Handling Missing Data Problems with Sampling Methods

Rima Houari[§], *Ahcène Bounceur*[†], *A-Kamel Tari*[§], *M-Tahar Kechadi*[‡]
[§]*LIMED laboratory - University of Abderrahmane Mira Bejaia.*
[†]*Lab-STICC Laboratory- University of Brest France.*
[‡] *PCRG laboratory, University College Dublin Ireland*
Email: {ri.houari, tarikamel59}@gmail.com
Email: Ahcene.Bounceur@univ-brest.fr
Email: Tahar.kechadi@ucd.ie

***Abstract*—Missing data cases are a problem in all types of statistical analyses and arise in almost all application domains. Several schemes have been studied in this paper to overcome the drawbacks produced by missing values in data mining tasks, one of the most well known is based on preprocessing, formerly known as imputation. In this work, we propose a new multiple imputation approach based on sampling techniques to handle missing values problems, in order to improve the quality and efficiency of data mining process. The proposed method is favorably compared with some imputation techniques and outperforms the existing approaches using an experimental benchmark on a large scale, waveform dataset taken from machine learning repository and different rate of missing values (till 95%).**

***Keywords*-Data mining, Data Pre-Processing, Missing values, Multidimensional Sampling, Copula.**

I. Introduction

Data pre-processing is an important step in data mining process, it is often loosely controlled, resulting in out-of-range values and representing in a format that is acceptable in the next phase. As we always say "no quality data no quality results", data cleaning, part of a pre-processing-step, is extremely important, it can be applied to remove noise, correct inconsistencies, missing values and its problems. Missing data could be caused by varied factors such as (1) faulty equipment or incorrect measurements, (2) a value is missing because it was forgotten or lost, (3) it can be generated from human errors on either forget to ask a question or forget to record the answer. However the most serious problems of missing values is that it can result in loss of efficiency, less information extracted from data or conclusions statistically less strong, complication in handling and analyzing the data where methods are in general not prepared to handle them and may have an impact on modeling and sometimes they can destroy it, furthermore missing data can introduce bias resulting from differences between missing and complete data. Many data mining algorithms and statistical techniques are generally tailored to draw inferences from complete datasets. It may be difficult or even inappropriate to apply these algorithms and statistical techniques on incomplete datasets. However, any missing data treatment method should not change the data distribution and the relationship among the attributes should be retained. In this paper, we propose a new approach which involves estimating missing values based on sampling methods and adresses some challenges mentioned. The paper is organized as follows: the basic problem statement are presented in Section 2, missing data techniques taxonomy and related works are discussed in Section 3. Section 4 describes the proposed method. The experimental results are given in Section 5, and finally Section 6 concludes the paper.

II. Problem statements

The missing data problem arises when values for one or more variables are missing from recorded observations. Common notation is following. A dataset is a number of observations, and is denoted as $X = (X_1, ..., X_m)$ where $X_{j,j=1,2,...,m}$ is the j^{th} column. We assume that X_j is a random variable, defined by the CDF (Cumulative Density Function) $F_j(.)$. $X_j \sim F_{j,j=1,2,...,m}$, and each X_j may have a different PDF (Probability Density Function) $f_j(.)$.

The CDF of $X_{j,j=1,2,...,m}$ is defined as : $F(X_j) = P(X_i \leq X_j)$. We assume that the data set X contains missing values. We define v_l^k as the missing value of the k^{th} row and the l^{th} column. We define $V_{(n_1,m),n_1<n} = \{v_l^k\}$, as the set of the data rows that contains missing values. Thus, $V_{(n_1,m)} \subset X$.

III. Missing data techniques taxonomy and Related work

There are several methods for handling missing data, described in a rich literature. According to [21], there are three possible strategies to deal with missing data as shown by Figure 1. The first one is based on missing data ignoring techniques [15][8][17]. The second one represents missing data imputation techniques [9][25]. The third one represents missing data based modeling techniques [6][22]. In this paper, we will focus our attention on the use of missing data imputation methods.

