




Extension of the Hybrid Method for Efficient Imputation of Records with Several Missing Attributes

Kone Dramane^(✉) , Kimou Kouadio Prosper, and Goore Bi Tra

Computer and Telecommunications Research Laboratory: LARIT, INP-HB, Yamoussoukro,
Côte d'Ivoire
dramane.kone18@inphb.ci

Abstract. The treatment of records with several discrete missing values present in the databases is still a delicate problem. Indeed, these records can bias the results of data mining algorithms, thus invalidating the results. In this paper, we present an extension of the Hybrid Method for Efficient Imputation of Discrete Missing Attributes (HMID) to effectively handle these records. The method consists of partitioning the database into two subsets, one containing complete records and the other incomplete records. From the complete set, decision trees for all missing discrete attributes are created. The multiple missing records can be in the same leaf or in different leaves. In the same leaf, they are estimated directly by the HMID method. Otherwise, the sheets containing them are merged into a horizontal segment to determine the dominant modality of the complete attributes. In which case, multiple records are estimated. We evaluate our algorithm using two databases. The Adult dataset extracted from the UCI Machine Learning database and SH_CDI_Single extracted from the World Bank database. Finally, we compare our algorithm with four imputation methods using the accuracy of missing value estimation and RMSE. Our results indicate that the proposed method performs better than the existing algorithms we compared.

Keywords: Correlation · Discretization · Classification · Data · Quality

1 Introduction

The treatment of data errors is essential for the efficiency of data mining algorithms. Indeed, a good treatment of errors ensures a good quality of the input data of these algorithms and improves the forecasting results. Among these treatments, we cite data cleaning, which consists in treating imperfections such as outliers, noise and missing data. These imperfections are mostly introduced during data manipulation or are caused by human error. Therefore, it is imperative to process all input data to these algorithms in order to minimize errors and improve their quality. Omitting this data preparation phase exposes users of data mining algorithms to biased predictions. Indeed, these algorithms work when all the attributes of all the records are completed. In this context, some

software simply deletes all records with missing attributes to comply with this rule [1, 2]. Thus, the sample size of the data is reduced, which negatively impacts the sampling of the data and the results of the data mining algorithms. At this point, it becomes natural to ask how to maintain the sample size while preserving the use of data mining algorithms? This question will be answered by proposing an estimation model for multi-attribute missing data.

Several estimation strategies exist in the literature. Among them, we find very popular methods such as those using the mean, the mode or linear regression. Unfortunately, the disadvantage of these easy-to-use methods is the loss of relationships between the attributes [3]. For example, an attribute estimated by the mean method has the same value everywhere it appears, to mention just one example. This situation promotes the development of new methods [4–10] which take into account the direct relationship between the attribute to be estimated and the surrounding data. Among these new methods, we have the Hybrid Method for Efficient Imputation of Discrete Missing Attributes (HMID) [4] which preserves the relationships between the attributes. Moreover, it allows to estimate records with only one missing attribute. In this paper, we will extend this method to efficiently estimate multi-attribute missing data (which is the highest cause of errors in the data). We propose the Extension of HMID for efficient estimation of records with several missing attributes (ExHMID). It constructs decision trees of all missing discrete attributes from a complete dataset. All leaves containing the same records with multiple missing attributes are selected and merged into a horizontal segment. In this segment, the missing attribute is estimated by the HMID model from the complete records of the dominant modality. Using real data, our experiments show improved model imputation accuracy using Root Mean Square Error (RMSE) and RV correlation tests. This paper is presented as follows: Sect. 2 presents a review of related work on imputation methods. Then, Sect. 3 describes our proposed model. Then, in Sect. 4, we validate our model by RMSE error tests. Finally, in Sect. 5, we analyze our results and draw consequences.

2 Review of Related Literature

2.1 Missingness Mechanism

The researcher needs to understand the mechanism behind the missing data. Since the effectiveness of methods for handling missing data is directly related to this mechanism [11–13]. Imbert [12] recommend its consideration in the development of new methods for handling missing data. The mechanism is first proposed by Rubin [14], the current standard. Indeed, it creates a variable R , indicator variable of the data. $R = 0$, if the data is observed and $R = 1$, if the missing data. When he considers R as a random variable, it obtains two distributions, complete data Y and missing data indicator R . Finally, he defines the mechanism by the conditional distribution of R according to Y defined by the function $f(R|Y, \theta)$, with θ the parameter of the missing data model. This function provides the three types of mechanism which are Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR), formalized as follows:

$$MCAR: f(R|Y, \theta) = f(R|\theta) \quad (1)$$

The probability of the missing data does not depend on either the observed or the missing data.

$$\text{MAR: } f(\mathbf{R}|\mathbf{Y}, \theta) = f(\mathbf{R}|Y^o, \theta) \quad (2)$$

Here, the probability of the missing value depends on the observed data.

$$\text{MNAR: } f(\mathbf{R}|\mathbf{Y}, \theta) = f(\mathbf{R}|Y^o, Y^m, \theta) \quad (3)$$

This expression clearly indicates that the probability of the missing value depends on both the observed and the missing value, respectively Y^o and Y^m .

