Soft Computing (2020) 24:555–569
https://doi.org/10.1007/s00500-019-03972-x

METHODOLOGIES AND APPLICATION

A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation

Samaher Al-Janabi[1] · Ayad F. Alkaim[2]

Published online: 11 April 2019

Abstract
One of the important trends in an intelligent data analysis will be the growing importance of data processing. But this point faces problems similar to those of data mining (i.e., high-dimensional data, missing value imputation and data integration); one of the challenges in estimation missing value methods is how to select the optimal number of nearest neighbors of those values. This paper, attempting to search the capability of building a novel tool to estimate missing values of various datasets called developed random forest and local least squares (DRFLLS). By developing random forest algorithm, seven categories of similarity measures were defined. These categories are person similarity coefficient, simple similarity, and fuzzy similarity (M1, M2, M3, M4 and M5). They are sufficient to estimate the optimal number of neighborhoods of missing values in this application. Hereafter, local least squares (LLS) has been used to estimate the missing values. Imputation accuracy can be measured in different ways: Pearson correlation (PC) and NRMSE. Then, the optimal number of neighborhoods is associated with the highest value of PC and a smaller value of NRMSE. The experimental results were carried out on six datasets obtained from different disciplines, and DRFLLS proves the dataset which has a small rate of missing values gave the best estimation to the number of nearest neighbors by DRFPC and in the second degree by DRFFSM1 when $r = 4$, while if the dataset has high rate of missing values, then it gave the best estimation to number of nearest neighbors by DRFFSM5 and in the second degree by DRFFSM3. After that, the missing value was estimated by LLS, and the results accuracy was measured by NRMSE and Pearson correlation. The smallest value of NRMSE for a given dataset is corresponding to DRF correlation function which is a better function for a given dataset. The highest value of PC for a given dataset is corresponding to DRF correlation function which is a better function for a given dataset.

Keywords Intelligent data analysis · Missing values · Imputation methods · Random forest · Local least squares

1 Introduction

Missing value problem is considered as one of the challenges in multi-applications, and it affects directory of taking the decision. Several solutions have been suggested to solve this problem. First, the simplest solution for this problem is the reduction in the dataset and the elimination of all samples with missing values. This is possible when large datasets are available, and missing values occur only in a small percentage of samples. Second, a data miner, together with the domain expert, can manually examine samples, which have no values and enter a reasonable, probable, or expected value, based on a domain experience. This solution is straightforward for small numbers of missing values and relatively small datasets. However, if there is no obvious or plausible value for each case, the minor is introducing noise into the dataset by manually generating a value. Finally, automatic replacement of missing values with some constants such as some of the solutions will be explaining later (Han and Kamber 2006; Graham 2012; Rubin 1976). Several solutions are possible here, which are:

- Replace all missing values with a single global constant
- Replace a missing value with its feature mean

Communicated by V. Loia.

✉ Samaher Al-Janabi
samaher@itnet.uobabylon.edu.iq

Ayad F. Alkaim
alkaim@iftc.uni-hannover.de

[1] Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq

[2] Department of Chemistry Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq

- Replace a missing value with its feature mean for the given class
- Replace a missing value with the nearest neighborhood from top or bottom.

In intelligent data analysis, the researcher is often interested in discovering knowledge, which has a certain predictive power. The basic idea is to predict the value that some attribute(s) will take on in "the future," based on previously observed data. However, the missing value problem always leads to some of the mistake in discovered knowledge process and then become very difficult to be comprehensible to the user and in many times results in incomplete intelligent analysis of dataset behaviors (Abualigah and Hanandeh 2015; Abualigah and Khader 2017; Abualigah et al. 2017; Al-Janabi 2018).

In real-world applications of data mining, even when there are huge amounts of data, the subset of cases with complete data may be relatively small. Some of data mining methods accept missing values and satisfactorily process data to reach a final conclusion. Other methods require the all values be available (Rubin 1996). The question is whether theses missing values can be filled in during data preparation, prior to the application of any data mining methods. Other, multivariate supervised classification methods such as support vector machines and multivariate statistical analysis methods such as principal component analysis, singular value decomposition and generalized singular value decomposition cannot be applied to data with missing values (Liew et al. 2010). Therefore, missing value estimation is an important preprocessing step.

Missing data present a problem in many fields, and different techniques have been developed to effectively handle this problem in datasets. The simplest way is to remove the entire raw that contains missing values and replace them with row average, medium or even with zero (Ali 2012a). However, this leads to biases since the correlation structure between the variables is ignored. Missing value methods are designed for continuous data and to those data that have a mixture of nominal and categorical variables, and implementation can break down in challenging datasetting. The inability to deal with complex interactions between variables prevents these techniques to tailor the missing data handling procedure and match a set of analysis goals. In addition, it is unclear how these methods appropriate for missing value problem because no universal tool exists for different types of datasets. For these reasons, there has been much interest in using machine learning (ML) methods to estimate missing data problems.

Random forest (RF) is considered to be a popular classifier that has shown significant attainments on a wide range of classification problems due to its ability to deal with high-dimensional data. Although RF has shown to have many good features, it has the potential to perform even better for estimating missing values (Rieger et al. 2010; Cutler et al. 2007; Hapfelmeier et al. 2012; Waljee et al. 2013; Stekhoven and Bühlmann 2012; Verikas et al. 2011; Aljarah et al. 2018). RF has offered different advantages for classification problems when applied on large datasets: it does not overfit, it has the ability to estimate feature importance during training with little additional time and most importantly, and it effectively accommodates nonlinear relations and interactions among variables (Mafarja et al. 2018; Genuer et al. 2010). Local least squares (LLS) technique has been employed, and its performance is proved to be effective and suitable for missing value imputation (Ali 2012b; Wasito and Mirkin 2006; Moorthy et al. 2014; Bose et al. 2013).

In this study, a new tool called DRFLLS based on the developed random forest (DRF) algorithm and LLS imputation strategy is applied to find the optimal estimation of missing values. Different equations as similarity measures inside the original RF were used rather than depending on the correlation only to estimate the optimal number of nearest neighbors (k). Then, we used LLS as the local strategy to estimate the missing values based on the number of neighbors. The Pearson correlation (PC) and normalized root mean squared error (NRMSE) are used to assess the imputation accuracy, where the value that gives the highest PC and the least NRMSE represents the optimal value. Thus, the similarity measure that generated the k is considered the best generator of the number of neighbors.

The rest of the paper includes the following sections. In Sect. 2, we summarize and analyze the work of previous related works. In Sect. 3, we describe materials and methods used in this work. In Sect. 4, the design of the proposed method is discussed. In Sect. 5, the results and evaluation measures are presented. This section is followed by the conclusion of the study.

2 Related works

In recent years, ML techniques were designed to estimate missing values. This section is centered upon the previous literature which is related to missing value estimation and the solutions that have been proposed.

Several recent works have been proposed to handle missing data. Al-Janabi (2017) presented a solution for missing value problem, which consists of many steps: first, dataset design: "Vertical" decomposition, "horizontal" decomposition; second, new constraint short hands: "distributed key" and "distributed foreign key"; third, new dataset updating construct: "multiple assignment"

(table level); fourth, decomposition by query to derive (an improved) PERS_INFO when needed. However, this technique cannot find the missing values effectively because it depends on finding substation values from the tables of the missing values based on propositional constructions.