The missing data ignoring techniques simply omits the cases that contain missing data. They are widely used [11] and tend to be the default method for handling missing data, but this may not lead to the most efficient use and they are

978-1-4799-5178-9/14 $31.00 © 2014 IEEE
DOI 10.1109/INDS.2014.25

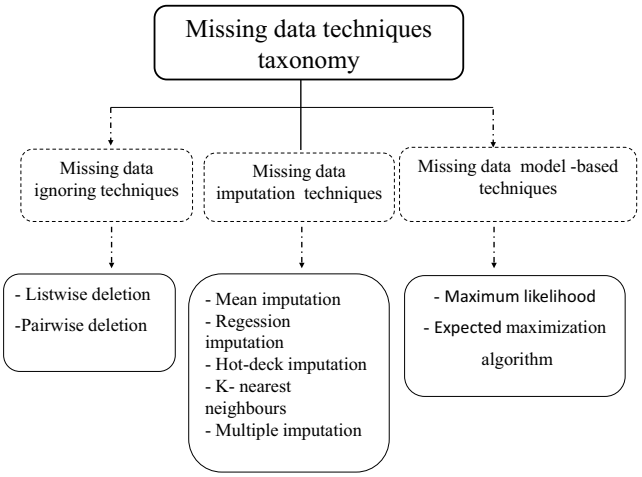

Figure 1. Missing data techniques taxonomy.

used in very small amount of missing values. Two general approaches of the missing data ignoring techniques are used to deal the problem of missing values: listwise deletion approaches [15][8] and pairwise deletion approaches [16][17].

Rather than removing variables or observations with missing data, another approach is to fill in or "impute" missing values. The missing data imputation techniques [6][12][8][24] are a strategy for completing missing value in the data with plausible value which is the estimation of the true value of the unobserved observation. These methods keep the full sample size, which can be advantageous for bias and precision [9].

- Mean imputation [6]: this is one of the most common used methods. This method involves replacing a missing value with the overall sample mean. It is easily implemented and simple, but there are some problems according to [12]. The drawbacks of mean imputation are (a) Sample size is overestimated, (b) variance is underestimated, (c) correlation is negatively biased, and (d) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean.
- Regression imputation [8][24]: a somewhat more sophisticated single-imputation technique is regression-based imputation, in which each missing value is replaced by a predicted data using multiple regression based on non missing data on other variables. This method depends on the assumption of linear relationship between attributes. But in the most cases, the relationship is not linear. Predict the missing data in a linear way will bias the model.
- K-nearest neighbors [10][14][26]: also is known as distance function matching. This method is an approach where a random selection is made from serval closest nearest neighbor or the suitable distance is defined, the observed unit with the smallest distance to the missing value. This approach preserves the sample distribution by substituting different observed values for each missing observation, takes into account the correlation structure of the data, treats instances with multiple missing values, but the dataset must be large enough to find appropriate donor cases, it is difficult to choose the distance function and the the number of neighbors [27].
- Multiple imputation [8][12]: is a strategy of replacing each missing value with a set of m plausible values drawn from their predictive distribution. The multiple imputed datasets can be analyzed by complete-data methods and the results from these analysis are combined and overall estimates are produced. This method avoids the problems of single imputation, can relieve the distortion of the sample variance and produces unbiased estimates, but data must meet normal distribution assumptions, as well as computing and storage requirements [1].

Missing values Model-based methods [6][22][10] define a model from available data and inferences based on the distribution. These methods assume multivariate normality or multi-normality of continuous outcome variables. A common problem with multivariate data is that different individuals are observed on different subsets of a complete set of variables. When the data are a sample from a multivariate normal population, there do not exist explicit closed-form expressions for the maximum-likelihood estimates [6] and Expected maximization method [22][10] of the means, variances and covariances of the normal population.

