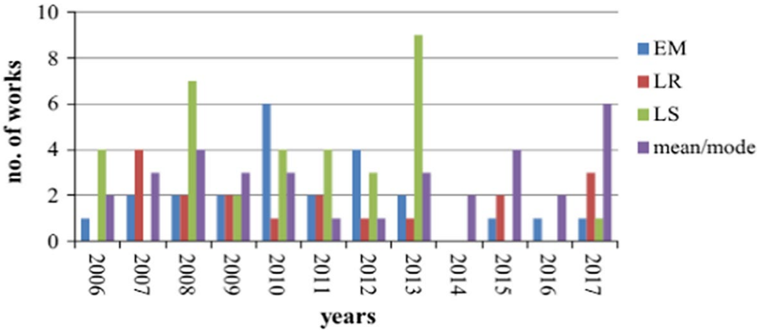


Fig. 2 Distribution in the number of studies using EM, LR, LS, and mean/mode techniques [44]

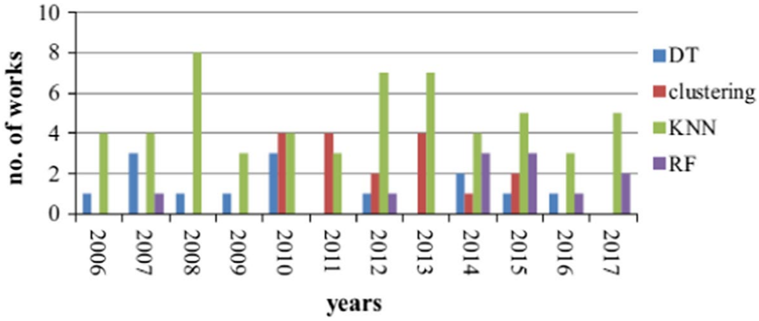


Fig. 3 Distribution in the number of studies using DT, clustering, KNN, and RF techniques [44]

Meanwhile, the commonly used machine learning methods are decision tree (DT), clustering, K-nearest neighbor (KNN), and random forest (RF) [44] are shown in Fig. 3.

Among the various machine learning algorithms, KNN is widely used to suggest missing data due to its simple implementation and relatively high accuracy [15, 48, 55–57]. It is possible to fill in missing data using nearest neighbor (NN) imputation algorithms, which are efficient methods. Each missing value on some records was replaced by a value derived from similar cases across the entire set of records [1]. The k-NN method imputed missing values using those from the k nearest neighbors. It was discovered by minimizing a distance function [58], usually the Euclidean distance [59, 60]. Some tree-based algorithms, like CART and C4.5, have built-in procedures for handling missing value. According to a target variable, a decision tree divides records into several leaves (end nodes) so that records in each leaf have similar values for the target variable. Recordings from a leaf are the best records for estimating missing values in the target variable. Based on the observed values in the leaf, a numerical missing value is imputed in decision tree-based missing value imputation [61].

Normalization and outlier removal in class center-based firefly algorithm for missing value imputation

This study proposed combination of normalization and outlier removals and normalization on class center-based Firefly Algorithm. In addition, it was based on preliminary studies conducted by the author [42]. The fireflies' pattern with a lower light intensity was used, while the group of fireflies with a lower intensity were approximated when the missing data was entered. Furthermore, fireflies with less light and those with brighter intensity were analogous to the missing and complete data attributes respectively.

In the firefly algorithm, the degree of brightness of firefly (I) and the distance r between the firefly i and j demonstrated its attractiveness [62]. The objective function of fireflies was related to their brightness. In this way, the brightness can be used to reveal the most recent position in relation to its objective function $f(x)$ [63].

$$I_i = f(x_i) \tag{1}$$

They were attracted to each other because of their respective attractiveness values β and their distances from each other.

$$\beta_r = \beta_0 e^{-\gamma r^2} \tag{2}$$

In this case, β_0 is the firefly attraction at $r = 0$, and γ is the medium-light absorption coefficient. A firefly i at position x_i which moved to a brighter firefly j at position x_j was described by

$$x_{i+1} = x_1 + \beta_0 e^{-\gamma r^2}(x_j - x_i) + \alpha\left(rand - \frac{1}{2}\right) \tag{3}$$

The pseudocode below summarized the Firefly Algorithm's main steps for handling missing values based on class center [42].

