Kurdistan Journal of Applied Research (KJAR) | Print-ISSN: 2411-7684 – Electronic-ISSN: 2411-7706 | kjar.spu.edu.iq
Volume 2 | Issue 3 | August 2017 | DOI: 10.24017/science.2017.3.30

Improving accuracy of missing data imputation in data mining

Nzar A. Ali
Computer Dept.
Cihan University
Sulaimani, Iraq
Nzar@mail.com

Zhyan M. Omer
Statistics and Informatics
University of Sulaimani
Sulaimani, Iraq
zhyan.omar16@gmail.com

Abstract: ***In fact, raw data in the real world is dirty. Each large data repository contains various types of anomalous values that influence the result of the analysis, since in data mining, good models usually need good data, databases in the world are not always clean and includes noise, incomplete data, duplicate records, inconsistent data and missing values. Missing data is a common drawback in many real-world data sets. In this paper, we proposed an algorithm depending on improving (MIGEC) algorithm in the way of imputation for dealing missing values. We implement grey relational analysis (GRA) on attribute values instead of instance values, and the missing data were initially imputed by mean imputation and then estimated by our proposed algorithm (PA) used as a complete value for imputing next missing value.***

We compare our proposed algorithm with several other algorithms such as MMS, HDI, KNNMI, FCMOCS, CRI, CMI, NIIA and MIGEC under different missing mechanisms. Experimental results demonstrate that the proposed algorithm has less RMSE values than other algorithms under all missingness mechanisms.

Keywords: Data mining; Missing value; Missing value ; Data preprocessing ,

1. INTRODUCTION

Data mining (DM) known as Knowledge Discovery in Databases (KDD) is *"the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* [1]. Databases in the world includes big amounts of data, every day millions of peoples send data over the web through social media, banks applications, governmental offices, mobile applications, university portals, etc. today with powerful storage devices and large network for computer connections, data growth exponentially. The organization of this big data and preprocessing them so that useful knowledge extracted automatically from them are lead to new branch of science known as Data Mining (DM).

In data mining process the quality of results depends on the quality of the data, for that the data pre-processing is one of the important steps to reach clean and quality data and has grate effect on the success of the mining results

Data pre-processing is one of the main step in the the Knowledge discovery in databases (KDD) process that decreases the complexity of the data and gives better conditions to subsequent analysis. Through this process the nature of the data is understood and the analysis of the data is performed more accurately and efficiently.

The next important step is the data itself. Input data must be suitable in structure and format that suit each DM task perfectly. It is unrealistic to expect that data will be perfect after they have been extracted. Since good models usually need good data, a thorough cleansing of the data is an important step to improve the quality of data mining methods. Not only is the correctness, also the consistency of values important. Missing data can also be a particularly pernicious problem. Especially when the number of missing data is large, not all attributes (instances) with missing values can be deleted from the sample [3]

2.LITERATURE REVIEW

- J. Tian , B. Yu , D. Yu , Sh. Ma , (2014) [4] ,In this paper a hybrid missing data completion method is proposed named "*Multiple Imputation using Gray-system-theory and Entropy based on Clustering (MIGEC)*". First step the non-missing data records are distributed on several clusters. Then, the imputed missing value is estimated after multiple calculations by utilizing the information entropy of the proximal category for each incomplete records in terms of the similarity metric based on Gray System Theory (GST).
- X.Y. Zhou , J. S. Lim , (2014) [5] , they studied a new method, the NB-EM (Naïve Bayesian-Expectation Maximization) algorithm, for handling missing values .The comparison between their method and traditional EM(Expectation Maximization) and non-substitution approaches for dealing with datasets containing randomly missing value is performed. They proof the most effective method, compared with the traditional EM algorithm, the NB-EM algorithm has a higher accuracy rate, which suggests that the NB-EM algorithm can obtain a better results on missing values in practice.
- O. B. Shukur , M.H. Lee , (2015) [6] , In this paper, the hybrid artificial neural network (ANN) and autoregressive (AR) method is studied for finding the missing values. They use ANN for finding the missing values in wind speed data with nonlinear characteristic and they use AR model for determining the structure of the input layer for the ANN. They use Lisewise deletion

before AR modeling to handle the missing values. A case study is carried out using daily Iraqi and Malaysian wind speed data. They compare their prosed imputation method with linear, nearest neighbor, and state space methods. The comparison has shown that AR-ANN outperformed the classical methods. As a result they conclude that the missing values in wind speed data with nonlinear characteristic can be impute more accurately using AR-ANN. Therefore, imputing the missing values using their algorithm tends to more accurate performance of time series modeling and analysis.