In another research, Bruggeman et al. (2009), suggested PhyloPars Web server to provide a statistically consistent technique. The authors merged an incomplete set of empirical records with species phylogeny to produce a complete set of estimates parameter for all species. Their model is extended to enable better handling of missing data. They stated that their method achieved optimal and accurate use of all available information than ad hoc alternatives. Another optimal method for replacing missing ensemble temperature data can be seen in the work of McCandless et al. (2011). The main objective of their work is to produce a consensus forecast through the use of statistical post-processing techniques to find out the results of replacing the missing data on these post-processing schemes. However, this method does not explain how to handle missing value problem in details.

Qi et al. (2005) presented a method to measure such similarities at task classifying pairs of proteins as they interact or not. They used direct and indirect information about interaction pairs to construct a RF from a training dataset. The resulted RF is employed to find the similarity between protein pairs using a modified k-nearest neighbor. Their final results demonstrate that the RF approach achieved a high level of accuracy compared to other proposed tools. Carranza and Laborte (2015) used RF to investigate its suitability for data-driven predictive model and to examine its ability to handle missing values using Abra data in Philippines. The analysis results indicate that RF is useful for both data-driven predictive and missing value handling. Pantanowitz and Marwala (2009) applied five methods to impute missing data using HIV seroprevalence dataset. The final results show that RF is a powerful and accurate method which can successfully be applied to handle missing values.

Golub et al. (2005) proposed the imputation method based on least squares formulation. The authors used local similarity structures in the data and least squares optimization process to estimate the missing values in gene expression dataset. In addition, the experiments show that their method achieved comparable results alongside with other approaches for missing value estimation on different datasets. In another work, four imputation approaches were assessed to handle missing values in Epistatic miniarray profiling (E-MAPs) data (Ryan et al. 2010). Three local (nearest neighbor-based) and one global (BPCA-based) techniques that adapted to work with symmetric pair-wise data were used. The experimental results prove that good missing value imputation can be achieved by using LLS. The work of Chiu et al. (2013) confirmed the use of LLS as the best and suitable technique for missing values handling.

3 Methods and material

3.1 Random forest

Random forest (RF) is a supervised learning technique (Breiman 2001) that is widely used to solve classification and regression problems in different domains (Heidari et al. 2018, 2019; Adam et al. 2014; Elyan and Gaber 2016; Ali 2013; Xie et al. 2009). RF ensembles of trees that have grown from bootstrapped training data for classification purposes, the trees are combined using majority voting with one vote per tree over all trees in the forest, while for regression purposes, forests are created by averaging over trees. The remaining samples that are not selected for training are collected to another subset called out-of-bag (OOB). This subset aims to assess generated decision trees and to estimate the classification or regression error rate in the RF (James et al. 2013).

3.2 Local least squares

Local least squares (LLS) imputation consists of two steps: the first is to use k similarity records, and the second is to utilize regression and estimation, regardless of how the k records are chosen. The traditional LLS is based on the L2-norm or Pearson correlation coefficients as methods to select the similarity records (number of neighbors k) and then recover missing values that depend on the k records with the largest Pearson correlation coefficients (Kumar et al. 2008).

3.3 Used datasets

In this work, six datasets were used which have different types and features as explained in Table 1. All of these datasets suffer from the missing value problem.

4 DRFLLS as a novel tool

In intelligent data analysis, we are often concerned to discover knowledge which has a certain predictive power. Intelligent data analysis goal is to predict the value that some attribute(s) will take on in 'the future' based on previously observed data. But missing value problem always leads to inaccurate conclusions that can be drawn from the data, and the process then becomes difficult for the users to understand. To overcome this problem, this work proposes a tool called developed random forest local least squares (DRFLLS) to estimate the number of neighbors and find the optimal

Table 1 Overview of the dataset information

Dataset name	# Records	# Features	# Missing values	Area
Communities and Crime (Redmond 2009)	766	128	26,850	Social
DNA (Genbank 1992)	2000	181	31,000	Bioinformatic
P53 Center for Machine Learning and Intelligent Systems, USA (2010a)	889	255	15,870	Life Sciences
URL reputation (Saul et al. 2009)	11,256	28,712	98,341	Computer
Splice (Genbank 2018)	3190	61	16,780	Life
GIS Center for Machine Learning and Intelligent Systems, USA (2010b)	1001	9	822	Bioinformatic

estimation of missing values. The DRFLLS design is divided into three parts: (1) the first part aim to generate k similarity records depending on seven different measures of similarity, where each of these measures uses a new correlation function of the random forest. As a result, this part generates seven different values of neighbors, i.e., solve the select k-nearest neighbor problem; (2) the second part takes the different values of k results from the previous step and applies LLS to estimate the missing values. This generates seven values for each missing value based on the number of similarity measures; (3) the third part evaluates which of these seven estimation values is optimal by applying two types of evaluation measures including Pearson correlation between the predicted and actual values and the normalized root mean squared error (NRMSE). Figure 1 shows the structure of DRFLLS.

Figures 2 and 3 illustrate the pseudo-code that is used to handle missing values and the procedure that is applied to build the tree in the training phase, respectively.

5 Results and performance evaluation

5.1 Estimating the number of nearest neighbors

One of the remaining problems in the various missing value methods is how to select the optimal local nearest base of the missing values (the number of nearest neighbors). Figure 4 shows the outline of the DRF pseudo-code that used to find the optimal nearest neighbors' numbers.

In this paper, different types of correlation functions or similarity functions among the trees to estimate the optimal number of nearest neighbors are used. For a given forest f, the similarity between target record and other record X_1 and X_j pairs is computed in the following way. For each of the two pairs, we first propagate their values down all trees within f. Next, the terminal node position for each pair in each of the trees is recorded. Let $Z_1 = (Z_{11}, Z_{1L})$ be these tree node positions for X_1 and similarly define Z_j. The similarity between X_1 and X_j pairs takes different forms that apply in parallel as follows:

- Random forest with Simple Similarity (RFSS)

  This measure is computed using Eq. (1), where I presents the indicator function as explained in Eq. (1)
- Random forest with Pearson correlation (RFPC)

  For this measure, Eq. (2) is used, where $\bar{Z}_j$ represents the average of values in $\boldsymbol{X}_{jj}$ and σ_j is the standard deviation. The components of $\boldsymbol{X}_1$ that correspond to missing values are not included in computing the coefficients, as explained in Eq. (2).
- Random forest with fuzzy similarity measure 1 (RFFSM1)

  This measure is based on fuzzy Minkowski distance. A smaller distance between X_1 and X_j is observed, and there is a greater similarity between them as explained in Eq. (3).

  RFFSM1 is computed using Eq. (3), where $r \to \infty$.
- Random forest with fuzzy similarity measure 2 (RFFSM2)

  It is computed using Eq. (4), where $i = 2, \ldots, L$.
- Random forest with fuzzy similarity measure 3 (RFFSM3)

  This measure needs to define the notion of cardinality of the fuzzy set. The cardinality is given by the number of elements in that set. This concept can be extended to fuzzy sets using the sigma count, which can be defined as Zhou et al. (2015), and Eq. (5) is used to compute this measure.
- Random forest with fuzzy similarity measure 4 (RFFSM4)

  This measure is computed using Eq. (6).
- Random forest with fuzzy similarity measure 5 (RFFSM5)

  This measure is calculated using Eq. (7).

