Cumulative bayesian ridge for handling missing data

Samih M. Mostafa [a,*] Abdelrahman S. Mohamed [a] and Safwat Hamad [b]
[a]*Mathematics Department, Faculty of Science, South Valley University, Qena, Eygpt*
[b]*Faculty of Computer and Information Sciences, Ain Shams University, Cairo, 11566, Egypt*
**samih_montser@sci.svu.edu.eg*

Abstract. Old approaches that manipulate missing values may lead to biased estimations. In addition, they may also decrease or magnify the statistical influence, which could result in unacceptable conclusions. The performance of various missing value imputation algorithms may depend on the amount of the missing values in the dataset and the dataset's dimension. In this paper, the authors proposed a new algorithm for handling missing data against some registered practical imputation methods. The proposed algorithm depends on the Bayesian Ridge technique, which operates in a cumulative order with the aid of gain ratio feature selection to select the candidate feature to be imputed. The imputed feature will be included in the Bayesian Ridge equation to impute missing values in the next candidate feature. Here, the authors are attempting to choose the best imputation method succeeded to give high imputation accuracy with less imputation time. Finally, we applied the proposed algorithm on eight datasets with various missing values proportions generated from the missingness mechanisms. The empirical study indicates the effectiveness of the proposed algorithm with any missingness mechanism and with any missing data percentage.

Keywords: missing value, imputation, missingness mechanism, Bayesian Ridge and Gain ratio.

1. Introduction

Avoiding missing data is the best approach for dealing with incomplete instances. All skilled researchers still encounter missing values that would happen for unpredictable reasons. In the data gathering stage, the researcher has the prospect to make decisions about what data to collect and the way to screen data gathering. Both, the distribution and scale of the features within the data along with the causes for missing data are two acute topics to select the right method for handling missing data[1].

Data preparation is considered the most important and time consuming task, which toughly impacts the success of the research. Feature selection lies in detecting a valuable subset of possible predictors from a huge set of candidates. Refusing features with a large proportion of missing values (e.g., >50 %) is often a good rule of thumb. Nevertheless, it is not a risk-free route. Refusing a feature may lead to a loss of analytical power and capability to observe statistically significant differences, and it is often a cause of bias, affecting the representativeness of the outcomes. For these reasons, feature selection needs to be custom-made to the missing data mechanism. Imputation is often completed beforehand or later of feature selection[2].

1.1. Handling missing data

To elect how to manage missing data, it is important to know the reason of the missingness. This paper takes into consideration three general missingness mechanisms[2–5]:

- *Missing completely at random (MCAR):* If the likelihood of missingness is identical for all instances. The cause for the missingness in a feature X does not depend on X itself or any other feature within the dataset. In MCAR, missing value deletion does not bias your conclusions[6].
- *Missing at random (MAR):* If the likelihood of missingness is identical only within the data that are observed. The cause for the Missingness in a feature X depends on other features within the dataset, not on the feature X itself. It is frequently rational to model this process as a logistic regression, where the resulting feature equals 1 for detected cases and 0 for missing cases[6].
- *Missing not at random (MNAR):* If the likelihood of missingness for a feature X depends on X itself or other features that already include missing values[6].

The best approach for handling missing data is to avoid it through careful data gathering and follow up, along with determining missing data after the fact (for example, by detecting missing data or re-contacting study members). However, it is commonly impossible to avoid missing data in total, therefore, statistical methods for dealing with missing data are required. Since missing data are exceptionally complex, statisticians cannot create a universal set of rules that manipulate all cases. So, they run emulation to detect the best method[7]. The approaches for handling missing data have to be tailored to the causes of missingness, the dataset, and the percentage of missing data[2]. The best approach to handle missing data is to get rid of instances that involve missing values. In general, case deletion methods result in valid conclusions just for MCAR[8]. Imputation is the alternative approach for handling missing data and overwhelmed the disadvantages of the deletion approach[4].

The proposed algorithm depends on Bayesian ridge regression, so it is a regression model with a regularization parameter for the coefficients. The model satisfies the following[9]:

$$y \sim N(\mu, \alpha) \tag{1}$$

Where:

$$\mu = \beta X = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

$$\beta \sim N\left(0, \lambda^{-1} I_p\right)$$

$$\alpha \sim \mathcal{g}(\alpha_1, \alpha_2)$$

$$\lambda \sim \mathcal{g}(\lambda_1, \lambda_2)$$

So y follows a normal distribution characterized by variance α and mean $\mu = \beta X$. λ and α are regularizing parameters that follow gamma distribution. $\alpha_1, \alpha_2, \lambda_1$ and λ_2 are hyper-parameters of the gamma prior distributions.

The rest of the paper is organized as follows: Section 2 presents the literature review of manipulating the missing data, Section 3 and 4 present the proposed algorithm and discuss the experimental implementation respectively, results and discussion are demonstrated in section 5. Finally, Section 6 concludes the paper.

2. Literature review

For manipulating a dataset that contains missing values, one can use deletion; Deletion might be 'complete deletion', 'list-wise deletion', or 'Complete Case Analysis', where all observations contain one or more of their feature values missing are removed[4,10]. There might also be 'specific deletion', where only those observations are removed, which contain more than a pre-specified percentage of their feature values that are missing[4,7,10,11]. There can also be 'pair-wise deletion' or 'variable deletion' where the observations contain missing values in the features involved in the running analysis removed. In the worst case, each feature contains missing values through many observations that may cause the deletion of the whole dataset[12].

Imputation approaches benefit from all the information available within the dataset to predict the missing value, in which an applicable value is imputed instead of the missed value[10,11,13]. The imputed value might be mean, mode, median or any pre-specified value of the feature that contains missing value[14–16], or might be acquired from case substitution. Imputed value can also be estimated using KNN(k nearest neighbors), regression models[3], cold deck imputation[17], EM (expectation maximization) imputation[18–20], hot deck imputation[21], etc. In techniques contain prediction models, a model is build based on available information within the dataset, then this model is used to predict values for the missing data[22]. Imputation methods are used to deal with missing values if the missing values are of MCAR or MAR type, and if each record or feature in the dataset is important and a single observation does not include missing values across many features[12]. For MCAR type, missing values in a dataset may be managed by using deletion, if possible listwise deletion or maximum likelihood methods. There are not any general approaches to deal with MNAR missing values type[4,23]. Imputation types are single imputation or multiple imputation. In single imputation, a single applicable value is imputed rather than missed value[17]. Multiple imputation, in which 'm' full datasets are acquired by imputing the missing values 'm' times, the finale imputed dataset being the analysis average of these 'm' datasets[24]. Although multiple imputation requires more resources[25], it has benefits over other approaches, namely maximum likelihood techniques, single imputation, and deletion[26].