The literature on imputation methods in data mining [2] employs well-known machine learning methods for their studies [7], in which the authors show the convenience of imputing the missing values for the mentioned algorithms, particularly for classification [12]. The vast majority of missing values studies in classification usually analyze and compare one imputation method against a few others under controlled amounts of missing values and induce them artificially with known mechanisms and probability distributions [3].

Imputation from conditional distribution [4][5] usually means simulation from the conditional distribution $f_{X_k|H}$. The goal of different approach base on conditional imputation [13][15] is to use the $\arg\max$ of conditional density in order to impute missing values, that means find the conditional mean as the imputed value using to maximum likelihood method. Thus conditional distribution consist to (1) construct the joint distribution function f_{X_k}, (2) the density function using marginal $F_1, ..., F_k$, then (3) find the conditional density function, and finally (4) find the

$\arg\max$ function to the imputation method [4]. conditional distribution contains all the informations about the history of data and about marginal distributions and it can estimate the parameters of distribution using the maximum likelihood method, but the main problem with this method is that that the joint distribution may be unknown and finding conditional distribution may therefore impossible [4].

[4][5] [13] present imputation method using the idea of imputing missing value based on conditional distribution and copula of dropout in repeated measurements. [4][5] shows how the copulas can be used to analyze repeated measurements with missing data. The main problem in practice is that there does not exist any theoretical joint distribution describing the repeated measurements [28], and then the copula function is a good tool to create the joint distribution. Imputations can be performed using conditional models. [28]

In the repeated measurements, in general the missingness matrix is monotone, each attribute are ordered [4]. where the monotonicity of the missingness matrix follows from two assumption: (a) subject witch drops out does not return. (b) order of subjects in sample does not matter, important is, that one dimension of data matrix (time) has the fixed ordering. All observations on a subject are obtained until a certain time point, after all measurement are missing, then the authors use ordered data $H = (X_1, X_2, ..., X_{k-1})$ called history for estimating missing values [15]. Thus proposed approach consist to: (1) to resolve the problem of the joint distribution and find the conditional density function based on copula theory, (2) find the $\arg\max$ function to the imputation method using different types of correlation (Compound Symmetry (CS), First Order Autoregressive (AR), Banded Toeplitz (BT)) [4]. [12] presents a comparison of two imputation methods: Markov Chain Monte Carlo (MCMC) and Copulas [4][5] to handle missing data in repeated measurements. The performance of each imputation method was evaluated using mean square error (MSE). [12] shows that the results from the simulation Copulas method was superior effective than the MCMC method.

Copula method can: (1) address the drawbacks of conditional distribution imputation, (2) use a good tool to create the joint distribution (copula theory). (3) Copula is easy for computation, but it is not easy to replace the values of missing data when there are lots of missing items. In addition, the data on repeated measurements those are prior to the observation of missing value must be complete. It also needs to check whether the correlation structure among repeated observations is under CS or AR [12].

For these raisons, in this paper we are interested to use the missing data imputation techniques based on copulas theory in order to model the joint distribution, solve the problem with multivariate data and propose a new multiple imputation approach to replace the missing values with a large database and on a large rate of missing values (till 95%).

IV. Proposed imputation approach

In this paper we propose a new approach based on sampling techniques to handle missing values. The main objective of the proposed method is: (a) to estimate missing values with the most effective method, (b) improve the quality and efficiency of data mining process that is ready for mining. The proposed method operates into two main steps. In the first step the raw datasets are compared by taking into account heterogeneous data, in other words the goal is to estimate the multivariate joint probability distribution without imposing constraints to specific types of marginal distributions of data using copulas. Furthermore in the second step, we want to estimate and complete the missing data.

A. Modeling using copula

If the joint distribution of data X is known then it is possible to find a method to impute the missing value v_l^k. The idea of using copulas is to create a joint distribution from marginal distribution of the 1^{st}, 2^{nd}, 3^{rd},..., $(m)^{th}$ columns. Then, we can define their joint multivariate distribution function as follows: [18]

$$F(x_1, ..., x_m) = C(F_1(x_1), ..., F_m(x_m)). \quad (1)$$

where: F is the CDF function on R^m, $F_1, ..., F_m$ are univariate marginal distributions, $C(F_1(x_1), ..., F_m(x_m))$ is a multivariate distribution function with marginals, and C is copula function.