Several methods for imputing missing data are proposed in the literature. However, few of these methods address the mechanism causing the missing data. While there is an interaction between the choice of the imputation method and the mechanism causing the missing data [11]. In this context, its consideration in the development of algorithms for handling missing data improves the estimation efficiency of imputation methods. Also, it allows a simple adaptation of Machine Learning algorithms to other existing methods for possible improvement. Audigier [15] show that the missingness mechanism has a direct impact on the quality of the predictions of the missing data processing methods. Ignorance of this mechanism appears to be one of the factors of inefficiency of existing methods. In this study, we exploit the MAR to propose an efficient method for handling missing data.

2.2 Discretization

The C4.5 supervised classification algorithm is one of the most widely used algorithms in machine learning [16]. The construction of the decision trees is top-down, by recursively partitioning the training records. At each iteration, the set of records is partitioned using the best (qualitative or quantitative) segmentation attribute [17]. However, quantitative attributes have a very wide domain, posing problems in determining their cut-off point. This is an important challenge in machine learning. To solve this problem, Yang and Garcia [18, 19] propose a discretization using C4.5 before or during the modelling process. Their discretization uses binarization, which consists in dividing the quantitative attributes into two intervals. The C4.5 algorithm has the advantage of fast sorting of quantitative attributes with a complexity $O(n \log(n))$. To take more advantage of quantitative attributes, it is possible to develop a process in the C4.5 binarization method [20]. However, one of the important problems of the C4.5 algorithm is the generalization limit of quantitative attributes. The selected threshold value often does not reflect the generalization of the quantitative attribute. Also, it does not judge its generalizability. In this case, the C4.5 algorithm uses quantitative attributes with low generalization performance for data segmentation. From this point of view, in this study we exploit an upstream discretization, taking into account the analysis context for the segmentation of quantitative attributes. The user himself defines the cut-off point of the quantitative attributes in order to optimize the accuracy of the model. The best cut-off point reflecting the data distribution improves the performance of the decision tree.

2.3 Imputation Methods Based on Partitioning

The existing imputation methods in the literature are classified into two approaches: deletion and imputation. Deletion consists of eliminating all records with missing attributes in the database. This way of dealing with missing data is called full case analysis [1]. Its advantage lies in its simplicity of implementation and use. However, it leads to loss of information and reduces the sample size considerably.

The imputation approach aims to correct the loss of information and preserve the original structure of the data set. It estimates the missing value in contrast to deletion methods [21]. However, popular imputation methods should be avoided because of their weaknesses [3]. Mainly the loss of relationships between variables and the underestimation of standard errors. In this context, these references [4–6, 8, 9, 15, 22] study the correlation between records to deal with missing data. Furthermore, they combine their correlation model with machine learning algorithms for the estimation of missing data. In this context, two approaches are used: the global approach [9, 22] and the local approach [4, 6, 7].

Expectation Maximization Imputation (EMI) [22] is a global approach that uses the EM algorithm. The EM algorithm relies on the mean and covariance matrix of the dataset to deal with the missing quantitative attributes. The EM algorithm calculates the mean and covariance matrix. This process continues recursively until the mean and covariance matrix differs from the previous mean and covariance matrix below user-defined thresholds. The method is effective in sets with more attributes than records. However, it only works for quantitative missing attributes with strong correlations in a random missing data set. Iterative Bi-Cluster based Local Least Square Imputation (IBLLS) [10] also handles quantitative missing attributes.

It combines both Local Least Square imputation (LLSImpute) and Local Weighted Linear Approximation imputation (LWLA) [23]. LWLA uses the Euclidean distance and correlation matrix. It then automatically identifies the most similar sets of records in the overall data set. Highly correlated attributes are taken into account, by applying the local least squares framework estimation technique to them. However, the IBLLS ignores the influence of low correlation attributes. The objective of the Framework for Imputing Missing values Using co-appearance, correlation and similarity analysis (FIMUS) [9] is to impute qualitative missing values in contrast to the EMI and IBLLS. For this purpose, the authors exploit the co-appearance of values belonging to different attributes, the similarity of values belonging to an attribute and the correlations of attributes. The use of these properties allows the estimation of quantitative and qualitative attributes. In terms of execution time, it is better than IBLLS. However, its main problem is its mathematical complexity. As the number of records in the dataset increases, the realization of similarity graphs becomes more and more tedious. Furthermore, the similarity values depend on the co-occurrence values. If there is no co-occurrence value, then the associated similarity value has no impact on imputing missing values. The segmentation done by FIMUS results in a loss of information.

For the development of a better segmentation method, Rahman [24] show that the correlations between attributes from horizontal segmentation of a dataset are higher than the correlations on the whole dataset, contrary to Rahman and Schneider [9, 22]. Their Decision tree based Missing value Imputation (DMI) is an extension of the EMI, FIMUS methods. It uses the *C4.5* decision tree algorithm and the EMI estimation technique. The quantitative attributes are estimated by the EMI technique. The qualitative missing value is estimated by majority voting. The DMI method is clearly superior to the EMI. However, it does not work if all records have the same value for a numerical attribute. Also, when all numerical values are missing in a record.