Single imputation approaches deal with all data values, even imputed ones as true values, which lead to inflated type I error as a result of not accounting uncertainty for a missing value. Inverse Probability Weighting (IPW) methods also good for dealing with missing data, which depend on the inverse of the detected probability to weight detected observations, in this way representing the entire data even the missing values. Nevertheless, the imputation method produces more reliable performance[4,27]. Ordinary least squares (OLS), partial least squares (PLS), and singular value decomposition (SVD) are also good choices for multiple imputation[4]. For datasets having binary and ordinal features, Multi Variate Normal Imputation (MVNI) and Fully Conditional Specification (FCS) approaches generate similar results and usually less biased. Model specification is easily provided by MVNI, but as a result, of its unrealistic nature, more or fewer people may have a problem with it. FCS tends to have a complex model specification as a result of requiring a single regression model for each feature whose missing values are going to be imputed[28]. KNN is used by FINNIM (which is an effective nonparametric iterative multiple imputation approach) to predict missing values[29]. Sequential regression trees used as a Multiple imputation conditional model, which has the facility to detect complex relation and need slight correction by the user[30]. Handling missing values with predictive mean matching (PMM) approach is accomplished using a randomly drawn observation from a set of detected observations whose predictive mean is near to predictive mean of missing values[4]. Local residual draw (LRD) approach handles missing value using the predictive mean as in the situation of PMM, with extra randomly drawn from the residuals of a set of detected instances with predictive means near to that of the missing value[31]. Reinforcement programming (RP), which depends on Reinforcement Learning gives, better performance over mean per category imputation, zero imputation, and genetic algorithm (GA) in terms of the sum of square error and computational time[32]. Cumulative linear regression, which depends on the linear regression algorithm to handle missing data, works well for small and large datasets[3]. Handling missing values in features of concern with the aid of detected values from other features depends on the similarities of the instance values within the donor feature to deal with the missing values within the recipient feature. It works well when the proportion of the missing values is large[33].

3. Proposed algorithm

Table 1 contains a list of terminologies that help in elaborating the proposed algorithm. For any dataset, two cases may happen; the first case when all features contain missing values, even the dependent feature, and the second case when there is at least one complete feature. Suppose that the dependent feature y does not involve any missing data and $X = \{X_i : i = 1, \dots, n\}$ is the set of all independent features. The proposed algorithm splits the input dataset into two separate datasets $X^{(mis)}$ and $X^{(comp)}$. The selection of the candidate features from $X^{(mis)}$ to be imputed, must exhibit the highest information gain ratio with the target feature y using Eq. (4). The candidate feature will be the dependent feature, and y and $X^{(comp)}$ will be independent. The model is fitted to impute the missing values in that feature, and then the imputed feature X_{Imp} will be removed from $X^{(mis)}$ and added to $X^{(comp)}$. Another feature from $X^{(mis)}$ will be selected to be the dependent feature, and the model will be fitted using $X^{(comp)}$, $X_{imp}^{(mis)}$ and, y as independent features. In the same way, another feature will be selected until $\mathrm{X}^{(\mathrm{mis})}$ be empty. The proposed algorithm is presented in Figure 1, and Figure 2 shows the algorithm flow chart.

- ***Initialization***
  - Identify features that contain missing values.
    - $X^{(comp)} = \{X_1^{(comp)}, X_2^{(comp)}, \cdots, X_c^{(comp)}\}$,and
    - $X^{(mis)} = \{X_1^{(mis)}, X_2^{(mis)}, \cdots, X_m^{(mis)}\}$.
  - $y \in \begin{cases} X^{(comp)} & if\ MissObs.y = \emptyset \\ X^{(mis)} & otherwise \end{cases}$

- ***Feature selection***
  - From $X^{(mis)}$, find $X_i^{(mis)}, i \in \{1, \ldots, \mathrm{m}\}$, select the feature with:
    - Higher Gain Ratio $(X_i^{(mis)}, y)$

- ***Imputation***
  - For each column in $X^{(mis)}$:
    - Fit the model with cumulative Bayesian Ridge Regression equation:

$$X_g^{(miss)} \sim N(\mu_g, \alpha_g)$$

$$g = 1,2, \ldots, m$$

Where:

$$\mu_g = \beta_g X_g = \beta_o + \sum_{i=1}^{c} \beta_i X_i^{(comp)} + \beta_{c+1} y + \sum_{imp=1}^{g-1} \beta_{imp+c+1} X_{imp}^{(mis)2}$$

$$\beta_g \sim N(0, \lambda^{-1}{}_g Ip)$$

$$\alpha_g \sim \mathcal{g}(\alpha_{1g}, \alpha_{2g})$$

$$\lambda_g \sim \mathcal{g}(\lambda_1 g, \lambda_{2g})$$

    - Impute missing values.
  - Repeat until all missing values in all columns are imputed.

Figure 1. Algorithm: CBRG (Cumulative Bayesian Ridge regression with Gain-ratio)

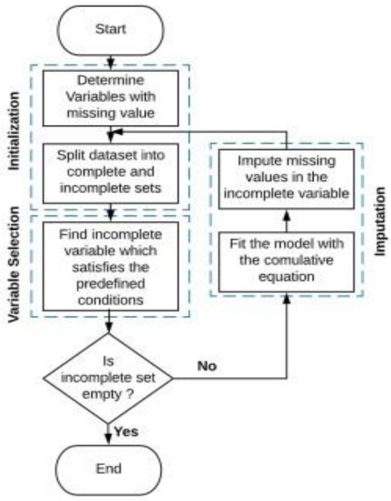

Figure 2. Algorithm flowchart

Table 1. List of terminologies

Terms	Description	Comments
n	number of all variables	Comp + Miss = n
$X^{(comp)}$	set of complete variables	
$X^{(mis)}$	set of incomplete variables	
$X^{(mis)}_{imp}$	imputed variable from X(Miss)	
c	number of complete independents	c +m = n
m	number of variables contain missing values	
$\mathrm{MissObs.\,y}$	set of missing observations in the dependent variable y	
$\mathrm{MissObs.\,X}_i^{(mis)}$	set of missing observations in the independent variable Xi	

4. Experimental implementation

4.1. Datasets

The authors perform an analysis on the effect of the proportion of missing values and dataset dimension on imputation time and the accuracy of such imputation. To accomplish this study the authors used eight different datasets that are usually used within the literature (Table 2).