1. Incomplete data sets are divided into complete and incomplete subsets.
2. The calculation of the class center, ($centD_i$), and standard deviation (std) for each class i of the complete subset.
3. Calculation of the distance between class center $centD_i$ and other data samples in class i using the Euclidean distance.

$$Dis(cent(D_i), j) = \sqrt{(x_i - cent(D_i))^2} \tag{4}$$

4. Computation of attribute correlations (R) for the complete subset.

$$R_{x_1x_2} = \frac{n\sum x_1x_2 - \left(\sum x_1\right)\left(\sum x_2\right)}{\sqrt{\left(n\sum x_1^2 - \left(\sum x_1\right)^2\right)\left(n\sum x_2^2 - \left(\sum x_2\right)^2\right)}} \tag{5}$$

5. For each attribute in the incomplete dataset, a value (x) is calculated based on the objective function $f(x)$ and class center values.

$$I(x) = \frac{1}{cent(D_i)} \tag{6}$$

6. The greatest value $I(x) = \frac{1}{centD_i}$ is determined. In cases of data with the largest $I(x)$, the data movement $x^k_{i\_new}$ can be updated using the following movement equation with the assumptions, $\beta_0 = 1, r = Dis(cent(D_i), j)$ and $\alpha \in [0, 1]$.

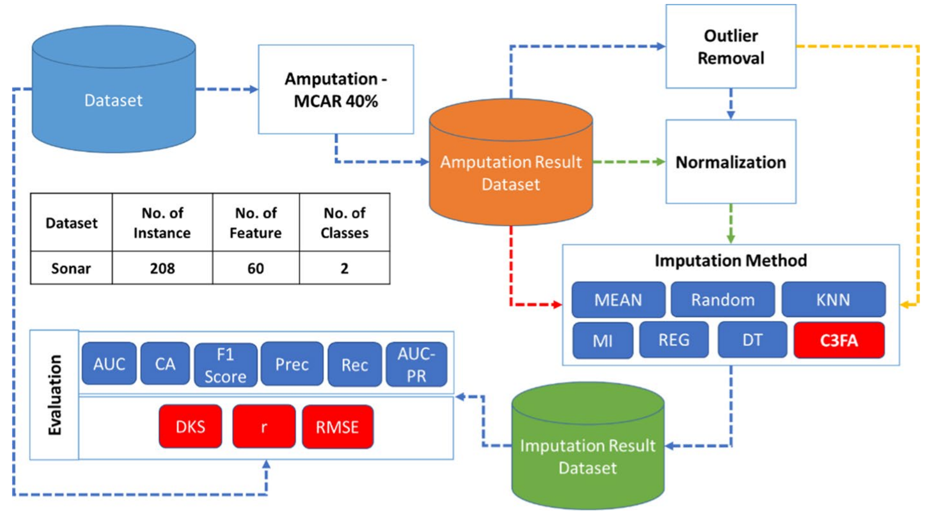
a. The formula below is used when the class center value ($centD_i$) of the attribute containing missing data is equal to the correlated attribute data's $centD_i$.

$$x^k_{i\_new} = x^k_{i\_old} + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \text{ with } \gamma = centD_i \tag{7}$$

b. In cases where the $centD_i$ of the attribute with missing data is below the correlated attribute data's $centD_i$, the formula below is used.

$$x^k_{i\_new} = x^k_{i\_old} + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \text{ with } \gamma = \left({}^{centD_i}/_{R}\right) + \left|diff\ of\ centD_i\right| \tag{8}$$

c. The formula below is used when the $centD_i$ of the attribute containing missing data is greater, compared to the correlated attribute data's counterpart.



Dataset	No. of Instance	No. of Feature	No. of Classes
Sonar	208	60	2

Fig. 4 Block diagram of the proposed method

$$x^{k}_{i\_new} = x^{k}_{i\_old} + \beta_0 e^{-\gamma r^2} \left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \text{ with } \gamma = (centD_i \times R) - \left|diff\ of\ centD_i\right| \tag{9}$$

7. Analysis of imputation result by comparing the distance between the data and class center obtained from the previous imputation value ± *stdev*. This result is determined based on the closest distance.

A class-centered firefly algorithm for handling missing values was previously developed by the author. However, data normalization and outlier removal were not considered before performing the imputation process. Normalization and outlier removal in class center-based Firefly Algorithm for missing value imputation was proposed as a development of previous studies. The block diagram of our imputation and performance evaluation is shown in Fig. 4.

Performance evaluation

Evaluation of machine learning models, for instance, AUC, precision, recall, and F1-score based on the confusion matrix were used to observe the effect of normalization and outlier detection on several imputation methods. This matrix is a very popular technique for solving classification problems, and is suitable for binary classification or classification problems with multiple classes as shown in Table 1 [64].

The confusion matrix represents the predicted and actual state of the data generated by the machine learning algorithm. Precision is the relationship between true positive prediction and overall positive prediction.

$$\text{Precesion} = \frac{TP}{TP + FP} \tag{10}$$

Meanwhile, recall (Sensitivity) is the relationship between the true positive prediction and the overall true positive data.

