Class Center-Based Firefly Algorithm for Handling Missing Data

Heru Nugroho (heru@tass.telkomuniversity.ac.id)
Telkom University https://orcid.org/0000-0002-7460-7687
Nugraha Priya Utama
Institut Teknologi Bandung
Kridanto Surendro
Institut Teknologi Bandung

Research

Keywords: Missing data, Correlation, Imputation, Firefly Algorithm, Class Center

Posted Date: January 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-107394/v2

Version of Record: A version of this preprint was published on February 23rd, 2021. See the published version at https://doi.org/10.1186/s40537-021-00424-y.

Class Center-Based Firefly Algorithm for Handling Missing Data

Heru Nugroho[1]*, Nugraha Priya Utama[2], Kridanto Surendro[3]

School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia

**Corresponding author: heru@tass.telkomuniversity.ac.id*

Abstract

A significant advancement that occurs during the data cleaning stage is estimating missing data. Studies have shown that improper data handling leads to inaccurate analysis. Furthermore, most studies indicate the occurrence of missing data irrespective of the correlation between attributes. However, an adaptive search procedure helps to determine the estimates of the missing data when correlations between attributes are considered in the process. Firefly Algorithm (FA) implements an adaptive search procedure in the imputation of the missing data by determining the estimated value closest to others' value. Therefore, this study proposes a class center-based adaptive approach model for retrieving missing data by considering the attribute correlation in the imputation process (C3-FA). The result showed that the class center-based firefly algorithm (FA) is an efficient technique for obtaining the actual value in handling missing data with the Pearson correlation coefficient (r) and root mean squared error (RMSE) close to 1 and 0, respectively. In addition, the proposed method has the ability to maintain the true distribution of data values. This is indicated by the Kolmogorov–Smirnov test, which stated that the value of D_{KS} for most attributes in the dataset is generally closer to 0. Furthermore, the accuracy evaluation results using three classifiers showed that the proposed method produces good accuracy.

Keywords: Missing data, Correlation, Imputation, Firefly Algorithm, Class Center

1. Introduction

Missing data is a general weakness capable of making the prediction system ineffective [1–3]. Therefore, ignoring it adversely affects the analysis [4–9], learning outcomes, prediction[10] and potentially weakens the validity of the results and conclusions [8,9], thereby leading to the estimation of biased parameters [7,11–14]. Prediction and classification are the principle obligations needed in many areas and spaces to obtain accurate data [15].

Data mining is the festivity commonly found among the most emerging fields of the current epoch. It is associated with extensive data generation, thereby leading to the need to mine interesting trends and patterns

[16]. Data mining methods only work with complete data [1,17–19], however this is dependent on the amount of missing data/ rate and the domain. The Do Not Impute (DNI) method is used to ensure all missing data remains unreplaced. Therefore, the networks need to use their default missing values strategies [20]. However, in practice, most data analysis techniques are not robust to missing values, hence, they are filled in advance [21]. The issues related to the missing data is the research opportunities used to obtain the correct procedures [22].

Class center missing value imputation (CCMVI) is a hybrid process that uses the statistical and machine learning approaches with the baseline imputation method using k-NN to combine the imputation method [23]. Many techniques do not correlate the data attributes due to their categorization suitability [24]. In previous studies, most of the data estimation methods were missing, with the inputs dependent on the attributes [25]. The performance of the imputation algorithm of the missing data is significantly influenced by the correlation structure in the data [1,4,26], as well as its missing distribution [27,28], and the proportion mechanisms [1].

The adaptive search procedure is possible for the new techniques to estimate missing values in accordance with the correlation between variables [29]. This procedure helps in determining the estimation of missing data by optimizing objective functions in each search problem given [30]. In late 2007 and early 2008, X.S Yang developed the Firefly Algorithm (FA), which implemented an adaptive search procedure [31]. Firefly Algorithm is an optimization algorithm inspired by nature which is considered matured due to its inception approximately ten years ago [32]. The behavior of fireflies during which the brighter ones comprises weaker flashes is applied within the imputation of missing data. It is carried out by determining the estimated value closest to the known and replacing the missing ones [9].

This research proposed the class center-based imputation method of missing data using the Firefly Algorithm in the imputation process. Furthermore, the research is developed from preliminary studies carried out by the author [33]. This research contributed to the imputation method based on the class center by considering the correlation of attributes with the imputation stage combined with the Firefly algorithm (C3-FA). In previous studies, the firefly algorithm was considered in the imputation process based on the class center that has never been used. Therefore, the advantage of this method over others is that it is able to reproduce the actual values and maintain adequate distribution without knowing the original data. The proposed technique is tested on the iris, wine, ecoli, and sonar datasets. In addition, the imputation performance is measured based on the predictive, distributional, and classification accuracies.

The remaining section of this research is organized as follows: In the 'Related work' section, a brief review of

some related work on missing data imputation was analyzed. The authors' detailed approaches for handling missing data are proposed in the section 'The Proposed Methods.' Also, the research experiments are presented in the section 'Experimental results.' Discussion, future work, and conclusion are presented in the section 'Discussion and Conclusions.'

2. Related Work

2.1. Methods for Handling Missing data

The methods used to handle missing data are dependent on the type of data and needs. Therefore, the imputation techniques are classified into two types, namely statistical-based and machine learning [34]. The statistical methods widely used in previous studies were Expectation maximization (EM), Linear/logistic regression (LR), Least squares (LS), and mean/mode, while machine learning includes Decision Tree (DT), clustering, k-Nearest Neighbor (k-NN), and Random Forest (RF) [35]. In 2018, Tsai proposed a class center-based missing data imputation method, which is a combination of statistical and machine learning techniques. Figure 1 shows a method for handling missing data.