As a result, we get seven forests, each one of which is built based on one of the above correlation equations and provides one predicate value to the number of neighbors. Table 2 explains DRF results for several datasets used in this paper.

The optimal numbers of nearest neighbors generated by DRF for a given dataset are shown in bold. The table above

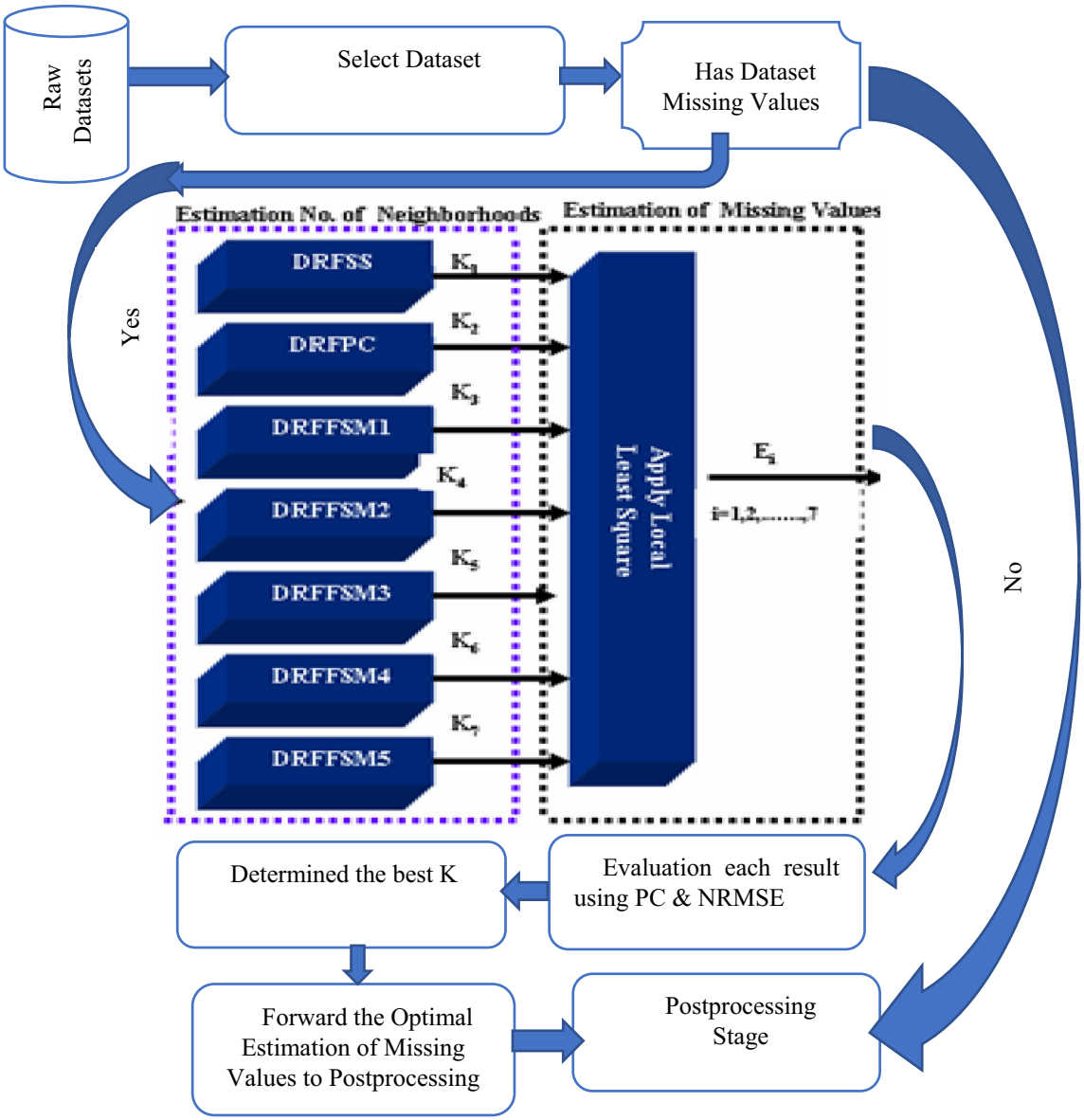


Fig. 1 Novel tool DRFLLS

Fig. 2 Missing value handling

Algorithm1: DRFLLS for Handle Missing Values

Input: *Dataset has a missingvalue*

Output: *Dataset without missing values*

1: *Set parameters; number of bootstrap samples, Max no of trees, Max, No. of level, Min no of node, no of terminal node, no of epoch*

2: *Call Build Tree Procedure*

3: *Estimation Number of Nearest Neighbors using (DRFSS, DRFPC, RFFSM1, DRFFSM2, DRFFSM3, DRFFSM4, DRFFSM5)*

4: *Estimation Missing Values using LLS*

5: *Validation of Results base on Pearson Correlation and NRMSE*

6: *End DRFLLS Algorithm.*

explains why the datasets have a low rate of missing values given that the best estimate of the number of nearest neighbors is given by DRFPC, followed by DRFFSM1 with $r = 4$. When the datasets have a high rate of missing values, the best estimate of the number of nearest neighbors is given by DRFFSM5, followed by DRFFSM3.

Fig. 3 Building tree in the training phase

Algorithm 2: Build Tree (Grow an un-pruned tree on training records)

```
While number of records in training set< >Null
    Read a new record
    Pass it down the tree
    If it reaches a terminal node
      If first record at this node
        Randomly choose n attributes
        Find intervals for each of the n attributes update counters.
        If node has seen n_m in records
          If similarity measure test is satisfied
            Save node split attribute
            Save corresponding split value
            If no more records in the training set
              If node records are highest similarity
                Take average response from all individually trained trees
                Assign the average response to number of nearest neighbors
              Else
                Save best split attribute seen so far
                Save corresponding split value
End while
```

5.2 Missing value estimation

We get seven nearest neighbors (*k1*, *k2*, *k3*, *k4*, *k5*, *k6*, *k7*) by the DRF method. Let *r1* be a record containing *n* features and *q* missing values. We deal with the case in which there is more than one missing value in a record vector by the local least squares method. In this case, recovering the total number of *q* in any location is as follows:

(A) The nearest neighbor record vectors for *r1* are first found. In this process of finding similar records, the *q* components of each record having the same location of the missing values in *r1* are ignored.

(B) Build a matrix, $\boldsymbol{A} \in R^{k*(n-q)}$, where *A* is a two-dimensional matrix in which the number of rows equals the number of nearest neighbors k_i {i = 1, …, 7}. The number of columns equals the number of total features n minus the number of columns containing missing values *q*.

(C) Build a matrix $\boldsymbol{B} \in R^{k*q}$, where *B* is a two-dimensional matrix in which the number of rows equals the number of nearest neighbors k_i {i = 1, …, 7}. The number of columns equals the number of columns containing missing values *q*.

(D) Build a vector $\boldsymbol{w} \in R(n-q)^{*1}$, where *w* is a one-dimensional matrix in which the number of columns equals the number of total features n minus the number of columns containing missing values *q*.