The i-Decision tree based Missing value Imputation (iDMI) [24] is an improvement of DMI and deals with the computational time complexity of a smaller data set. The k-Decision tree based Missing (kDMI) [25] not work properly if all attributes of a record are qualitative. Also, it does not handle missing value records that fall into more than one leaf. The SCMI [26] method uses the correlation Interest factor and Support count (IS) [8] to address the weaknesses of the kDMI method. The correlation is calculated using the IS measure. The final estimation value is obtained by random sampling according to the weighted correlation distribution.

Decision tree and Sampling based Missing value Imputation (DSMI) [8] is an extension of the SCMI method in large data sets. DSMI exploits intra- and inter-record correlations for the estimation of missing values. For this purpose, two correlation measures are used, a direct one called 1st level similarity and a transitional one called 2nd level similarity. DSMI outperforms DMI, iDMI, k-Nearest Neighbor Imputation (KNNI) [21], FIMUS in data sets with qualitative missing values. However, the method does not work for records containing at most one missing attribute.

The Model based Missing value Imputation using Correlation (MMIC) [6] processes all records with at most one missing attribute to improve the DSMI. It exploits the IS correlation by determining a correlation index before and after imputation. For this purpose, three correlation index models are used (MMIC1, MMIC2 and MMIC3). The maximum correlation index value is selected for the imputation of the qualitative attribute and the average for the quantitative attribute. MMIC increases the accuracy of classifiers in the classification domain. However, it introduces bias in the imputation of quantitative attributes.

Sefidian [5] improve the imputation accuracy of previous work. For this purpose, they use ten estimation models based on the Correlation Maximization-based Imputation Methods (CMIM). The CMIM approach uses correlation directly for the estimation of missing data. In this context, it estimates missing values by applying one of the regression models to the discovered segments. Unlike conventional regression-based imputation methods that apply regression models to the whole data set. However, it has difficulties in selecting the best subset of highly correlated data.

The HMID [4] estimates the discrete attributes that the DSMI method ignores during its processing. It combines the decision tree and the minimum distance estimation model between the mean and non-missing values of the same modality. The method outperforms methods such as KNNI and DSMI. However, it does not handle records with multiple missing values. Furthermore, it overestimates or underestimates records with missing values at the ends of the modalities.

3 Presentation of Our Method

Our model is an extension of the HMID [4] model. Indeed, HMID retains monotony, reduces the high variability of the dispersion and efficiently handles discrete missing values that fall within narrowly reduced segments. We retain these advantages in order to estimate records with multiple quantitative missing values. The double segmentation contributes to an optimization of the inter-attribute and intra-attribute correlation in the same or different modalities. Our estimation model is based on minimizing the distance between the available variables and their mean, incorporating a majority modality score calculation. The original dataset noted D_o , represented in Table 1, is partitioned into two subsets of data. The first subset of data, denoted D_c , with complete values is represented by Table 2. The second subset denoted D_m , with missing values is represented in Table 3.

Table 1. Original data set, D_o .

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

Table 2. Complete records set, D_c .

Records	Age	Salary	DF	Sex	DTS
R1	39	77516	13	Male	40
R5	28	338409	13	Female	40
R8	52	209642	9	Male	45
R9	31	45781	14	Female	50
R10	42	159449	13	Male	40

Table 3. Incomplete records set, D_m .

Records	Age	Salary	DF	Sex	DTS
R2	?	83311	?	Male	13
R3	38	215646	?	Male	40
R4	?	234721	7	Male	40
R6	?	284582	14	Female	40
R7	?	160187	5	Female	?

Vertical segmentation is carried out upstream according to the reality of the field of study. This process improves the construction of decision trees. Moreover, it promotes the determination of the intra-attribute correlation of the different modalities. All quantitative attributes of the complete data subset, D_c , are selected for this segmentation. We vertically segment each quantitative attribute into two modalities from the D_c dataset in Table 2. In this case, the values of each of these attributes are filtered in ascending order and removed from redundant values. This segmentation technique is shown in Fig. 1. For each quantitative random variable $X = \{x_1, x_2, \dots, x_n\}$; $C_{1X} = [X_{min}, b]$; $C_{2X} =]b, X_{max}]$. The variable b is a cut point depending on the context of the study, perhaps a closed or open interval. The variables X_{min}, X_{max} represent respectively the lower bound of the first modality and the upper bound of the second modality.

In our case, the attributes Age, Duration of training (DF) and Duration of work per week (DTS) are segmented as follows respectively: $C_{1Age} = [18; 35]$ and $C_{2Age} =]35; 65]$; $C_{1DF} = [0; 5]$ and $C_{2DF} =]5; 15]$; $C_{1DTS} = [0; 40]$ and $C_{2DTS} =]40; 80]$.

