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

Estimating missing data in a dataset is a significant advance during the data cleaning stage. Improper data handling can make inaccurate results when conducting data analysis. Most of the research about missing data estimation is irrespective of the correlation between attributes. However, an adaptive search procedure helps find 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 finding the estimated value that is closest to the value in other data known. Therefore, this study proposes a class center-based adaptive approach model for missing data by considering the attribute correlation in the imputation process (C3-FA). Based on the experiment, the general result find that the class center-based firefly algorithm is an efficient technique for getting the actual value in handling the missing data. This can be seen on the value of Pearson correlation coefficient (r) that close to 1 and the root mean squared error (RMSE) value is generally closer to 0. In addition, the proposed method can maintain the true distribution of data values. This is indicated by the Kolmogorov–Smirnov test that value of D_{KS} for most of the attributes in the dataset is generally closer to 0. Also, the results of the accuracy evaluation 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 that can influence the consequences of the prediction system to be ineffective [1–3]. Ignoring the missing data has an impact on the results of the analysis [4–9], learning outcomes, predictive results [10] and potentially weakens the validity of the results and conclusions [8,9] and leads to estimation of biased parameters [7,11–14]. Prediction and classification are the principle obligations required in many areas and spaces that expect admittance to finish and accurate data [15]. Until now, data mining methods can only

work with complete data [1,16–18]. The issues related to the missing data become research opportunities to get the correct procedures to overcome them [19].

Class center missing value imputation (CCMVI) is a hybrid method that combines the imputation method through a statistical approach and machine learning with the baseline imputation method using k-NN [20]. There are many techniques ignoring the correlation between data attributes or somehow it is only suitable for category data [21]. In previous studies, most of the data estimation methods were missing, imputing data regardless of the dependence between attributes [22]. The performance of imputation algorithm of the missing data is significantly influenced by correlation structure in the data [1,4,23], missing data mechanism, distribution of missing data [24,25], and the proportion of the missing data within the data [1].

Adaptive search procedure is possible for the new techniques to estimate missing velue when between variables are considered [26]. This procedure can help the researcher to find the estimation of missing data by optimizing objective functions in each search problem given [27]. In late 2007 and early 2008, Xin-She Yang developed Firefly Algorithm (FA). This method implemented an adaptive search procedure [28]. Firefly Algorithm is kind of optimization algorithm that was inspired by nature and has mature considerably since its look around ten years ago [29]. The behavior of fireflies during which the brighter fireflies attract fireflies with weaker flashes may be applied within the imputation of missing data. It is done by finding the estimated value that is closest to the value known and then replacing the missing data [9].

This research proposed class center-based imputation method of missing data by utilizing Firefly Algorithm in the imputation process. This research is a development from previous research that has been done by the author [30]. The contribution of this research is the method of imputation of data based on the class center by considering the correlation of attributes with the imputation stage combined with the Firefly algorithm (C3-FA). In the previous studies, the firefly algorithm and consideration of correlation in the imputation process which is based on the class center has never been used. Therefore, the advantage of this method over others is that it's not only able to reproduce the actual values, but also maintains the distribution without knowing the original data. The proposed technique is going to be tested on the iris, wine, Ecoli, and sonar datasets. The imputation performance is going to be measured based on predictive accuracy, distributional accuracy, and classification accuracy.

2. Related Work

2.1. Methods for Handling Missing data

In handling missing data, the method to use would be dependent on the type of data and needs. Therefore, the missing data imputation techniques could be classified into two types, namely statistical-based and machine learning [31,32]. The statistical methods that were widely used in previous studies were EM, LR, LS, and mean / mode, while that of machine learning includes DT, clustering, k-NN, and RF [33]. In 2018, Tsai proposed a class center-based missing data imputation method which is a combination of statistical and machine learning technique. Figure 1 shows a method for handling missing data.