Table 2. Datasets specifications

Dataset name	Instances	Features	References
graduate admissions	500	8	[34]
profit estimation of companies	1000	6	[35]
red & white wine dataset	4898	12	[36]
California	20640	9	[37]
diamonds	53940	10	[38]
BNG_heart_statlog	1,000,000	14	[39]
Poker Hand Dataset	1,025,010	11	[40]
diabetes	442	11	[41]

The missing values were generated from the three missingness mechanisms with different proportions 10%, 20%, 30%, and 40% missingness ratios for each type. Analysis for BNG_heart_statlog and Poker Hand datasets were completed on randomly sampled sub datasets of 10000, 15000, 20000, and 50000 of instances[4].

4.2. Feature selection

Information gain's feature selection main disadvantage is that it is biased towards selecting features with lots of values, which encouraged Quinlan to define the Gain Ratio given by Eq. (4), which reduces this bias[42].

$$IG(A_k,X) = I(X) - E(A_k,X) \tag{2}$$

$$IV(A_k) = -\sum_{i=1}^{n} \frac{|X_i|}{|X|}.\,log_2 \frac{|X_i|}{|X|} \tag{3}$$

$$G_R = IG(A_k,X)/IV(A_k) \tag{4}$$

Where:

- *IG* is the information gain *X* is the set of examples.
- *I* specifies the entropy, and *E* is the expected information of the attribute A_k.
- *n* is the number of possible values of attribute A_k, and *IV* is the intrinsic value.
- G_R is the Gain ratio.

In the feature selection stage, the proposed algorithm depends on selecting the feature that presents the greatest information gain ratio with the target feature. The proposed algorithm named CBRG (Cumulative Bayesian Ridge regression with Gain-ratio).

4.3. Performance evaluation

The performance of the proposed algorithm evaluated using RMSE, MAE, and R^2, and time of imputation in seconds (t)[3].

$$RMSE = \sqrt{\frac{\sum_{l=1}^{n}(y_l - \hat{y_l})^2}{n}} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{l=1}^{n}|y_l - \hat{y_l}| \tag{6}$$

$$R^2(y_l - \hat{y_l}) = 1 - \frac{\sum_{l=1}^{n}(y_l - \hat{y_l})^2}{\sum_{l=1}^{n}(y_l - \bar{y})^2}\,; \tag{7}$$

$$\overline{y} = \frac{1}{n}\sum_{l=1}^{n} y_l$$

Where y_l and $\hat{y}_l$ are the real and predicted values of the *lth* observation, respectively, and *n* is the number of samples.

5. Results and discussion

The imputation method is considered to be effective if it imputes missing data in a short time and exhibits high accuracy. The results evaluated by taking the average performance of missing values imputation for every missingness mechanism for the four generated proportions. Figures 3 to 16 exhibit the results.

From the perspective of imputation time: The selection of the candidate feature to be imputed by the CBRG depends on the feature selection technique of information gain ratio. Calculating the information gain ratio depends on calculating the entropy, which is computationally expensive. Therefore, the CBRG consumes time during the imputation. However, CBRG exhibits a good imputation time with samples data and small size datasets. The MICE (Multivariate Imputation by Chained Equations) is not efficient in imputation time. LeastSquares and Norm offer the best imputation time with all datasets. Fast KNN and EMI offer a good imputation time with small datasets but consume big imputation time when dealing with large datasets.

From the perspective of performance: The proposed algorithm selects the most relevant feature (that offers higher information gain ratio) with the output feature, to be imputed and involving the imputed feature within the Bayesian ridge equation to impute another candidate feature, and so on until the imputation of all missing values within the dataset. Besides, the CBRG depends on the Bayesian Ridge, which in turn, is a probabilistic model with a ridge parameter. So the CBRG exhibits good prediction accuracy and low error. LeastSquares and MICE exhibit better performance against other mentioned methods. LeastSquares and Stochastic depend on the least squares approach. LeastSquares impute the missing data via the best fit line from a set of predictors. Stochastic adds a random draw to the prediction from samples from the regression's error distribution. Norm constructs a Gaussian distribution from the sample mean besides variance of the detected data, and then random samples from this distribution are used to impute missing data.

The results also revealed that the proposed method works well with any missingness mechanism and with any amount of missing data percentage. The CBRG, LeastSquares, and MICE present the best performance among other mentioned methods.

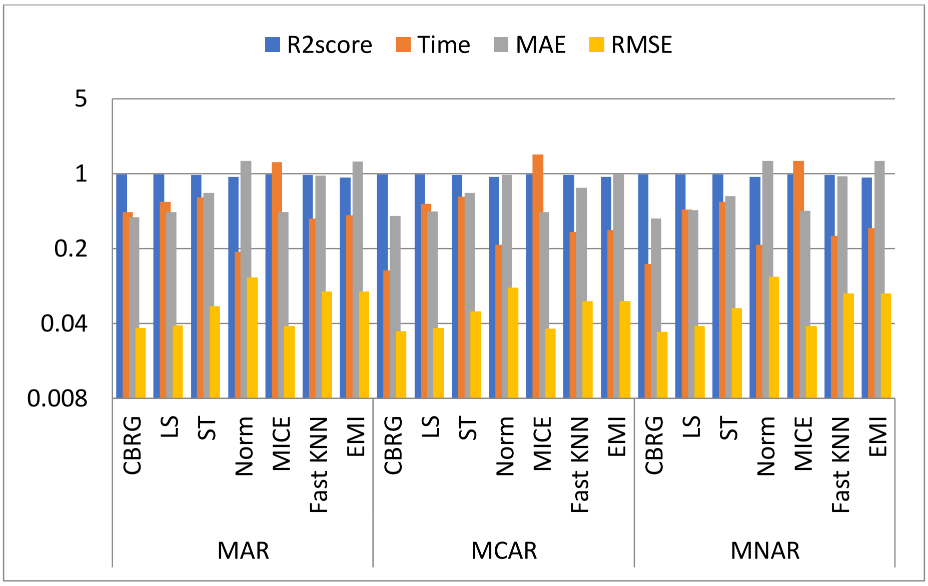

Figure 3. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (graduate admission dataset).