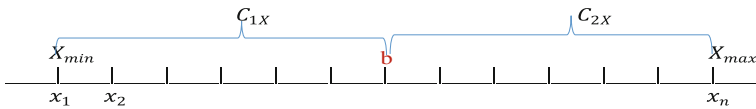


Fig.1. Vertical segments of missing and non-missing attributes

The records R2, R3, R4, R6 and R7 (see Table 3) each contain one or more attributes with missing values. Decision trees are constructed based on the missing attributes from D_c (see Table 2). In our case, the missing attributes are Age, DF and DTS. To do this, we need to construct three decision trees Age, DF and DTS. The plausible estimation values of the missing attributes in the different leaves of the trees are predicted by the formulas Age~Salary+DF+Sex+DTS, DF~Salary+Age+Sex+DTS and DTS~Salary+DF+Sex+Age respectively. The different decision trees obtained during this prediction are respectively Age (Fig. 2), DF (Fig. 3) and DTS (Fig. 4). Finally, each record with missing attributes is assigned to the appropriate leaves. The pairs (R6, R7) and (R2, R4) of missing attribute

records Age are in sheet 2 and sheet 5 respectively, see Fig. 3. The R2 record with the missing attribute DF is found in sheets 6 and 7. R3 is found in sheet 7. The record R7 with the missing attribute DTS is in sheet 10 (see Fig. 4). Each of these sheets constitutes the horizontal segment. In addition, these different sheets serve as estimation samples of the missing attributes. The different sample estimates of the attributes (Age, DF and DTS) are represented in Tables 4 and 5 respectively. R2 for the attribute DF is found in several sheets. These different sheets are selected and then merged into a single sheet (horizontal segment see Table 6). This merged sheet constitutes the estimation sample.

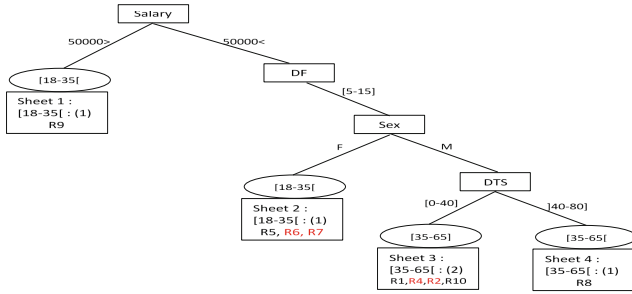


Fig. 2. Age decision tree

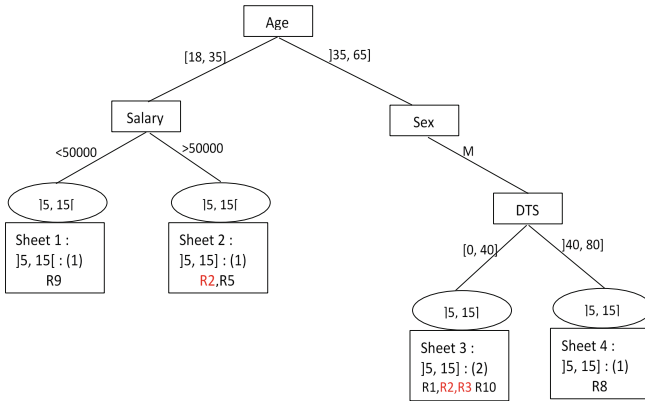


Fig. 3. DF decision tree

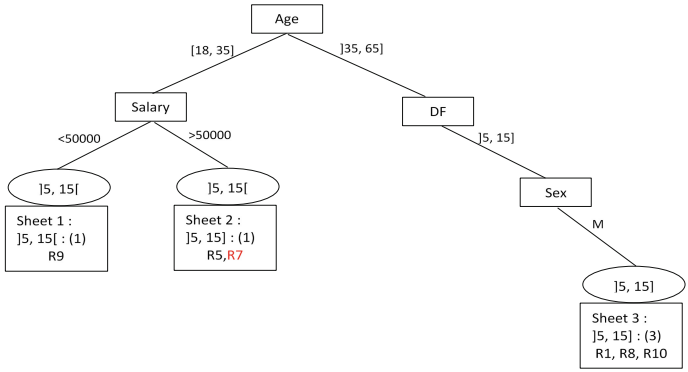


Fig. 4. DTS decision tree

Table 4. Missing attribute estimation segments Age.

Records	Age	Salary	DF	Sex	DTS
R1	39	77516	13	Male	40
R2	?	83311	?	Male	13
R3	38	215646	?	Male	40
a-Segment 1					
R5	28	338409	13	Female	40
R6	?	284582	14	Female	40
R7	?	160187	5	Female	?
R8	52	209642	9	Male	45
b-Segment 2					

We obtain two sample estimates for the Age attribute (see Table 4). In the first sample (Table 4, b-Segment 1), the Age attribute is estimated by the same value 28 for records R6 and R7. The estimation is done by majority voting according to the principle of the HMID model. This same principle allows the missing attribute DTS (Table 5, c-segment 5) to be estimated by 40 for the record R7. The R2 record is found in both sample 3 (Table 5, a-segment) and sample 4 (Table 5, b-segment). The estimation of the DF attribute of record R2 consists of the merging of samples (a-Segment) 3 and (b-Segment) 4 from Table 5. In contrast to method [8], we consider the merge (see Table 6) as an estimation sample for records with multiple missing attributes: the case of R2 for the attribute DF.

Table 5. Missing attribute estimation segments DF and DTS.