A general schema to use copula in the first step of the proposed approach requires following steps.

1) Modeling univariate marginal distributions either. The copula method works with any given marginal distributions i.e. it does not restrict the choice of margins.
2) Determining empirical copula: empirical Copula is useful for examining the dependence structure of multivariate random vectors. Formally, the empirical Copula is given by the following equation: [20]

$$(C_{ij}) = \frac{1}{m}(\sum_{k=1}^{m} \mathbf{1}_{(v_{kj} \leq v_{ij})}). \quad (2)$$

Where the function $\mathbf{1}_{(arg)}$ is the indicator function, which equals to 1 if arg is true and 0 otherwise. Here, m (the number of observations) is used to keep the empirical CDF less than 1.
3) Specifying a multivariate distribution function with copula. some copulas for the joint distribution can be applied. For selecting the best-fitting copula some Goodness-of-Fit statistics can be used. In the following the main Copulas will be described.

- Gaussian Copula: the Gaussian Copula is defined as follows:[19]

$$C(\Phi(x_1), ..., \Phi(x_m)) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp(\frac{-1}{2} X^t (\Sigma^{-1} - I) X). \quad (3)$$

where $f_i(x_i)$ is the standard Gaussian distribution, i.e. $X_i \sim N(0,1)$, and Σ is the correlation matrix. The resulting Copula $C(u_1, ..., u_n)$ is called Gaussian Copula. The density associated with $C(u_1, ..., u_n)$ is obtained using Equation (4), and it is written as follows:

$$c(u_1, ..., u_m) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp\left[\frac{-1}{2} \xi^t (\Sigma^{-1} - I) \xi\right]. \quad (4)$$

Where $u_i = \Phi(x_i)$ and $\xi = (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m))^T$.

- Student Copula: the student Copula is extracted from the multivariate Student distribution which is given by the following equation:
$\forall (u_1, ..., u_m) \in [0,1]^m$,

$$C(u_1, ..., u_m) = \frac{(f_{(v,\Sigma)}(t_v^{(-1)} u_1, ..., t_v^{(-1)} u_1))}{(\prod_{i=1}^{(n)} (f_{(v)} t_v^{(u_i)}))}. \quad (5)$$

where $t_v^{(-1)}$ is standard student distribution to univariate degrees of freedom. $f_{(v,\Sigma)}$ is the probability density function of the standard student distribution. Σ is the correlation matrix and $f_{(v)}$ is the univariate density of the student distribution.

- Archimedean Copulas: the Archimedean Copula, is defined as follows:

$$C(u_1, ..., u_n) = \begin{cases} \varphi^{-1}(\varphi(u_1, ..., u_n)) & \text{if } \sum_{\varphi(u_i) \leq \varphi(0)} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Let the generator φ be a continuous, strictly decreasing function from $[0,1]$ to $[0, \infty[$ such that $\varphi(0) = \infty$ and $\varphi(1) = 0$, and let φ^{-1} be the inverse of φ. Then $\varphi = -\ln(t)$. The three main families of Archimedean copulas is: Frank copula, Gumbel copula and finally Clayton copula.

4) Determining theoretical sample of the appropriate copula:
To generate a large sample of the theoretical copula of the appropriate copula, we start with an uniformly distributed sample of U that has a uniform distribution in $[0,1]$, using a standard pseudo-random generator. $C(U_1, ..., U_m) = F(F_1^{-1}(U_1), ..., F_m^{-1}(U_m))$. For each sampled value of U, we can calculate a value of X using the inverse CDF given by $X = F^{-1}(U)$.

B. Imputation approach

After Specifying the appropriate copula and generating the theoretical sample having the same parameters of empirical sample in the previous step, we will illustrate the proposed imputation approach shown by Figure 2.

