ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(14), DOI: 10.17485/ijst/2019/v12i14/141616, April 2019*

A New Method of Imputation for the Missing Value in the IN/OUT Procedure of the Random Forest (RF)

Amjad Ali* and Qamruz Zaman

Department of Statistics, University of Peshawar, Pakistan;
safwan_state@yahoo.com, ayanqamar@gmail.com

Abstract

Objective: The main objective of study is to propose a new method of imputation for missing data. The study discuss misclassification rate, out of bag error for simulated and real data. **Method**: In this article, a new imputation method has been proposed for IN/OUT procedure of Random Forest (RF). The proposed method does not depend on the missing data mechanisms which are the principal advantages of this method. The method was evaluated and compared with non-missing data sets. **Findings and Conclusion**: It is concluded that the proposed method reduced the Out-of-Bag error and also the misclassification error rates in case of missing values using IN/OUT Procedure of RF and Conventional RF procedure at the different level of missing percentages. The proposed method gives interesting results in case of (5-15)% missing data and after that, the rest of the results are same therefore no need to compute the results for this percentage % of missing values. The most important is that this method was first time developed in the IN/OUT procedure of RF and conventional RF. **Novelty/Motivation:** Missing values a serious problem for all statistical problems. RF and IN/OUT RF are not exception. Therefore a bootstrap based method to impute missing value in the IN/OUT RF was developed.

Keywords: Classification and Regression Tree (CART), Misclassification, Out of Bag (OoB), Random Forest (RF)

1. Introduction

Intelligent data analysis techniques are useful for better investigate real-world data sets. However, the real-world data sets such as industrial, research and survey are always plagued by missing data this is one major factor affecting the quality in a typical data set[1]. In many applications, missing data is an issue which greatly affects large number of databases[2].

The missing values are occurred due to some unknown causes i.e. errors and lapses when the set of data recorded and also transferred.

Normally the assumption about the process that ever causes missing data seems to be that each value in the dataset is equally likely to be missing[3–5].

Most of the statistical and learning tools cannot handle missing values and hence these are needed to be deleted. However, this deletion process may produce biased estimates as well as loss of huge amount of precious information. Further, the remaining data set will be no more representative for the entire universe[6]. Instead of removing cases or variables having missing values, the alternative approach is to "impute" missing values. A lot of imputation methods are generally available which

**Author for correspondence*

are simple to more complex in range. By the use of these methods, entire sample can be kept to avoid biasness and to obtain more precise estimates.

Generally, imputation means to substitute some reasonable guess for each missing value and then proceed to do the analysis[7].

The imputation procedure is further divided into several procedures like mean imputation, regression method stochastic regression method, Hot-deck imputation method, all possible value imputation etc. These procedures effectively applied with distinct parametric models such as Gaussian regression and log-linear models. But, their usefulness has yet to be demonstrated for tree-based models, such as Classification and Regression Trees (CART) and Random Forest which is usually considered as non-parametric methods.

A CART model has two types: Regression type, in which the researcher predicts the value of continuous variables from a list of continuous and/or categorical predictor variables. On the other hand, if CART models predict the value of the categorical variable from a list of continuous and/or categorical predictor variables, is called classification-type model. The main advantage of tree-based models, such as the CART, is that these are scalable to large problems and can tackle smaller data set[8].

CART functions in linear as well as non-linear situations while no assumption is made about the distribution of data. The CART has some drawbacks the most important being is its' over fitting, high misclassification rate and data noise. To overcome these problems researchers have to introduce ensemble classifiers by learning more than one CART with some randomness steps. These ensemble classifiers include Boosting, Bagging and Random Forest (RF). In 1996, Boosting was primarily used to develop the precision of Classification/Regression Trees developed[9].

Boosting as an effective ensemble classifier uses a weighted average of results. The samples used at each step are given enlarged weight depending on the misclassified or incorrectly predicted cases. In[10] proposed bagging which is an ensemble technique. Its algorithm is based on drawing of several bootstrap samples from the existing training data sets, a classification or regression tree is learned for each bootstrap sample, and finally, the results are joint to obtain the overall prediction, by averaging for regression Tree and majority voting rule for the classification tree[11] demonstrated that Bagging always minimizes the classification error rate as compared to the conventional method. One of the drawbacks of Bagging is that it produces diverse classifiers due to small changes in the training data set. In[12] proposed RF, which has an extra layer of randomness. Besides, taking a different bootstrap sample of the training data set, RF also takes different set of variables.