(E) After *A* and *B* matrices and a vector *w* are formed, the least squares problem is formulated as:

$$\min_{x} \left\| A^{\mathrm{T}} X - W \right\|_2 \tag{8}$$

(F) The vector $u = (\alpha_1, \alpha_2, \ldots, \alpha_q)^{\mathrm{T}}$ of *q* missing values can be calculated as

$$\mathbf{u} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{pmatrix} = B^{\mathrm{T}}\mathrm{X} = B^{\mathrm{T}}(A^{\mathrm{T}})^{t}\mathbf{w} \tag{9}$$

where $(\boldsymbol{A}^{\mathrm{T}})^{t}$ presents pseudo-inverse of $\boldsymbol{A}^{\mathrm{T}}$ and the pseudo-inverse $\boldsymbol{A}^{t}$ of $\boldsymbol{A}$ can be computed by:

$$\begin{aligned} A^{t} &= [V_1 \;\; V_2] \begin{bmatrix} \sum_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} [U_1 \;\; U_2]^{\mathrm{T}} \\ &= V_1 \sum_1^{-1} U_1{}^{\mathrm{T}} \end{aligned} \tag{10}$$

Fig. 4 DRF used to find the optimal nearest neighbors' numbers

Algorithm 3: DRF to predict the optimal nearest neighbours' numbers

Input: *Dataset have N Records.*

Output: *Predicted Optimal Number of Nearest Neighbors.*

1: *Divided Dataset to n tree samples, one for each tree.*

2: *For each tree t, do following steps:*

3: *For each node n in t, do the following:*

4: *Choose random mtry set of variables X.*

5: *Choose best split variable x to split node n.*

6: *If splitting condition is True, then split n.*

7: *Else set n as a terminal node.*

8: *make pruning for tree t.*

9: *After Radom Forest of n -tree trees is completed then predicate value by find Similarity functions among the trees to estimate the optimal number of nearest neighbors as follow of n-tree predication.*

10: *Random Forests with Simple Similarity as explained in eq*

11:
$$S(X_1, X_j) = \frac{1}{(L-1)} \sum_{i=2}^{L} I(Z_{1i} == Z_{ji}) \quad (1)$$

12: *Random Forests with Pearson Correlation as explained in eq 2*

13:
$$S(X_1, X_j) = \frac{1}{(L-1)} \sum_{i=2}^{L} \left(\frac{Z_{1i} - \bar{Z}_1}{\sigma_1} \right) \left(\frac{Z_{ji} - \bar{Z}_j}{\sigma_j} \right) \quad (2)$$

14: *Random Forests with Fuzzy Similarity Measure 1 as explained in eq 3*

15:
$$S(X_1, X_j) = 1 - \left(\frac{1}{(L_1 * L_j)} \left| \sum_{i=2}^{L} Z_{1i} - Z_{ji} \right|^r \right)^{\frac{1}{r}} \quad (3)$$

16: *Random Forests with Fuzzy Similarity Measure 2 as explained in eq 4*

17:
$$S(X_1, X_j) = 1 - Max \left| Z_{1i} - Z_{ji} \right| \quad (4)$$

18: *Random Forests with Fuzzy Similarity Measure 3 as explained in eq 5*

19:
$$|X_1| = \sum_{i=2}^{L} Z_{1i}$$

20:
$$S(X_1, X_j) = \frac{\left| Z_{1i} \cap Z_{ji} \right|}{Z_{1i} \cap Z_{ji}}$$

21:
$$= \frac{\sum_{i=2}^{L} Min \quad (Z_{1i}, Z_{ji})}{\sum_{i=2}^{L} Max \quad (Z_{1i}, Z_{ji})} \quad (5)$$

22: *Random Forests with Fuzzy Similarity Measure 4 as explained in eq 6*

23:
$$S(X_1, X_j) = \frac{\left| Z_{1i} \cap Z_{ji} \right|}{Max \left(|Z_{1i}|, |Z_{ji}| \right)}$$

24:
$$= \frac{\sum_{i=2}^{L} Min \quad (Z_{1i}, Z_{ji})}{Max \left(\sum_{i=2}^{L} Z_{1i}, \sum_{i=2}^{L} Z_{ji} \right)} \quad (6)$$

25: *Random Forests with Fuzzy Similarity Measure 5 as explained in eq 7*

26:
$$S(X_1, X_j) = \frac{1}{(L_1 * L_j)} \sum_{i=2}^{L} \left[\frac{Min \ (Z_{1i}, Z_{ji})}{Max \ (Z_{1i}, Z_{ji})} \right] \quad (7)$$

27: *End.*

Table 2 Number of nearest neighbor estimation by the developed random forests method

Correlation function	Communities and crime	DNA	P53 mutants	URL reputation	SPLICE	GIS
DRFSS	17	30	22	31	14	13
DRFPC	21	16	**24**	14	**18**	**11**
DRFFSM1	23	24	20	13	15	7
DRFFSM2	16	18	21	15	12	21
DRFFSM3	19	23	13	28	10	18
DRFFSM4	23	25	5	22	26	19
DRFFSM5	**24**	**28**	11	**35**	22	13

where $V_1 \in R^{n-1*\text{rank}}, \sum_1^{-1} \in R^{\text{rank}*\text{rank}}, U_1^{\text{T}} \in R^{\text{K}*\text{rank}}$.

The known elements of $\boldsymbol{w}$ can be presented by

$$\mathbf{w} \simeq x_1 a_1 + x_2 a_2 + \cdots + x_k a_{ki} \tag{11}$$

where the x_i refers to the coefficients of the linear combination found from the least squares formulation given by Eq. (9).

(G) As a result, the multiple regressions represent a target record, i.e., the record has multiple missing value features, as a linear combination of its nearest neighbors as:

$$\text{Target} = x_1 b_1 + x_2 b_2 + \cdots + x_k b_k \tag{12}$$

where b_k is the kth nearest neighbor and x_k represents the regression coefficient corresponding to that neighbor.

5.3 Performance evaluation

To assess the performance and accuracy of the DRFLLS method, two measures were used. Pearson correlation and NRMSE served to evaluate the differences between predicted and actual values

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left[\left(ij_{\text{answer}} - ij_{\text{guess}}\right)^2\right]}{\text{variance}[ij_{\text{answer}}]}} \tag{13}$$

where the mean and variance are calculated over the missing elements in the whole matrix, the set of known values are ij_{answer}, while ij_{guess} are the set of predicted values.

In each column of Table 3, the smallest value of NRMSE for a given dataset is presented in the bold font. This means that corresponding DRF correlation function is better function for a given dataset. In addition, the error generated of each database based on the number of nearest neighbors is yielded by different types of DRF correlation measures as shown in Table 4.

In each column of the Table 4, the highest value of the correlation for a given dataset is presented in bold font. This means that the corresponding DRF correlation function is the best function for a given dataset.

The results from the correlation measures (DRFSS, DRFPC, DRFFSM1, DRFFSM2, DRFFSM3, DRFFSM4 and DRFFSM5) and the imputation accuracy measures on the six datasets are provided. Higher correlation and a lower NRMSE scores will result in more accurate imputations. The results from the NRMSE method are provided in Table 3, while those from the Pearson correlation are provided in Table 4. As it can be seen in each column in Table 3, smaller values of NRMSE for a given dataset are shown in bold, while in Table 4, the correlation for a given dataset has the highest values. From these experiment results, we can find that the corresponding DRF correlation function is a better function for the datasets used in this study.

5.4 Relative importance of each similarity measure

To further illustrate the importance of each similarity measures in the estimation process, the generated error by the DRF is provided

$$\text{Generation Error} \leq \frac{S_i^{-}(X1, X_j)[\text{ST}^2]}{\text{ST}^2!} \tag{14}$$

where S_i^{-} is the average correlation among trees. ST^2 is the 'strength' of the tree classifiers, i.e., the average performance of the classifiers (Waske et al. 2010).