3.Missing Value

Many databases in the world such as governmental and non- governmental contain missing values (MVs) in their attribute values. MVs is a value for attribute that was lost in the recording process. There are various reasons for their lost, such as manual data entry errors, equipment fail and incorrect measurements. The process of preparing clean data usually requires a preprocessing stage in which the data is prepared and cleaned, in order to be useful for the knowledge extraction process. The simplest way of dealing with MVs is to delete the record that contains them from the data set. However, this method is not practical when the data contains a large number of records with MVs which make bias during the inference. MVs make data analysis difficult. The occurrence of MVs can also lead to serious problems for researchers. In fact, unsuitable handling of the MVs in data analysis may found bias and can result in ambiguous conclusions being drawn from a research study, and can also limit the generalizability of the research findings [5] .

Three types of problems are usually associated with MVs in DM [7] :

a) Efficiency loss.
b) Complexity in handling and analyzing the data.
c) Unfairness resulting from differences between missing and complete data.

2.1: Missing Data Mechanisms

The algorithm of missingness describes the relationship between the likelihood of a value being missing and the other variables in the data set. If Y perform the complete data that can be partitioned as (Y_{obs}, Y_{mis}) where Y_{obs} is the observed part of Y and Y_{mis} is the missing part of Y, and R be an indicator random variable (or matrix) indicating whether or not Y is observed or missing. Let R = 1 present a value which is observed and let R = 0 present a value which is missing. The statistical model for missing data is P (R\Y, Ø)
where Ø is the parameter for the missing data process. The mechanism of missingness is determined by the dependency of R on the variables in the data set [8] .

The following are different mechanisms of missingness [9] .

i- Missing Completely At Random (MCAR)

The first mechanism of missingness is a special case of MAR known as missing completely at random (MCAR). In this case, the mechanism of missingness is given by:

P (R\Y, Ø) = P (R, Ø) ………………... (1)

That is, the probability of missingness is not conditional on any observed or unobserved values in Y. One example of MCAR might be a computer malfunction that arbitrarily deletes some of the data values.

ii- Missing At Random (MAR)

The second mechanism of missingness is missing at random (MAR), this mechanism of missingness is given by:

P (R\Y, Ø) = P (R\Y_{obs}, Ø) ………… (2)

That is, the probability of missingness is only conditional on observed values in Y and not on any unobserved values in Y. A simple example of MAR is a survey where subjects over a certain age refuse to answer a particular survey question and age is an observed covariate.

iii- Not Missing At Random (NMAR)

The third mechanism of missingness is referred to as missing not at random (MNAR). This mechanism of missingness is given by:

P (R\Y, Ø) = P (R\Y_{obs},Y_{mis}, Ø) …….. (3)

This mechanism observed when the conditions of MAR are break so that the likelihood of missingness is dependent on Y_{mis} or some unobserved covariate. One instance of MNAR might be subjects who have an income above a certain value refusing to report an income in the survey. Here the missingness is dependent on the unobserved response, income.

2.2- Methods for Handling Incomplete Data (Missing Data)

There are different methods and strategies exists to handle missing-data. Managing missing data can be classified into three categories: tolerance, ignoring and imputation-based procedures.

a. Tolerance

The simplest method point to preserve the source entries in the incomplete mode. It may be a functional and computationally low cost solution, intime it requires the techniques to work strongly even if the data quality stays low .

b. Ignoring

Missing data obscurity often refers to “Case Deletion”. It is the most repeatedly applied procedure nowadays, this method undergo from a loss of information in the insufficient cases and risk of alignment if the missing data is not MCAR and it is lying down to reduce the data quality. The strength lies in the ease of application: deleting the records with missing values is done in two ways [10]

(i) List-wise/Case-wise Deletion(complete-case analysis):

Delete the entire records including missing values. The main hitch of this method is that the dataset may lead to large loss of data, which may result in high inexactness in particular if the main dataset is itself too small or the number of records that contain missing value is too large.