Records	Age	Salary	DF	Sex	DTS
R2	?	83311	?	Male	13
R5	28	338409	13	Female	40
a- Segment 3 DF					
R1	39	77516	4	Male	40
R2	?	83311	?	Male	13
R3	38	215646	?	Male	40
R5	28	338409	2	Female	40
R10	42	159449	9	Male	40
b- Segment 4 DF					
R5	28	338409	2	Female	40
R7	?	160187	5	Female	?
c-Segment 5 DTS					

Table 6. DF fusion segments.

Records	Age	Salary	DF	Sex	DTS
R1	39	77516	4	Male	40
R2	?	83311	?	Male	13
R3	38	215646	?	Male	40
R5	28	338409	2	Female	40
R10	42	159449	9	Male	40

The merger has two modalities C_{1DF} and C_{2DF} . The modality C_{1DF} represents the majority modality. The estimation is done in the majority score modality, $s = 2$. Once this modality C_{1DF} is determined, the average of the observed attributes is calculated using the following formula:

$$m_j = \frac{\sum_{i=1}^{n_j} a_{ij}}{n_j}, \text{ With } n_j > 0 \text{ } i \text{ record index, and } j \text{ variable index} \quad (4)$$

In this case, $m_4 = \frac{2+4}{2} = 3$. From this average, the following distance matrix is calculated:

Available attributes a_{ij}	2	4
$d_{ij} = m_j - a_{ij} $	1	1

Here $d_{14} = d_{24}$, due to this equality the center $C = \frac{b_f + b_s}{2} = \frac{0+5}{2} = 2,5$ of the modality is calculated. If $m_j < C$, in this case, the estimation value e is included in the interval $[b_f, m_j]$ otherwise, it is in the interval $[m_j, b_s[$. In our case, $C < m_j$ from which the estimation value is $e = 4$. We illustrate the process of our model and its architecture by a diagram (Fig. 5).

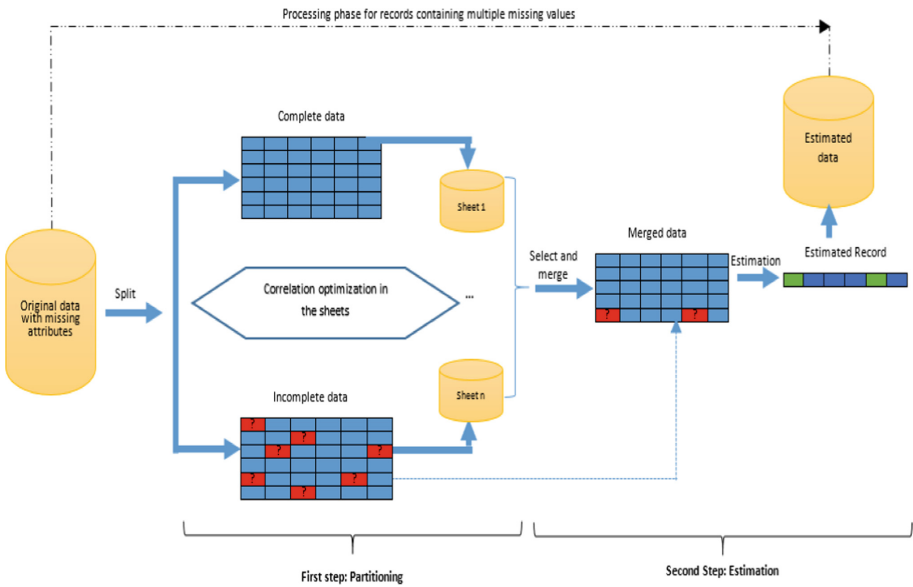


Fig. 5. Process and architecture diagram

We propose the algorithm of our method entitled Extension of HMID for efficient estimation of records with several missing attributes (ExHMID) as follows:

```

Step 1 : Partition  $D_o$  into two subsets complete  $D_c$  and incomplete  $D_m$ 
 $D_o \leftarrow \{R | R \in D_c \cup R \in D_m\}$ 

Step 2 : Vertical partitioning of attributes into modality segments
 $A_j' \leftarrow C_{1j}$  and  $C_{2j}$  // with  $j=1,2,\dots,p$  variable index
 $D_c' \leftarrow D_c \cup A_d'$  // add to the set  $D_c$  the new matrix  $A_j'$ 

Step 3: Generate the set of decision trees using C4.5 from  $D_c'$ 
Step 4: Assign records from  $D_m$  records in the sheets and create the horizontal segments  $F_n$ 
Step 5: Impute missing values
    FOR each Table  $S_n$  DO
        Determine the majority modality  $C_p$  of the missing attribute,
        Select the complete set of attributes  $S_n' = \{a_{1j}, \dots, a_{ij}\} \in C_p$ 
        Determine the number of occurrences  $a_{ij}$  of  $s_n'$  and let N be the number.
        IF  $N \leq 2$  THEN
             $M \leftarrow \frac{a+b}{2}$ 
             $m \leftarrow \frac{a_{1j} + a_{2j}}{2}$ 
        IF  $m < M$  THEN
             $e \leftarrow a_{ij} \in [a, M]$  ELSE
        IF  $N > 2$  THEN
            FOR Each  $a_{ij} \in C_p$  DO
                 $m \leftarrow \frac{1}{N} \sum_{i=1}^N a_{ij}$ 
                Calculate  $\Delta_{ij} = \inf |m - a_{ij}|$ 
                Generate the distance matrix T
            END FOR
             $e \leftarrow \text{Min}(\Delta_{ij}) \text{ de } T$ 
        END IF
    END IF
END
```