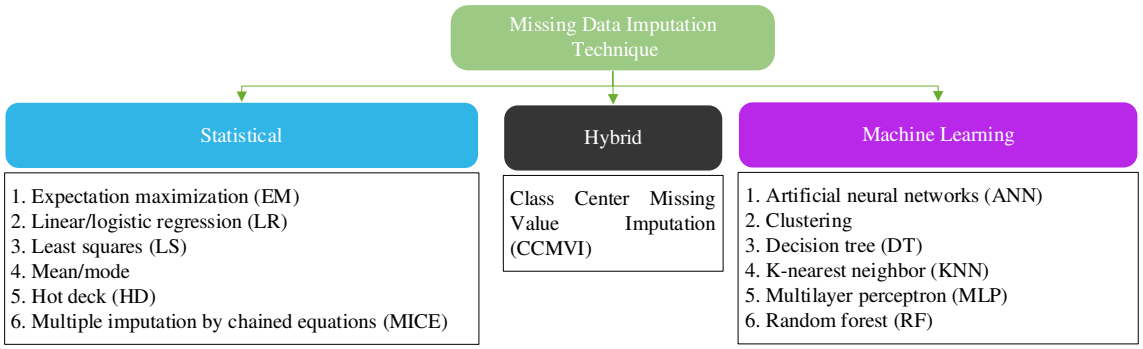

Fig.1 Methods for Handling Missing data

2.2. Class Center Based Missing Value Imputation (CCMVI) Algorithm

The CCMVI is a hybrid method that combines imputations through statistical approaches and machine learning [23]. This study comprises of two modules, namely threshold identification (Module A) and imputation (Module B). The algorithm of each stage is shown in algorithms 1 and 2.

Algorithm 1. Threshold identification [23]

Input: Incomplete dataset D containing M feature dimensions, N classes, and Num data samples
Output: *N* threshold values for *N* classes

```
For j = 1 to Num
If D(j,) has missing value(s) then
    Get the class label of D(j,) and set this class label to variable i
    Put D(j,) to D_i_incomplete
Else
    Get the class label of D(j,) and set this class label to variable i
    Put D(j,) to D_i_complete
End
For i = 1 to N
  Avg(i,)=Average(D_i_complete)
  Std(i,)=Standard Deviation(D_i_complete)
  Get the number of rows in D_i_complete and set to variable Num
End
For j = 1 to Num
    Distance(j)=Euclidean distance(D_i_complete(j,), Avg(i,))
End
Threshold(i) = Median(Distance)
```

Algorithm 2. Imputation [23]

```
Input: D_i_incomplete containing M feature dimensions and N classes
Output: imputed dataset for D_i_incomplete
For i = 1 to N
  Get the number of rows in D_i_incomplete and set to variable Num
  For j = 1 to Num
    If D_i_incomplete(j,) has one missing value
      Get attribute index with the missing value and set to variable miss_attr
      D_i_incomplete(j, miss_attr) = Avg(i, miss_attr)
      Distance = Euclidean distance(D_i_incomplete(j,), Avg(i,))
      If Distance > Threshold(i)
        D_i_incomplete(j, miss_attr) = D_i_incomplete(j, miss_attr) +/- Std(i, miss_attr)
    Else
      Get attribute index with the missing value and set to variable array miss_attr
      Get miss_attr length and set to variable size
      For s = 0 to size-1
        D_i_incomplete(j, miss_attr(s)) = Avg(i, miss_attr(s))
      Distance = Euclidean distance(D_i_incomplete(j,), Avg(i,))
      If Distance > Threshold(i)
        Missing_array = array[size]
        For s = 0 to size-1
          Missing_array(s,)= D_i_incomplete(j, miss_attr(s)) +/- Std(i, miss_attr(s))
            Distance_array(s) = Euclidean distance(Missing_array(s,), Avg(i,))
        Find minimum Distance_array and set index to variable index
        D_i_incomplete(j,) = Missing_array(index,)
```

Some of the issues associated with the CCMVI include the proposed method only uses a simple distance function, namely Euclid, and does not compare performance based on the MAR and MNAR mechanisms. In addition, the simulation results showed that the imputation performance of the proposed CCMVI is not better than the statistical approach for categorical data.

2.3. Effect of Attribute Correlation on Missing data Algorithm

A lot of techniques ignore the correlation between attributes, despite their suitability for categorical data [24]. The missing data imputation algorithm's performance is significantly influenced by factors, such as the correlation structure, missing mechanism, entries distribution, and the values percentage [1].

In a study entitled, "Missing data imputation for the analysis of incomplete traffic accident data," Deb and Liew proposed an algorithm (see in algorithm 3) using a decision tree to determine the correlated attributes. The measure used was the IS and the weighted similarity. Therefore, the algorithm outperforms several popular imputation methods on the traffic accident dataset with the condition that a large number of attributes are categorical [5].

Algorithm 3. Decision Tree and Sampling-Based Missing Value Imputation [5]

Step I: Decompose full dataset into complete and missing values sub-datasets: $D_{Full} = D_{Complete} + D_{Miss}$
Step II: Generate a set of decision trees using C4.5 from $D_{Complete}$ where each missing attribute in D_{Miss} produces a tree
Step III: Assign the records in D_{Miss} into leaves of the decision trees and create tables of related records
Step IV: Impute missing values
  FOR each table T DO
    FOR each missing record R in T DO
      Find records in T that match with the maximum number of non-missing attribute(s) in the missing record R, and let N be the number of such records
      FOR $k = 1$ to N determine
        O_k = possible imputed value(s) from the kth matched record
        IS_k = IS measure computed for O_k
        S_k = weighted similarity measure between the kth matched record and missing record R
        ϑ_k = affinity degree for O_k
      END FOR
      Imputed value(s) is obtained by random sampling from the set of possible imputed values $\{O_1, ..., O_N\}$ based on the sampling probabilities specified by the set of affinity degrees $\{\vartheta_1,...,\vartheta_N\}$
    END FOR
  END FOR

For high dimensional cases, the nearest neighbor approach's imputation method proposes a new distance that explicitly uses the correlation between variables. This method automatically selects the relevant variables contributing to distance, while the simulation results showed that performance depends on their correlation [26]. In addition, the accuracy of the imputation is based on the dependency level of the missing data with other variables in the dataset. This can be ignored for completely uncorrelated attributes [4].