(i) Pairwise Deletion (available-case analysis):

insufficient records are deleted on an analysis-by-analysis basis, Unlike list wise deletion which deletes records that have missing values on any of the variables under analysis, pair wise deletion only deletes the specific missing values from the analysis (not the entire records) such that any given record may participate to some analyses but not to others.

c. Imputation

"Imputation is the process of replacing missing data with substituted values". Missing data create problems for analyzing data; imputation is seen as a way to avoid difficulty involved with list wise deletion of records that have missing values. That is, when one or more observations are missing for a case, statistical applications default to discard any records that has a missing value, which may introduce partiality in the results. Imputation preserves all records by replacing missing values with an estimated value based on other information. When all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data

4: Proposed Algorithm

Our proposed algorithm depends on improving existing algorithm (MIGEC) proposed by [4] after adding the following steps:

1- Converting incomplete data set to binary dataset.
2- GRA based on attribute instead of instance.
3- Attribute merging instead of instance merging.
4- After each missing elements of attributes imputed

By mean imputation, next times we use the result of new imputation (imputation by PA) instead of mean to calculate imputation of reminders missing values of specific attribute.

The procedure of the proposed algorithm is schematized in Figure (1).

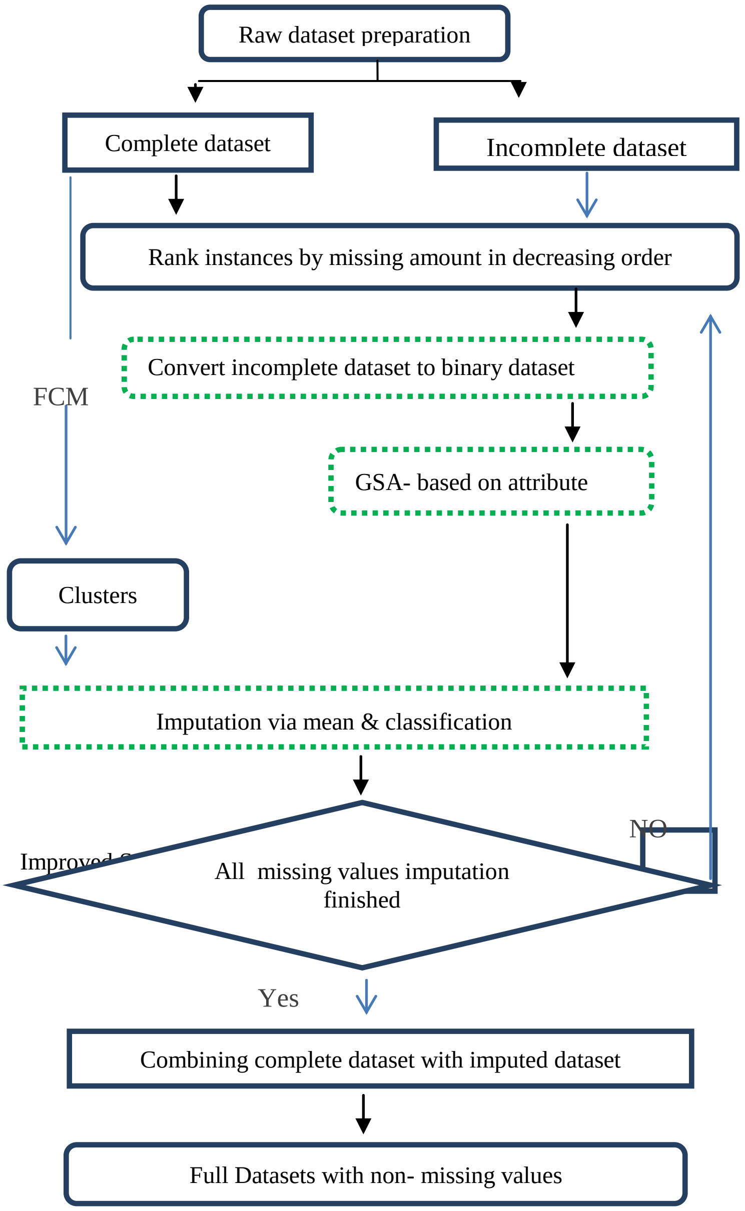

Raw dataset preparation
Complete dataset
Incomplete dataset
Rank instances by missing amount in decreasing order
Convert incomplete dataset to binary dataset
FCM
GSA- based on attribute
Clusters
Imputation via mean & classification
Improved G
All missing values imputation finished
NO
Yes
Combining complete dataset with imputed dataset
Full Datasets with non- missing values