Moreover, some practical classification problem has to deal with imbalanced data; i.e. the data set having at least one of the classes constitutes only a very small minority of the data, while the other classes is/are in majority. For such problems, the researcher is also interested in correct classification of the "rare" class (which is usually referred to as the "positive" class). For example, data related to fraud in debit card, network interruption, rare disease diagnosis, etc. Although, Random Forest minimizes the overall error rate but tends to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class. It may be possible that the rare cases (positive class cases) are not selected during the training phase, but selected during the testing phase, which consequently will be misclassified.

RF is constructed to decrease the overall error rate, but it tends to focus more on the prediction accuracy of the majority class, which is regularly a result in reduced accuracy for the minority class. It may be possible that during the training phase, these rare cases are not selected and therefore during testing phases they are misclassified.

In[13] proposed two ways to deal with the classification of imbalanced data using RF: 1. Weighted RF; and 2. Balanced RF. In Weighted RF, a heavier penalty was placed on misclassifying the minority class, by assigning a weight to each class, with the minority class given larger weight. On the other hand, balanced RF used a sampling technique: whether it is down-sampling the majority class either over-sampling the minority class or both of them. They showed that both methods gave improvement to the prediction accuracy of the minority class and had satisfac-

tory performance as compared to the existing algorithms. In[11] also used the down-sampling and over-sampling techniques in combination with RF for the imbalanced data of sentence boundary detection in the speech. They concluded that the sampling approaches outperformed than the conventional approach in terms of misclassification error.

They have shown that both approaches give improvement to the prediction accuracy of the minority class and have favorable performance compared to the conventional algorithm, but no one is clear winner. The reason was that by the down-sampling loss of information was occurred; while by over-sampling bias was introduced.

In proposed a new algorithm based on existing RF technique. In his technique, the conventional/classical RF procedure is implemented with two additional steps: 1. By randomly adding *j* cases of minority class in the Training data set; and 2. In the second step, removing first *j* cases and applying the RF algorithm to this new data set. Then removing the next *j* cases and this process is repeated until no more removal of cases is possible (just like jackknifing technique). The name for this algorithm was proposed as "IN/OUT algorithm for RF". Note that by adding and removing the same number of cases, the training data set becomes equal to the original size and also in this way all data is utilized, means no loss of information is done.

2. Missing Values Imputation in Random Forest

RF can deal with missing values in two ways to impute the missing values. In the library of RF the choice of "na. roughfix" is simply apply i.e. the column median is used for missing values for the numerical type of variable, on the other hand, the most frequent levels are to be used for the missing values in case of factor type variable[14].

The more advanced procedure for missing values is ("rfImpute()" on the proximity matrix in the RF library.

Proximity matrix is the symmetric matrix of n × n order. Where n is the total number of cases in the dataset. Run all cases in the training set are dropped down the tree .Now if both i and j cases lies in the identical terminal nodes then add to the proximity among i and j by one. Finally of the run, the proximities are divided by the total number of trees in the run[15]. The proximity matrix from the random forests is used to revise the imputations of the NAs. For continuous predictors, the imputed value is the weighted average of the non-missing observations, where the weights are the proximities. So, cases that are more like the cases with the missing data are given greater weight. For categorical predictors, the imputed value is the category with the major average proximity. Again, cases more like the case with the missing data are given larger weight. This process is relatively slow and requires a large number of iterations of forest growth. And the use of imputed values "tends to make the OOB measures of fit too optimist[12]. The computational demands are also quite daunting and may be impractical for many data sets until more efficient ways to handle the proximities are found.

In[16] proposed studied the performance of different algorithms for dealing with missing data in Random Forest. Their findings revealed that RF imputation performance is good under moderate to high missingness, especially in case of missing not at random.