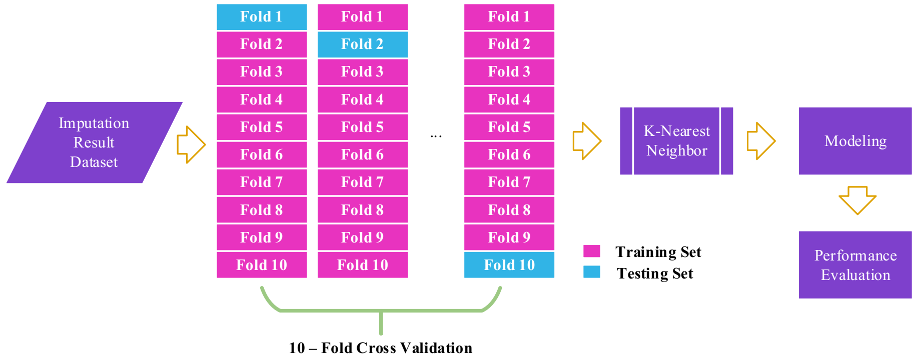


Fig. 5 Method of evaluation for getting AUC, AUC-PR, accuracy, F1-score, precision, and recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

The F1 score is a weighted average comparison of precision and recall.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

AUC (area under the curve) is the ROC (receiver operating characteristic), a curve depicting probability with sensitivity and specificity variables, with a limit value between 0 and 1. This area provided an overview of the model's overall top measurement suitability and was a standard measure used to indicate the prediction result's quality [65]. Various additional metrics used for evaluating the final model or numerous plots, such as ROC and Precision-Recall (PRC)/AUC-PR plots, gave visual representations [66]. The precision-recall (PRC) curve indicated precision values with related sensitivity (recall) levels [67]. Classifier evaluation on imbalanced data sets is more informative when using the precision–recall plot than ROC [68]. AUC, AUC-PR, accuracy, F1-score, precision, and recall can be calculated using machine learning evaluation methods as shown in Fig. 5.

In addition to the utilization of these methods, for instance, models based on the confusion matrix, the proposed imputation method was evaluated based on two approaches, namely predictive accuracy (PAC), which is concerned with the imputation technique's efficiency in obtaining the true data value, and distributional accuracy (DAC). Furthermore, Pearson correlation coefficient (r) and root mean-squared error (RMSE) were used for the PAC assessment [69, 70]. The Pearson correlation coefficient r can be used to measure the correlation between the imputed and actual values through variance, and the degree to which data points tend to deviate from the mean [71]. An effective imputation method should be close to 1 [69, 70]. In cases where x and $\hat{x}$ is the attribute value in the complete and incomplete data, the correlation coefficient r can be calculated according to the formula (13).

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)(\hat{x}_i - \overline{\hat{x}}_i)}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2 \sum_{i=1}^{n} (\hat{x}_i - \overline{\hat{x}}_i)^2}} \tag{13}$$

The root mean square error (RMSE) is a well-known main criterion used to compare the performance of prediction methods by measuring the difference between the estimated value of a given characteristic and the baseline value. In this case, a value closer to 0 results in better imputation [69, 70].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{x}_i)^2} \tag{14}$$

In addition, DAC represents the technical ability to maintain the actual distribution of data values and was assessed in this study using the Kolmogorov–Smirnov Distance (D_{KS}). Given that F_x and $F_{\hat{x}}$ are the empirical cumulative distribution function of x and $\hat{x}$, D_{KS} was calculated using Eq. (15), where a smaller distance value indicated a better interpolation result [69, 70].

$$D_{KS} = \left\| F_x - F_{\hat{x}} \right\| \tag{15}$$

Subsequently, the complete dataset results of classification accuracy were analyzed using algorithms like k-Nearest Neighbor (KNN). As shown in Fig. 6, the KNN algorithm was selected based on previous results, where a classifier was frequently used to evaluate the effectiveness of the interpolation algorithm.

Experimental results

In this study, the first stage was the selection of Sonar dataset obtained from the UCI Machine Learning Repository (www.arsip.Ics.uci.edu/ml) and Kaggle Datasets (www.kaggle.com/datasets). Terry Sejnowski, currently at the Salk Institute and the University of California, San Diego, contributed the dataset to the benchmark collection. The dataset was developed in collaboration with R. Paul Gorman of Allied-Signal Aerospace Technology Center. Furthermore, various aspect angles were used to obtain the signals in the dataset with 90 and 180 degrees for the cylinder and rock respectively. Moreover,

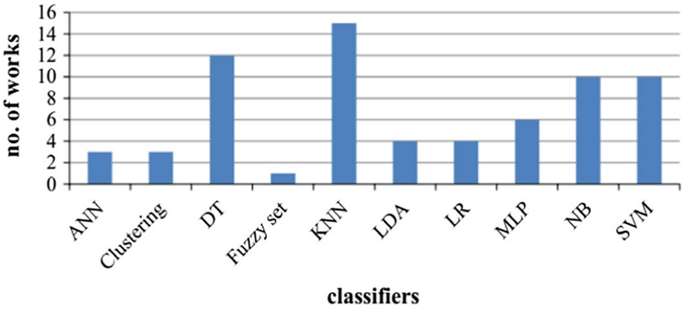
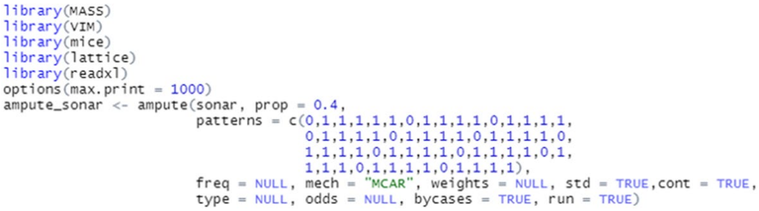