Figure (1): The Diagram of the Proposed Algorithm

4.1 Steps of Algorithm

Let X_{ic} denote an incomplete dataset with n attribute $X_{ic} = \{x_1, x_2, \ldots\ldots, x_n\}$ and m instances. For each elements of incomplete dataset is defined by A_{ij} $i = 1,2, \ldots..m$ and $j = 1,2, \ldots..n$, it contains two parts: $A_{ij} = \{a_{ij}^{obs}, a_{ij}^{mis}\}$ where a_{ij}^{obs} is observed values and a_{ij}^{mis} is missing values.

$$\begin{array}{ccc} & \text{n Attributes} & \\ A = & \begin{bmatrix} a_{11} & a_{12} & \ldots a_{1n} \\ a_{21} & a_{22} & \ldots a_{2n} \\ \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} \ldots & a_{mn} \end{bmatrix} & \text{m Instances} \end{array}$$

A binary matrix (R) from incomplete dataset (X_{ic}) in which converting each observed values (a_{ij}^{obs}) to one and each missing values (a_{ij}^{mis}) to zero is produced, in this case (R) becomes a matrix of missing data indicators, when this R matrix has the same number of rows and columns as the data matrix (A) .

$$R_{ij} = \begin{cases} 1 \text{ if } a_{ij} \text{ is observed} \\ 0 \text{ if } a_{ij} \text{ is missing} \end{cases}$$

For example:

$$A= \begin{bmatrix} 9 & NA & 2 \\ NA & 10 & 8 \\ 3 & NA & NA \end{bmatrix} \rightarrow R= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

- NA=Missing Value (Not Available)

After each time that one attribute has been assigned to the most nearest cluster via (GST) in our proposed improvement, finally one instance inserted to binary matrix and called it class (target): $R = (r_{ij})_{m\times n}$ associates with the data matrix of the cluster.

Then imputation technique starts as follows:

Step 1. Calculate Expected information (Entropy) after partitioning each instance due to class.

$$\text{Enropy}(m_j) = -\sum_{i=1}^{k} p_i \log_2 p_i \text{ ………. (4)}$$

$$= -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots\ldots p_k \log_2 p_k)$$

Where :

- p_i is the likelihood of event i occurring .
- m is number of records
- k is number of clusters ($k \geq 2$)

Information needed after split m due to j

$$I_f = \text{Info}(m = 0|1) = \sum_{j=1}^{b} \frac{|D_j|}{|D|} * \text{Enropy(mj)} \qquad (5)$$

- $b = 2$ (b is represent 0&1)
- $f = 1,2,3, \ldots m$ is No. of instances

Step 2. The coefficient of difference for the f^{th} records computed :

$$t_f = 1 - I_f \qquad f = 1,2, \ldots., m \text{ ……….. (6)}$$

t_f Performs the attendant contrast intensity of the f^{th} parameter. The greater value of t_f express the more important parameter.

Step 3. Extract the coefficient of weight for the f $^{\text{th}}$ copy:

$$w_f = \frac{t_f}{\sum_{f=1}^{m} t_f} \quad \text{…………………..… (7)}$$

Step 4. The mean mode substitution (MMS) is appoint to set missing values in the first imputation. The straightforward technique could implement well only when the data is normally separated.

Then, estimate the jth attributive missing value of x_i^{mis} :

$$x_{ij}^{mis} = \sum_{q=1,q\neq j}^{m} w_q x_{iq}^{mis} \qquad \text{……..……..(8)}$$

After each missing elements of attributes imputed by mean imputation, next times the result of new imputation (imputation by PA algorithm) used instead of mean to calculate imputation of reminders missing values of specific attribute.

5: Experimental Results

In this section Experimental results of proposed algorithm for both Wine and simulated dataset are displayed and discussed, also comparison between both proposed algorithms with other previous techniques for dealing with missing data is described.