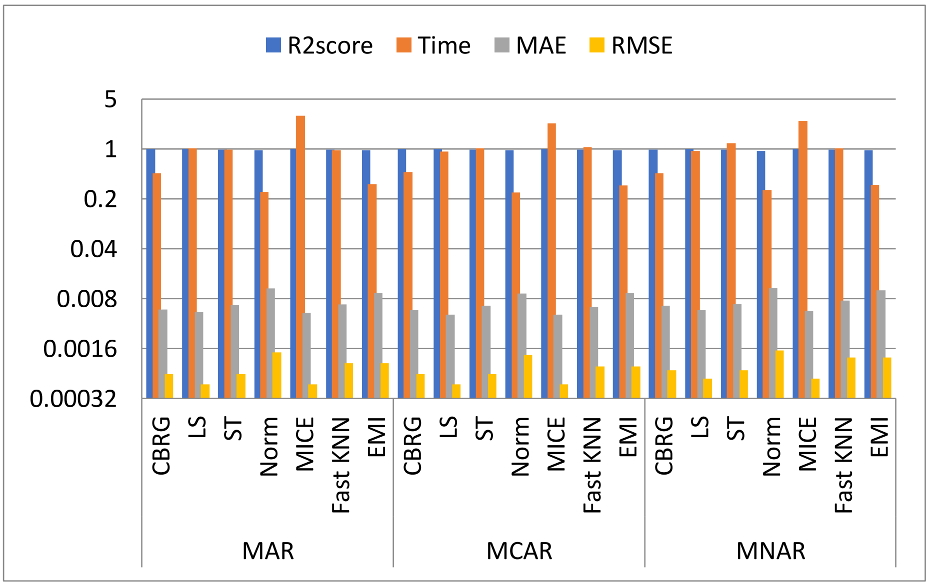

Figure 4. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (diabetes dataset).

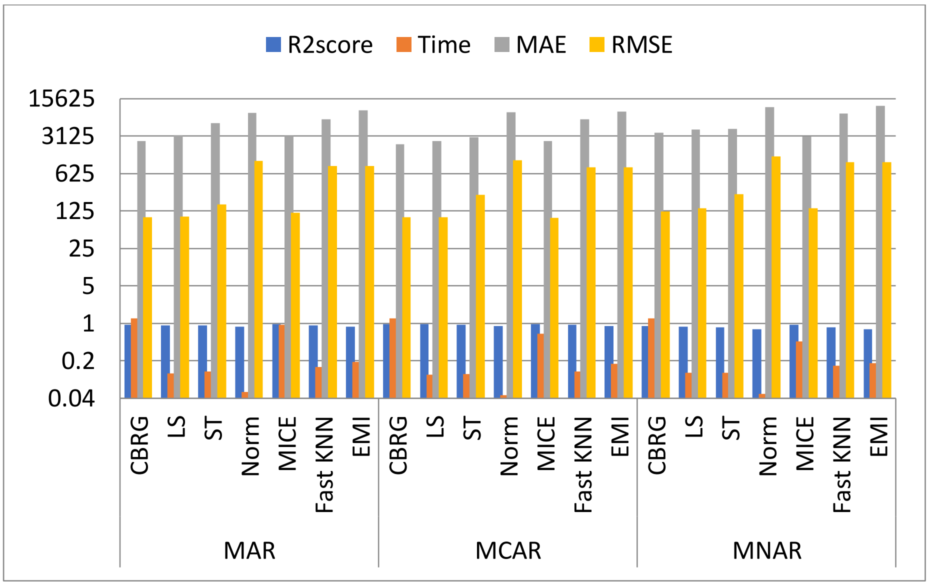

Figure 5. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (Profit dataset).

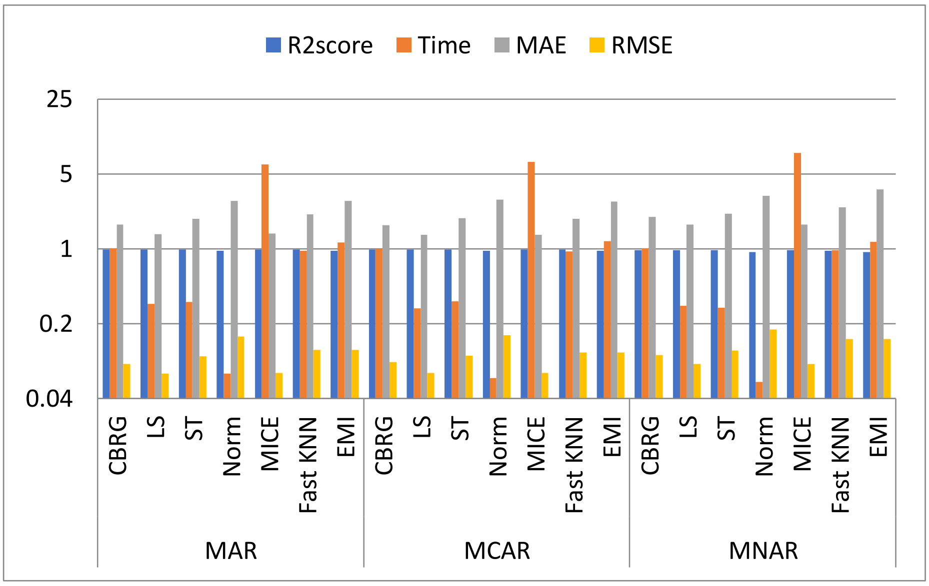

Figure 6. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (wine dataset).