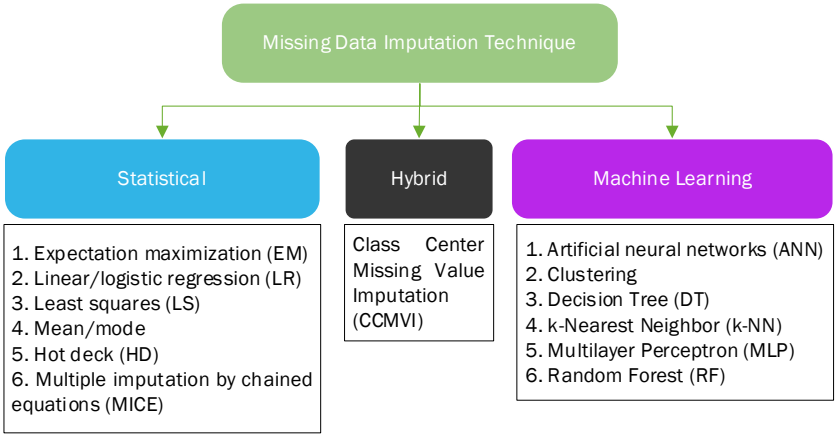

Fig.1 Methods for Handling Missing data

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 (Tsai et al., 2018). Also, in this study, there are two modules which include, threshold identification (Module A) and imputation (Module B). The algorithm of each stage can be seen in Figure 2 and Figure 3 below.

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

01. For j = 1 to Num
02. If D(j,) has missing value(s) then
03.         Get the class label of D(j,) and set this class label to variable i
04.         Put D(j,) to D_i_incomplete
05. Else
06.         Get the class label of D(j,) and set this class label to variable i
07.         Put D(j,) to D_i_complete
08. End
09. For i = 1 to N
10.   Avg(i,)=Average(D_i_complete)
11.   Std(i,)=Standard Deviation(D_i_complete)
12.   Get the number of rows in D_i_complete and set to variable Num
13. End
14. For j = 1 to Num
15.         Distance(j)=Euclidean distance(D_i_complete(j,), Avg(i,))
16. End
17. Threshold(i) = Median(Distance)
```

Fig.2 Threshold identification pseudo-code [20]