Fig. 6 The Classifier used in previous studies [44]



```
library(MASS)
library(VIM)
library(mice)
library(lattice)
library(readxl)
options(max.print = 1000)
ampute_sonar <- ampute(sonar, prop = 0.4,
                       patterns = c(0,1,1,1,1,1,0,1,1,1,1,0,1,1,1,1,
                                    0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,0,
                                    1,1,1,1,0,1,1,1,1,0,1,1,1,1,0,1,
                                    1,1,1,0,1,1,1,1,0,1,1,1,1),
                       freq = NULL, mech = "MCAR", weights = NULL, std = TRUE,cont = TRUE,
                       type = NULL, odds = NULL, bycases = TRUE, run = TRUE)
```

Fig. 7 The pseudo-code for generate missing values with R

Table 1 Confusion matrix for binary classification

Predictions	Actual	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

there were 60 numbers in each pattern, which ranged from 0.01 to 1.0, therefore, an integrated energy value for a particular frequency band was represented by each number [72].

The experiment was continued by an amputation process of sonar dataset using R programming, where 40% of the data was removed using the MCAR mechanism. Schouten et al. [73] showed that an important aspect of missing data research produced missing values in the complete data set, through the amputation procedure.

1. feed the function of complete dataset,
2. define the missing proportion,
3. specify the missing data patterns,
4. specify the relative occurrence of these patterns,
5. choose between MCAR, MAR, or MNAR mechanism.

The pseudo-code for generating missing values with R is shown in Fig. 7.

The next stage is the imputation process, using seven (7) methods, mean imputation, random value imputation, multiple imputations, regression linear imputation, KNN imputation, decision tree (DT)-based missing value imputation, and proposed technique (conducted in the previous study), class center-based firefly algorithm (C3-FA) [24]. To observe normalization and outlier detection's effect on the missing data imputation method, the simulation process was conducted in 4 ways as follows, (i) imputation, (ii) normalization + imputation, (iii) outlier removal's + imputation, and (iv) outlier removal's + normalization + imputation.

Imputation

At this stage, several imputation methods, mean imputation (Mean), random value imputation (Rand), linear regression imputation (Reg), multiple imputation (MI), KNN imputation, decision tree (DT) imputation, and C3FA, were compared. Table 2 showed the evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for

Table 2 Evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for imputation

Imputation method	Performance evaluation					
	AUC	CA	F1-score	Precision	Recall	AUC-PR
Mean	0.884	0.793	0.791	0.798	0.793	0.69
Rand	0.868	0.764	0.762	0.767	0.764	0.71
Reg	0.908	0.798	0.797	0.801	0.798	0.7
MI	0.899	0.798	0.797	0.801	0.798	0.7
KNN	0.894	0.808	0.807	0.808	0.808	0.67
DT	0.914	0.822	0.821	0.825	0.822	0.7
C3FA	0.927	0.861	0.86	0.862	0.861	0.67

Table 3 Evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for normalization + imputation

Imputation method	Performance evaluation					
	AUC	CA	F1-score	Precision	Recall	AUC-PR
N + Mean	0.922	0.837	0.835	0.841	0.837	0.68
N + Rand	0.906	0.803	0.802	0.805	0.803	0.66
N + Reg	0.926	0.812	0.811	0.815	0.812	0.68
N + MI	0.923	0.832	0.831	0.832	0.832	0.69
N + KNN	0.929	0.822	0.821	0.822	0.832	0.65
N + DT	0.926	0.822	0.821	0.826	0.822	0.66
N + C3FA	0.97	0.909	0.909	0.909	0.909	0.66

seven imputation methods, where k-nearest neighbor (KNN) was the most widely used classifier according to the study.

Based on Table 2, the previous study showed that C3-FA was superior in terms of AUC, accuracy, F1-score, precision, and recall.

Normalization + imputation

One of the contributions of this study is to evaluate the impact of normalization on several missing data imputation methods. Before the imputation process, the values in a dataset were normalized between 0 and 1 using the following formula.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{16}$$

where z_i: ith normalized value in the dataset; x_i: ith value in the dataset; $min(x)$: minimum value in the dataset; $max(x)$: maximum value in the dataset.

Table 3 showed the evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for seven imputation methods with k-nearest neighbor (KNN) classifier after the dataset values were normalized.