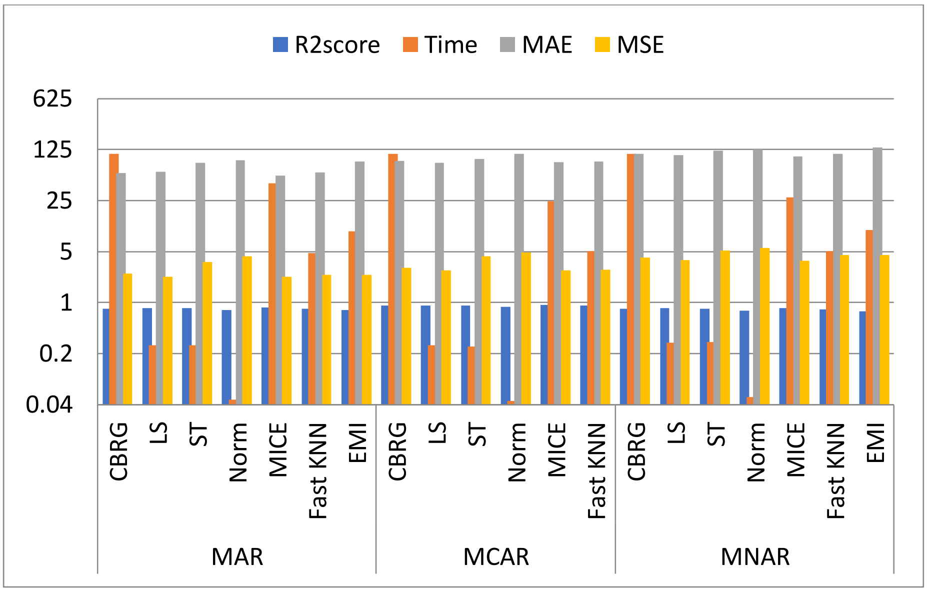

Figure 7. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (California dataset).

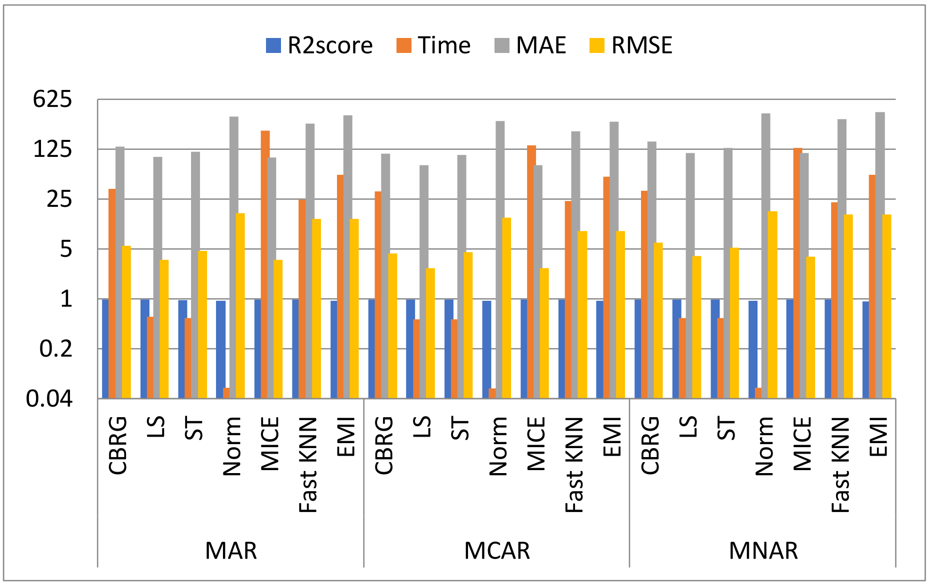

Figure 8. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (diamond dataset).

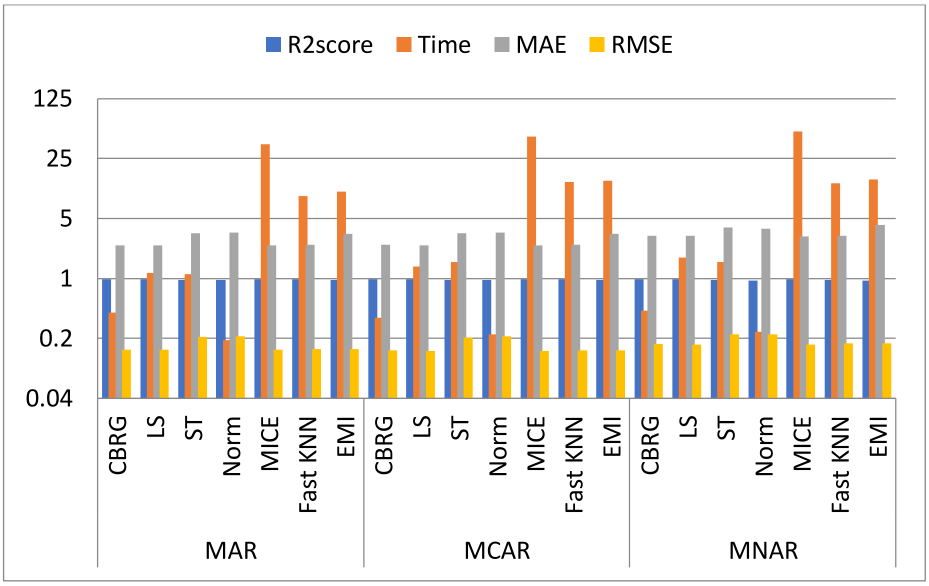

Figure 9. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (BNG (10000) dataset).

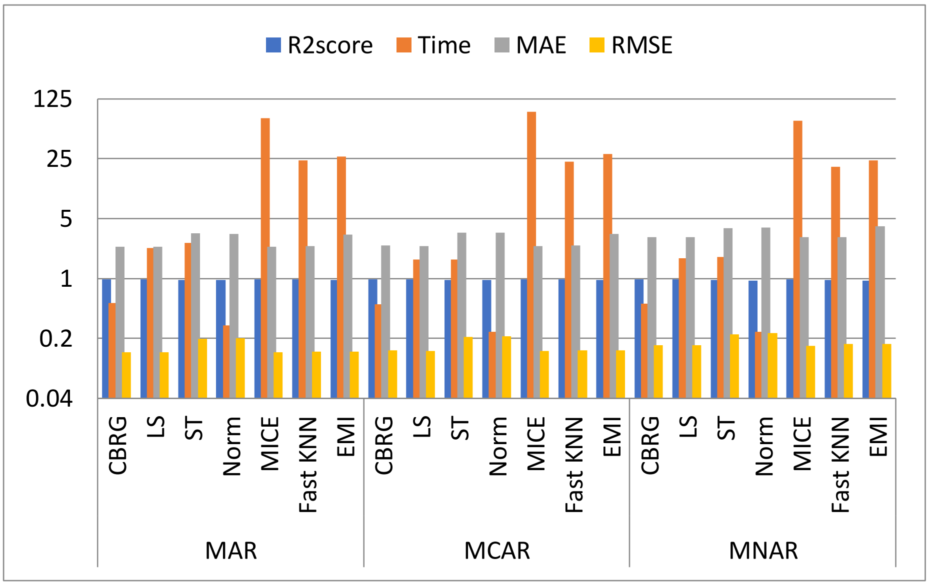

Figure 10. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (BNG (15000) dataset).