```
Input: D_i_incomplete containing M feature dimensions and N classes
Output: imputed dataset for D_i_incomplete
01. For i = 1 to N
02.   Get the number of rows in D_i_incomplete and set to variable Num
03.   For j = 1 to Num
04.     If D_i_incomplete(j,) has one missing value
05.       Get attribute index with the missing value and set to variable miss_attr
06.       D_i_incomplete(j, miss_attr) = Avg(i, miss_attr)
07.       Distance = Euclidean distance(D_i_incomplete(j,), Avg(i,))
08.       If Distance > Threshold(i)
09.         D_i_incomplete(j, miss_attr) = D_i_incomplete(j, miss_attr) +/- Std(i, miss_attr)
10.     Else
11.       Get attribute index with the missing value and set to variable array miss_attr
12.       Get miss_attr length and set to variable size
13.       For s = 0 to size-1
14.         D_i_incomplete(j, miss_attr(s)) = Avg(i, miss_attr(s))
15.       Distance = Euclidean distance(D_i_incomplete(j,), Avg(i,))
16.       If Distance > Threshold(i)
17.         Missing_array = array[size]
18.         For s = 0 to size-1
19.           Missing_array(s,)= D_i_incomplete(j, miss_attr(s)) +/- Std(i, miss_attr(s))
20.             Distance_array(s) = Euclidean distance(Missing_array(s,), Avg(i,))
21.         Find minimum Distance_array and set index to variable index
22.         D_i_incomplete(j,) = Missing_array(index,)
```

Fig.3 Imputation Pseudo-code[20]

However, some issues from 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, even when they are only suitable for categorical data [21]. In addition, the performance of the missing data imputation algorithm is significantly influenced by factors, which include the correlation structure in the data, the mechanism of the missing data, the distribution of the entries, and the percentage of the values [1].

Deb and Liew in a study entitled, Missing data imputation for the analysis of incomplete traffic accident data, proposed an algorithm using a decision tree to find correlated attributes. The measure used was the IS and the weighted similarity, as a result, the algorithm outperforms several popular imputation methods on the traffic accident dataset with the condition that a large number of attributes are categorical [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 $\{O_1, \ldots, O_N\}$ based on the sampling probabilities specified by the set of affinity degrees $\{\vartheta_1,\ldots,\vartheta_N\}$
    END FOR
  END FOR

Fig.4 DSMI Algoritm [5]

For high dimensional cases, the imputation method with the nearest neighbor approach proposes a new distance that explicitly uses the correlation between variables. This method automatically selects the relevant variables that contribute to distance while the simulation results showed that performance depends on the correlation between them [23]. Also, the accuracy of the imputation is based on the dependency level of the missing data with other variables in the dataset, and 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 Xin-She Yang in late 2007 and early 2008 [28]. It is an optimization algorithm inspired by nature and has developed significantly for about 10 years [29]. As a result, to properly design the Firefly algorithm, two important issues need to be defined, which includes attraction and variation in light intensity. This intensity is influenced by the objective function, and its level for a firefly minimizing problem *x* can be expressed as $I(x) = \frac{1}{f(x)}$. The value of $I(x)$ is the level of light intensity on the firefly *x* which is inversely proportional to the solution of the objective function of the problem to be sought $f(x)$.

The Attractiveness β is of relative value because the light intensity must be seen and judged by other fireflies. However, the results of the assessment will differ depending on the distance between one firefly and another rij. In addition, the intensity will decrease from the source because it is absorbed by the media, such as air. As a result, Attractiveness (β) with a distance $r$can be calculated by equation (1).

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

Where $\beta 0$ is 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 two fireflies i and j at positions x_i and x_j is the Cartesian distance calculated by 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 because of the attraction to j with a brighter light intensity, will change 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 firefly, the second term occurs because of attraction, the third term is a firefly random movement where α is the random parameter coefficient and rand is a real random 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 that can be used in various data loss cases. As a result, this model is proposed to overcome the missing data by considering the attribute relationship/correlation, which is a development of class center-based imputation. Also, the consideration was done due to one of the weaknesses from the previous method in the category data. However, other studies with categorical and attribute correlation yielded very good accuracy. Furthermore, the results of the previous study showed that the class center-based imputation method by considering correlation is an efficient imputation technique to get the actual value in the data.