According to the experimental results, normalizing the dataset before imputation showed a significant effect. Meanwhile, Table 3 compared the AUC, accuracy, F1-score,

Table 4 Evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for outlier removal's + imputation

Imputation method	Performance evaluation					
	AUC	CA	F1-score	Precision	Recall	AUC-PR
O + Mean	0.893	0.826	0.823	0.829	0.826	0.68
O + Rand	0.904	0.826	0.825	0.825	0.826	0.72
O + Reg	0.921	0.865	0.865	0.866	0.865	0.68
O + MI	0.909	0.86	0.859	0.859	0.86	0.72
O + KNN	0.915	0.854	0.853	0.853	0.854	0.7
O + DT	0.937	0.879	0.878	0.879	0.879	0.71
O + C3FA	0.922	0.83	0.828	0.833	0.83	0.69

precision, and recall results. However, these results showed the proposed method (N + C3FA) outperformed the others.

Outlier removal's + imputation

Another contribution of this study is the effect evaluation of outliers on imputation results. The local outlier factor (LOF) method was used to detect outliers at this stage. It was also calculated using the five-step method.

1. Calculation of the distance.
2. Kth-nearest neighbor distance calculation.
3. Calculation of the K-nearest neighbor.
4. Calculation of the local reachability density (LRD).
5. Analysis.

In this stage, LOF was an algorithm that identified outliers in the dataset before imputation. Meanwhile, KNN classifier was used to evaluate seven imputation methods after the removal of an outlier from the dataset. The evaluation results are presented in Table 4.

Based on these experiments, the removal of an outlier (O) before the imputation process was effective on the evaluation results. Table 4 compares the Accuracy, F1-score, precision, and recall results and the comparison results show the combination outlier removals and decision tree (DT) imputation (O + DT) is outperforms the others.

Outlier removal's + normalization + imputation

The main contribution of this study is the proposed combination of normalization and outlier removals before imputing missing values in the class center-based firefly algorithm method (ON + C3FA). Local outlier factors (LOF) were used in this stage to identify outliers in a dataset, and values were normalized between 0 and 1 as they were in the previous stage of the analysis process. Furthermore, KNN classifier was used to evaluate seven imputation methods after an outlier was removed from the dataset and the values were normalized between 0 and 1. Table 5 summarized the evaluation's results and conclusions.

Table 5 Evaluation results using AUC, accuracy, F1-score, precision, recall, and AUC-PR for outlier removal's + normalization + imputation

Imputation method	Performance evaluation					
	AUC	CA	F1-score	Precision	Recall	AUC-PR
ON + Mean	0.947	0.843	0.843	0.843	0.843	0.66
ON + Rand	0.921	0.831	0.831	0.831	0.831	0.66
ON + Reg	0.955	0.883	0.882	0.884	0.883	0.67
ON + MI	0.942	0.877	0.876	0.878	0.877	0.66
ON + KNN	0.949	0.877	0.877	0.877	0.877	0.67
ON + DT	0.955	0.867	0.866	0.867	0.867	0.55
ON + C3FA	0.972	0.906	0.906	0.908	0.906	0.61

Table 6 Comparison result of RMSE, DKS, and r (ON + C3FA vs C3FA)

	C3FA	ON+C3FA
RMSE	0.026	0.02
D_{ks}	0.04	0.04
R	0.932	0.935

Experimental results showed the combination of normalization and outlier removal effect in the imputation methods. Table 5 compared the AUC, accuracy, F1-score, precision, and recall results. However, the comparison results indicated that the proposed method (ON + C3FA) outperformed the others.

This technique was also evaluated based on RMSE, D_{KS}, and r values. Table 6 showed the result of this evaluation and comparison with the previous method (C3FA).

The value of RMSE and r was better, compared to the previous study [42]. Based on the simulation results, the Pearson correlation coefficient (r) value was close to 1, while root mean squared error (RMSE) was close to 0. The result showed that combining normalization and outlier removals in the C3-FA (ON + C3FA) method was an efficient technique for obtaining the actual data in handling missing values. Meanwhile, the D_{KS} result obtained from this technique was 0.04. This indicated that the proposed method was able to maintain the distribution of the values or the distribution accuracy.

Discussion

The experimental results show that normalizing and removing outliers separately or simultaneously affects the performance of the imputed results. Previous studies have also shown combining normalization and imputation using the mean, produces more accurate than traditional mean and median methods [30, 31]. The effect normalization on imputation can be seen from Tables 2, 3. There was an improvement in the values of AUC, accuracy, F1-score, precision, and recall after the data normalization process was carried out before imputation.