The error generation for each dataset based on the number of nearest neighbors yielded by different types of DRF similarity measures is provided. In Table 5 and Fig. 5, we can find that if the dataset has a small number of missing values, then the best estimation of the number of nearest neighbors is obtained from DRFPC, followed by DRFFSM1 with $r = 4$ in GIS dataset since it has both the smallest missing values and error generation. If the dataset has a large number of missing values, then the best estimation number of nearest neighbors is obtained from DRFFSM5, followed by DRFFSM3 as in the DNA dataset which includes the highest missing values.

Table 3 Accuracy, as measured by NRMSE for all datasets

Correlation function	Communities and crime	DNA	P53 mutants	URL reputation	SPLICE	GIS
DRFSS	0.65	0.72	0.74	0.60	0.76	0.92
DRFPC	0.71	0.76	**0.71**	0.65	**0.60**	0.67
DRFFSM1	0.87	0.80	0.72	0.68	0.63	**0.54**
DRFFSM2	0.75	0.74	0.83	0.60	0.73	0.77
DRFFSM3	0.64	0.67	0.85	0.59	0.82	0.83
DRFFSM4	0.85	0.69	0.81	0.71	0.78	0.95
DRFFSM5	**0.61**	**0.60**	0.88	**0.58**	0.65	0.82

Table 4 Accuracy, as measured by correlation for all datasets

Correlation function	Communities and crime	DNA	P53 mutants	URL reputation	SPLICE	GIS
DRFSS	0.74	0.81	0.79	0.59	0.63	0.67
DRFPC	0.76	0.77	**0.86**	0.78	**0.73**	**0.78**
DRFFSM1	0.66	0.68	0.81	0.68	0.68	0.72
DRFFSM2	0.71	0.65	0.33	0.48	0.51	0.49
DRFFSM3	0.80	0.83	0.45	0.79	0.49	0.39
DRFFSM4	0.78	0.74	0.60	0.76	0.63	0.32
DRFFSM5	**0.85**	**0.88**	0.71	**0.91**	0.44	0.56

Table 5 Error generation for each dataset based on the number of nearest neighbors yielded by different types of DRF correlation measures

DRF correlation measures	Communities and crime		DNA		P53 mutants		SPLICE		GIS	
	# Nearest neighbors	Error generation	# Nearest neighbors	Error generation	# Nearest neighbors	Error generation	# Nearest neighbors	Error generation	# Nearest neighbors	Error generation
DRFSS	17	0.892	30	0.063	22	0.587	14	0.174	13	0.179
DRFPC	21	0.59	16	0.073	24	0.479	18	0.159	11	**0.092**
DRFFSM1	32	0.589	24	0.065	20	0.487	15	0.169	7	**0.104**
DRFFSM2	16	0.58	18	0.071	21	0.586	12	0.194	21	0.172
DRFFSM3	19	0.489	23	**0.064**	13	0.588	10	0.222	18	0.174
DRFFSM4	23	0.575	25	0.065	5	0.586	26	0.132	19	0.176
DRFFSM5	24	0.078	28	**0.058**	11	0.589	22	0.195	13	0.179

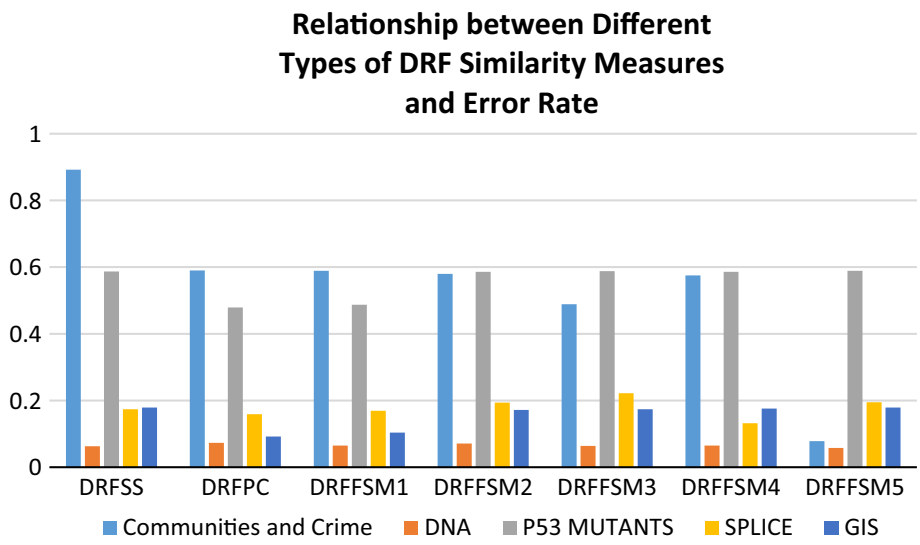


Fig. 5 Relationship between different types of DRF similarity measures and error rate

Table 6 Error generation by different numbers of trees in DRF

# Trees	Communities and crime	DNA	P53 mutants	SPLICE	GIS
10	0.981	0.074	0.980	0.222	0.395
20	0.973	0.062	0.887	0.174	0.388
30	0.952	0.063	0.884	0.115	**0.388**
40	0.911	0.082	0.858	0.102	0.376
50	0.850	0.081	0.782	0.091	0.367
60	0.490	0.086	0.762	**0.082**	0.358
70	0.215	0.068	0.715	0.076	0.357
80	**0.112**	0.072	0.483	0.074	0.366
90	0.110	0.068	0.398	0.070	0.354
100	0.108	**0.064**	0.377	0.066	0.354
110	0.108	0.073	0.279	0.065	0.350
120	0.105	0.075	**0.279**	0.062	0.359

In this study, we have used between 10 and 120 different DRF trees, based on the trial-and-error principle to decide the optimal number. Table 6 shows the error generation by different numbers of trees used in DRF.

In Table 6, the values of trees in the range of [10, 120] were used; at each stage, 10 trees were added. By another word, each dataset was tested 12 times starting from 10 to 120. In each test, the cumulative error was calculated according to Eq. (14). As the number of decision trees is increased, the required time for calculating missing values by the model will be increased, while the error rate will be decreased. The main aim is to reach a state of stability between the number of trees used in constructing the model and the percentage of generated errors. In Table 6, it can be seen that in communities and crime dataset when the number of trees reached 80, the cumulative error was 0.112, when the number of trees increased to 90, the error was 0.110; and the error became 0.108 when the number of trees reached 100. This indicates that the error is decreased by 0.002. For this, the number 80 was selected because it represents the state of stability and the changes in the error rate after this number were relatively small compared to the used time. The optimal number of trees for communities and crime, DNA, P53 MUTANTS, SPLICE and GIS datasets are 80, 100, 120, 60 and 25, respectively, as shown in Fig. 6. The value of each parameter used in DRF for each dataset and the relative importance that is given for each similarity measure are summarized in Table 7.

As the results showed in above figures and tables, the proposed DRFLLS method provides a great improvement in both accuracy and stability over the different types of dataset used in this paper for missing value estimation.

5.5 Comparing DRFLLS with other proposed methods

The result obtained with the DRFLLS for the datasets used in this study was compared among the proposed methods reported by other researchers. The compression of DRFLLS with other proposed methods with respect to the employed tools, datasets, structure and method of determine nearest neighbor is illustrated in Table 8.