When the estimation of missing data considers the correlation and associations between variables, an adaptive search procedure can be used. The Firefly algorithm implements the search procedure such that, as long as the bright fireflies attract the lower ones, the behavior can be applied in the imputation of missing data. The bright fireflies represent the non-missing data, while the lower ones represent the missing in dataset. The equations related to light intensity, attractiveness, distance, and firefly movement was applied in the imputation process using the class center approach, by considering the correlation between the attributes. The result was therefore analyzed by comparing the distance gotten from the imputation ± std result.

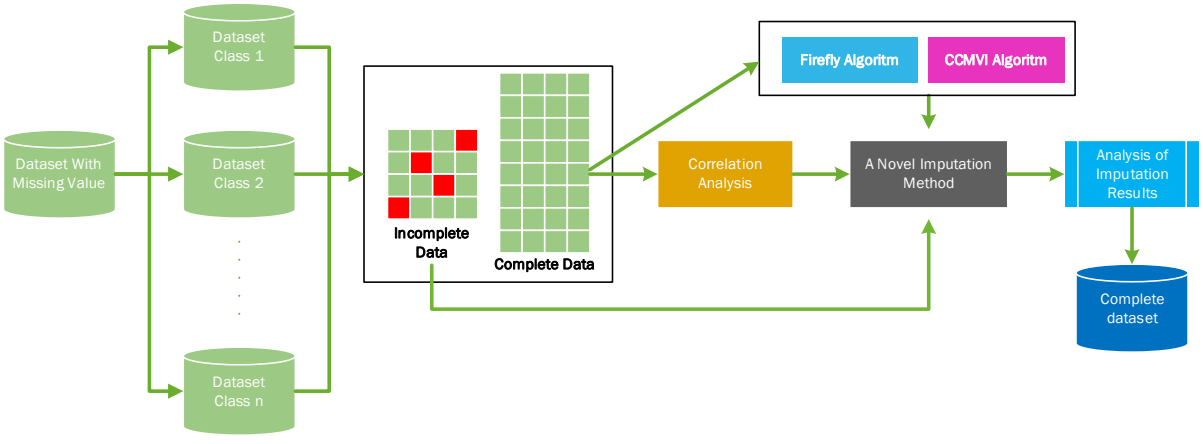

Fig.5 The Overall Architecture Novel Framework for Missing data Imputation (C3-FA)

In general, this research is divided into three parts, Data Collection, Data Imputation, Evaluation of Imputation Performance (Fig 6).

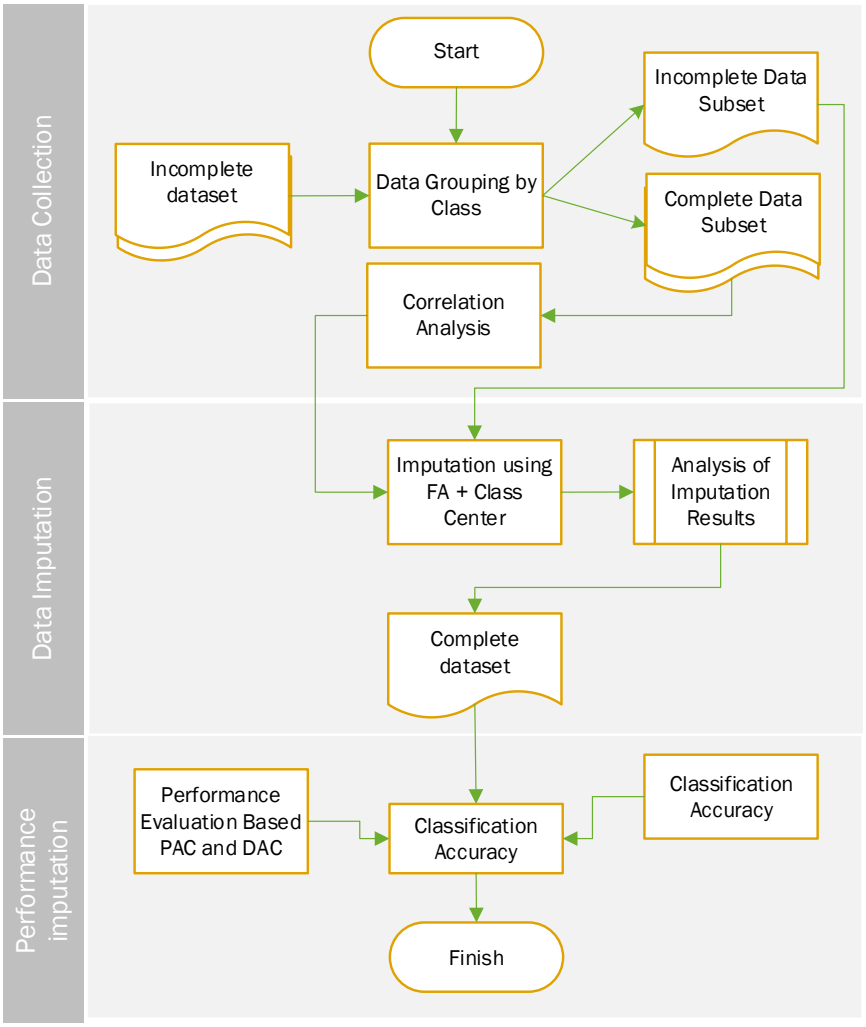

Fig.6 The Research Step

2.1 Data Collection

The beginning of this work consists of choosing accessible datasets like iris, wine, Ecoli, and sonar. All datasets were gathered from the Kaggle Datasets, UCI Machine Learning Repository, and Knowledge Extraction dataset based on Evolutionary Learning. Table 1 illustrates the characteristics of the dataset to be used in the simulation. The relationships between variables and the probability of missing data is show 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 of data collection can be summarized as the following pseudocode:

1. Incomplete dataset divided into two part, 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 $cent(D_i)$ find from average value of each data attribute for each class *i* of complete subset.
3. Calculate the distances between $cent(D_i)$ and other data samples in class *i* using Euclidean distance formula (4)

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

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

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

2.2 Data Imputation

In firefly algorithm, the intensity (I) of light in fireflies is influenced by the objective function. The firefly pattern in which the fireflies that have a dimmer light intensity approach the group of fireflies with a brighter light intensity used in the process of imputation of the missing data. Dim light fireflies are analogous to missing data attributes, while the fireflies with brighter light intensity are analogous to the complete data attribute.

The class center as the basis of imputation will be used as the objective function $f(x)$. Therefore, the value of the class center will be use as the first step in determining the value of $I(x)$ where x is the attribute value in the data. The steps of data imputation can be summarized as the following pseudocode:

1. For each attribute in the complete subset, calculate the value $I(x)$ based on the value of the objective function $f(x)$ which is the value of 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 $I(x) = \frac{1}{CentD_i}$. If there is data that $I(x)$ is bigger, the data movement $x_{i\_new}^{k}$ will be updated using the following movement equation (4) 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].

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

   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, using $\gamma = centD_i$

   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, using $\gamma = \left(centD_i \Big/ R_{x_1 x_2} \right) + \left| diff\ of\ centD_i \right|$

   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 $\gamma = \left(centD_i \times R_{x_1 x_2} \right) - \left| diff\ of\ centD_i \right|$

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 that given the shortest distance from class center are used to imputation the missing data of data j.

2.3 Imputation Performance

Evaluation of imputation performance is based predictive accuracy (PAC), distributional accuracy (DAC), and Classification Accuracy. The efficiency of imputation techniques is PAC concerned to get the real value in the data. Pearson Correlation Coefficient (*r*) and Root Mean-Squared Error (RMSE) are two measures for evaluation of PAC [24,25]. Pearson correlation coefficient provides a measure of the correlation between the

value of the imputation results with the actual value. An imputation technique is efficient when the correlation value close to 1 [24,25]. If x are the attribute values in the complete data and $\hat{x}$ are the attribute values in the incomplete data then the correlation coefficient is calculated by formula (8)

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

Root Mean Squared Error (RMSE) has been called the most benchmark for comparison the performance of prediction strategies by measuring the distinction between the imputation value and the real value of a given feature. In this case, a value closer to 0 results in a better imputation [24,25]. Fit to model is calculated using the RMSE so that the RMSE describes how closely the predicted value is related to the true value. If the RMSE value is getting lower (error value), the predictions of an algorithm will be better. To calculate RMSE, we use 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. DAC was assessed using Kolmogorov-Smirnov distance (D_{KS}). If F_x and $F_{\hat{x}}$ are the empirical cumulative distribution functions of x and $\hat{x}$ then D_{KS} is calculated by the formula (5). Smaller distance values represent the better imputation results [24,25].