2.4. Firefly Algorithm - Adaptive Search Procedure

The Firefly algorithm applies the adaptive search procedure developed by X.S Yang in late 2007 and early

2008 [31]. It is an optimization algorithm inspired by nature significantly developed for approximately 10 years [32]. Therefore, to properly design the Firefly algorithm, two important issues need to be defined: attraction and light intensity variation. The objective function influences this intensity, and the level for a firefly minimizing problem x is expressed as $I(x) = \frac{1}{f(x)}$. The value of (x) is the level of light intensity on the firefly x, which is inversely proportional to the solution of the problem's objective function to be searched $f(x)$.

The Attractiveness β is of relative value because the light intensity needs to be determined and judged by other fireflies. However, the assessment results tend to differ depending on the distance between one firefly and another rij. Meanwhile, the intensity decreases from the source because it is absorbed by the media, such as air. The Attractiveness (β) with a distance r is calculated in equation (1).

$$\beta = \beta_0 e^{-\lambda r^2} \quad (1)$$

Where β_0 denotes the attraction when there is no distance between the fireflies i.e $(r = 0)$ and $\gamma \in [0, \infty)$ is the light absorption coefficient. The distance between the two fireflies i and j at positions x_i and x_j is the Euclidean distance calculated in equation (2).

$$r_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{n} (x_i^k - x_j^k)} \quad (2)$$

where x_{ik} is the *k-th* component of x_i in firefly.

The movement carried out by firefly i, due to its attraction to j with brighter light intensity, changes its position according to equation (3).

$$x_{i\_new}^k = x_{i\_old}^k + \beta_0 e^{-\lambda r^2} (x_{j\_old}^k - x_{i\_old}^k) + \alpha (rand - \frac{1}{2}) \quad (3)$$

The first term is the old position of the firefly, the second occurs due to attraction, while the third term is a random firefly movement where α is the parameter coefficient and rand is a real number in the interval [0,1]. Most of the Firefly Algorithm implementations use $\beta_0 = 1$, $\alpha \in [0,1]$ and $\gamma \in [0, \infty)$.

3. The Proposed Methods

Basically, there is no generic imputation method used in determining data loss. Therefore, this research is carried out to obtain the missing data by considering the attribute relationship/correlation, which is a development of class center-based imputation. Furthermore, the consideration was carried out in accordance with the weaknesses of the previous method in the category data. Preliminary studies with categorical and attribute correlation yielded very good accuracy, which showed that the class center-based imputation method is an efficient imputation technique used to obtain the data's actual value.

An adaptive search procedure can be used to estimate the missing data that correlates with the associated variables. The Firefly algorithm implements the search procedure in such a way that the bright fireflies attract the weak ones, with the behavior applied in the imputation of missing data. The bright fireflies represent the non-missing data, while the weaker ones denote the missing dataset. The equations related to light intensity, attractiveness, distance, and firefly movement were applied in the imputation process using the class center approach and considering the correlation between the attributes. The result was therefore analyzed by comparing the distance obtained from the imputation ± std result. The overall architecture novel framework for missing data imputation (C3-FA) can be seen in Figure 2.

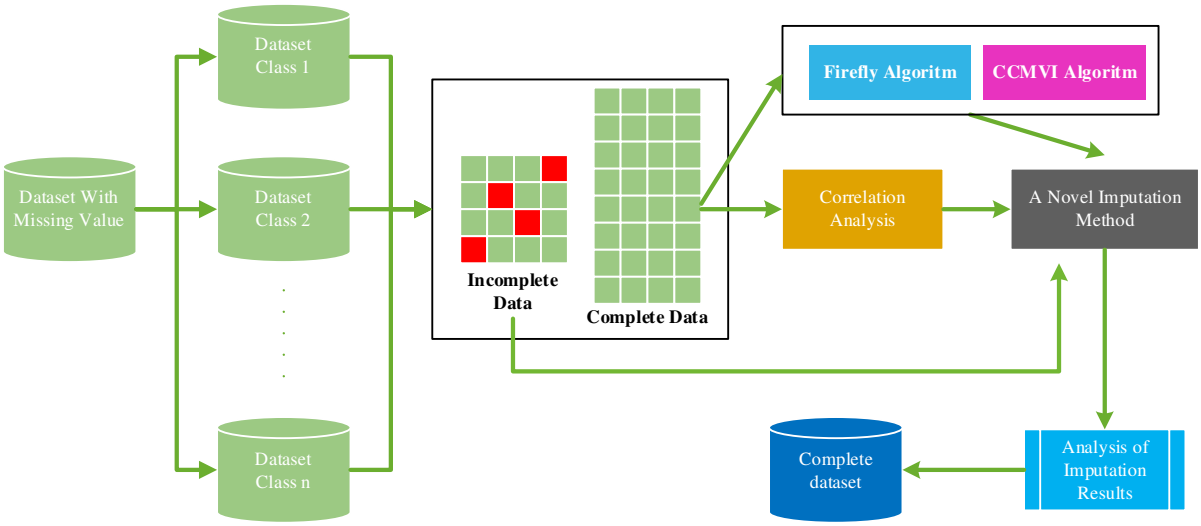

Fig.2 The Overall Architecture Novel Framework for Missing data Imputation (C3-FA)

In general, this research is divided into three parts, data collection, data imputation, and performance imputation, as shown in Figure 3.

