International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 31 No. 4 Jan. 2021, pp. 763-775
© 2021 Innovative Space of Scientific Research Journals
http://www.ijias.issr-journals.org/

New Hybrid Method for Efficient Imputation of Discrete Missing Attributes

Kone Dramane, Goore Bi Tra, and Kimou Kouadio Prosper

UMRI 78- Electronics and Electricity, LARIT: Computer and Telecommunications Research Laboratory, EDP/INP-HB, Yamoussoukro, Cote d'Ivoire

ABSTRACT: In this paper, we present a hybrid method for efficiently estimating missing discrete attributes appearing in data manipulation or processing. The principle of the method consists first of all in determining the segment to which the missing value belongs and then estimating it by majority vote when possible. Otherwise, the average of the missing attribute is determined from the complete data of the segment. Several cases may arise. The case where the non-missing attributes have the same modality (they are in the same interval) is dealt with by calculating the centre of the missing attribute. ***M*** of the class and the average ***m*** attributes that are not missing. If ***m*** is less than ***M*** then the value ***e*** of the missing attribute is estimated by the value of the non-missing attribute within the interval [***a***, ***M*** [where ***a*** is the lower bound of the modality. Otherwise, the value of the other non-missing attribute is used for estimation. The second case, where the non-missing attributes have different modalities, is treated by calculating the average ***m*** attributes that are not missing and then estimate the missing value. ***e*** by the not-missing attribute having the same modality as ***m***. Finally, an error test based on RMSE demonstrates the effectiveness of our method.

KEYWORDS: Cleaning, Estimation, Segmentation, Classification, MAR, Data Mining.

1 INTRODUCTION

The data preparation phase is a crucial step in knowledge discovery [1]. Data cleaning is one of its most important tasks. It consists of dealing with imperfections such as inconsistencies, outliers or missing data. These imperfections are due to machine errors and human carelessness or forgetfulness. These imperfections, especially missing data, prevent the direct use of data mining algorithms for the discovery of prediction models in the collected data [2], [3]. Their presence in databases also reduces the performance of data mining algorithms [4].

The processing of missing data has a twofold interest. The first interest is to improve the quality of the data. The second is to improve the performance of the data mining algorithms, thus favouring good prediction results. To this end, the authors [5] propose the CCA (Complete Case Analysis) and LOCF (Last Observation Carried Forward) methods. Their methods are the main methods for processing missing data in nutritional trials. However, they introduce biases in the results. The authors [6] propose the DMI (Decision tree based Missing value Imputation) method to improve the quality of road accident data. They develop their method using the decision tree and the EM (Expectation Maximization) algorithm. The DMI imputes missing numerical values using the EM algorithm and missing qualitative values using the values of the majority class in the sheets.

In addition, the authors [7] propose the DSMI (Decision tree and Sampling based Missing value Imputation) based on the sampling of distributions obtained from correlation measures to impute missing values. The method also uses the decision tree and correlation. Their method focuses on the treatment of qualitative missing values. Recently, the authors [8] develop the Correlation Maximization-based Imputation Methods (CMIM) based on correlation and regression. Their method first looks for highly correlated data segments. Then, they apply linear estimation models in these segments. Applying regression models to the highly correlated data segments instead of the data set thus improves the performance of both quantitative and qualitative missing data imputation methods. However, the CMIM ignores discrete quantitative missing values and does not correctly identify highly correlated segments. The majority of recent methods do not consider the correlation between quantitative attributes and do not clearly define majority voting. If they do, they are more suitable for qualitative data. Furthermore, they do not work properly for records with at most one missing attribute [7]. Thus, imputation methods can be improved. Our aim is to propose an efficient method for imputation of quantitative missing data. The authors [6] show that the use of subsets of data in the imputation process improves the performance of the imputation methods.

In this study, we propose a new algorithm to solve some of the problems of recent methods. Our method addresses the problem of ignoring discrete missing values of records with more than one missing attribute, in addition to the qualitative values treated by the authors [7]. The method also addresses the problem of segmentation. Our method is hybrid. It combines the decision tree and an estimation based on double segmentation. The minimum distance (distance between the mean and the non-missing values of the same modality) is used as the estimation value in the horizontal segments. Our method first performs a vertical segmentation to group the attributes into classes or modalities. The objective of the vertical segmentation is to optimize the construction of decision trees and the estimation of discrete missing values. Then, it proceeds to a horizontal segmentation to obtain highly correlated data subsets. These subsets constitute the estimation segments. Finally, estimation is performed vertically in these segments attribute by attribute.

Our experiments show that the proposed algorithm has a better imputation accuracy compared to the average, KNNI, DSMI. This paper is presented as follows: Section 2 presents an analysis of related recent work on methods for processing missing data. Section 3 then describes our new model for processing missing data. Then in section 4, we present the estimation results of the model which we validate by RMSE and MAE error tests. Finally, in section 5, we analyse our results to draw the consequences.

2 LITERATURE REVIEW

There are several methods for imputing missing data in the literature. They are classified into two approaches: suppression and imputation [5]. Deletion involves eliminating all records with missing attributes from the database. This way of dealing with missing data is called full case analysis [9, pp. 35-38], [10]. It therefore leads to loss of information and significantly reduces the sample size when the proportion of missing data is large [11] - [13]. The second method of suppression is the analysis of available cases [14] - [16]. It is an improvement on full case analysis to avoid loss of information and to preserve the original data structure. However, it ignores missing values. The advantages of both methods are their simplicity of implementation and use. In addition, they are implemented by default in learning machine and statistics applications.

The imputation approach is an improvement of the suppression methods. It consists of preserving the structure of the data and therefore the sample size. In addition to the preservation, it estimates the missing value unlike the suppression methods [17], [18]. It is the best approach for dealing with missing values [18], [19]. Several imputation methods are proposed in the literature. These include mean, KNNI (k-Nearest Neighbour Imputation) [20], [21], regression [22], [23], MI (Multiple Imputation) [24], SVMI (Support Vector Machine Imputation) [25]. Imputation by mean or mode. This is one of the most frequently used methods. It consists of replacing missing data for a given quantitative attribute with the average of the complete set of values. For the qualitative attribute the mode is used [21]. More robust methods are based on the relationships between attributes. There are also two conventional approaches based on relationships between attributes. These two approaches are the global approach [26], [27] and the local approach [7], [8], [28].

The author [27] develops the Expectation Maximization Imputation (EMI) method which deals with quantitative missing values. Missing values are imputed on the basis of the matrix of mean and covariance. The EMI method begins with an initial estimation of the mean and covariance matrix. It then proceeds through one iteration until the imputed values and the matrix are approximately equal from the previous iteration. For each record, missing values are estimated based on the relationship between the attributes. EMI outperforms conventional EM methods in datasets with more records than attributes. The EMI method only works for quantitative missing values and applies to a random data set. In addition, it applies to datasets with high correlations between attributes.