The removal of outliers from the data has an impact on the performance of the imputed results in addition to normalizing the data. This is in line with the facts that outliers play an important role in the imputation method's performance [47] and the

Table 7 Comparison result of AUC, accuracy, F1-score, precision, recall, and AUC-PR (ON + C3FA vs C3FA)

Imputation method	Performance evaluation					
	AUC	CA	F1-score	Precision	Recall	AUC-PR
C3FA	0.927	0.861	0.86	0.862	0.861	0.67
ON + C3FA	0.972	0.906	0.906	0.908	0.906	0.61

classical method is unable to accurately conduct imputation in the presence of outliers [38]. The effect outlier on imputation can be seen from Tables 2, 4. There was an improvement in the values of AUC, accuracy, F1-score, precision, and recall after outlier removals process was carried out before imputation for all methods.

When we compare the experimental results in Tables 2, 5, we get new knowledge that removing outliers and normalizing the data before the imputation process produces better values of AUC, accuracy, F1-score, precision, and recall. However, the comparison results indicated that our proposed method (ON + C3FA) outperformed the others. This research is a continuation of our previous research in this area. Table 7 shows the results of a comparison of the AUC, accuracy, F1-score, precision, recall, and AUC-PR values for the proposed method (ON + C3FA) and our previous method (C3FA). The proposed method is superior to the value of AUC, accuracy, F1-score, precision, and recall.

Imputation methods should ideally reproduce actual values in data, or predictive accuracy (PAC), while also preserving their distribution, or distributional accuracy (DAC). The Pearson correlation coefficient (*r*) and the root mean squared error (RMSE) are two metrics used to assess the PAC. The ability to keep the true distribution of data values is represented by DAC. The Kolmogorov–Smirnov distance (D_{KS}) was used to evaluate it. According to the results of the evaluation for PAC and DAC in Table 6, the proposed method (ON + C3FA) outperforms our previous method (C3FA).

Conclusion

Incomplete research data is frequently caused by missing values. Almost all studies, including those that are well-designed and controlled, have missing value. This can reduce the statistical power of a study, resulting in erroneous estimations and conclusions. In the data pre-processing step, data normalization and missing value, handling were deemed critical, while classification techniques were used to handle numerical features. In addition, when observed data contains outliers, the estimated missing values can be incorrect or possibly significantly different from the actual values. Adaptive search methods like those used in the Firefly algorithm should be implemented to avoid missing values in a dataset. Furthermore, the class center's initial objective feature helped to determine the best imputation value. Therefore, the Firefly algorithm can determine the closest approximation to the actual value. Combining normalization and imputation techniques has been shown in previous studies to improve accuracy values [30]. Meanwhile, others emphasized the significance of detecting outliers in the observed dataset before missing values imputation [32].

This study proposed combination of normalization and outlier removals as well as normalization on class center-based Firefly Algorithm. Furthermore, it was developed from preliminary studies conducted by the author [42]. This study contributed an evaluation of normalization and outlier detection combination's impact on several missing data imputation methods. Based on simulation results using the sonar dataset and seven (7) imputation methods, it was concluded that outlier removal (O) and normalization (N) conducted before the imputation process affected the results. The simulation results demonstrated that the AUC, accuracy, precision, F1-score, with KNN classifier recall, were superior to the imputations of mean, random, linear regression, multiple, KNN, and C3-FA methods without prior outlier removal and normalization (see Tables 3, 4, 5). However, the comparison results showed that the proposed method (ON + C3-FA) outperformed the others (see Table 5).

The area under the receiver operating characteristic curve was denoted by the AUC. The proposed method had the highest AUC, 0.972 compared to others. This indicated that the model was more accurate at classifying instances. However, in the evaluation using AUC-PR, the proposed method gave the smallest value compared to the other methods. When evaluating classifiers on imbalanced datasets, the precision-recall plot or AUC-PR is more informative than the ROC plot [67, 68]. However, this result was obtained from the current study since the experimental data used were not imbalanced.

Integrating outlier identification and normalization into the Firefly Algorithm for managing missing data based on the class center is an excellent method for obtaining the actual data value. This was indicated by the Pearson's correlation coefficient (r) and root mean squared error (RMSE) values that were close to 1 and 0. In addition, the proposed technique preserved the actual distribution of data values, as shown by D_{ks} average value close to 0. Several issues are needed to be addressed in the future as it would be useful to compare the performance of this approach with tMAR and MNAR mechanisms, while other distance functions can also be utilized.

Abbreviations

C3-FA: Class center-based firefly algorithm; RMSE: Root mean squared error; EM: Expectation maximization; LR: Linear regression; LS: Least squares; DT: Decision tree; KNN: K-nearest neighbor; RF: Random forest; PAC: Predictive accuracy; DAC: Distributional accuracy; AUC: Area under the curve; ROC: Receiver operating characteristic.

Acknowledgements

We would like to thank Institut Teknologi Bandung and Telkom University for supporting this research.

Authors' 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

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

Availability of data and materials

The original dataset used for this study is available in: UCI Machine Learning Repository (www.arsip.lcs.uci.edu/ml); Kaggle Datasets (www.kaggle.com/datasets).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author reports no potential competing interests.