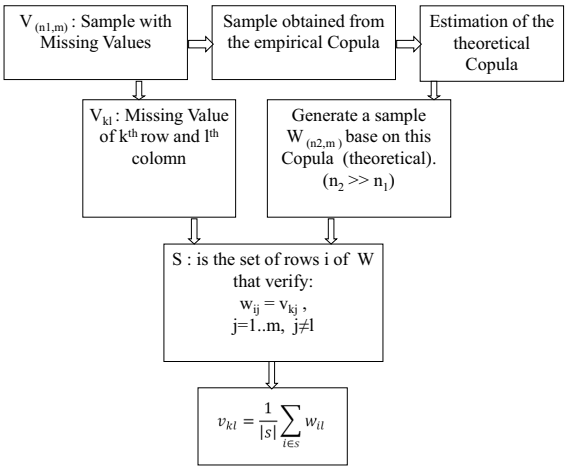

Figure 2. Proposed imputation approach

In order to indicate the known variable positions in the set of data rows contains missing values $V_{(n_1,m),n_1<n}$, with missing value v_l^k of k^{th} row and l^{th} columns, we will take the indicator ind which elements take the value 1. Let $W_{(n_2,m)}$ is set of data rows generated by the theoretical sample, with $n_2 >> n$. According to the known variable positions in the i^{th} row of $V_{(n_1,m)}$, we will determine the subset of rows in the theoretical sample $W_{(n_2,m)}$ whose have the same position of the known variable v_l^k, verifying $w_j^i = v_j^k$.

If the missing component is null, then the data set is complete, otherwise it contains missing values. Let $W_{(n_2,m)}$ is the results subset obtained. For each column of $W_{(n_2,m)}$, we will compute the mean of variables $W_{(n_2,m)}$, $v_j^k = \sum_{i \in |S|} w_j^i$ and we impute the variable v_j^k in i^{th} row of $V_{(n_1,m),n_1<n}$.

V. EXPERIMENTAL RESULTS

In this section, we experimentally investigate the performance and effectiveness of the proposed methods using real-world datasets, Waveform database taken from machine learning repository.

- Waveform Database Generator was obtained from a study on Waveform Generator from Wadsworth International Group Belmont, California in 11/10/1988.

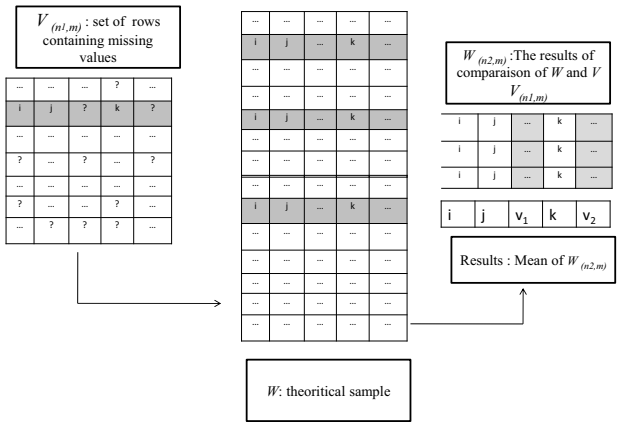

Figure 3. Overview illustrated of the proposed method.

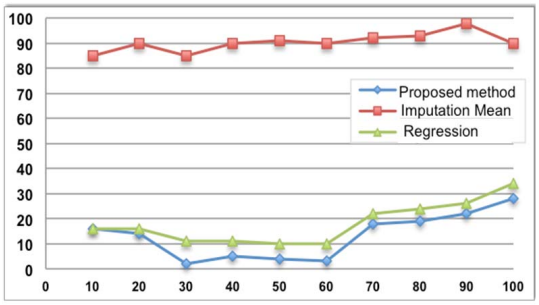

Figure 4. Imputation results obtained with proposed approach, mean and regression techniques.

The Waveform Database, includes 33367 rows and 21 attributes [29].