5.1 :Wine Data Set

The first dataset that we depend on to implement our algorithm is the Wine dataset that we used in this paper is achieved from The UCI (University of California, Irvine) Machine Learning Repository database, this dataset contain 13 attributes and 178 instances. The purpose of selecting this dataset is to compare the efficiency of our algorithm with previous (MIGEC) algorithm implemented by (J. Tian , B. Yu , D. Yu , Sh. Ma) [4] .

5.2: Simulated data

Data mining already work with massive quantities of data. For this reason we simulated data to know the performance of (proposed algorithm) with large amount of data. We used simple random samples of size 1000. We consider simulations under a normal distribution, this dataset contain 13 attributes and 1000 instances and all of them are numeric. The computer limitation (Intel(R) Core(TM) i5 CPU M 430 @ 2.27GHz) for implementing our algorithm doesn't allow us to increase the amount of simulated data.

We used (rnorm) command from (R programing language) to generate 1000 sample of data under normal distribution based on mean and standard deviation of Wine dataset .

5.3 Generating missingness:

To introduce artificial missingness, we look for two important factors which may affect the imputation results: missing rate and missing data mechanism. Three different levels of missing rate were considered, i.e., 5%, 10% and 20% and three missing mechanisms were taken into consideration, namely MCAR, MAR and NMAR.

For MCAR , In order to simulate missing values on attributes, the original datasets are run using a random generator and every data in the dataset have the same likelihood α to be missing, where α was the missing rate . "Nonparametric Missing Value Imputation using Random Forest" package from R programing language used to generate MCAR.

Simulating MAR was more challenging and it worked as follow: In case there is a complete dataset with two attributes (A_j, A_s), where A_j was the attribute in to which missing values were introduced, and A_s was the attribute that affected the missingness of A_j . Given a pair of attributes(A_j, A_s), and missing rate α , First split the instances into two equal-sized subsets according to their values at A_s and find the median a_s^{med} of A_s and then assigned all the instances into two subsets according to weather the instances have lower (or higher) values than the median a_s^{med} at A_s , $\mathrm{pr}\big(A_j = \mathrm{missing}\backslash A_s \leq \mathrm{median}(A_s)\big) = 4 \propto$. After the splitting of instances, randomly selected one subset of instances and let their values at A_j to be missing with the probability of 4α. The probability of 4α will result in a missing rate of 2α on the whole variable A_j which is equivalent to have a missing rate of α on the two variables (A_j, A_s) . For multi-attributes pair selection of attributes was based on high correlations among the attributes, different pairs of attributes were used to generate the missingness. Each attribute is paired with the one it is highly correlated to.

The process of generating missing values by NMAR was similar to MAR. The only difference was that there was no need to split variables into pairs, NMAR produced missingness on every variable directly. For a given variable A_j and specified missing rate α , first calculated the median a_j^{med} of A_j and then randomly let the values that are lower (or higher) than a_j^{med} to be missing with probability of 2α .

5.4 :Performance Measure

To evaluate the precision of various data imputation algorithms the Root Mean Square Error (RMSE) used in this paper .

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i - \hat{e}_i)^2} \quad \ldots\ldots\ldots\ldots\ldots (9)$$

Where e_i is the original value, $\hat{e}_i$ is the predicted plausible value, m is the total number of estimations and SD(e) is standard deviation. The larger value of RMSE suggests the less accuracy that the algorithm holds.

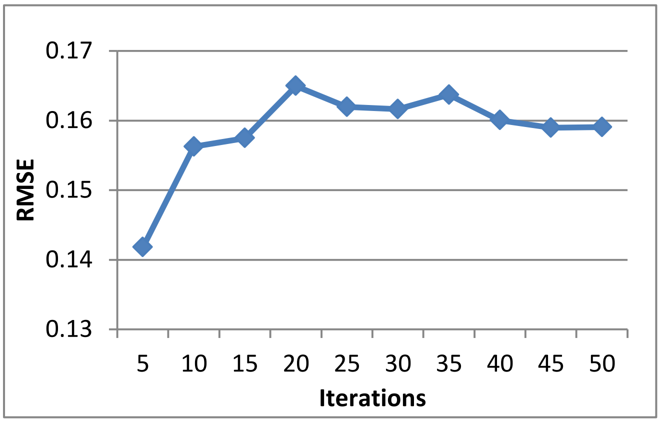
Before the comparative demonstrations, to capture the result of the data imputation accurately, it is requisite to select the optimum values for number of iteration (number of imputations) and clusters, by another expression mean that which clusters or iteration gives us minimum RMSE.