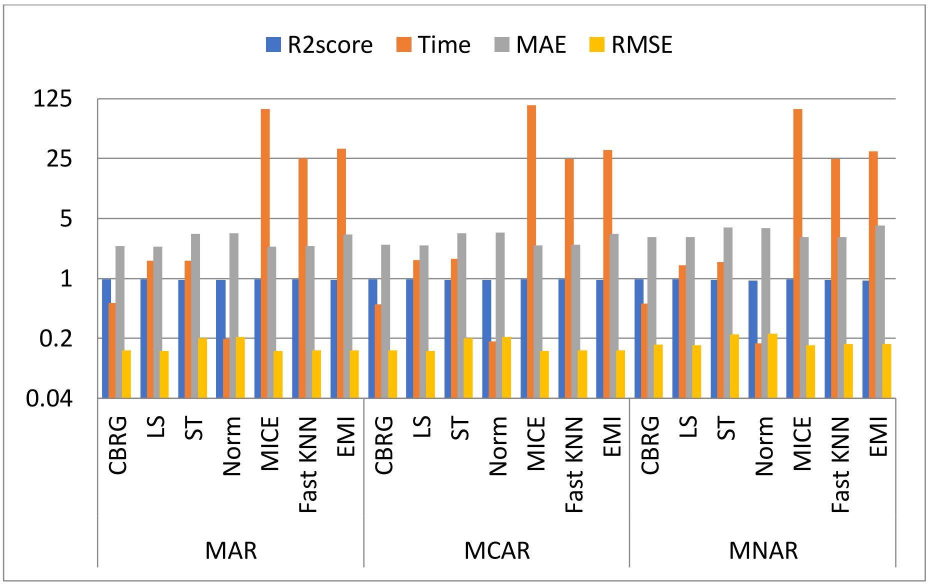

Figure 11. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (BNG (20000) dataset).

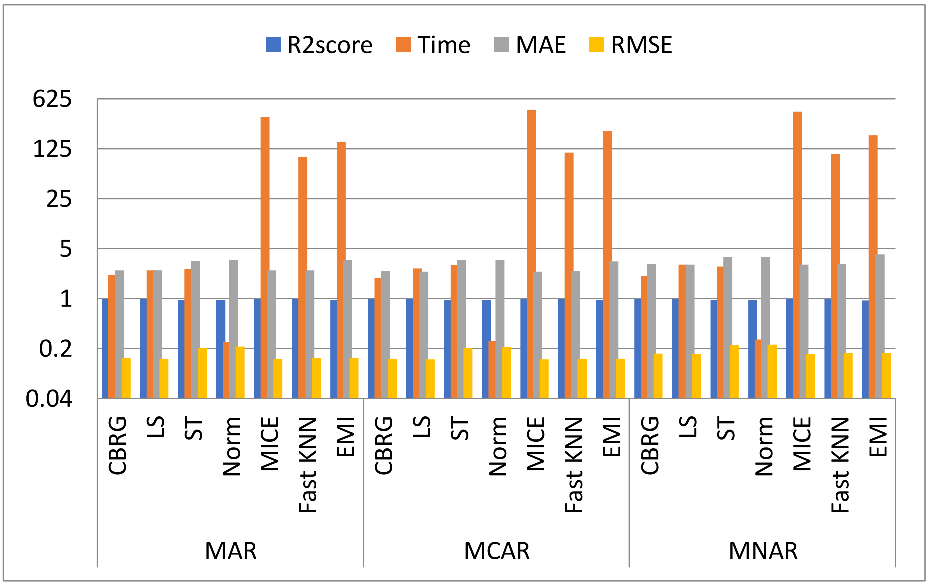

Figure 12. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (BNG (50000) dataset).

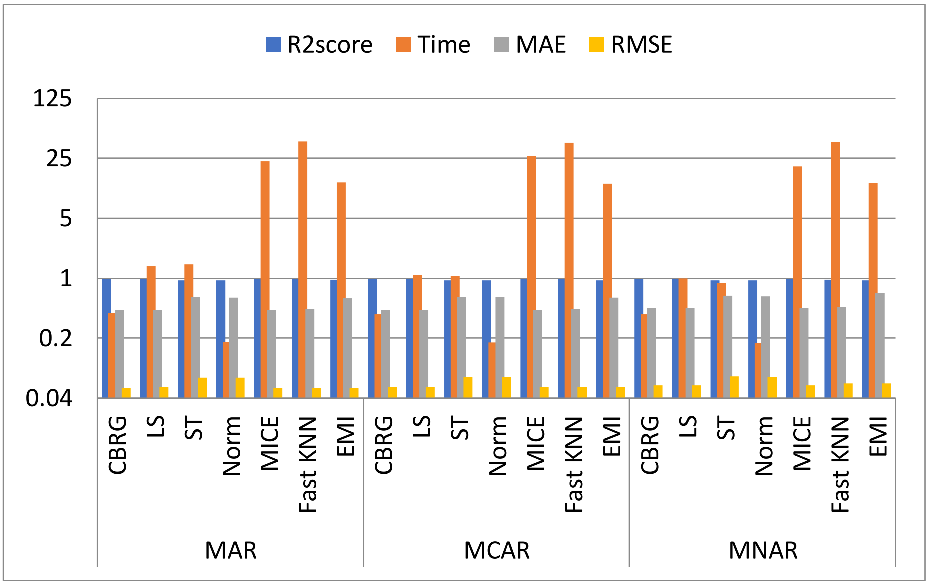

Figure 13. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (Poker (10000) dataset).