3. Material and Methods

To achieve the research objectives of this study, a method proposed for imputation for missing values in IN/OUT procedure of RF. The proposed method is used to compute the misclassification error rate and Out of Bag estimates for misclassification. The proposed methodology applied on real balance data set and Haber man's data set. Also, a simulation study conducted using normal, exponential and hypergeometric probability distribution for study the validity of the proposed imputation procedure.

The proposed method for imputation of missing values in IN/OUT Procedure:

A new algorithm is proposed to deal with missing values in the RF procedure. The proposed method does not depend on the missing data system which is the principal advantages of this method. This rectifies disadvantages of all other imputation methods. It also requires no knowledge of either the probability distribution or model

structure and successfully incorporates the estimates of uncertainty associated with the imputed data.

3.1 Proposed Imputation Procedure

The proposed method is described as follows:

Dataset consists of a set variable having no missing observation in each every variable whether the variable is of attribute type or continues type, produce the percentage of missing in the training dataset.

Draw the random sample of size n with replacement by not assuming any kind of missing mechanism from the data by ignoring the missing observation.

Compute the descriptive from the drown sample and consider the estimated value as observation again draw the random sample from the training data by including the estimated value as observed value and repeat the process until the missing values are covered and this iterative procedure will reduce the biasness in the estimated variance of the estimate as in the single value imputation only one value is replaced whether it is mean, median, arbitrary value (in case of numerical data) or mode (in case of attribute data set) for all missing value which underestimate the variance of the data in this method for each value which is imputed by applying the proposed method. Now the IN/OUT procedure apply to the new data set which consists of the missing value estimated values and the non-missing values the above discussion can be written in the form algorithms as:

- Take the complete set of data which consist of the categorical variables as well as continuous variables to produce different percentage amount of missing values in the complete data set.
- Now apply the proposed algorithm to the incomplete data.
- In steps three the data will be become complete consisting of the estimated value of missing values.
- Apply the IN/OUT procedure of the Random Forest.
- Compute the OOB for each Random Forest of the IN/OUT procedure.
- Construct the confusion matrix.

Mathematically the method can be explained as. Let a complete dataset consist of the different variable X_1, X_2, X_3, X_4, … ,X_n having different number of observation in each variable a small portion of observation are hypothesis missing now for IN/OUT procedure the data should be complete for this purpose initially a random sample of size of the total of non-missing values were drawn and computed the descriptive depend upon the nature of the variable say $\hat{o}_1$ for the first variable at same missing position and similarly for all other missing values in all other variable were computed next the same procedure were repeated by considering the $\hat{o}_1$ as non-missing observation for the first variable and similar to other variables by not disturbing the same position at which it was missing.

3.2 Simulation Study of the Proposed Method

In this part of the study, a simulation study is planned in statistical software R for comparing the new proposed method of imputation. In the first phase, data has been generated from, Normal distribution Exponential distribution and Hypergeomatrice distribution.

Further, model is initially fitted by using 5 variables (4 independent variables and 1 dependent variable) and then it is extended for 6 variables (5 independent variables and 1 dependent variable). The data sets generated from the above distributions with different number of minority classes and majority classes.

3.3 Application of the Proposed Method on a Real Dataset

To concentrate on the application of the proposed algorithms (in case of missing values), the two data set Haber man's Survival Data Set, Balance Data Set, were down-

loaded from UCI Machine Learning Repository website (http://archive.ics.uci.edu/ml/datasets.html).

The four approaches are applied to the mentioned data set. These are:

Conventional Random Forest (without missing).

In this approach, we have utilized the RF technique on the original data set (i.e. having no missing value) and results are obtained.

IN/OUT algorithm (without missing).

In this approach, IN/OUT algorithm is utilized using the original data set.

Conventional RF (with missing).

In this approach, first a specific percentage of missing values are produced and imputation method is applied to impute the missing values.

Then the conventional RF is applied to get the results.

IN/OUT algorithm (with missing).

Similar to the approach mentioned in (c), first a specific percentage of missing values are produced and imputation method is applied to impute the missing values. Then the In/out algorithm is applied to get the results.