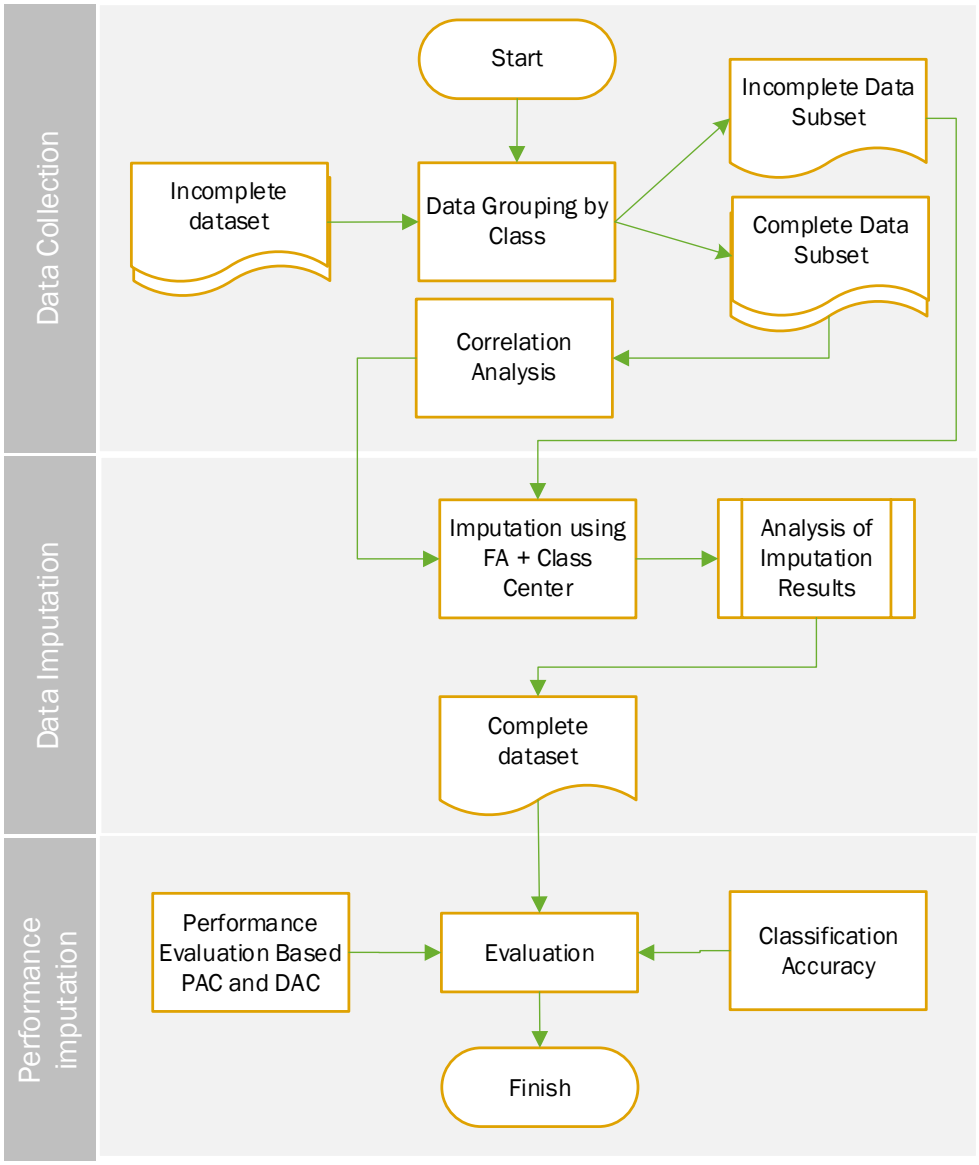

Fig.3 The Research Step

2.1 Data Collection

The beginning of this work consists of accessible datasets like iris, wine, Ecoli, and sonar. All datasets were gathered from the Kaggle, UCI Machine Learning Repository, and Knowledge Extraction based on Evolutionary Learning. Table 1 shows the characteristics of the dataset to be used in the simulation, with the relationships between variables and the probability of missing data indicated by MCAR missing mechanisms.

Table 1 Summary of Dataset Characteristics

Dataset	Sample Size	Number of Features	Missing Mechanism
Iris	135	4	MCAR
Wine	160	13	MCAR
Ecoli	302	7	MCAR
Sonar	208	50	MCAR

The steps associated with data collection are summarized in the following pseudocode:

1. The dataset is divided into two parts, namely complete ($D_{i\_complete}$) and incomplete subsets ($D_{i\_incomplete}$)
2. For the *i*th class, calculate its class center ($cent(D_i)$), and variance/standard deviation (std_i) of $D_{i\_complete}$. The $cent(D_i)$ is used to determine each data attribute's average value for class *i* of the complete subset.
3. Calculate the distances between $cent(D_i)$ and other data samples in class *i* using Euclidean distance formula (4)

$$Dis(cent(D_i), i) = \sqrt{\left(x_i - cent(D_i)\right)^2} \quad (4)$$

4. Calculate attribute correlations (*R*) of complete subset using formula (5).

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

2.2 Data Imputation

In the firefly algorithm, the intensity (I) of light is influenced by the objective function. Furthermore, the firefly pattern in which those with dimmer light intensity approach the brighter group is determined in the missing data's imputation. Dim and brighter light fireflies are analogous to the missing and complete data attributes.