5.4.1 Number of Iteration

Firstly we tested for number of iterations for each missingness mechanism (MCAR, MAR, NMAR) with 10 % missing rate and five clusters as initial value (cluster), for wine dataset.

For MCAR:

As seen in Figure (2), the RMSE declines to the least which is (0.1418) when number of iteration is 5 times. So it is the optimum iteration for MCAR.



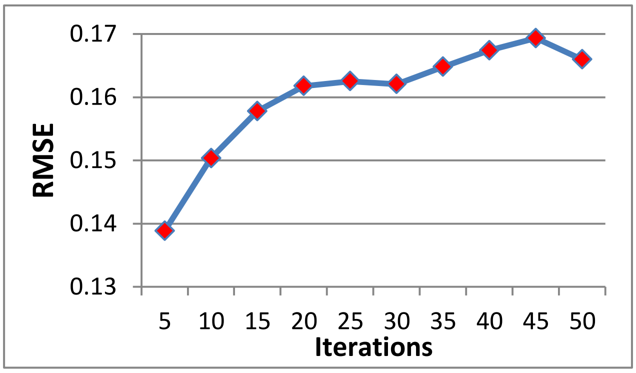


Figure (2): Checking optimality by number of Iterations (for MCAR)

For MAR

Figure (3), illustrate that the best number of iteration for MAR also is 5 which give minimum RMSE which is (0.1388).

Figure (3): Checking optimality by number of Iterations (for MAR)

For NMAR:

Finally, for NMAR as appeared in Figure (4), Iteration (10) give us lower RMSE which is (0.1347) and the worst iteration, which yield the maximum RMSE is iteration (15).

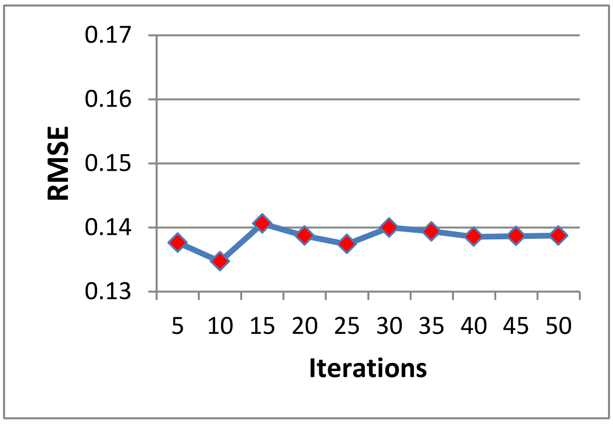


Figure (4): Checking optimality by number of Iterations (for NMAR)

5.4.2 Number of Clusters

Second step for checking optimality is number of clusters, we obtained it via (FCM) by applied it to complete dataset. we can match number of clusters that is affect the accuracy of our algorithm for imputation directly, since we used it for calculating classification therefore classification is a main part of imputing incomplete data in proposed algorithm.

For each three types of missing mechanism (MCAR, MAR, NMAR), we checked for optimal number of clusters from 2 to 10 by using %10 missing rate.

For MCAR:

Figure (5), shows that when the whole data is mass into 7 groups, the RMSE fall to the minimum that is (0.1330). In contrast the worst value obtained when 2 clusters exist.

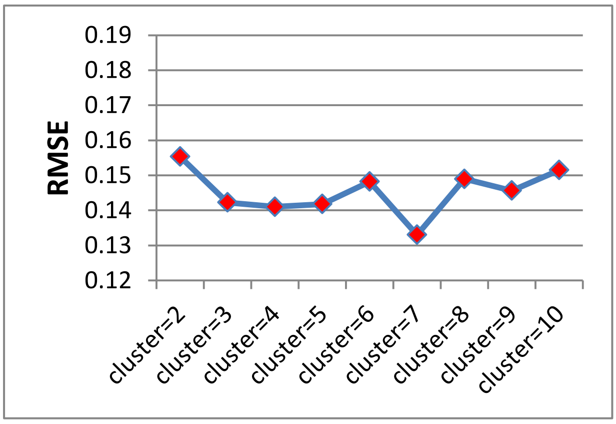


Figure (5): Checking optimality by number of clusters (for MCAR)

For MAR:

In Figure (6), MAR performs best when data partitions into 2 groups, it`s RMSE is (0.1321).