5.6 Assumptions and limitations

The main assumption of this paper is one of processing original datasets that suffer from many records that have missing values in different locations of the record but not processing the missing values that may occur in the post-processing stages, i.e., the clustering, association rules and decision stages. RF is considered as one of the statistical tools that perform well in many fields. However, our experiments found that the combination of RF and similarity measures to design DRF methods leads to an increase in the time complexity. Because it requires performing many of mathematical operations as explained in the above 2–8 equations; also when the number of trees in the RF is increased, the required time to handle missing values in the training data will be increased. While the main advantages of DRFLLS are ability to find the missing values for small and huge dataset, using different measures to determine the number of nearest neighbors, these measures including simple similarity, Pearson similarity coefficient and fuzzy similarity (M1, M2, M3, M4 and M5) make the tool suitable to deal with the dataset that differs in their natural, amount of missing values, and the type features that contain missing values. Combination between DRF and LLS leads to the optimal estimation of missing values by LLS.

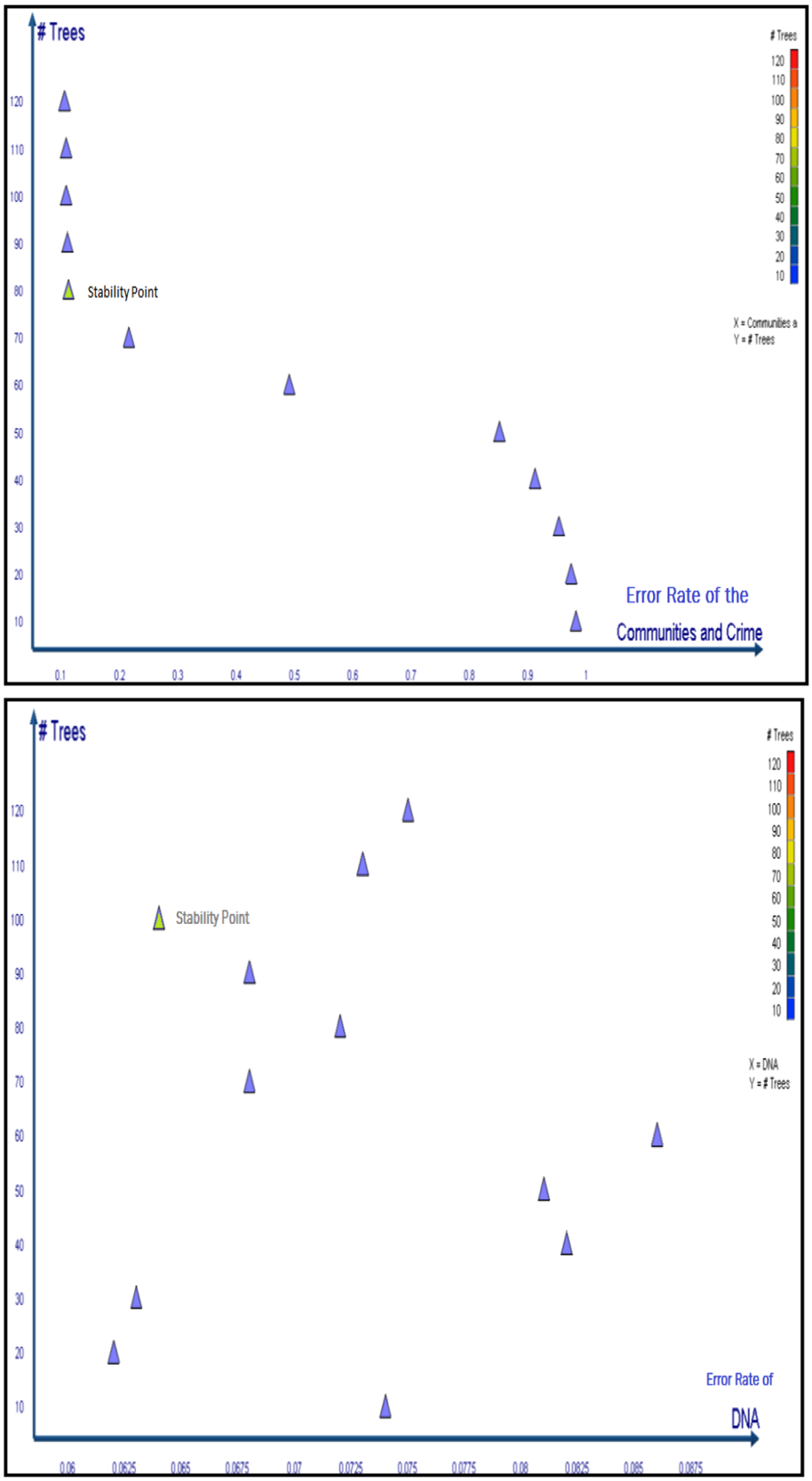
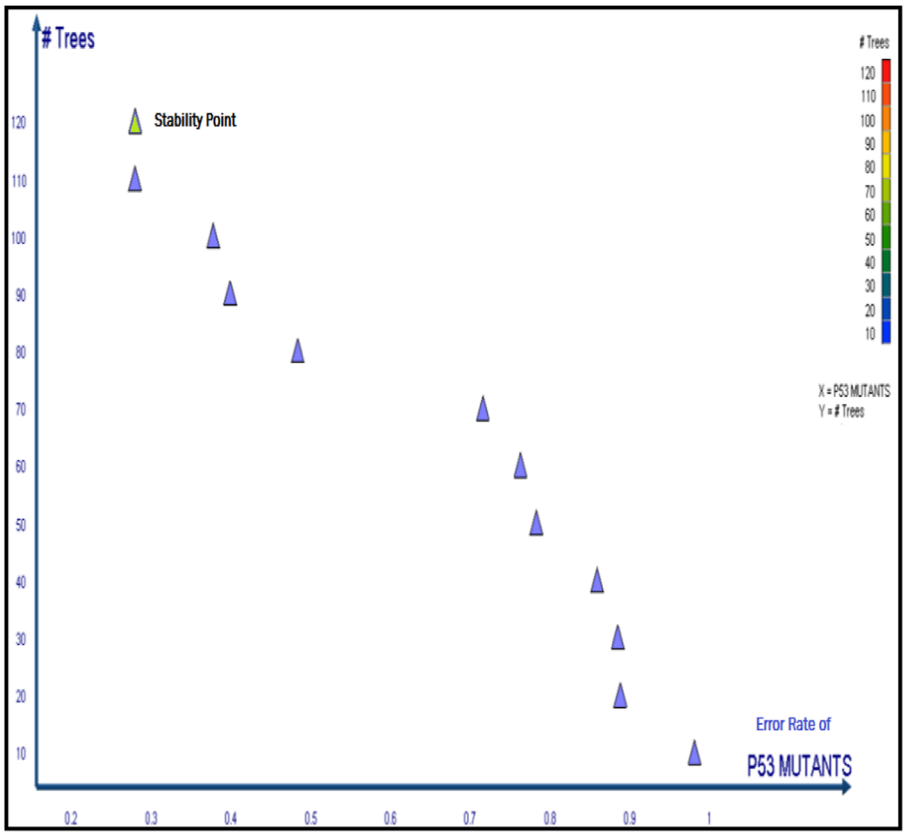


Fig. 6 Relationship between different numbers of trees used in DRF and error rate for each dataset