The EMI imputes only quantitative missing values. The basic idea of the FIMUS (Framework for Imputing Missing values Using co-appearance, correlation and similarity analysis) method [26] is to impute qualitative missing values. The authors use co-appearance, correlation and similarity of values of an attribute. These three parameters are used to impute qualitative missing values. FIMUS uses similarity and co-appearance at the same time. Two levels of similarity are calculated using the co-appearance of attribute values of records in a dataset. The first level of similarity is calculated using the co-appearance of attribute values of records. The second level uses the direct neighbour method. FIMUS also takes into account all the records in a dataset for the imputation of missing values. However, its main problem is its mathematical complexity. For imputation, similarity values depend on co-occurrence values. Indeed, FIMUS multiplies the similarity value by the co-appearance value. If there is no co-appearance value, then the associated similarity value has no impact to impute the missing values. In addition, it assigns a massive calculation to the similarity graphs when the number of records in the dataset is large. Validation of the accuracy of the imputation is done using RMSE (Root Mean Square Error) and the concordance index.

The EMI and FIMUS methods use the full set of data. Their imputation accuracy is better in a dataset with higher correlations than in a dataset with lower correlations. The authors [29] show that Correlations between attributes within a horizontal partition of a dataset can be higher than correlations across the dataset. Their method called DMI is an extension of the EMI method and deals with data sets with low correlations. It uses the *C4.5* decision tree algorithm and the EMI estimation technique. First, DMI divides the data set into a number of horizontal segments obtained from the decision tree leaves. The correlations between the attributes of the records in a leaf are higher than the correlations of the attributes in the dataset. Thus, the DMI estimates missing values for records in a leaf using the EMI rather than for all records in the dataset [26], [27]. The DMI impute numeric missing values using the EMI. While the majority vote of the class value is used to estimate the qualitative missing value. The DMI method is significantly superior to EMI. However, it suffers from various problems. First, it does not work if all records have the same value for a numeric attribute. Second, it is useless when all numerical values are missing from a record. A more serious problem is that the authors do not handle the imputation of records with missing values that are found in more than one sheet. In addition, it suffers from a complexity of computation time. This problem is due to its technique of estimating EMI to a small data set.

The authors [29] maintain the advantages of the DMI method and propose the iDMI method. Their study focuses on one of the problems of the DMI method. They deal with the complexity of the computational time of the DMI method in a smaller data set. The authors show that iDMI requires less computing time than DMI and IBLLS (Iterative Bicluster-based Least Square) [30]. Since it uses a sheet, instead of the data set, for the calculation of the mean value.

The authors [7] propose a further refinement of the iDMI method on more real road accident data sets in order to achieve greater imputation accuracy. This method is called DSMI. It imputes both quantitative and qualitative attributes unlike EMI. Missing attributes are imputed by calculating the IS **(**Interest factor and Support count) correlation between the missing attributes and those observed in a record. For this purpose, two correlation measures are used, a direct one called 1st level similarity and a transitory one called 2nd level similarity or weighted similarity measure. To account for the uncertainty inherent in actual data, the DSMI imputes missing attributes prior to correlation by sampling from a list of potential imputed values based on the degree of affinity. Random sampling by affinity helps to reduce systematic bias in the imputed dataset. Experience shows that DSMI outperforms DMI, iDMI, KNNI, FIMUS methods on qualitative data sets. However, the method does not work for records containing at most one missing attribute and requires a longer computation time. DSMI suffers a loss of performance when faced with a large number of records with a quantitative missing attribute.

The authors [28] divide the data set like DSMI into two subsets, MissA and NonMissA, in order to improve imputation of quantitative missing attributes, so that the method works for all records with at most one missing attribute. The Model based Missing value Imputation using Correlation (MMIC) method uses the same IS correlation measure as the DSMI method [7], this time determining a correlation index before and after imputation. Thus, three correlation index models MMIC1, MMIC2 and MMIC3 are used. The KNN algorithm applied to the MMIC model imputes both categorical and numerical attributes. For each missing attribute, it generates a table T containing the closest K neighbours to the record containing the missing value. The MMIC calculates the correlation index from the T-table using IS correlation and weighted similarity measures. After calculating the correlation index for each value, the maximum correlation index value is selected for the qualitative attribute imputation and the mean for the quantitative attribute. The MMIC generates an attribute ranking indicating the very first attribute to be used for imputation of missing data and provides an offset for the numerical values. In this way, the MMIC increases the accuracy of classifiers in the classification domain. However, it introduces biases in the imputation of quantitative attributes and decreases accuracy by increasing the average error when the rate of missing attributes is large.

The authors [8] improve the accuracy of imputation from previous work with ten imputation methods based on maximising CMIM correlation. The method consists in finding data segments with strong linear relationships between their characteristics using a well-known criterion. Thus, unlike the DMI method, the CMIM approach directly uses correlation for the estimation of missing data and also accurately measures the degree of correlation. Unlike the FIMUS method which does not use any advanced technique. The CMIM approach estimates missing values by applying regression models to discovered segments. Unlike conventional regression-based imputation methods that apply regression models to the entire data set, the proposed approach applies regression models to highly correlated data segments in order to obtain lower prediction errors. CMIM does not require complex non-linear models to estimate missing values. A ranking system is also used to determine the priority of imputation for each missing characteristic. However, the CMIM ignores discrete missing attributes. Also, the CMIM method has difficulties in selecting the best sub-set of highly correlated data. In addition, it requires more computation time and the ranking of characteristics is only performed once during the entire imputation process.

3 PROPOSED METHOD

3.1 RATING

Our study exploits two segments of modality represented by $\boldsymbol{C_n}.\, with\ n \in \{1,2\}$ or $C_1 = \ [18, 35\ [\ and\ C_2 = [35, 65]$. We refer to $\boldsymbol{a_{ij}}$ the occurrence of an attribute located in the missing or complete data table at the i^{eme} line and the j^{eme} column. i designating the registration number and j the attribute number. We introduce the following quantitative quantities. Either $\boldsymbol{m}$ the average number of occurrences a_{ij}, $\boldsymbol{M}$ the centre of the modality and $\boldsymbol{e}$ estimation of the missing value. In addition, we refer to $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ the limits of the different modalities. Δ is the deviation from the mean $\boldsymbol{m}$ and each of the non-missing values $\boldsymbol{a_{ij}}$. D_c. the complete data set. D_m, the missing data set. D_o, the original data set. Ri the records of these different datasets.

3.2 OUR METHOD

A novelty of our approach is the combination of vertical and horizontal segmentation.

The first segmentation is vertical. Its main interest is the creation of the modalities $\boldsymbol{C_n}$ upstream to allow the processing of discrete attributes. The modalities $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ created allow the grouping of similar or correlated records. The values $\boldsymbol{a_{ij}}$ correlated values resulting from the pooling represent the set of plausible estimation values.