The class center as the basis of imputation is used as the objective function $f(x)$. Therefore, the class center's value is used as the first step in determining $I(x)$ where x is the attribute in the data. The steps of data imputation are summarized as follows:

1. For each attribute in the complete subset, calculate $I(x)$ based on the value of the objective function $f(x)$, which is the class center, where

$$I(x) = \frac{1}{CentD_i} \quad (6)$$

2. Find the value $I(x)=\frac{1}{x_i}$ that is greater than $I(x)=\frac{1}{CentD_i}$. When a data greater than $I(x)$ is obtained, the data movement $x_{i\_new}^{k}$ is updated using equation (7) with $\beta_0=1$ based on previous studies, $r=Dis(cent(D_i),j)$, $\alpha\in[0,1]$, and *rand* is random numbers whose range is between [0,1].

$$x_{i\_new}^{k}=x_{i\_old}^{k}+\beta_0 e^{-\gamma r^2}\left|centD_i-x_{i\_old}\right|+\alpha\left(rand-\frac{1}{2}\right) \quad (7)$$

   a. If the $cent(D_i)$ value of the attribute that contains missing data is the same as the $cent(D_{i)}$ of correlated attribute data, use $\gamma=centD_i$

   b. If the $cent(D_i)$ value of the attribute that contains missing data is smaller than the $cent(D_i)$ of correlated attribute data, use $\gamma=\left(centD_i \middle/ R_{x_1x_2}\right)+\left|diff\ of\ centD_i\right|$

   c. If the $cent(D_i)$ value of the attribute that contains missing data is greater than the $cent(D_i)$ of attribute data that is correlated, using $\gamma=\left(centD_i\times R_{x_1x_2}\right)-\left|diff\ of\ centD_i\right|$

3. Analyse the imputation results by comparing the distance of the data with the class center generated from the previous imputation value +- std_i. The imputed values with the shortest distance from the class center are used to replace the missing data j.

2.3 Imputation Performance

Evaluation of imputation performance is based on predictive accuracy (PAC), distributional accuracy (DAC), and Classification Accuracy. The efficiency of imputation techniques is PAC, which aims at obtaining the real value in the data. Pearson Correlation Coefficient (*r*) and Root Mean-Squared Error (RMSE) are two measures for evaluation of PAC [27,28]. Furthermore, the Pearson correlation coefficient provides a measure of the correlation between the value of the imputation results with the actual. An imputation technique is efficient when the correlation value close to 1 [27,28]. Therefore, when x and $\hat{x}$ are the attribute values in the complete and incomplete data, then the correlation coefficient is calculated using the following formula (8)

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x}_i)(\hat{x}_i-\overline{\hat{x}}_i)}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x}_i)^2\sum_{i=1}^{n}(\hat{x}_i-\overline{\hat{x}}_i)^2}} \quad (8)$$

The Root Mean Squared Error (RMSE) is the most benchmark for comparison and performance of prediction strategies by measuring the distinction between the imputation and real values. In this case, a value closer to 0 results in a better imputation [27,28], with the fit to model calculated using the RMSE, which describes the close relationship of the predicted value to the true value. When the RMSE value is lower (error value), the predictions become better. Therefore, the RMSE is calculated using the following formula (9).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \quad (9)$$

where $\hat{x}_i$ is the actual value, $\hat{x}_i$ is the predicted value, and n is the total number of missing data.

DAC represents the technical ability to maintain the true distribution of data values. It was assessed using the Kolmogorov-Smirnov distance (D_{KS}). Therefore, when F_x and $F_{\hat{x}}$ are the empirical cumulative distribution functions of x and $\hat{x}$ then D_{KS} is calculated using formula (10). Smaller distance values represent better imputation results [27,28].

$$D_{KS} = \left\| F_x - F_{\hat{x}} \right\| \quad (10)$$

Another strategy used to evaluate the performance imputation is to look at the classification accuracy of some chosen classifiers trained by the imputed datasets. Decision Tree (DT), k-nearest neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM) are the top classifiers constructed for evaluating the imputation results [35].

4. Experiment Result

Evaluation of Predictive Accuracy, Distributional Accuracy, and classification accuracy on class center-based firefly algorithm for missing data imputation is determined based on each dataset's experiments.

4.1 Predictive Accuracy (PAC)

The correlation coefficient provides a measure of the correlation between the worth of the imputation results with the value. Therefore, an increase in the value of *r* indicates that the imputation method used is efficient. Root Mean Square Error (RMSE) is the most benchmark for comparing the performance of methods by measuring the difference between the imputation and original values. Table 2 lists the result for *correlation coefficient and RMSE* with different datasets.

Table 2 Correlation Coefficient and RMSE over datasets

Dataset	*r*	RMSE
Iris	0.978	0.19
Wine	0.948	23.98
Ecoli	0.960	0.06
Sonar	0.999	0.014

The average of *r* in the iris, wine, Ecoli, and sonar dataset is close to 1, indicating a correlation between the value of the imputation results and the missing data. In addition, the average value of RMSE is closer to 0, which means that there is no difference between imputation and real values. Furthermore, there is an increase in the value of RMSE in the wine dataset because it comprises three attributes with a standard deviation. The comparison of RMSE values with another imputation method such as SVM based on RBF kernel, KNNI, weighted voting random forests (WRF), feature weighted grey KNN (FKKNI), and class center missing data imputation (CCMVI) in previous research [23] are shown in Table 3.

Table 3 RMSE for different imputation Methods Vs Proposed Method

Dataset	Imputation Method					
	SVM	KKNI	WRF	FKKNI	CCMVI	C3-FA
Iris	0.4	1.16	1.02	1.02	0.42	**0.19**
Wine	71.43	81	57.6	55.76	50.39	**23.98**
Ecoli	0.14	0.16	0.18	0.17	0.11	**0.06**
Sonar	0.21	0.25	0.79	0.64	0.24	**0.014**

Table 3 shows that the proposed method has a smaller RMSE value.

4.2 Distributional Accuracy (DAC)

Imputation methods need to be able to maintain the distribution of these values through the evaluation of distributional accuracy (DAC). Therefore, this shows that the distribution of data after imputation does not change with the original data distribution using The Kolmogorov–Smirnov test. This statistic test quantifies a distance between the empirical distribution performed by a dataset once imputation and, therefore, the original dataset's cumulative distribution function as a reference distribution. Table 4 shows the simulation results of Kolmogorov-Smirnov distance (D_{ks}).

Table 4 The value of D_{ks} over datasets

Dataset	D_{KS}
Iris	0.032
Wine	0.040
Ecoli	0.039
Sonar	0.0004

Based on the result in Table 4, the class center-based firefly algorithm imputation can maintain the missing data distribution.