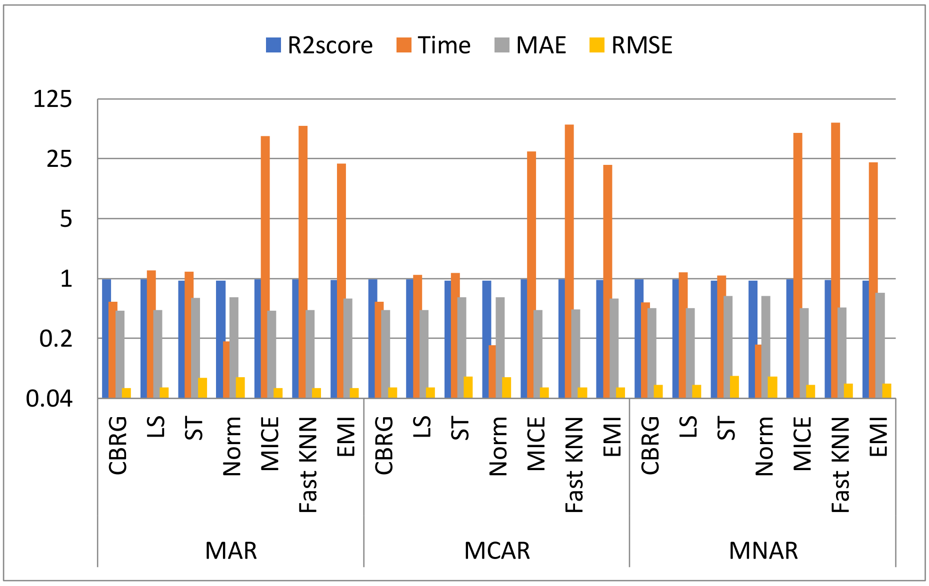

Figure 14. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (Poker (15000) dataset).

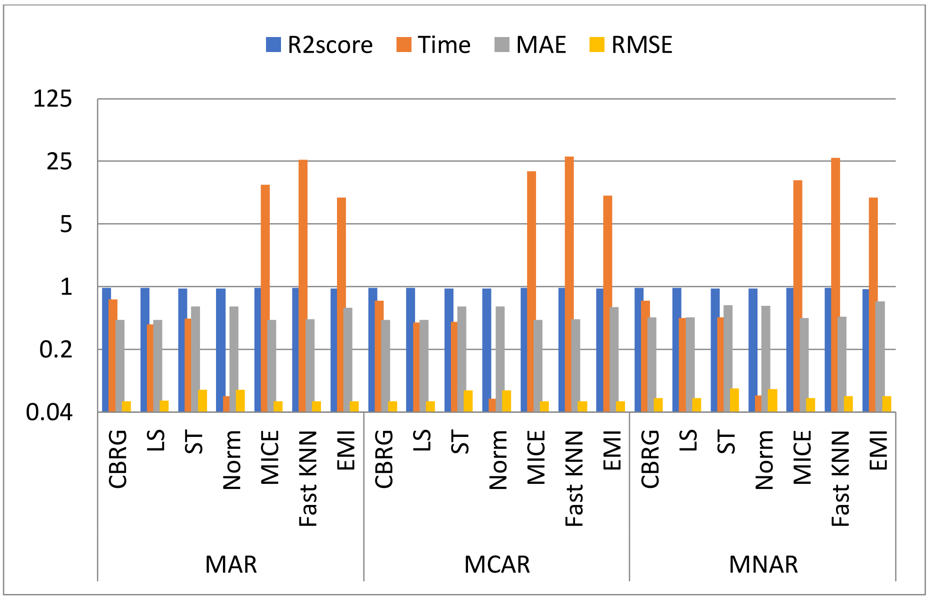

Figure 15. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (Poker (20000) dataset).

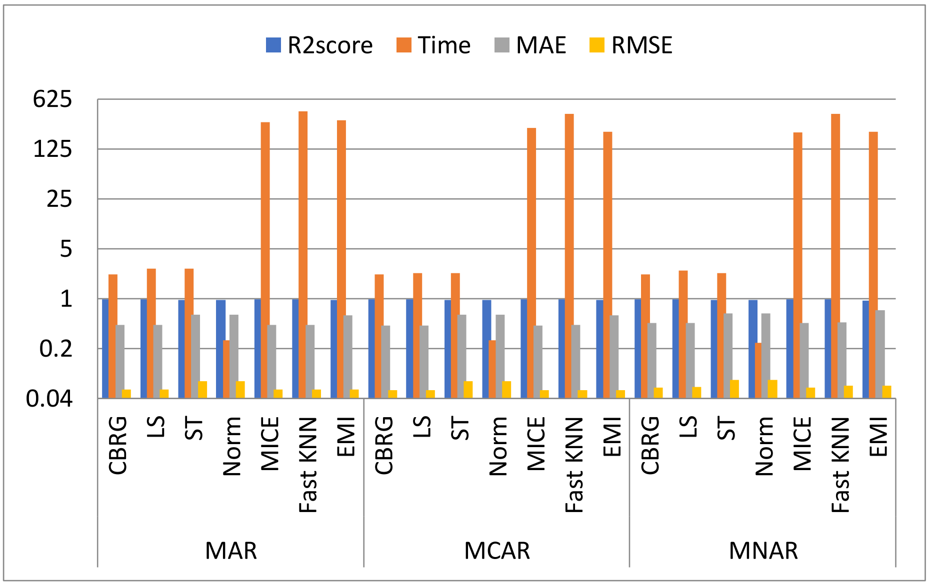

Figure 16. Comparison between the proposed method, LeastSquares, Stochastic, Norm, MICE, Fast KNN, and Expectation-Maximization Imputation (Poker (50000) dataset).

6. Conclusion

It is essential to handle missing data, as it happens in nearly all real-world research. Handling incomplete instances is very significant for observational analyses with several predictors.

In this paper, we have for a short time studied, a set of old published approaches that are to deal with missing data, reviewed their implementation on different datasets with different proportions of missing values generated from the three missingness mechanisms. Proposing a new algorithm Cumulative Bayesian Ridge Regression works in a cumulative order to impute all missing values in all features one after one. The candidate feature to be imputed is selected based on the information gain ratio. The proposed algorithm gives a better performance against the stated approaches even when its accuracy is sometimes a little worse than some packages but very close to them, in some cases, it is better than all of them. The proposed algorithm shows an acceptable running time and considered a fast method but not the fastest, as a result of calculating the gain ratio is computationally expensive.