The goal of this section is to test the effectiveness of the proposed imputation method with some well-known methods in the case of different rate (%) of missing values using real-world dataset Waveform. Figure 3 shows an illustration of the proposed method. At the first we generated nineteen situations of missing values rates (%). The missing values have been imputed arbitrarily in order to not interfere with our experiment, then we estimated the error rate. The performance of each imputation method was evaluated by dividing the data sum generated from the square difference between the original data and the impute data with the size of data multiplied by the rate of missing values; using the equation denoted by:

$$error = \sqrt{\frac{\sum_{i=1}^{n}(d_i)^2}{nb}} \tag{7}$$

$(d_i)^2$ is the square difference between real-world data and simulated data; nb = missing values rate $\times$ the length of data column $\times$ the length of data column.

By implementing different imputation techniques (means technique and regression) and the proposed approach. Figure 4 shows the numerical results obtained to test the performance of proposed method.

1) Discussion: The comparison of the different graphs in Figure 4, shows the correspondence with the results obtained with the waveform database for the same evaluation criteria. The increase in missing values by 5% to 95% caused a decrease in the minimum accuracy of 88% for the mean imputation method, 14% by the regression technique and a 1% by our method, for against the maximum error is 98% for the imputation method, and 36% by the regression technique and the average and 30% for the proposed method. Degradation of the error is always evident when increasing missing values. However our strategy (blue curve) based on Copula is much better than the mean imputation (red curve) and also superior to regression technique (green curve). The difference is most sensitive for all values. Our approach should be a practical solution to estimate missing values for a very large database because it overcomes the missing values using Copulas with small errors even with a very large percentage of missing values which is contrary to the mean imputation if is much less conclusive and may be better on small databases which is not the case in the data mining.

VI. Conclusion

In this paper, we have proposed a new approach based on sampling techniques to predict missing values which addresses the weakness of existing approaches. The proposed approach is compared with imputation techniques (means technique and regression) with an experimental benchmark and on a large scale and large rate of missing values (till 95%). The experimental study outperforms the existing approaches and provide very good results those that show the effectiveness of the proposed method. Future work will focus to develop a system for preprocessing data in data mining techniques which integrates our approach.

References

[1] A.Farhangfar, L.Kurgan, W.Pedrycz, A novel framework for imputation of missing values in databases, In IEEE Trans Syst Man Cybern Part, pp 692 – 709, 2007.

[2] A.Farhangfar , L.Kurgan , Impact of imputation of missing values on classification error for discrete data, Pattern Recognit, pp 3692 – 3705, 2008.

[3] E.Acuna, C.Rodriguez, clustering and data mining applications, In 2004 Proceedings of springer, pp 639 – 648, Berlin, 2004.

[4] E.Kaarik, M.Kaarik, Modelling Dropouts by Conditional Distribution a Copula-Based Approach, Journal of Statistical Planning and Inference, 2009.

[5] E.Kaarik, Imputation by conditional distribution using Gaussian copulas, In Springer Verlag Proceedings in Computational Statistics, 17th Symposium Copmstat 2006, pp $1447-1454$, Rome Italy,2006.

[6] Gabriel, L. Schlomer, Sheri Bauman and Noel A. Card , Best Practices for Missing Data Management in Counseling Psychology, Journal of Counseling Psychology, Volume $57(1):1$, 2010.

[7] G.Batista , M.Monard, An analysis of four missing data treatment methods for supervised learning, Appl Artif Intell 17(5),pp $519-533$, 2003.

[8] Jason Van Hulse, Taghi M. Khoshgoftaar, Chris Seiffert, A Comparison of Software Fault Imputation Procedures, In IEEE Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06), 2006.

[9] Julian Luengo, Salvador Garcia, Francisco Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods, Knowl Inf Syst in Springer-Verlag, London , 2011.

[10] Jason Van Hulse, Taghi M. Khoshgoftaar, Chris Seiffert, A Comparison of Software Fault Imputation Procedures in IEEE: 2012 11th International Conference on Machine Learning and Applications (ICMLA), Volume 1, pp$281-287$, Boca Raton FL, 2012.