Fig. 6 continued



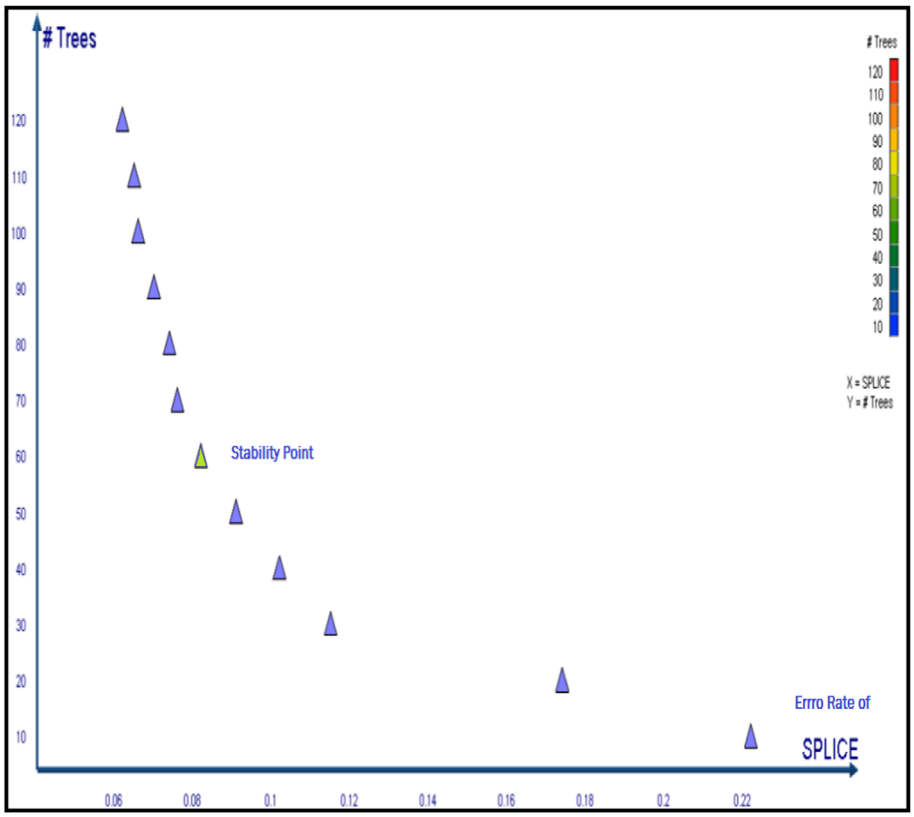


Fig. 6 continued

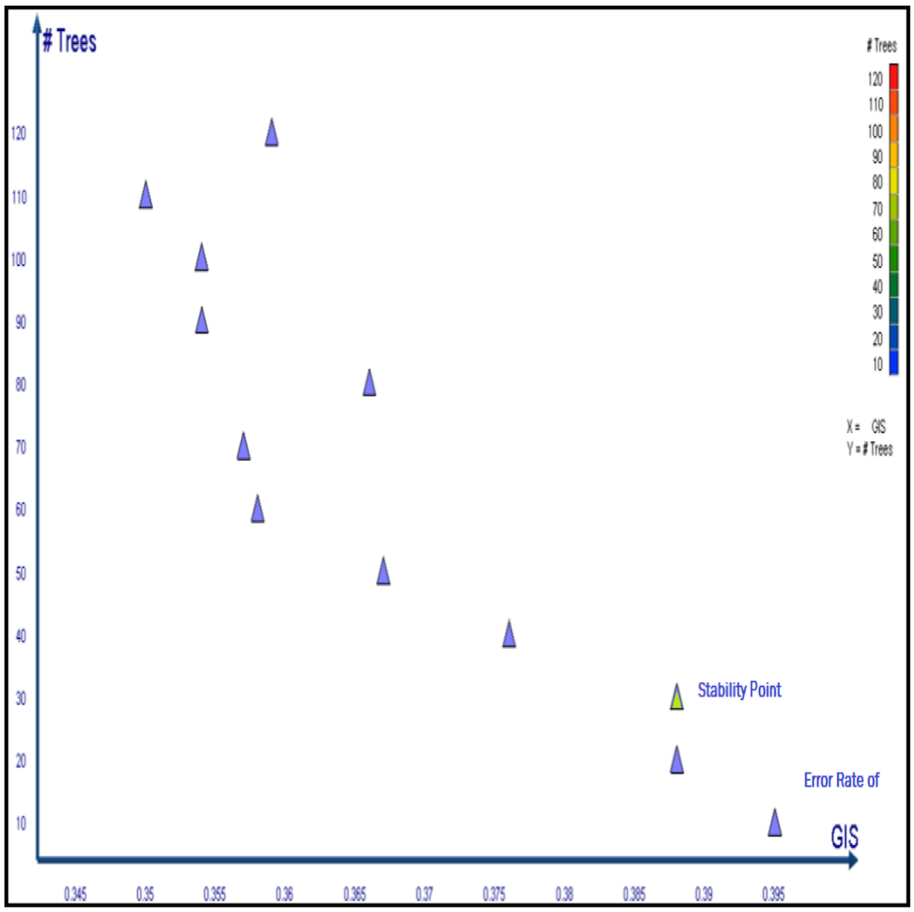


Table 7 Relative importance of each similarity measure

Database name	Total # trees	Max level of trees	Max # nodes in tree	RFSS[a]	RFPC[a]	RFFSM1[a]	RFFSM2[a]	RFFSM3[a]	RFFSM4[a]	RFFSM5[a]
Communities and crime	80	128	98,048	1.13	1.07	0.75	0.94	1.25	0.91	**1.39**
DNA	100	181	362,000	1.12	1.01	0.85	0.87	1.23	1.07	**1.46**
P53 mutants	120	255	226,695	1.06	**1.21**	1.12	0.39	0.52	0.74	0.88
SPLICE	60	61	194,590	0.82	**1.21**	1.07	0.69	0.59	0.80	0.67
GIS	25	9	9009	0.75	**1.36**	1.33	0.63	0.46	0.33	0.68

Max level of trees = No. of features of that dataset

Max number of nodes in tree = No. of records × No. of features in each dataset

*The relative importance of each similarity measure

6 Conclusion and future directions

In this study, a new DRFLLS approach for missing values and problem of estimating the optimal number of nearest neighbors is proposed. To attain this goal, RF was developed through replacement of the correlation function of the original RF with seven different types of similarity measures. These measures include simple similarity, Pearson similarity coefficient and fuzzy similarity (M1, M2, M3, M4 and M5). We obtained the optimal estimation of missing values by LLS that depends on the values of nearest neighbors generated by developed random forest (DRF). We investigated the feasibility of the new method using six datasets obtained from different disciplines. The DRFLLS method is evaluated using two imputation accuracy measures: PC and NRMSE where the value that yields

Table 8 Comparison among the previous studies and the current study for handling missing values