4.3 Classification Accuracy

Classification accuracy is additionally examined to measure the variations between the initial values and those imputed by the proposed method. This accuracy testing uses several classification algorithms, Decision Tree (DT), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN). Table 5 lists the classification performance for various datasets. The results showed that, on average, the proposed technique makes the SVM classifier offer the best classification accuracy rate.

Table 5 Classification Accuracy Test Result

Dataset	Classifier		
	DT	SVM	KNN
Iris	97%	98.5%	95.6%
Wine	90.6%	97.5%	96.9%
Ecoli	97%	98.5%	95.6%
Sonar	99.5%	97.1%	88.9%

5. Discussion and Conclusions

The three problems considered in the experimental procedure for missing data imputation include the selection of dataset, imputation methods, and evaluation results (Lin and Tsai, 2020). The selection of dataset for experiment relates to the problem domain (general or specific), the completeness of trial data (complete or incomplete), the type of test (numeric, categorical, or mixed), the scenario of data loss (MCAR, MAR, MNAR), and its percentage (missing rate). Meanwhile, regarding the imputation method, there is no comprehensive study that compares missing data imputation techniques in different dataset domains, with various rates and mechanisms [35]. These findings make it possible to understand the most suitable technique for an incomplete dataset type.

The adaptive search procedure performed in the Firefly algorithm is used to overcome the missing data in a dataset. In addition, the use of the class center as an initial objective function helps to determine the most optimal imputation value. Therefore, based on the simulation results from the datasets used, the general result showed that the class center-based firefly algorithm is an efficient technique for determining the actual value in handling the missing data. This is indicated by the values of the Pearson Correlation Coefficient (*r*) and Root Mean Squared Error (RMSE), which are closer to 1 and 0, respectively. In addition, the proposed method tends to maintain the true distribution of data values. This is indicated by the Kolmogorov–Smirnov test, which stated that the value of D_{KS} for most of the attributes in the dataset is generally closer to 0. Both results are in line with the fact that the imputation method can ideally reproduce actual values in data or Predictive Accuracy (PAC) and maintain the distribution of those values or Distributional Accuracy (DAC) [36]. However, some findings were obtained from the simulation results of wine datasets. Therefore, when the value of the standard deviation is high (more than 1), RMSE tends to increase.

Furthermore, previous studies' imputation methods were only tested on one missing data mechanism (MCAR/MAR/ MNAR). Therefore, further studies need to conduct tests by classifying the dataset based on the missing rate, starting from 10%, 20%, 30%, 40%, and 50% with the three mechanisms. Also, this follow-up study is expected to produce some rules from the different classes in the dataset with varying number of samples (S), Attribute (A), SA ratio, type of dataset, and percentage of missing data (missing rate). In addition, studies need to be conducted to determine the best imputation method for missing data on attributes that the value of standard deviations is high.

ABBREVIATIONS

FA: Firefly Algorithm

RMSE: Root Mean Squared Error

CCMVI: Class Center Missing Value Imputation

EM: Expectation maximization

LR: Linear/logistic regression

LS: Least squares

DT: Decision Tree

k-NN: k-Nearest Neighbor

RF: Random Forest

MAR: Missing at Random

MNAR: Missing Not at Random

IS: Information Similarity

DSMI: Decision Tree and Sampling Based Missing Value Imputation

MCAR: Missing Completely at Random

PAC: Predictive Accuracy

DAC: Distributional Accuracy

DECLARATIONS

Acknowledgments

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. The author 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:

1. UCI Machine Learning Repository (www.arsip.Ics.uci.edu/ml)
2. Kaggle Datasets (www.kaggle.com/datasets)
3. Knowledge Extraction based on Evolutionary Learning (https://sci2s.ugr.es/keel/missing.php#sub2b)

Competing interests

The author reports no potential conflict of interest.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Armina R, Mohd Zain A, Ali NA, Sallehuddin R. A Review on Missing Value Estimation Using Imputation Algorithm. Journal of Physics: Conference Series. 2017;892:012004.

2. Jugulum R. Importance of Data Quality for Analytics. In: Sampaio P, Saraiva P, editors. Quality in the 21st Century [Internet]. Cham: Springer International Publishing; 2016 [cited 2019 Apr 8]. p. 23–31. Available from: http://link.springer.com/10.1007/978-3-319-21332-3_2

3. Wazurkar P, Bhadoria RS, Bajpai D. Predictive analytics in data science for business intelligence solutions. 2017 7th International Conference on Communication Systems and Network Technologies (CSNT) [Internet]. Nagpur: IEEE; 2017 [cited 2019 Apr 8]. p. 367–70. Available from: https://ieeexplore.ieee.org/document/8418568/

4. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making [Internet]. 2016 [cited 2019 Apr 3];16. Available from: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0318-z

5. Deb R, Liew AW-C. Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences. 2016;339:274–89.

6. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008;41:3692–705.

7. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research & Method in Education. 2016;39:19–37.

8. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology. 2017;Volume 9:157–66.

9. Agbehadji IE, Millham RC, Fong SJ, Yang H. Bioinspired Computational Approach to Missing Value Estimation. Mathematical Problems in Engineering. 2018;2018:1–16.

10. 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.

11. Malarvizhi R, S. Thanamani A. K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation. IOSR Journal of Computer Engineering. 2012;6:12–5.

12. Marlin BM. Missing Data Problems in Machine Learning. [nadaCaa]: Department of Computer Science, University of Toronto; 2008.

13. Ng CG, Yusoff MSB. Missing Values in Data Analysis: Ignore or Impute? Education in Medicine Journal [Internet]. 2011 [cited 2019 Apr 8];3. Available from: http://eduimed.usm.my/EIMJ20110301/EIMJ20110301_02.pdf

14. Salleh MNM, Samat NA. FCMPSO: An Imputation for Missing Data Features in Heart Disease Classification. IOP Conf Ser: Mater Sci Eng. 2017;226:012102.