The second segmentation is horizontal. It consists in selecting the terminal nodes or leaves from the construction of decision trees. The end nodes constitute the horizontal segments. In these segments, we estimate the missing occurrences by the minimum distance Δ calculated between the values $\boldsymbol{a_{i1}}$ and the average $\boldsymbol{m}$. The different decision trees are constructed from the complete data set (Table 2).

To do this, the original data set (Table 1) is partitioned into two sets D_c and D_m. Each missing record is assigned to an appropriate sheet (Fig. 2). Thus, the horizontal segments contain both records with complete and incomplete correlated attributes. We present our model and implement it from the original data set D_o of Table 1.

Table 1. Original data set D_o

$\boldsymbol{Ri}$	Age	Salary	DF	Sex	DTS
R1	39	77516	13	Male	40
R2	50	83311	13	Male	13
R3	38	215646	9	Male	40
R4	?	234721	7	Male	40
R5	28	338409	13	Female	40
R6	?	284582	14	Female	40
R7	49	160187	5	Female	16
R8	52	209642	9	Male	45
R9	31	45781	14	Female	50
R10	42	159449	13	Male	40

The original data set D_o is partitioned into D_c and D_m. These sets are represented by Table 2 and 3 respectively.

Table 2. Complete data set D_c

$\boldsymbol{Ri}$	Age	Salary	DF	Sex	DTS
R1	39	77516	13	Male	40
R2	50	83311	13	Male	13
R3	38	215646	9	Male	40
R5	28	338409	13	Female	40
R8	52	209642	9	Male	45
R7	49	160187	5	Female	16
R9	31	45781	14	Female	50
R10	42	159449	13	Male	40

Table 3. Missing data set D_m

$\boldsymbol{Ri}$	Age	Salary	DF	Sex	DTS
R4	?	234721	7	Male	40
R6	?	284582	14	Female	40

The Age attribute contains missing values for records R4, R6. For their processing, we first segment the Age attribute vertically into two modalities. In this case, the Age attribute values in the set D_c (Table 2) are filtered in ascending order. Then, duplicates are removed in order to proceed with segmentation according to the reality of the study. This segmentation is repeated for all quantitative attributes. It is represented in the following Fig.1:

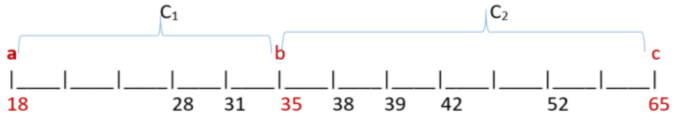


Fig. 1. Vertical segmentation of modalities

We construct the decision tree for the Age attribute using the C4.5 algorithm [7]. The tree is constructed from D_c (Table 2). It is also a matter of predicting sets of plausible values for estimating the Age attribute using the formula Age~Salary+DF+Sex+DTS. The ends of the tree represent leaves or horizontal segments (Fig. 2). In our case, we have three leaves. We assign the missing records R4 and R6 to the appropriate leaves. Thus, they are found in the leaves Leaf 2 and Leaf 3. Sheets 2 and 3 are represented by Tables 4 and 5 respectively.

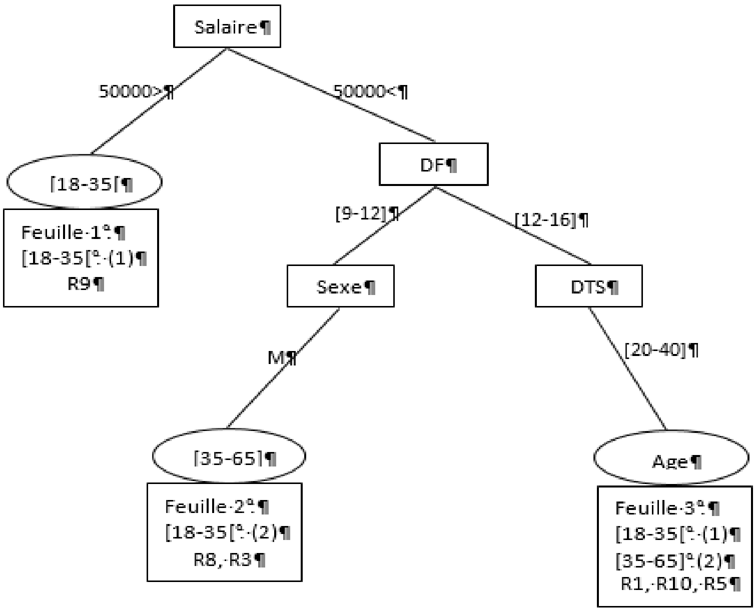


Fig. 2. Age attribute tree

Table 4. Recording Sheet 2 or Horizontal Segment 1

Ri	Age	Salary	DF	Sex	DTS
R3	38	215646	9	Male	40
R8	52	209642	9	Male	45
R4	?	234721	7	Male	40

Table 5. Recording Sheet 3 or Horizontal Segment 2

Ri	Age	Salary	DF	Sex	DTS
R1	39	77516	13	Male	40
R5	28	338409	13	Female	40
R10	42	159449	13	Male	40
R6	?	284582	14	Female	40

The processing of missing data by our model is done in four cases as follows:

- **First case: the two not-missing attributes belong to the same modality**

The modality represents the numerical segment obtained during the vertical segmentation of numerical or discrete attributes. Occurrences 38 and 52 belong to the same modality. $C_2 = [35;\ 65]$. In this case, we calculate the centre of the modality $\boldsymbol{M} = \frac{35+65}{2} =$

50, then the average of the non-missing values $\boldsymbol{m} = \frac{38+52}{2} = 45$. If $\boldsymbol{m} < \boldsymbol{M}$ then the estimate of the missing value **e** is equal to the value of the non-missing attribute in the segment [35, 50 [i.e. 38]. Illustrated in Fig. 3.

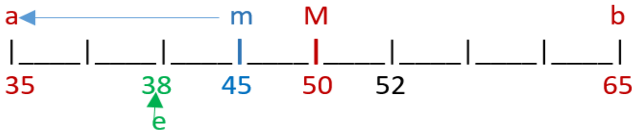


Fig. 3. Diagrams of case 1 estimation

- **Second case: non-missing attributes belong to different modalities**

The average of the values of the Age attribute in Table 4 is given by the formula $m = \frac{1}{i}\sum_{1}^{i} a_{ij}$. This results in $m = \frac{30+50}{2} = 40$. Then, we deduce the missing value estimate by choosing the non-missing value belonging to the segment containing the mean value. m (see Fig. 4).

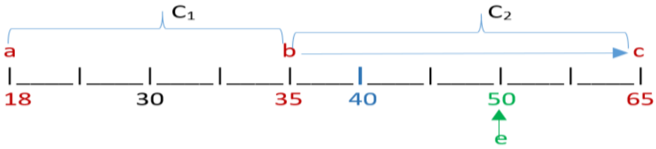


Fig. 4. Diagrams of case 2 estimation

- **Third case: several attributes that are not missing in the same modality**.