Researcher	Tools	Dataset	Structure	Method of determining nearest neighbors
Golub et al. (2005)	LLS	Gen expression	Local similarity	Randomization method
Qi et al. (2005)	Random forest	Pairs of proteins	Modify KNN	Randomization method
Al-Janabi (2017)	Four stages	Multivariate	Substations structure	Propositional method
Bruggeman et al. (2009)	PhyloPars	Large dataset (web dataset)	State-of-the art evolutionary	Combines an incomplete set
Rieger et al. (2010)	KNN	Multivariate	Correlation structure	Randomization method
McCandless et al. (2011)	Statistical post-processing	Temperature dataset	Replacement structure	Not handle
Ryan et al. (2010)	Three local (nearest neighbor based) and one global (BPCA based)	E-MAPs	Local similarity	Not handle
This study	DRFLLS	Multivariate	Local and fuzzy similarity	Development of random forest

the highest PC and the least NRMSE represents the optimal value. The DRFLLS considerably showed a high performance and accuracy with regard to missing value problem. The experimental results show that an improvement in estimating missing values can be achieved by the proposed DRFLLS tool in comparison with other proposed methods.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abualigah LMQ, Hanandeh ES (2015) Applying genetic algorithms to information retrieval using vector space model. Int J Comput Sci Eng Appl 5(1):19

Abualigah LM, Khader AT (2017) Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J Supercomput 73(11):4773–4795

Abualigah LM, Khader AT, Al-Betar MA, Alomari OA (2017) Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst Appl 84:24–36

Adam E, Mutanga O, Odindi J, Abdel-Rahman EM (2014) Land-use/cover classification in a heterogeneous coastal landscape using Rapid Eye imagery: evaluating the performance of random forest and support vector machines classifiers. Int J Rem Sens 35(10):3440–3458

Ali SH (2012a) Miner for OACCR: case of medical data analysis in knowledge discovery. In: IEEE, 2012 6th international conference on sciences of electronics, technologies of information and telecommunications (SETIT), Sousse pp 962–975. https://doi.org/10.1109/setit.2012.6482043

Ali SH (2012b) A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining. In: IEEE, 2012 6th international conference on sciences of electronics, technologies of information and telecommunications (SETIT), Sousse, pp 951–961. https://doi.org/10.1109/setit.2012.6482042

Ali SH (2013) Novel approach for generating the key of stream cipher system using random forest data mining algorithm. In: IEEE, 2013 sixth international conference on developments in e-systems engineering, Abu Dhabi, pp 259–269 (2013). https://doi.org/10.1109/dese.2013.54

Al-Janabi S (2017) Pragmatic miner to risk analysis for intrusion detection (PMRA-ID). In: Mohamed A, Berry M, Yap B (eds) Soft computing in data science. SCDS 2017. Communications in Computer and Information Science, vol 788. Springer, Singapore. https://doi.org/10.1007/978-981-10-7242-0_23

Al-Janabi S (2018) Smart system to create optimal higher education environment using IDA and IOTs. Int J Comput Appl. https://doi.org/10.1080/1206212X.2018.1512460

Aljarah I, Mafarja M, Heidari AA, Hossam F, Yong Z, Mirjalili S (2018) Asynchronous accelerating multi-leader Salp chains for feature selection. Appl Soft Comput 71:964–979. https://doi.org/10.1016/j.asoc.2018.07.040

Bose S, Das C, Chakraborty A, Chattopadhyay S (2013) Effectiveness of different partition based clustering algorithms for estimation of missing values in microarray gene expression data. In: Advances in computing and information technology. Springer, Berlin, pp 37–47

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bruggeman J, Heringa J, Brandt B (2009) PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Res 37(2):W179–W184

Carranza EJM, Laborte AG (2015) Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Comput Geosci 74:60–70

Center for Machine Learning and Intelligent Systems, USA (2010a) http://archive.ics.uci.edu/ml/datasets/p53+Mutants

Center for Machine Learning and Intelligent Systems, USA (2010b). https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis

Chiu CC, Chan SY, Wang CC, Wu WS (2013) Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Syst Biol 7(6):S12

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792

Elyan E, Gaber MM (2016) A fine-grained Random Forests using class decomposition: an application to medical diagnosis. Neural Comput Appl 27(8):2279–2288

Genbank 64.1 (1992) http://archive.ics.uci.edu/ml/machine-learning/datasets/DNA/

Genbank 64.1 (2018). http://idke.ruc.edu.cn/news/2008/dataset.htm

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31(14):2225–2236

Golub GH, Kim H, Park H (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2):187–198

Graham JW (2012) Missing data: analysis and design. Springer, New York

Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. University of Illinois at Urbana-Champaign. San Francisco. Elsevier 2006. www.books.elsevier.com

Hapfelmeier A, Hothorn T, Ulm K (2012) Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Comput Stat Data Anal 56(6):1552–1565

Heidari AA, Faris H, Aljarah I, Mirjalili S (2018) An efficient hybrid multilayer perceptron neural network with grasshopper optimization. Soft Comput. https://doi.org/10.1007/s00500-018-3424-2

Heidari AA, Aljarah I, Faris H, Chen H, Luo J, Mirjalili S (2019) An enhanced associative learning-based exploratory whale optimizer for global optimization. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04015-0

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Kumar V, Wu X, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37

Liew AWC, Law NF, Yan H (2010) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform 12(5):498–513

Mafarja M, Aljarah I, Heidari AA, Faris H, Fournier-Viger P, Li X, Mirjalili S (2018) Binary dragonfly optimization for feature selection using time-varying transfer functions. Knowl Based Syst 161:185–204. https://doi.org/10.1016/j.knosys.2018.08.003

McCandless T, Haupt SE, Young G (2011) Replacing missing data for ensemble systems. J Comput 6(2):162–171

Moorthy K, Saberi Mohamad M, Deris S (2014) A review on missing value imputation algorithms for microarray gene expression data. Curr Bioinform 9(1):18–22

Pantanowitz A, Marwala T (2009) Missing data imputation through the use of the random forest algorithm. In: Yu W, Sanchez EN (eds) Advances in computational intelligence. Advances in Intelligent and Soft Computing, vol 116, Springer, Berlin, pp 53–62

Qi Y, Klein-Seetharaman J, Bar Z (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. Pac Symp Biocomp 10:531–542

Redmond M (2009) Center for machine learning and intelligent systems. Computer Science, La Salle University, Philadelphia, PA

Rieger A, Hothorn T, Strobl C (2010) Random forests with missing values in the covariates. Technical Report Number 79, Department of Statistics, Ludwig-Maximilians-Universität, Munich

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (1996) Multiple imputation after 18 + years. J Am Stat Assoc 91(434):473–489

Ryan C, Green D, Cagney G, Cunningham P (2010) Missing value imputation for epistatic MAPs. Bioinformatics 11:197

Saul LK, Savage S, Ma J, Voelker GM (2009) Identifying suspicious URLs: an application of large-scale online learning. In: 26th annual international conference on machine learning (ICML), Montreal (2009) pp 681–688

Stekhoven DJ, Bühlmann P (2012) MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. Pattern Recognit 44(2):330–349

Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, Higgins PD (2013) Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 3(8):e002847

Wasito I, Mirkin B (2006) Nearest neighbours in least-squares data imputation algorithms with different missing patterns. Comput Stat Data Anal 50(4):926–949

Waske B, Chi M, Benediktsson JA, van der Linden S, Koetz B (2010) Algorithms and applications for land cover classification—a review. In: Li D, Shan J, Gong J (eds) Geospatial technology for earth observation. Springer, Boston, MA, pp 203–233

Xie Y, Li X, Ngai EWT, Ying W (2009) Customer churn prediction using improved balanced random forests. Expert Syst Appl 36(3):5445–5449

Zhou Z, Zhang R, Lin Y, Wang R (2015) A comparison of similarity measures of intuitionistic fuzzy sets. In: LISS 2014, pp 1237–1242