Received: 18 May 2021 Accepted: 12 September 2021
Published online: 09 October 2021

References

1. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. 2016. https://doi.org/10.1186/s12911-016-0318-z.
2. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. 2015. https://doi.org/10.1186/s12874-015-0022-1.
3. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64:402–6.
4. Ma Z, Chen G. Bayesian methods for dealing with missing data problems. J Korean Stat Soc. 2018;47:297–313.
5. Malarvizhi R, Thanamani SA. K-NN classifier performs better than K-means clustering in missing value imputation. IOSR J Comput Eng. 2012;6:12–5.
6. Marlin BM. Missing data problems in machine learning. Toronto: University of Toronto; 2008.
7. Ng CG, Yusoff MSB. Missing values in data analysis: ignore or impute? Educ Med J. 2011. https://doi.org/10.5959/eimj.3.1.2011.or1.
8. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. Int J Res Method Educ. 2016;39:19–37.
9. Rahman MdG, Islam MZ. Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems. 2016;46:389–422.
10. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. Stat Methods Med Res. 2016;25:2021–35.
11. Gupta V, Singh VK, Ghose U, Mukhija P, Pinto D, Singh V. A quantitative and text-based characterization of big data research. IFS. 2019;36:4659–75.
12. Armina R, Mohd Zain A, Ali NA, Sallehuddin R. A review on missing value estimation using imputation algorithm. J Phys Conf Ser. 2017;892:012004.
13. Cao L. Data science thinking. New York: Springer Science + Business Media; 2018.
14. Nishanth KJ, Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing. 2016;218:17–25.
15. Van Hulse J, Khoshgoftaar TM. Incomplete-case nearest neighbor imputation in software measurement data. Inf Sci. 2014;259:596–610.
16. Nugroho H, Surendro K. Missing data problem in predictive analytics. 8th International Conference on Software and Computer Applications—ICSCA '19. Penang: ACM Press; 2019. p. 95–100.
17. Jugulum R. Importance of data quality for analytics. In: Sampaio P, Saraiva P, editors. Quality in the 21st century. Cham: Springer International Publishing; 2016. p. 23–31.
18. Wazurkar P, Bhadoria RS, Bajpai D. Predictive analytics in data science for business intelligence solutions on communication systems and network technologies (CSNT). IEEE: Piscataway; 2017. p. 367–70.
19. Deb R, Liew AW-C. Missing value imputation for the analysis of incomplete traffic accident data. Inform Sci. 2016;339:274–89.
20. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recogn. 2008;41:3692–705.
21. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66.
22. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009;72:1483–93.
23. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus. 2013;2:222.
24. Bhati S, Kumar Gupta MKG. Missing data imputation for medical database: review. Int J Adv Res Comput Sci Softw Eng. 2016. https://doi.org/10.21203/rs.3.rs-538193/v1.
25. Alizadeh NA, Babadi M, Homayouni S. Assessment Of normalization techniques on the accuracy of hyperspectral data clustering. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2017;XLII-4/W4:27–30.
26. Huang H-C, Qin L-X. Empirical evaluation of data normalization methods for molecular classification. PeerJ. 2018;6:e4584.
27. KumarSingh B, Verma K, Thoke SA. Investigations on impact of feature normalization techniques on classifier & performance in breast tumor classification. IJCA. 2015;116:11–5.
28. Rozenstein O, Paz-Kagan T, Salbach C, Karnieli A. Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements. IEEE J Sel Top Appl Earth Observ Remote Sens. 2015;8:2393–404.
29. Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, El-Salhi SMFS. The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data. 2021;6:11.
30. Madhu G, Lalith BB, Sai VK, Naga CG. A normalized mean algorithm for imputation of missing data values in medical databases. Innov Electron Commun Eng. 2020. https://doi.org/10.1007/978-981-15-3172-9_72.
31. Christobel A, Prakasam S. The negative impact of missing value imputation in classification of diabetes dataset and solution for improvement. IOSRJCE. 2012;7:16–23.