4 Results

We implement our ExHMID method and three other existing methods, DSMI [8], KNNI [21] and Mean, to compare the accuracy of ExHMID with these methods on two real datasets. One extracted from the UCI Machine Learning database, Adult [27]. And the other SH_CDI_Single extracted from the World Bank database via <https://microdata.worldbank.org/>. This last includes 930 variables and 2950 records which aim is to study the financial inclusion of rural agricultural households in Côte d’Ivoire via mobile currencies. For experimental purposes of studying financial inclusion, we constitute three subsets (Single1 to Single3), limited to fifteen variables relevant to the impact of mobile money. The second dataset (Adult) concerns the US Census of Population whose main objective is to predict the age groups with an annual salary gain of more than 50,000 euros. This data set contains the US population aged 16 to 100, with at least one year of education (DF) and working hours per week (DTS). In this set, five attributes (Age, DF, DTS, Sex and Salary) are used out of the twelve. In all, we obtain four simulation datasets (Adult, Single1, Single2 and Single3), in each of which the missing data MAR [12, 28] of 5 to 40% are inserted and distributed in a multivariate model. These missing values are estimated by the ExHMID, DSMI, KNNI and Mean methods. Knowing the real values of the randomly created missing values, we evaluate the imputation accuracy of the techniques using the root mean square error (RMSE) [1, 8], by comparing the real values and the imputed values. The experiment is repeated

48 times (number of proportions*number of subsets*number of mechanisms*number of structures). We present the overall average performances of the four methods on the four data sets in terms of RMSE in Table 7. The RMSEs range from 0 to ∞ , where a lower value indicates a better imputation. The used correlation coefficient RV [27, 28] measures the ratio of the initial data set to the estimated data set to know that the natural property (correlation) of the data set does not change. In our case, the calculated RV is equal to one ($RV = 1$). It is between 0 and 1, the more the RV tends towards the value 1, the better the imputation performance. Values in bold indicate the best results. It is clear from Table 7 that in all four datasets, ExHMID outperforms the other methods with respect to the RMSE evaluation criterion. The results of the performance of our model in the different data sets are shown in Fig. 6 and the comparison in Fig. 7.

Table 7. RMSE results

DataSets	Missing ratio	Methods			
		ExHMID	DSMI	Mean	KNNI
Adult	5%	0,1671	0,1904	0,2906	0,1906
	10%	0,1685	0,1918	0,2921	0,1922
	15%	0,1679	0,1912	0,2927	0,1925
	20%	0,1681	0,1914	0,2958	0,1927
	30%	0,1695	0,1928	0,3094	0,1934
	40%	0,1696	0,1929	0,3201	0,1935
Single 1	5%	0,1679	0,2624	0,3856	0,2724
	10%	0,1703	0,2713	0,3871	0,3023
	15%	0,1709	0,2666	0,3877	0,3106
	20%	0,1713	0,2639	0,3908	0,3149
	30%	0,1727	0,2652	0,4044	0,3162
	40%	0,1665	0,2683	0,4151	0,3193
Single 2	5%	0,1679	0,2744	0,3756	0,2884
	10%	0,1693	0,2833	0,3771	0,3173
	15%	0,1687	0,2786	0,3778	0,3226
	20%	0,1689	0,2759	0,3808	0,3398
	30%	0,1703	0,2772	0,3944	0,3412
	40%	0,1704	0,2803	0,4051	0,3443
Single 3	5%	0,1676	0,2734	0,3851	0,2928
	10%	0,1690	0,2823	0,3866	0,3167
	15%	0,1684	0,2776	0,3873	0,3213
	20%	0,1686	0,2749	0,3903	0,3296
	30%	0,1700	0,2762	0,4039	0,3312
	40%	0,1701	0,2793	0,4146	0,3343
Global average		0,1691	0,2534	0,3688	0,2863