This is the same principle as the first case. But this time, we introduce the distance notedΔ. It is defined by $\Delta = \inf\left|m - a_{ij}\right|$. It represents the minimum distance between the average m and the non-missing attributes that are in the same segment as m.

Either $m = \frac{37+38+45+55+60+50}{6} = 48$. The estimated value lies in the range of $[35, 50[$. The following distance table can be used to determine the final estimated value.

Table 6. Minimum distance case 3

Attributes aij	37	38	45	48	50
Distance Δ	11	10	3	0	2

$\Delta = 0$, is the smallest of the distances to the average. It corresponds to the value of the not-missing attribute 48. We deduce from this that the estimate of the missing value e is 48 ($e = 48$). Diagram shown in Fig. 5 :

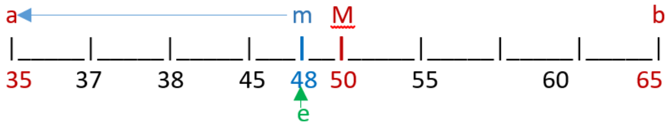


Fig. 5. Diagrams of case 3 estimation

- **Fourth case : Several not-missing attributes in different modalities**

This is the same principle as the second case. We also use the distance Δ which is the minimum of the distances between the average and the data with not missing values of the same modality as m. Then the estimate of the third case is applied.

$a_{i1} = \{25, 27,28,39,40,42,47,48\}$; $\boldsymbol{m} = \frac{25+27+28+39+40+42+47+48}{8} = 36$; $36 \in C_1$ *then* $\boldsymbol{e} \in \{39,40,42,47,48\}$

The distance table below is used to determine the final estimated value.

Table 7. Minimum distance case 4

Attributes aij	39	40	42	47	48
Distance Δ	3	4	6	11	12

$\Delta = 3$ is closer to $m = 36$ then $e = 39$ This is shown in Fig. 6 :

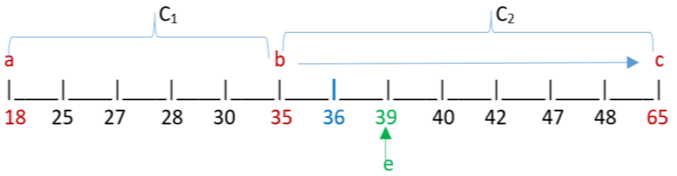


Fig. 6. Diagrams of case 4 estimation

We formalise our model for the treatment of discrete attributes by means of the flowchart in Fig. 7. This flowchart defines the main steps of our model. We name our method HMID (Hybrid Method Imputation of Discrete Missing Attributes).

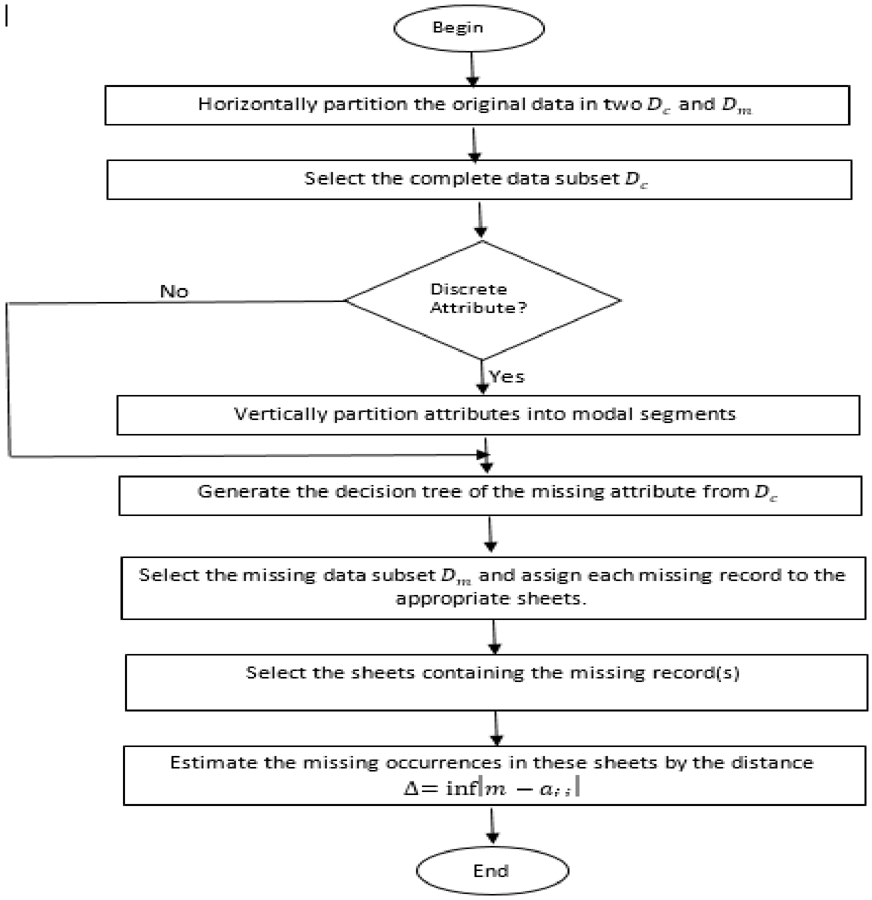
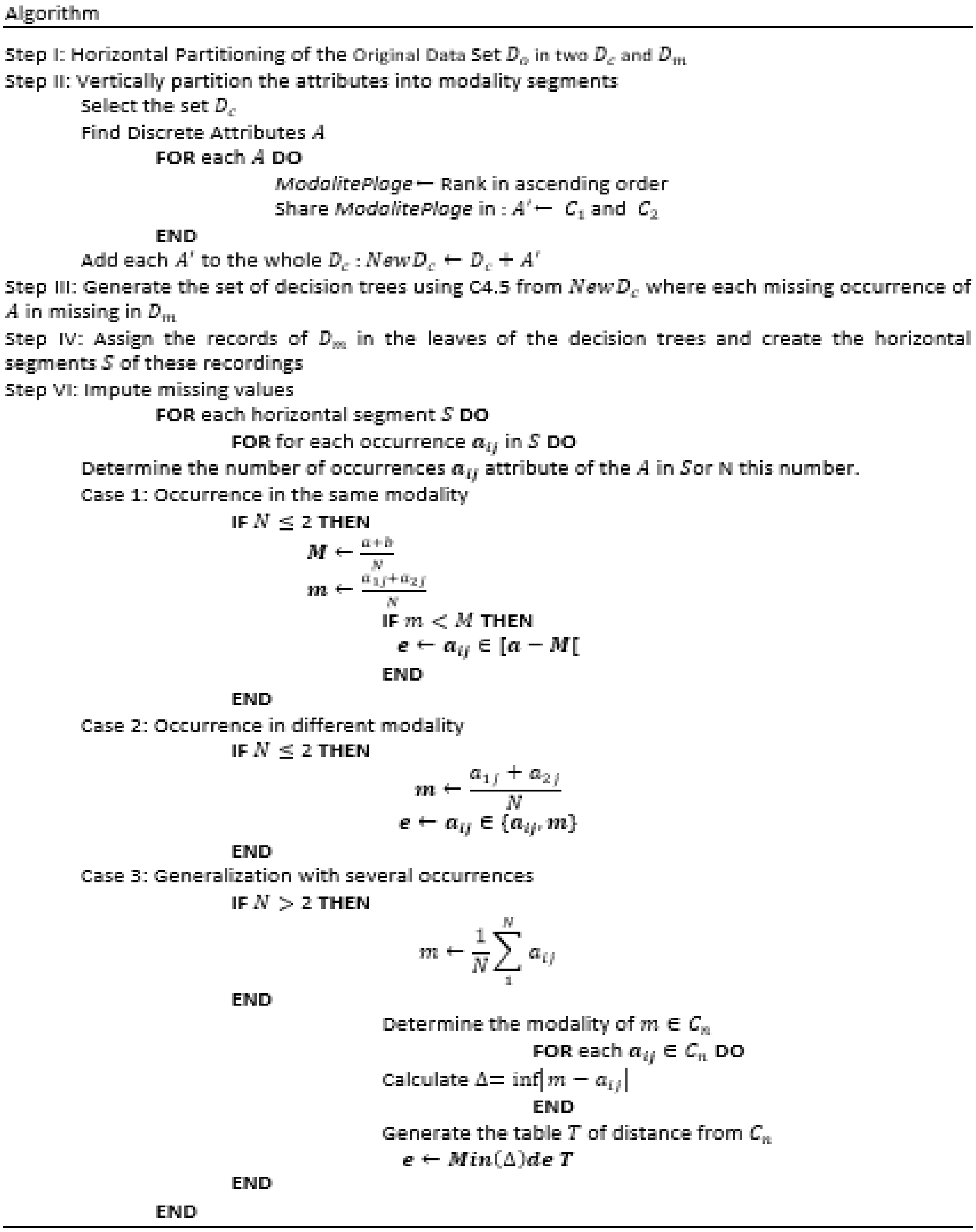