15. Leke C, Twala B, Marwala T. Modeling of missing data prediction: Computational intelligence and optimization algorithms. 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC) [Internet]. San Diego, CA, USA: IEEE; 2014 [cited 2019 Sep 30]. p. 1400–4. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6974111

16. Nazir S, Asif M, Ahmad S. The Evolution of Trends and Techniques used for Data Mining. 2019 2nd International Conference on Advancements in Computational Sciences (ICACS) [Internet]. Lahore, Pakistan: IEEE; 2019 [cited 2020 Dec 26]. p. 1–6. Available from: https://ieeexplore.ieee.org/document/8689125/

17. Cao L. Data science thinking. New York, NY: Springer Science+Business Media; 2018.

18. Nishanth KJ, Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing. 2016;218:17–25.

19. Van Hulse J, Khoshgoftaar TM. Incomplete-case nearest neighbor imputation in software measurement data. Information Sciences. 2014;259:596–610.

20. Grzymala-Busse JW, Hu M. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In: Ziarko W, Yao Y, editors. Rough Sets and Current Trends in Computing [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001 [cited 2020 Dec 26]. p. 378–85. Available from: http://link.springer.com/10.1007/3-540-45554-X_46

21. Ryu S, Kim M, Kim H. Denoising Autoencoder-Based Missing Value Imputation for Smart Meters. IEEE Access. 2020;8:40656–66.

22. Nugroho H, Surendro K. Missing Data Problem in Predictive Analytics. 8th International Conference on Software and Computer Applications (ICSCA 2019). Penang: ICSCA 2019; 2019.

23. 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.

24. Zahin SA, Ahmed CF, Alam T. An effective method for classification with missing values. Applied Intelligence. 2018;48:3209–30.

25. Nekouie A, Moattar MH. Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization. Journal of King Saud University - Computer and Information Sciences. 2019;31:287–94.

26. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Computational Statistics & Data Analysis. 2015;90:84–99.

27. Pompeu Soares J, Seoane Santos M, Henriques Abreu P, Araújo H, Santos J. Exploring the Effects of Data Distribution in Missing Data Imputation. In: Duivesteijn W, Siebes A, Ukkonen A, editors. Advances in Intelligent Data Analysis XVII [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 May 29]. p. 251–63. Available from: http://link.springer.com/10.1007/978-3-030-01768-2_21

28. Santos MS, Soares JP, Henriques Abreu P, Araújo H, Santos J. Influence of Data Distribution in Missing Data Imputation. In: ten Teije A, Popow C, Holmes JH, Sacchi L, editors. Artificial Intelligence in Medicine [Internet]. Cham: Springer International Publishing; 2017 [cited 2019 May 29]. p. 285–94. Available from: http://link.springer.com/10.1007/978-3-319-59758-4_33

29. Leke CA, Marwala T. Deep Learning and Missing Data in Engineering Systems [Internet]. Cham: Springer International Publishing; 2019 [cited 2019 Oct 18]. Available from:

http://link.springer.com/10.1007/978-3-030-01180-2

30. Abdella M, Marwala T. The use of genetic algorithms and neural networks to approximate missing data in database. Mauritius: IEEE; 2005 [cited 2019 Oct 22]. p. 207–12. Available from: http://ieeexplore.ieee.org/document/1511574/

31. Yang X-S. Nature-inspired metaheuristic algorithms. 2. ed. Frome: Luniver Press; 2010.

32. Yang X-S, He X-S. Why the Firefly Algorithm Works? In: Yang X-S, editor. Nature-Inspired Algorithms and Applied Optimization [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 Sep 27]. p. 245–59. Available from: http://link.springer.com/10.1007/978-3-319-67669-2_11

33. Nugroho H, Utama NP, Surendro K. Performance Evaluation for Class Center-Based Missing Data Imputation Algorithm. Proceedings of the 2020 9th International Conference on Software and Computer Applications [Internet]. Langkawi Malaysia: ACM; 2020 [cited 2021 Jan 15]. p. 36–40. Available from: https://dl.acm.org/doi/10.1145/3384544.3384575

34. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications. 2010;19:263–82.

35. 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.

36. Chambers R. Evaluation Criteria for Statistical Editing and Imputation [Internet]. Department of Social Statistics University of Southampton; 2001. Report No.: 28. Available from: https://www.researchgate.net/publication/246110442_Evaluation_Criteria_for_Statistical_Editing_and_Imputation