In future research, it is advisable to study the proposed imputation algorithm in additional datasets; additional units of standard error (like T-value and P-value) will be taken into mind when picking the candidate feature. The best future trend is to take the help of algorithms that deal with optimization problems with mixed features such as the GSA-GA algorithm[43].

References

[1] Pigott TD. A Review of Methods for Missing Data. Educ Res Eval. 2001;7(4):353–83.

[2] Salgado CM, Azevedo C, Proença H, Vieira SM. Missing data. Secondary Analysis of Electronic Health Records. 2016. p. 143–62.

[3] Mostafa SM. Imputing missing values using cumulative linear regression. CAAI Trans Intell Technol. 2019 Sep;4(3):182–200.

[4] Yadav ML, Roychoudhury B. Handling missing values: A study of popular imputation packages in R. Knowledge-Based Syst. 2018 Nov;160(November 2018):104–18.

[5] Albayati MB, Altamimi AM. An empirical study for detecting fake facebook profiles using supervised mining techniques. Inform. 2019;43(1):77–86.

[6] Gelman A, Hill J. Missing-data imputation. In: Gelman A, Hill J, editors. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2006. p. 529–44. (Analytical Methods for Social Research).

[7] Sainani KL. Dealing With Missing Data. PM R. 2015 Jul;7(9):990–4.

[8] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999 Feb;8(1):3–15.

[9] Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. J Mach Learn Res. 2011 Oct;12(1):2825–30.

[10] Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2019 Apr;53(2):1487–509.

[11] Croiseau P, Génin E, Cordell HJ. Dealing with missing data in family-based association studies: A multiple imputation approach. Hum Hered. 2007 Mar;63(3):229–38.

[12] Royston P. Multiple imputation of missing values. Stata J. 2004 Aug;4(3):227–41.

[13] Batista GEAPA, Monard MCC. A study of k-nearest neighbour as an imputation method. Front Artif Intell Appl. 2002 Oct;87:251–60.

[14] Bertsimas DJ, Pawlowski C, Zhuo Y. From predictive methods to missing data imputation: An optimization approach. J Mach Learn Res. 2018 Apr;18(1):1–39.

[15] Li Y, Parker LE. Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks. Inf Fusion. 2014 Jan;15(1):64–79.

[16] Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. 2016 Jul;16(3):198–208.

[17] Farhangfar A, Kurgan LA, Pedrycz W. A Novel Framework for Imputation of Missing Values in Databases. IEEE Trans Syst Man, Cybern - Part A Syst Humans. 2007 Sep;37(5):692–709.
[18] Razavi-Far R, Cheng B, Saif M, Ahmadi M. Similarity-learning information-fusion schemes for missing data imputation. Knowledge-Based Syst. 2020 Jan;187(January 2020):104805.
[19] Rahman MG, Islam MZ. Missing value imputation using a fuzzy clustering-based EM approach. Knowl Inf Syst. 2016 Feb;46(2):389–422.
[20] Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol. 2008 Aug;26(8):897–9.
[21] Ma Z, Chen G. Bayesian methods for dealing with missing data problems. J Korean Stat Soc. 2018 Sep;47(3):297–313.
[22] Jiang H. Defect features recognition in 3D Industrial CT Images. Informatica. 2018 Sep;42(3):477–82.
[23] Sciences F. Working With Missing Values. J Marriage Fam. 2005 Nov;67(4):1012–28.
[24] Horton NJ, Lipsitz SR. Multiple imputation in practice : Comparison of software packages for regress ... Sci York. 2001 Aug;55(3):244–54.
[25] Fichman M, Cummings JN. Multiple Imputation for Missing Data: Making the Most of What You Know. Organ Res Methods. 2003 Jul;6(3):282–308.
[26] Graham JW. Missing Data Analysis: Making It Work in the Real World. Annu Rev Psychol. 2009 Jan;60(1):549–76.
[27] Chen Q, Paik MC, Kim M, Wang C. Using link-preserving imputation for logistic partially linear models with missing covariates. Comput Stat Data Anal. 2016 Sep;101(September 2016):174–85.
[28] Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. Am J Epidemiol. 2010 Jan;171(5):624–32.
[29] Sahri Z, Yusof R, Watada J. FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset. IEEE Trans Ind Informatics. 2014 Nov;10(4):2093–102.
[30] Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol. 2010 Sep;172(9):1070–6.
[31] Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. Comput Stat Data Anal. 1996 Aug;22(4):425–46.
[32] Rachmawan IEW, Barakbah AR. Optimization of missing value imputation using Reinforcement Programming. In: 2015 International Electronics Symposium (IES). 2015. p. 128–33.
[33] Mostafa SM. Missing data imputation by the aid of features similarities. Int J Big Data Manag. 2020 Jan;1(1):81.
[34] Acharya MS. Graduate admissions-1-6-2019 [Internet]. [cited 2019 Jun]. Available from: https://www.kaggle.com/mohansacharya/graduate-admissions
[35] Stephen B. profit estimation of companies [Internet]. [cited 2019 Aug]. Available from: https://github.com/boosuro/profit_estimation_of_companies
[36] Kartik P. Red & White wine Dataset [Internet]. [cited 2019 Feb]. Available from: https://www.kaggle.com/numberswithkartik/red-white-wine-dataset
[37] Cam N. California Housing Prices [Internet]. [cited 2019 Jul]. Available from: https://www.kaggle.com/camnugent/california-housing-prices
[38] S. Magrawal. Diamonds [Internet]. [cited 2019 Aug]. Available from: https://www.kaggle.com/shivam2503/diamonds
[39] Geoffrey Holmes,Bernhard Pfahringer,Jan van Rijn JV. BNG_heart_statlog [Internet]. [cited 2019 Sep]. Available from: https://www.openml.org/d/267
[40] Robert Cattral FO. Poker Hand Dataset [Internet]. [cited 2019 Nov]. Available from: https://archive.ics.uci.edu/ml/datasets/Poker+Hand
[41] Bradley Efron, Trevor Hastie IJ and RT (2004). Diabetes Data [Internet]. [cited 2019 Jun]. Available from: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html
[42] De Mántaras RL. A Distance-Based Attribute Selection Measure for Decision Tree Induction. Mach Learn. 1991 Jan;6(1):81–92.
[43] Garg H. A hybrid GSA-GA algorithm for constrained optimization problems. Inf Sci (Ny). 2019 Apr;478(April 2019):499–523.