Begin
Horizontally partition the original data in two D_c and D_m
Select the complete data subset D_c
Discrete Attribute?
No
Yes
Vertically partition attributes into modal segments
Generate the decision tree of the missing attribute from D_c
Select the missing data subset D_m and assign each missing record to the appropriate sheets.
Select the sheets containing the missing record(s)
Estimate the missing occurrences in these sheets by the distance $\Delta = \inf|m - a_{ij}|$
End

Fig. 7. Flow chart of our method

3.3 THE MODEL ALGORITHM



```
Algorithm
Step I: Horizontal Partitioning of the Original Data Set $D_o$ in two $D_c$ and $D_m$
Step II: Vertically partition the attributes into modality segments
        Select the set $D_c$
        Find Discrete Attributes $A$
                FOR each $A$ DO
                        $ModalitePlage \leftarrow$ Rank in ascending order
                        Share $ModalitePlage$ in : $A' \leftarrow C_1$ and $C_2$
                END
        Add each $A'$ to the whole $D_c$ : $NewD_c \leftarrow D_c + A'$
Step III: Generate the set of decision trees using C4.5 from $NewD_c$ where each missing occurrence of
$A$ in missing in $D_m$
Step IV: Assign the records of $D_m$ in the leaves of the decision trees and create the horizontal
segments $S$ of these recordings
Step VI: Impute missing values
                FOR each horizontal segment $S$ DO
                        FOR for each occurrence $a_{ij}$ in $S$ DO
        Determine the number of occurrences $a_{ij}$ attribute of the $A$ in $S$or N this number.
        Case 1: Occurrence in the same modality
                        IF $N \leq 2$ THEN
                                $M \leftarrow \frac{a+b}{N}$
                                $m \leftarrow \frac{a_{1j}+a_{2j}}{N}$
                                        IF $m < M$ THEN
                                          $e \leftarrow a_{ij} \in [a - M[$
                                        END
                        END
        Case 2: Occurrence in different modality
                        IF $N \leq 2$ THEN
                                        $m \leftarrow \frac{a_{1j} + a_{2j}}{N}$
                                        $e \leftarrow a_{ij} \in \{a_{ij}, m\}$
                        END
        Case 3: Generalization with several occurrences
                        IF $N > 2$ THEN
                                        $m \leftarrow \frac{1}{N}\sum_{1}^{N} a_{ij}$
                        END
                                Determine the modality of $m \in C_n$
                                        FOR each $a_{ij} \in C_n$ DO
                                Calculate $\Delta = \inf|m - a_{ij}|$
                                        END
                                Generate the table $T$ of distance from $C_n$
                                  $e \leftarrow Min(\Delta) de\ T$
                        END
                END
```

4 RESULTS

The algorithm is developed with the programming language R using the VIM (Visualization and Imputation of Missing values) library [31]. Each experiment is repeated six times. The average performance is presented as the final result. All missing values are inserted according to the random missing data (MAR) mechanism [13], [16]. In this mechanism, the probability that a value is missing is independent of the missing values. However, it depends on the complete values. We also use a uni-varied structure. In this structure, only the missing values belong to one and only one attribute [16].

Our data is extracted from the UCI Machine Learning database [32]. These data relate to the US population census in 1994. Its main objective was to predict the age groups with an annual wage gain of more than 50,000 euros. These data contain the US population group between 16 and 100 with at least one year of training (DF) and non-zero hours worked per week (DTS). This database contains twelve attributes such as age, salary, weekly working hours, to name but a few. This database is used to test our missing data processing model. We limit our study to the following five attributes Age, duration of training (DF), weekly working time (DTS), Sex and Salary. Also, there are no missing data in the existing database. For the purposes of applying our model, we delete data in order to obtain missing data. The missing data are introduced according to these six missing data rates 5%, 10%, 15%, 20%, 30% and 40%. The missing data thus obtained are estimated by our model. The results of these estimates are given in Fig. 8. After their estimation, the values obtained are compared with the initial suppressed data (see Fig. 9).