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

Another strategy for evaluation the performance of 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 [34].

4. Experiment Result

Evaluation of Predictive Accuracy, Distributional Accuracy, and classification accuracy on class center-based firefly algorithm for missing data imputation can be seen based on the experiments of each datasets.

4.1 Predictive Accuracy (PAC)

Correlation coefficient provides a measure of the correlation between the worth of the imputation results with the value. The greater value of *r* indicates that the imputation method used is efficient. Root Mean Square Error (RMSE) has been called the most benchmark for comparing the performance of methods by measuring the difference between the imputation value and the original value. 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 real values of missing data. In addition, the average of RMSE value is closer to 0 which means that there is no difference between the imputation value and the real value. The value of RMSE in wine dataset is very high because in the wine dataset there are three attributes that have a high standard deviation. The comparison of RMSE values with another imputation method which are SVM based on the radial basis function (RBF) kernel, KNNI, weighted voting random forests, feature weighted grey KNN, and class center missing data imputation (CCMVI) in previous research [20] with the same dataset can be seen in the 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**

The results in Table 3 show that the proposed method has a smaller RMSE value than another one.

4.2 Distributional Accuracy (DAC)

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

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

4.3 Classification Accuracy

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

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 needed to be considered in the experimental procedure for missing data imputation includes, the selection of dataset, the methods used in imputation process, and the evaluation of the 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 [33]. These findings make it possible to understand which technique is more suitable for an incomplete dataset type.

The adaptive search procedure was performed in the Firefly algorithm can be used to overcome the missing data in a dataset. The use of class center as an initial objective function helps the researcher to find the most optimal imputation value. Based on the simulation results from datasets used, the general result find that the class center-based firefly algorithm is an efficient technique for getting the actual value in handling the missing data. This can be seen on the value of Pearson Correlation Coefficient (*r*) that close to 1 and the Root Mean Squared Error (RMSE) value is generally closer to 0. In addition, the proposed method can maintain the true distribution of data values. This is indicated by the Kolmogorov–Smirnov test that value of D_{KS} for most of the attributes in the dataset is generally closer to 0. Both of these result is 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) [35].However, some findings were obtained from the simulation results of wine datasets. If the value of standard deviation is high (more than 1), the value RMSE does not close to 0 (tends to be high).

Furthermore, most imputation methods in previous studies are only tested on one missing data mechanism (MCAR/MAR/ MNAR). Therefore, further study should 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 number of samples (S), Attribute (A), SA ratio, type of dataset, and percentage of missing data (missing rate). Other further research is how to get 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.

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

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

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

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

20. 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.

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

22. 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.

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

24. 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

25. 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

26. 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

27. 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/

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

29. 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

30. Nugroho H, Utama NP, Surendro K. Performance Evaluation for Class Center-Based Missing Data Imputation Algorithm. Langkawi; 2020.

31. 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.

32. Peng L, Lei L. A Review of Missing Data Treatment Methods. Int Journal of Intel Inf Manag Syst Tech. 2005;8.

33. 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.

34. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review [Internet]. 2019 [cited 2019 Apr 17]; Available from: http://link.springer.com/10.1007/s10462-019-09709-4

35. Chambers R. Evaluation Criteria for Statistical Editing and Imputation. Department of Social Statistics University of Southampton; 2001. Report No.: 28.