3.3.1 Balance Data Set

This data set was first used to model psychological experimental results (Klahr and Seiglar 1978) (Lichman, 2013). Each example/unit was classified into one of the three classes according to balance scale tip: Tip to the Right (R), tip to the Left (L) or be Balanced (B). Total numbers of instances/cases were 625 having 576 majority cases (L and R) and 49 minority cases (B). There were five categorical type variables in this data set, whose detail is mentioned below:

S#	Variables	No. of Categories	Description of categories
Class Name		3	L, B, and R
Left-Weight		5	1, 2, 3, 4 and 5
Left-Distance		5	1, 2, 3, 4 and 5
Right-Weight		5	1, 2, 3, 4 and 5
Right-Distance		5	1, 2, 3, 4 and 5

3.3.2 Haber Man's Survival Data Set

This data set relates to a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer Haber man (Lichman, 2013). Total numbers of instances/cases were 306 having 225 majority cases and 81 minority cases. There were total 4 variables, whose detail is as under:

S#Variables	Description of categories
1. Age of patient at a time of operation	numerical type
2. Patient's year of operation	year in two digits, numerical type
3. Number of positive auxiliary nodes	numerical type
4. Survival status	= the patient survived 5 years or longer = the patient died within 5 year

4. Discussions on Simulation Results

The results of misclassifications rates and OOB estimate for data set having different % of missing values generated

Table 1. Misclassification rates and OOB rates considering normal distribution

Missing %	Classifier used	OOB	Misclassification rates	
			Minority	Majority
0%	Conventional RF IN/OUT Algorithm	53.51 47.32	0.61 0.47	0.45 0.98
5%	Conventional RF IN/OUT Algorithm	35.47 28.24	0.51 0.39	0.21 0.14
10%	Conventional RF IN/OUT Algorithm	38.28 33.38	0.40 0.40	0.36 0.21
15%	Conventional RF IN/OUT Algorithm	40.08 30.27	0.46 0.35	0.33 0.10
20%	Conventional RF IN/OUT Algorithm	43.09 43.56	0.58 0.43	0.29 0.44

Table 2. Misclassification rates and OOB rates considering hypergeomatrice distribution

Missing %	Classifier used	OOB	Misclassification rates	
			Minority	Majority
0%	Conventional RF IN/OUT Algorithm	28.06 27.34	0.01 0.04	0.97 0.70
5%	Conventional RF IN/OUT Algorithm	23.65 21.17	0.08 0.09	0.57 0.50
10%	Conventional RF IN/OUT Algorithm	24.65 24.01	0.02 0.05	0.90 0.57
15%	Conventional RF IN/OUT Algorithm	25.65 26.22	0.09 0.07	0.61 0.67
20%	Conventional RF IN/OUT Algorithm	27.45 27.13	0.01 0.04	0.95 0.82

from normal distribution using the proposed method of imputation are presented in Table 1. The OOB estimate for Conventional RF (having no missing value) is 53.51, which is dropped to 35.47 after applying the proposed imputation method on a data set of 5% missing values. Similarly, the OOB estimate for IN/OUT algorithm for data set having no missing value is 47.32, while for data set having 5% missing value, the OOB estimate is 28.24 (i.e. less than conventional RF approach).

On the other hand, the misclassification rate for minority class are also favorable for IN/OUT algorithm i.e. the misclassification rates of minority class through conventional RF approach on original data set and on missing data set are 67% and 51% respectively, while the misclassification rates of minority class through In/Out algorithm on non-missing data set and missing data set are 47% and 39% respectively.

The OOB estimate for Conventional RF (having 10% missing value) is 38.32 after applying the proposed imputation method on missing values data set. Similarly, the OOB estimate for IN/OUT algorithm for data set having 10% missing value, the OOB estimate is 33.38 (i.e. less than conventional RF approach).

On the other hand, the misclassification rate for minority class are also favorable for IN/OUT algorithm i.e. the misclassification rates of minority class through conventional RF approach on original data set and on missing data set are 61% and 40% respectively, while the misclassification rates of minority class through In/Out algorithm on non-missing data set and missing data set are 47% and 40% respectively.