Using two evaluation criteria and a correlation coefficient, we show the effectiveness of our method. These criteria are Root Mean Square Error (RMSE) [7], [10], Mean Absolute Error (MAE) [8] and correlation coefficient RV [33], [34]. RMSE is the most widely used performance indicator to measure the accuracy of predictions. These results are given in Table 7. The MAE assesses the closeness of the estimated values to the initial values (see Table 6). The RV measures the ratio of the initial data set to the estimated data set. The calculated RV coefficient is equal to one (RV=1). It is between [0, 1]. The higher the RV; the lower the MAE and RMSE values (see Tables 6 and 7 respectively), the better imputation performance. These criteria are formalised as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varepsilon_{ij}}^{2}} \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \boldsymbol{\varepsilon_{ij}}| \qquad (2)$$

$\boldsymbol{a_{ij}}$ represents the original value and $\boldsymbol{e_{ij}}$ the estimated value. Error $\boldsymbol{\varepsilon_{ij}},\ \boldsymbol{\varepsilon_{ij}} = \boldsymbol{e_{ij}} - \boldsymbol{a_{ij}}$

$$RV\left(X^{(i)}, X^{(j)}\right) = \frac{tr\,(S_{ij}S_{ji})}{\sqrt{tr\left(S_{ii}^{2}\right)tr\left(S_{jj}^{2}\right)}} \qquad (3)$$

With $S_{ij} = \frac{1}{n-1}\sum\left(X_{\alpha}^{(i)} - \bar{X}^{(i)}\right)\left(X_{\alpha}^{(j)} - \bar{X}^{(j)}\right)'; i, j = 1;\ 2$

In addition, we compare our HMID (Hybrid Method Imputation of Discrete Missing Attributes) with the DSMI method, the KNNI (see Fig. 10). Our HMID outperforms these methods.