32. Huang M-W, Lin W-C, Tsai C-F. Outlier removal in model-based missing value imputation for medical datasets. J Healthc Eng. 2018;2018:1–9.
33. Garcia S, Derrac J, Cano JR, Herrera F. Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Trans Pattern Anal Mach Intell. 2012;34:417–35.
34. Leyva E, González A, Pérez R. Three new instance selection methods based on local sets: a comparative study with several approaches from a bi-objective perspective. Pattern Recogn. 2015;48:1523–37.
35. Wada K. Outliers in official statistics. Jpn J Stat Data Sci. 2020;3:669–91.
36. Kim M-G, Shin K-I. A multiple imputation for reducing outlier effect. Korean J Appl Stat. 2014;27:1229–41.
37. Branden KV, Verboven S. Robust data imputation. Comput Biol Chem. 2009;33:7–13.
38. Toka O, Çeťin M. Imputation and deletion methods under the presence of missing values and outliers: a comparative study. Gazi Univ J Sci. 2016;29:799.
39. Cheng T-C, Victoria-Feser M-P. High-breakdown estimation of multivariate mean and covariance with missing observations. Br J Math Stat Psychol. 2002;55:317–35.
40. Hubert M, Rousseeuw PJ, Vanden BK. ROBPCA: a new approach to robust principal component analysis. Technometrics. 2005;47:64–79.
41. Kumar N, Hoque MdA, Shahjaman Md, Islam SMS, Mollah MdNH. A new approach of outlier-robust missing value imputation for metabolomics data analysis. CBIO. 2018;14:43–52.
42. Nugroho H, Utama NP, Surendro K. Class center-based firefly algorithm for handling missing data. J Big Data. 2021;8:37.
43. Nugroho H, Utama NP, Surendro K. Performance evaluation for class center-based missing data imputation algorithm. Proceedings of the 9th International Conference on 2020 Software and Computer Applications. Langkawi: ACM; 2020. p. 36–40.
44. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53:1487–509.
45. Pires IM, Hussain F, Garcia NM, Zdravevski E. Improving human activity monitoring by imputation of missing sensory data: experimental study. Future Internet. 2020;12:155.
46. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017;70:407.
47. Quintano C, Castellano R, Rocca A. Influence of outliers on some multiple imputation methods. Metodološki Zvezki. 2010;7:16.
48. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Comput Appl. 2010;19:263–82.
49. Peng L, Lei L. A review of missing data treatment methods. Int J Intel Inf Manag Syst Tech. 2005;8:412.
50. Khan SI, Hoque ASML. SICE: an improved missing data imputation technique. J Big Data. 2020;7:37.
51. He Y. Missing data imputation for tree-based models. Los Angeles: University of California; 2006.
52. Wasito I. Least squares algorithms with nearest neighbour techniques for imputing missing data values. London: University of London; 2003.
53. van Buuren S. Flexible imputation of missing data. US: CRC Press Taylor & Francis Group; 2012.
54. Chen J, Rao JNK, Sitter RR. Efficient random imputation for missing data in complex surveys. Stat Sinica. 2000;10:1153–69.
55. Hu L-Y, Huang M-W, Ke S-W, Tsai C-F. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus. 2016;5:1304.
56. Nugroho H, Utama NP, Surendro K. Comparison method for handling missing data in clinical studies. 9th International Conference on Software and Computer Applications (ICSCA). Langkawi: ICSCA; 2020. p. 6.
57. 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;43:614–32.
58. Wilson DR, Martinez TR. Improved heterogeneous distance functions. Jair. 1997;6:1–34.
59. Strike K, El Emam K, Madhavji N. Software cost estimation with incomplete data. IIEEE Trans Softw Eng. 2001;27:890–908.
60. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520–5.
61. Nikfalazar S, Yeh C-H, Bedingfield S, Khorshidi HA. Missing data imputation using decision trees and fuzzy clustering with iterative learning. Knowl Inf Syst. 2020;62:2419–37.
62. Yang X-S. Firefly algorithm, lévy flights and global optimization. In: Bramer M, Ellis R, Petridis M, editors. Research and development in intelligent systems XXVI. London: Springer, London; 2010. p. 209–18.
63. Farahlina JN, Mohd ZA, Haszlinna MN, Udin A. Machining parameters optimization using hybrid firefly algorithm and particle swarm optimization. J Phys Conf Ser. 2017;892:012005.
64. Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. Data Democr. 2020. https://doi.org/10.1016/B978-0-12-818366-3.00005-8.
65. Yuliansyah H, Othman ZA, Bakar AA. Taxonomy of link prediction for social network analysis: a review. IEEE Access. 2020;8:183470–87.
66. Haibo HE, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263–84.
67. Saito T, Rehmsmeier M, Brock G. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10:e0118432.
68. Saito T, Rehmsmeier M. Precrec: fast and accurate precision–recall and ROC curve calculations in R. Bioinformatics. 2017;33:145–7.
69. Pompeu Soares J, Seoane Santos M, Henriques Abreu P, Araújo H, Santos J. Exploring the effects of data distribution in missing data imputation. Advances in intelligent data analysis XVII. Cham: Springer; 2018. p. 251–63.
70. Santos MS, Soares JP, Henriques AP, Araújo H, Santos J. Influence of data distribution in missing data imputation. Artificial intelligence in medicine. Cham: Springer; 2017. p. 285–94.

71. Oytun M, Tinazci C, Sekeroglu B, Acikada C, Yavuz HU. Performance prediction and evaluation in female handball players using machine learning models. IEEE Access. 2020;8:116321–35.
72. Gorman RP, Sejnowski TJ. Analysis of hidden units in a layered network trained to classify sonar targets. Neural Netw. 1988;1:75–89.
73. Schouten R. Generating missing values for simulation purposes: a multivariate amputation procedure. J Stat Comput Simul. 2018;88:2909–30.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.