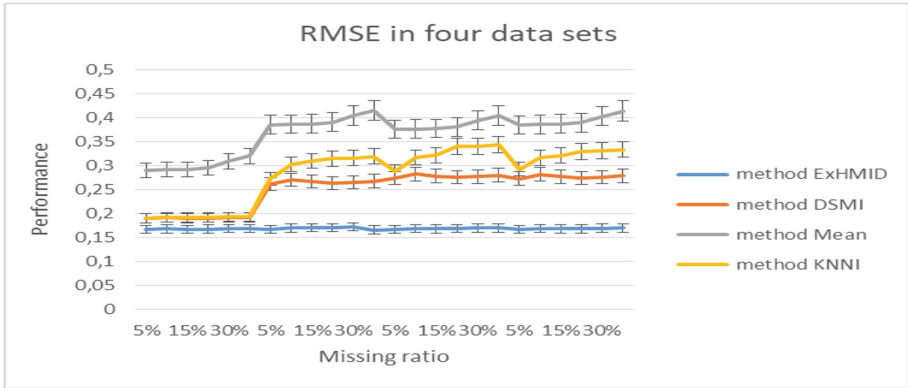


Fig. 6. Results of the performance in the different data sets

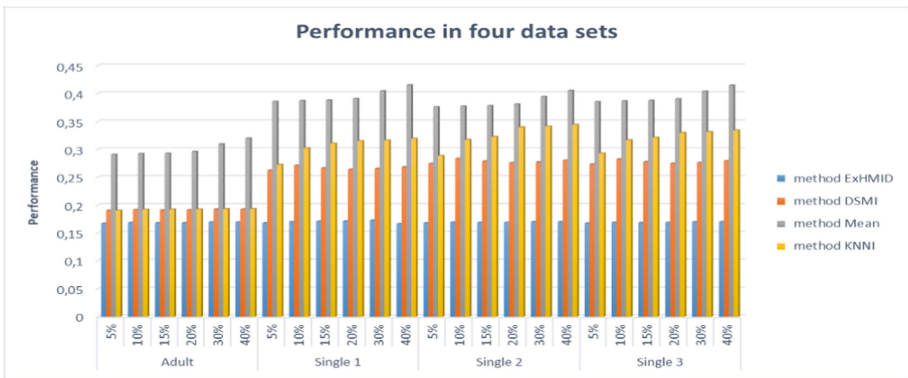


Fig. 7. Performance comparison on four data sets

5 Discussion

It is clear from our experiments (see Table 7) that in all four datasets, ExHMID outperforms the other methods regarding the RMSE criterion, with a global RMSE of 0.1691. Furthermore, its strong performance is observed in the Adult and Single 3 datasets, respectively with a global average of 0.1685 and 0.1690. The ExHMID approach uses HMID [4] (as described in Sect. 3) to estimate multiple missing records in the merged horizontal segment, unlike other methods using the C4.5 algorithm which avoid them [8]. The efficiency of ExHMID in the Adult dataset is a consequence of the efficiency of the HMID [4] estimation technique in a dataset with a uniform distribution. The prediction performance of the ExHMID method is tested against three different imputation methods, namely mean imputation, KNNI, DSMI, see Fig. 6 and 7. Mean imputation produces the worst estimation results, see Fig. 7. The ExHMID method outperforms the other imputation methods, regardless of the missing rate and valuation measure in all four datasets. The KNNI and DSMI methods overall in the Adult dataset, with the different rates of missing data (5 to 40%) have an approximately similar performance, see

Fig. 6. When the rate of missing data is high in the uniform distribution, its performance is reduced while it is better in the skewed distribution. ExHMID sometimes performs as well as HMID [4], when the database size is small. The use of user-defined attribute cut points reflects the good distribution of the data and improves the performances of the decision tree. Therefore the performance of the ExHMID is better. Sometimes the ExHMID provides a multiple estimate of the same attribute with missing values as well for the KNNI. This multiple estimation is the result of maximizing the intra-attribute correlation (during the creation of the decision trees) as opposed to the mean imputation method. The mean imputation method estimates a single value for all records with missing values for the same attribute. This promotes the loss of relationship between attributes when the rate of missing data is significant. An important advantage of ExHMID is the approximate determination of all forms of horizontal estimation segments (estimation subsets) of the missing values in contrast to some methods like CMIM [5]. This is also a better solution to the discretization of quantitative attributes. Since this procedure provides an appropriate threshold value that perfectly defines the limits of the attribute interval. In practice, this technique allows for data reduction by matching data from a wide spectrum of quantitative values to a strongly reduced subset of discrete values, hence to a set of highly correlated records [8, 9]. In addition to the RMSE measure, the RV correlation is calculated between the original data set and the estimated data set, equal to one ($RV = 1$). This result shows that the natural properties (correlation) of the data set after estimation have not changed. Our model provides better results for the imputation of multiple missing attribute records, an extension of the HMID model [4], dealing with multiple missing value records falling in multiple leaves. The model applies to both high and low correlation datasets.

6 Conclusion

Our ExHMID method is an extension of the HMID [4] method. It deals with records with multiple missing attributes. It works by combining the C4.5 algorithm and the estimation model minimizing the dispersion around the mean. The estimation model uses the score of the majority modality to estimate the missing attributes. It works by vertically partitioning the quantitative attributes into modalities. These modalities are defined according to the context of the study. In addition, it performs a horizontal partitioning. This partitioning provides a set of records with highly correlated attributes. In each subset the HMID estimation model is applied. An important advantage of the method is the ease of adding more records to the missing attributes. These added records are estimated correctly without affecting the functioning of the model. The method is effective in real data sets. The experimental results (using $RV = 1$ and $RMSE = 0.161$) show a better performance of our method for different real data sets. Its performance in a symmetric or uniformly distributed dataset is significantly better than in a skewed distribution dataset. Moreover, it remains effective for qualitative as well as quantitative attributes. In the current literature, several methods are developed but few studies propose their evaluation. We are planning an evaluation study of some existing methods in the literature.