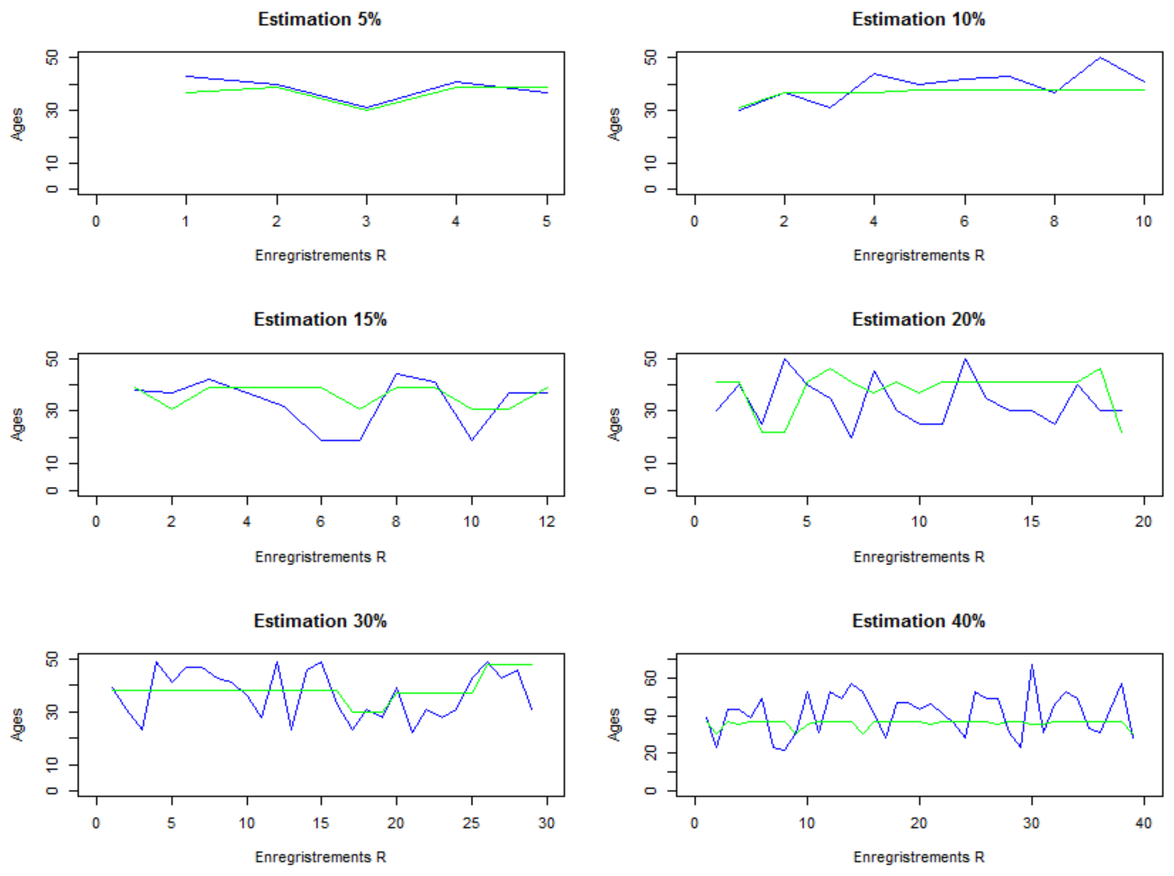


Fig. 8. Results of the estimation of missing values

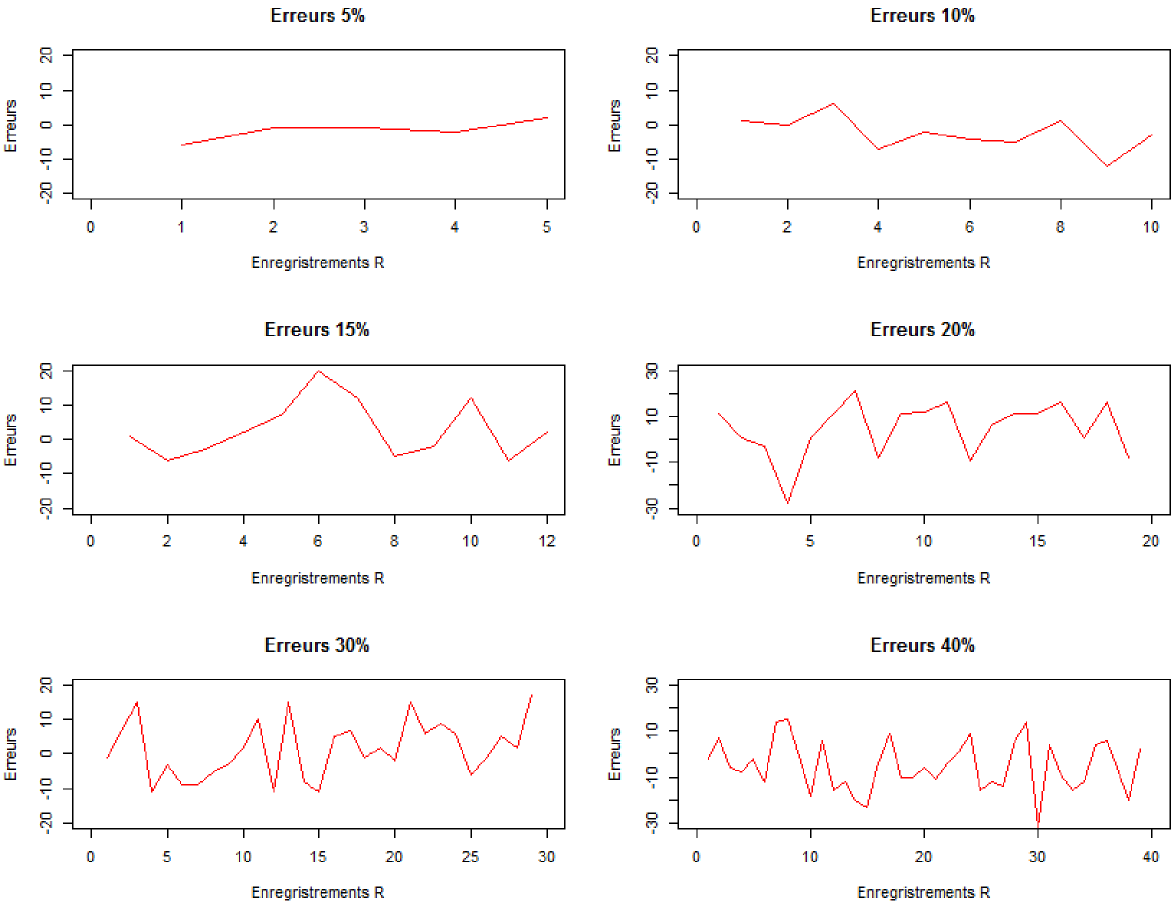


Fig. 9. Estimation errors according to the missing rate

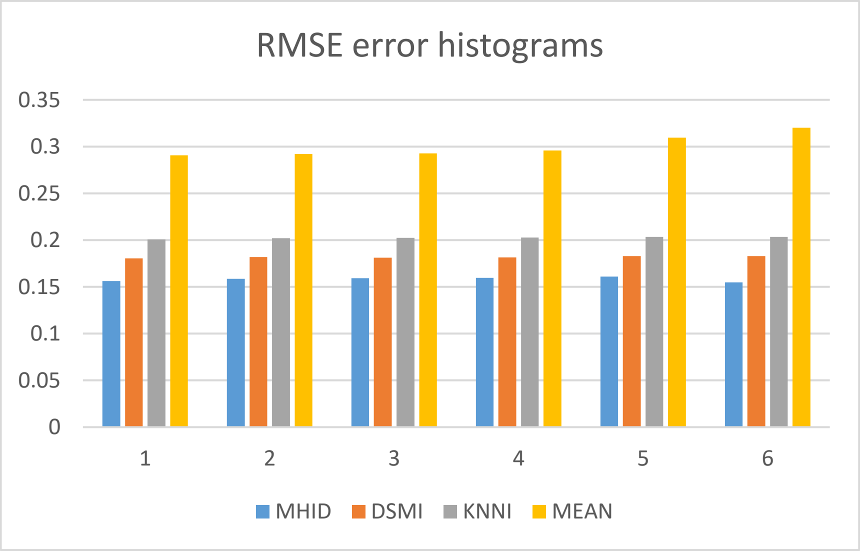


Fig. 10. Overall RMSE error rates

Table 8. MAE results

Rates	MAE			
	HMID	***DSMI***	***KNNI***	***MEAN***
5%	**0,1662**	0,1704	0,1706	0,2706
10%	**0,1686**	0,1718	0,1722	0,2721
15%	**0,1692**	0,1712	0,1725	0,2727
20%	**0,1696**	0,1714	0,1727	0,2758
30%	**0,1710**	0,1728	0,1734	0,2894
40%	**0,1648**	0,1729	0,1735	0,3001

Table 9. RMSE results

Rates	RMSE			
	HMID	***DSMI***	***KNNI***	***MEAN***
5%	**0,1862**	0,1904	0,1906	0,2906
10%	**0,1886**	0,1918	0,1922	0,2921
15%	**0,1892**	0,1912	0,1925	0,2927
20%	**0,1896**	0,1914	0,1927	0,2958
30%	**0,1910**	0,1928	0,1934	0,3094
40%	**0,1848**	0,1929	0,1935	0,3201

5 Discussion

Our experiments show that our model is effective in dealing with discrete missing data as shown in Figure 8. Indeed, it retains monotony. Also, when the database size is large, it has similar properties to the KNNI algorithm [31]. Decision tree induction can handle continuous values. Furthermore, it allows multiple estimation of missing values of the same attribute, thus reducing the high variability of the dispersion, unlike the mean imputation method. The mean imputation method estimates a single value for all records with missing values. This single value is the average of all complete values of the missing attribute.

Our model approximately reproduces all shapes of horizontal segments unlike CMIM [8]. The CMIM which has difficulty in finding highly correlated horizontal segments. One of the advantages of the HMID is its handling of missing values that fall into narrowly reduced segments. We favour its use in domains where the number of descriptors is high in large sample sizes. However, HMID has areas of instability due to extreme values of the modalities (see Fig. 8). Consequently, occurrences closer to the lower end are overestimated. While those at the upper end are underestimated. This under- or over-estimation induces noisy data in the dataset. The HMID outperforms the DSMI, KNNI and mean methods with an average RMSE rate of 0.1682 (see Table 7). Our model outperforms DSMI, KNNI and the mean with missing data proportions ranging from 5% to 40%. The DSMI, like ours, uses the decision tree and majority vote to impute missing values. KNNI uses the complete occurrences of the nearest K neighbours. In the case of qualitative occurrences, KNNI uses the majority vote among the closest K neighbours, otherwise the average. The results (see Table 6) show the increase in the RMSE and MAE error rate as a function of the rate of missing values. When the RV correlation between the original data set and the estimated data set tends towards 1, the lower the RMSE and MAE decreases. Attribute correlations are natural properties of a data set. These correlations cannot be improved or changed for the dataset [6]. Thus, our calculated correlation coefficient (RV=1) shows that our method does not change the structure of the dataset unlike the mean imputation method. Our model provides better results for the imputation of discrete data.

6 Conclusion

In this paper, we use a hybrid approach to estimate missing values of discrete attributes. The case where several records have at most one missing attribute. In this context, we implement the method in different cases. The first case where only two complete occurrences fall in a horizontal segment, the estimation is done with non-missing attributes belonging to the same modality and then with different modalities. The second case deals with the estimation of missing attributes from several data that may belong to different modalities. The model has been validated using MAE, RMSE and RV correlation measurements with very good accuracy. Our method HMID outperforms methods such as mean, KNNI, DSMI in a uni-varied structure data set. In the future, we plan to extend our method to the processing of records with several missing attributes. The difficulty with these records is that they fall into several sheets. Thus, their estimation is ignored by the majority of current methods. We also plan to propose awareness raising tools upstream of data collection. This practice could reduce the proportion of missing data in order to improve both the performance of Data Mining algorithms and data cleaning methods.

REFERENCES

[1] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, et F. Herrera, " A survey on data preprocessing for data stream mining: Current status and future directions ", Neurocomputing, vol. 239, p. 39-57, mai 2017, doi: 10.1016/j.neucom.2017.01.078.