[11] Little and Rubin, Statistical Analysis with Missing Data, New York: John Wiley and Sons, pp 11-13, New York, 2002.

[12] Lily Ingsrisawang, Duangporn Potawee, Multiple Imputation for Missing Data in Repeated Measurements Using MCMC and Copulas, Proceedings of the International MultiConference of Engineers and computer scients (IMECs), Hong kong, 2012.

[13] Meelis Kaarik, Ene Kaarik, Imputation by Gaussian Copula Model with an Application to Incomplete Customer Satisfaction Data, on Proceedings of COMPSTAT'2010, $pp485-492$,2010.

[14] Poolsawad, C. Kambhampati, and J. G. Cleland, Feature Selection Approaches with Missing Values Handling for Data Mining A Case Study of Heart Failure Dataset, in World Academy of Science, Engineering and Technology, pp. $828-837$, 2011.

[15] C.Y. J.Peng, M.Harwell, S.M.Liou, L. H.Ehman, Advances in missing data methods and implications for educational research, In S. Sawilowsky (Ed.), Real data analysis, pp $31-78$, Greenwich, 2006.

[16] P.D. Allison, Missing Data (Quantitative Applications in the Social Sciences), Sage University Papers Series on Quantitative Applications in the Social Sciences, $07-136$. Sage, Thousand Oaks CA, 2001.

[17] Roth, Missing data: a conceptual review for applied psychologists, Personnel Psychology 47, $537-560$, 1994.

[18] R. Houari, A. Bounceur, T. Kechadi, A New Approach for Pretreatment of Large Multi-Dimensional Data using Sampling Methods, Colloque sur l'Optimisation et les SystÃ©mes d'Information COSI 2013, Algeria, 2013.

[19] R. Houari, A. Bounceur, T. Kechadi, A-K. Tari and R. Euler, A New Method for Estimation of Missing Data Based on Sampling Methods for Data Mining, In springer Proceedings of the Third International Conference on Computational Science, Engineering and Information Technology (CCSEIT-2013), Volume 225, pp 89-100, Turkey, 2013.

[20] R. Houari, A. Bounceur, and T. Kechadi, A New Method for Dimensionality Reduction of Multi-Dimensional Data using Copulas, In IEEE 11th International Symposium on Programming and Systems ISPS, Algeria, 2013.

[21] R.B Kline, Principles and Practice of Structural Equation Modelling, Guilford Press, New York, 1998.

[22] Shichao Zhang , Xindong Wu , Manlong Zhu, Efficient Missing Data Imputation for Supervised Learning, In proceeding of: Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010, Beijing China, 2010.

[23] S.Shirani, F.Kossentini, R.Ward, Reconstruction of baseline JPEG Coded Images in error prone environments, in IEEE Transactions on Image Processing, $1292-1298$, 2000.

[24] Sathit Prasomphan, Imputing Landsat7 ETM with SLC-off Image Using the Similarity Measurement between Two Clusters, in IEEE 2012 International Conference on Future Generation Communication Technology (FGCT), $190-195$, London, 2012.

[25] Tabachnick, Fidell, Using Multivariate Statistics, fourth ed. Allyn and Bacon, Needham Heights, 2001.

[26] V.Kumutha, S. Palaniammal, An Enhanced Approach on Handling Missing Values Using Bagging k-NN Imputation, in IEEE International Conference on Computer Communication and Informatics ICCCI 2013, INDIA, 2013.

[27] W.Ling, F.Dong Mei, Estimation of missing values using a weighted k nearest neighbors algorithm, In IEEE Proceedings of the 2009 International Conference on Environmental Science and Information Application Technology, vol.3, pp.$660-663$, 2009.

[28] Ene Kaarik, Handling Dropouts by Copulas, Institute of Mathematical Statistics University of Tartu Estonia, $http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.330.3886\&rep=rep1\&type=pdf$.

[29] http://archive.ics.uci.edu/ml/