References

1. McMahon, P., Zhang, T., Dwight, R.A.: Approaches to dealing with missing data in railway asset management. *IEEE Access* **8**, 48177–48194 (2020). <https://doi.org/10.1109/ACCESS.2020.2978902>
2. Li, P., Stuart, E.A.: Best (but oft-forgotten) practices: missing data methods in randomized controlled nutrition trials. *Am. J. Clin. Nutr* **109**(3), 504–508 (2019). <https://doi.org/10.1093/ajcn/nqy271>
3. Khan Faizan, U.F., Khan Kashan, U.Z., Singh, S.K.: Is group means imputation any better than mean imputation: a study using C5.0 classifier. *J. Phys. Conf. Ser.* **1060**, 012014 (2018). <https://doi.org/10.1088/1742-6596/1060/1/012014>
4. Dramane, K., Tra, G.B., Prosper, K.K.: New hybrid method for efficient imputation of discrete missing attributes. *Int. J. Innov. Appl. Stud.* **31**(4), 763–775 (2021)
5. Sefidian, A.M., Daneshpour, N.: Estimating missing data using novel correlation maximization based methods. *Appl. Soft Comput.* **91**, 106249 (2020). <https://doi.org/10.1016/j.asoc.2020.106249>
6. Zahin, S.A., Ahmed, C.F., Alam, T.: An effective method for classification with missing values. *Appl. Intell.* **48**(10), 3209–3230 (2018). <https://doi.org/10.1007/s10489-018-1139-9>
7. Huang, J., et al.: Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. *J. Syst. Softw.* **132**, 226–252 (2017). <https://doi.org/10.1016/j.jss.2017.07.012>
8. Deb, R., Liew, A.W.-C.: Missing value imputation for the analysis of incomplete traffic accident data. *Inf. Sci.* **339**, 274–289 (2016). <https://doi.org/10.1016/j.ins.2016.01.018>
9. Rahman, Md.G., Islam, M.Z.: FIMUS: a framework for imputing missing values using co-appearance, correlation and similarity analysis. *Knowl.-Based Syst.* **56**, 311–327 (2014)
10. Cheng, K.O., Law, N.F., Siu, W.C.: Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. *Pattern Recognit.* **45**(4), 1281–1289 (2012). <https://doi.org/10.1016/j.patcog.2011.10.012>
11. Garcarena, U., Santana, R.: An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Syst. Appl.* **89**, 52–65 (2017). <https://doi.org/10.1016/j.eswa.2017.07.026>
12. Imbert, A.: Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques: une revue des approches existantes. *J. Soc. Française Stat.* **159**(2), 1–55 (2018)
13. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, p. 15 (2002)
14. Rubin, D.B.: Inference and missing data. *Biometrika* **63**(3), 581 (1976)
15. Audigier, V., et al.: Multiple imputation for multilevel data with continuous and binary variables. *Stat. Sci.* **33**(2), 160–183 (2018)
16. Lu, Z., Wu, X., Bongard, J.C.: Active learning through adaptive heterogeneous ensembling. *IEEE Trans. Knowl. Data Eng.* **27**(2), 368–381 (2015)
17. Patel, N., Singh, D.: An algorithm to construct decision tree for machine learning based on similarity factor. *Int. J. Comput. Appl.* **111**(10), 22–26 (2015)
18. Yang, Y., Chen, W.: Taiga: performance optimization of the C4.5 decision tree construction algorithm. *Tsinghua Sci. Technol.* **21**(4), 415–425 (2016)
19. Garcia, S., et al.: A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **25**(4), 734–750 (2013) <https://doi.org/10.1109/TKDE.2012.35>
20. Cherfi, A., Nouira, K., Ferchichi, A.: Very fast C4.5 decision tree algorithm. *Appl. Artif. Intell.* **32**(2), 119–137 (2018)

21. Batista, G.E.A.P.A., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* **17**(5–6), 519–533 (2003)
22. Schneider, T.: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* **14**, 853–871 (2001)
23. Liu, C.-C., Dai, D.-Q., Yan, H.: The theoretic framework of local weighted approximation for microarray missing value estimation. *Pattern Recognit* **43**(8), 2993–3002 (2010)
24. Rahman, M.G., Islam, M.Z.: iDMI: a novel technique for missing value imputation using a decision tree and expectation-maximization algorithm. In: 16th International Conference on Computer and Information Technology, Khulna, pp. 496–501 (2014). <https://doi.org/10.1109/ICCITechn.2014.6997351>
25. Rahman, M.G., Islam, M.Z.: kDMI: a novel method for missing values imputation using two levels of horizontal partitioning in a data set. In: Motoda, H., Wu, Z., Cao, L., Zaiane, O., Yao, M., Wang, W. (eds.) ADMA 2013. LNCS (LNAI), vol. 8347, pp. 250–263. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-53917-6_23
26. Deb, R., Wee-Chung Liew, A., Oh, E.: A correlation based imputation method for incomplete traffic accident data. In: Pham, D.-N., Park, S.-B. (eds.) PRICAI 2014. LNCS (LNAI), vol. 8862, pp. 905–912. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13560-1_77
27. Ahmad, M.R.: A significance test of the RV coefficient in high dimensions. *Comput. Stat. Data Anal.* **131**, 116–130 (2019). <https://doi.org/10.1016/j.csda.2018.10.008>
28. Robert, P., Escoufier, Y.: A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Stat.* **25**(3), 257–265 (1976). <https://doi.org/10.2307/2347233>