The OOB estimate for Conventional RF (having 15% missing value) is 40.08 after applying the proposed imputation method on missing values data set. Similarly, the OOB estimate for IN/OUT algorithm for data set having 15% missing value, the OOB estimate is 30.27 (i.e. less than conventional RF approach). On the other hand, the misclassification rate for minority class are also favorable for IN/OUT algorithm i.e. the misclassification rates of minority class through conventional RF approach

Table 3. Misclassification rates and OOB rates considering exponential distribution

Missing %	Classifier used	OOB	Misclassification rates	
			Minority	Majority
0%	Conventional RF IN/OUT Algorithm	5.21 1.83	0.82 0.09	0.006 0.009
5%	Conventional RF IN/OUT Algorithm	3.01 0.64	0.56 0.02	0.004 0.004
10%	Conventional RF IN/OUT Algorithm	3.41 0.75	0.53 0.01	0.002 0.004
15%	Conventional RF IN/OUT Algorithm	4.01 1.45	0.60 0.06	0.006 0.008
20%	Conventional RF IN/OUT Algorithm	5.01 1.52	0.80 0.07	0.008 0.005

Table 4. Misclassifications rates and OOB estimate for balance data set

Missing %	Classifier used	OOB	Misclassification rates		
			B	L	R
0%	Conventional RF IN/OUT Algorithm	15.84 11.25	100 3	7.6 12.25	9.7 12.7
5%	Conventional RF IN/OUT Algorithm	9.6 8.58	58 23.4	5.2 7.5	4.8 4
10%	Conventional RF IN/OUT Algorithm	10.08 9.95	74 25.4	2.7 8.6	6 5.7
15%	Conventional RF IN/OUT Algorithm	11.2 10.06	69.2 25.4	4 6.6	7.2 7.6
20%	Conventional RF IN/OUT Algorithm	11.36 10.71	76.4 30.6	6.1 7.6	5 6.4

Table 5. Misclassifications rates and OOB estimate Haber man's data set

Missing %	Classifier used	OOB	Misclassification rates	
			Died	Survived
0%	Conventional RF IN/OUT Algorithm	32.03% 16.03%	0.68 0.25	0.74 0.03
5%	Conventional RF IN/OUT Algorithm	16.65% 15.1	0.09 0.15	0.39 0.14
10%	Conventional RF IN/OUT Algorithm	17.01 % 15.907	0.08 0.18	0.36 0.12
15%	Conventional RF IN/OUT Algorithm	18.63 16.25	0.09 0.15	0.42 0.17
20%	Conventional RF IN/OUT Algorithm	20.26 16.83	0.102 0.18	0.48 0.14

Figure 1. Line graph for misclassification rate.

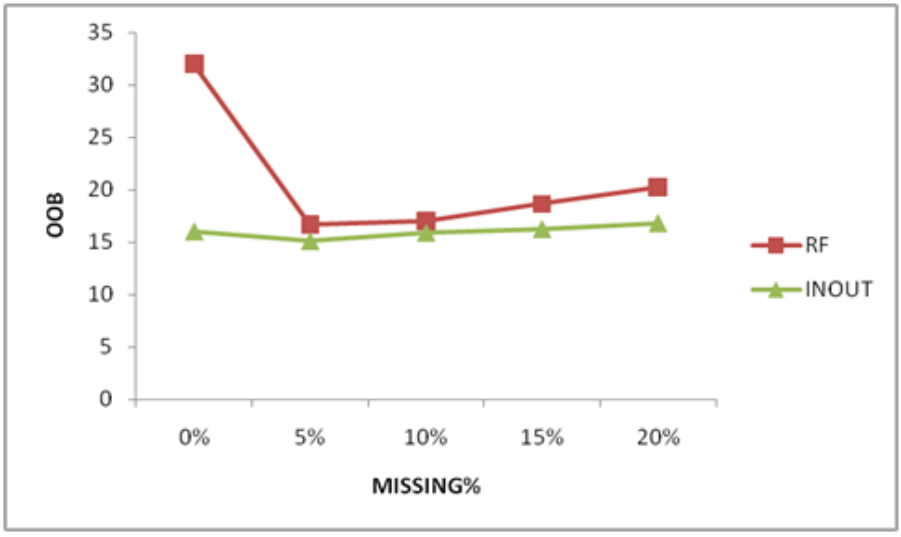

Figure 2. Line graph of missing data.

on original data set and on missing data set are 61% and 46% respectively, while the misclassification rates of minority class through IN/OUT algorithm on non-missing data set and missing data set are 47% and 35% respectively.