[2] H. Nugroho et K. Surendro, " Missing Data Problem in Predictive Analytics ", in Proceedings of the 2019 8th International Conference on Software and Computer Applications - ICSCA '19, Penang, Malaysia, 2019, p. 95-100, doi: 10.1145/3316615.3316730.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, et P. Smyth, " From data mining to knowledge discovery: an overview ", in Advances in knowledge discovery and data mining, USA: American Association for Artificial Intelligence, 1996, p. 1-34.

[4] U. Garciarena et R. Santana, " An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers ", Expert Syst. Appl., vol. 89, p. 52-65, déc. 2017, doi: 10.1016/j.eswa.2017.07.026.

[5] P. Li et E. A. Stuart, " Best (but oft-forgotten) practices: missing data methods in randomized controlled nutrition trials ", Am. J. Clin. Nutr., vol. 109, no 3, p. 504-508, mars 2019, doi: 10.1093/ajcn/nqy271.

[6] Md. G. Rahman et M. Z. Islam, " Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques ", Knowl.-Based Syst., vol. 53, p. 51-65, nov. 2013, doi: 10.1016/j.knosys.2013.08.023.

[7] R. Deb and A. W.-C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data", Inf. Sci. Vol. 339, pp. 274-289, Apr. 2016, doi: 10.1016/j.ins.2016.01.018.

[8] A. M. Sefidian et N. Daneshpour, " Estimating missing data using novel correlation maximization based methods ", Appl. Soft Comput., vol. 91, p. 106249, juin 2020, doi: 10.1016/j.asoc.2020.106249.

[9] V. H. Bousquet, "Traitement des données manquantes en épidémiologie: application de l'imputation multiple à des données de surveillance et d'enquêtes", p. 339, 2012.

[10] P. McMahon, T. Zhang, et R. A. Dwight, " Approaches to Dealing With Missing Data in Railway Asset Management ", IEEE Access, vol. 8, p. 48177-48194, 2020, doi: 10.1109/ACCESS.2020.2978902.

[11] J. L. Peugh et C. K. Enders, " Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement ", Rev. Educ. Res., vol. 74, no 4, p. 525-556, déc. 2004, doi: 10.3102/00346543074004525.

[12] J. L. Schafer et J. W. Graham, " Missing data: Our view of the state of the art. ", Psychol. Methods, vol. 7, no 2, p. 147-177, 2002, doi: 10.1037/1082-989X.7.2.147.

[13] R. J. A. Little et D. B. Rubin, Statistical analysis with missing data, Third edition. Hoboken, NJ: Wiley, 2019.

[14] J. Huang, Y.-F. Li, et M. Xie, " An empirical analysis of data preprocessing for machine learning-based software cost estimation ", Inf. Softw. Technol., vol. 67, p. 108-127, nov. 2015, doi: 10.1016/j.infsof.2015.07.004.

[15] R. J. A. Little, "Regression With Missing X's : A Review," J. Am. Stat. Assoc., p. 22, 1992.

[16] A. Imbert, "Describing, taking into account, imputing and evaluating missing values in studies statistiques : a review of existing approaches", J. Soc. Franc¸aise Stat. Stat., vol. 159, no. 2, p. 55, 2018.

[17] Q. Song, M. Shepperd, X. Chen, et J. Liu, " Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation ", p. 10, 2008.

[18] G. E. A. P. A. Batista et M. C. Monard, " An analysis of four missing data treatment methods for supervised learning ", Appl. Artif. Intell., vol. 17, no 5-6, p. 519-533, mai 2003, doi: 10.1080/713827181.

[19] P. Madley-Dowd, R. Hughes, K. Tilling, et J. Heron, " The proportion of missing data should not be used to guide decisions on multiple imputation ", J. Clin. Epidemiol., vol. 110, p. 63-73, juin 2019, doi: 10.1016/j.jclinepi.2019.02.016.

[20] J. Huang et al., " Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study ", J. Syst. Softw., vol. 132, p. 226-252, oct. 2017, doi: 10.1016/j.jss.2017.07.012.

[21] T. Aljuaid et S. Sasi, " Proper imputation techniques for missing values in data sets ", in 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, août 2016, p. 1-5, doi: 10.1109/ICDSE.2016.7823957.

[22] Andrew Gelman & Jennifer Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models. Columbia University, New York, 2006.

[23] Nurzaman, T. Siswantining, S. M. Soemartojo, et D. Sarwinda, " Application of Sequential Regression Multivariate Imputation Method on Multivariate Normal Missing Data ", in 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, oct. 2019, p. 1-6, doi: 10.1109/ICICoS48119.2019.8982423.

[24] D. B. Rubin, Éd., Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1987.

[25] J. Wang et al., " An Improvement of Support Vector Machine Imputation Algorithm Based on Multiple Iteration and Grid Search Strategies ", in 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, févr. 2020, p. 538-543, doi: 10.1109/ICIoT48696.2020.9089571.

[26] Md. G. Rahman et M. Z. Islam, " FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis ", Knowl.-Based Syst., vol. 56, p. 311-327, janv. 2014, doi: 10.1016/j.knosys.2013.12.005.

[27] T. Schneider, " Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values ", J. Clim., vol. 14, p. 19, 2001.

[28] S. A. Zahin, C. F. Ahmed, et T. Alam, " An effective method for classification with missing values ", Appl. Intell., vol. 48, no 10, p. 3209-3230, oct. 2018, doi: 10.1007/s10489-018-1139-9.

[29] M. G. Rahman et M. Z. Islam, " iDMI: A novel technique for missing value imputation using a decision tree and expectation-maximization algorithm ", in 16th Int'l Conf. Computer and Information Technology, Khulna, mars 2014, p. 496-501, doi: 10.1109/ICCITechn.2014.6997351.

[30] K. O. Cheng, N. F. Law, et W. C. Siu, " Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data ", Pattern Recognit., vol. 45, no 4, p. 1281-1289, avr. 2012, doi: 10.1016/j.patcog.2011.10.012.

[31] A. Kowarik and M. Templ, "Imputation with the R Package VIM", J. Stat. Softw. Vol. 74, No. 7, 2016, doi: 10.18637/jss.v074.i07.

[32] Ronny Kohavi and Barry Becker, "UCI Machine Learning Repository: Adult Data Set", 1994. https: //archive.ics.uci.edu/ml/datasets/Adult (accessed Sept. 08, 2020).

[33] M. Rauf Ahmad, " A significance test of the RV coefficient in high dimensions ", Comput. Stat. Data Anal., vol. 131, p. 116-130, mars 2019, doi: 10.1016/j.csda.2018.10.008.

[34] P. Robert and Y. Escoufier, "A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient", Appl. Stat. Appl. Stat., vol. 25, no. 3, p. 257, 1976, doi: 10.2307/2347233.