Figures

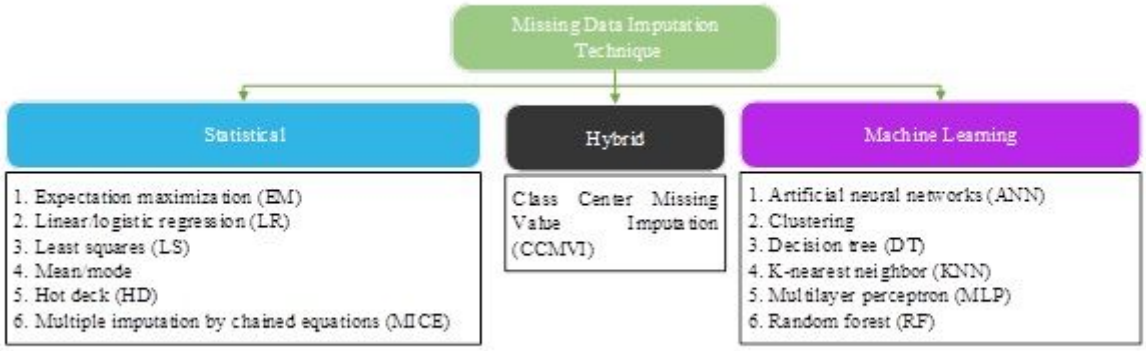

Figure 1

Methods for Handling Missing data

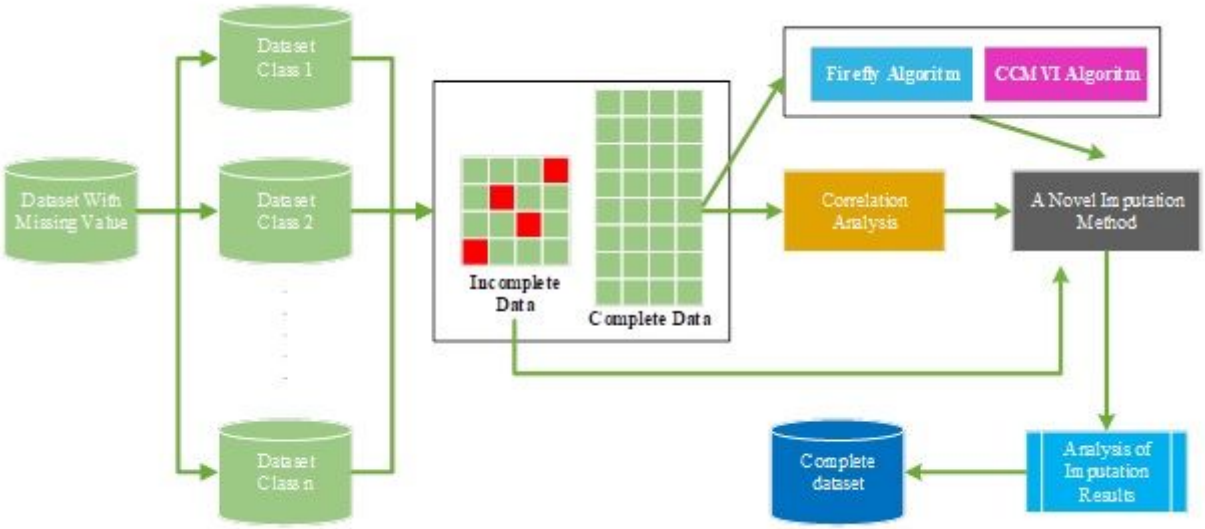

Figure 2

The Overall Architecture Novel Framework for Missing data Imputation (C3-FA)

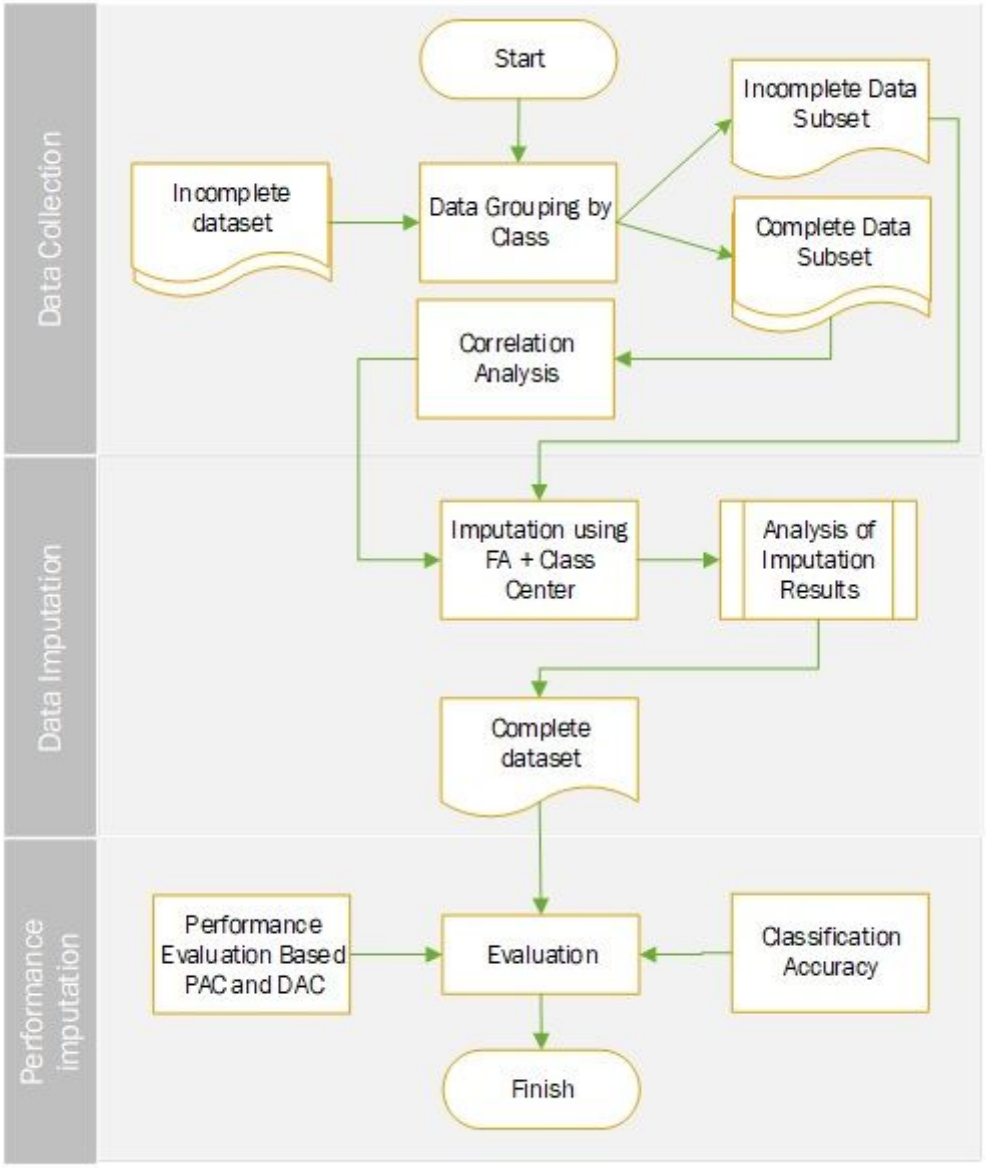

Figure 3

The Research Step