The OOB estimate for Conventional RF (having 20% missing value) is 43.09 after applying the proposed imputation method on missing values data set. Similarly, the OOB estimate for IN/OUT algorithm for data set having 20% missing value, the OOB estimate is 43.56.

On the other hand, the misclassification rate for minority class are also favorable for In/out algorithm i.e. the misclassification rates of minority class through

conventional RF approach on original data set and on missing data set are 61% and 58% respectively, while the misclassification rates of minority class through IN/OUT algorithm on non-missing data set and missing data set are 47% and 43% respectively. All these discussions are summarized and depicted in Figures 1 and 2.

The rest of the table numbers 2, 3, 4 and 5 all gives the results in the same pattern even though the data were generated from the different sources i.e. for Table 2 and Table 3 the data from exponential and hypergeomatrice distribution while the Table 4 and Table 5 are showing the results from the real data sets.

5. Conclusion

A new method was developed for the imputation of the missing values and the performance of the method was compared with non-missing values by using the conventional Random Forest as well as IN/OUT the procedure of the random forest. The comparisonsare based on the simulation study, the misclassification rate and the Out of Bag error were minimum when the proposed method was applied with different % of missing values were produced. The similar results were found of the proposed method based on the Conventional Random Forest and IN/OUT procedure of Random Forest, when some real data set was used after producing the 5%, 10%, 15% and 20% of the data were missing. It can be concluded that the proposed method does not depend on the missing data system which is the main advantages of this method on the other common methods of imputation. It also requires no knowledge of either the probability distribution or model structure and successfully incorporates the estimates of uncertainty associated with the imputed data.

6. References

1.Graham JW. Missing data analysis: Making it work in the real world. Annual Review of Psychology. 2009; 60:549–76. PMid: 18652544. https://doi.org/10.1146/annurev.psych.58.110405.085530.

2. Data Mining: Practical machine learning tools and techniques. 2005. file:///C:/Users/a/Downloads/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf.

3. Afifi AA, Elashoff RM. Missing observations in multivariate statistics I. Review of the literature. Journal of the American Statistical Association. 1966; 61(315): 595–604.

4. Anderson TW. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association. 1957; 52(278):200–3. https://doi.org/10.1080/01621459.1957.10501379.

5. Hartley H, Hocking R. The analysis of incomplete data. Biometrics. 1971; 27(4):783–823. https://doi.org/10.2307/2528820.

6. Myers TA. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. Communication Methods and Measures. 2011; 5(4):297–310. https://doi.org/10.1080/19312458.2011.624490.

6. Myers TA. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. Communication Methods and Measures. 2011; 5(4):297–310. https://doi.org/10.1080/19312458.2011.624490.

7. Statistical analysis with missing data. 2002. https://www.wiley.com/en-us/ Statistical+Analysis+with+Missing+Data%2C+2nd +Edition-p-9780471183860

8. Markham IS, Mathieu RG, Wray BA. Kanban setting through artificial intelligence: A comparative study of artificial neural networks and decision trees. Integrated Manufacturing Systems. 2000; 11(4):239–46. https://doi.org/10.1108/09576060010326230.

9. Game theory, on-line prediction and boosting. 1996. http://www.cs.cmu.edu/~ninamf/LGO10/wm-minimax.pdf

10. Random forests-random features. 1999. https://www.stat.berkeley.edu/~breiman/random-forests.pdf

11. Liu Y. A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer Speech and Language. 2006; 20(4):468–94. https://doi.org/10.1016/j.csl.2005.06.002.

12. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

13. Using Random Forest to learn imbalanced data. 2004. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

14. Troyanskaya O. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–5. PMid: 11395428. https://doi.org/10.1093/bioinformatics/17.6.520
15. Impute: impute: Imputation for microarray data. 2011. https://rdrr.io/bioc/impute/.
16. Tang F, Ishwaran H. Random Forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2017; 10(6):363–77. PMid: 29403567 PMCid: PMC5796790. https://doi.org/10.1002/sam.11348