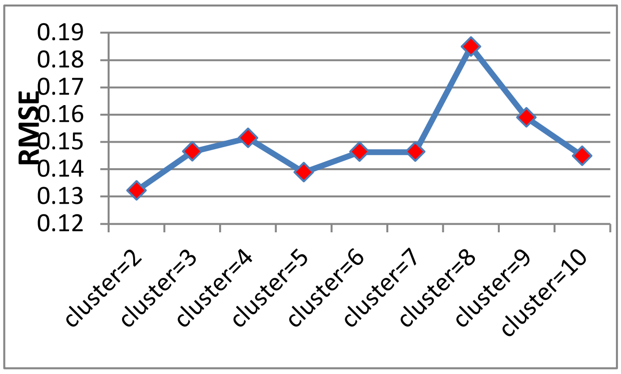


Figure (6): Checking optimality by number of clusters (for MAR)

For NMAR:

From Figure (7), results of RMSE for NMAR are between [0.12 - 0.15], minimum RMSE (0.12) yield when number of clusters is 8.

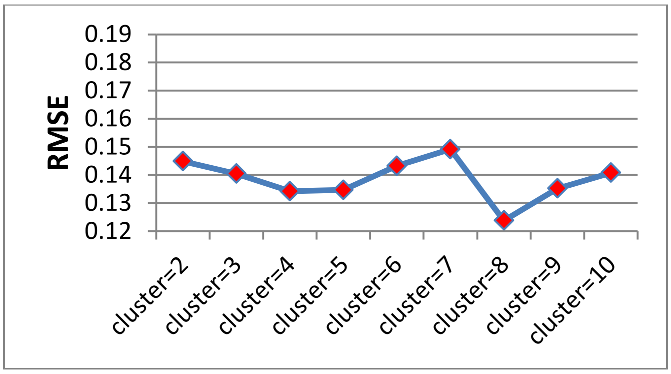


Figure (7): Checking optimality by number of clusters (for NMAR)

6 Comparative experiments

In investigation of making comparisons as extensively as possible, we select eight other approaches, which are MMS (Mean Mode Substitution), HDI (Hot Deck Imputation), KNNMI (K Nearest Neighbour Imputation with Mutual Information), FCMOCS (Fuzzy C-Mean based on Optimal Completion Strategy), CRI (Clustering-based Random Imputation) , CMI (Clustering-based Multiple Imputation) , NIIA (The Non Parametric Iterative Imputation) and MIGEC (Multiple Imputation algorithm using Gray System Theory and Entropy based on Clustering) .

After selecting optimum number of clusters and iterations for all three types of missingness, we compared proposed algorithm with various methods with varying missing rates by using RMSE (average of each RMSE) as displayed in Figures (8), (9) and (10).

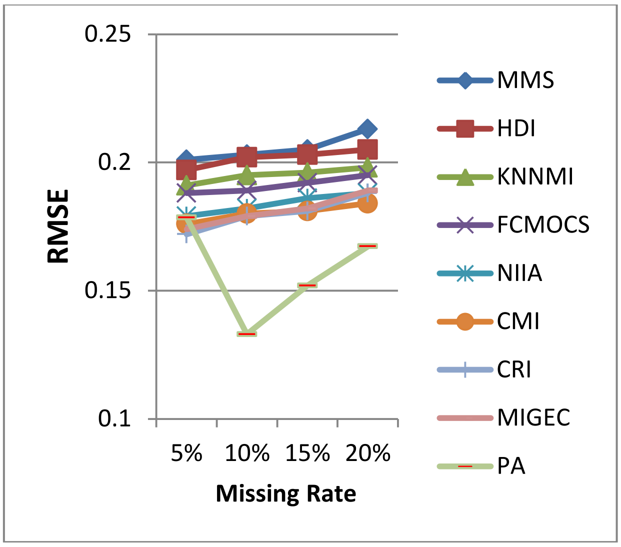


Figure (8): Comparison between proposed algorithm (PA) and other methods for imputation (MCAR)

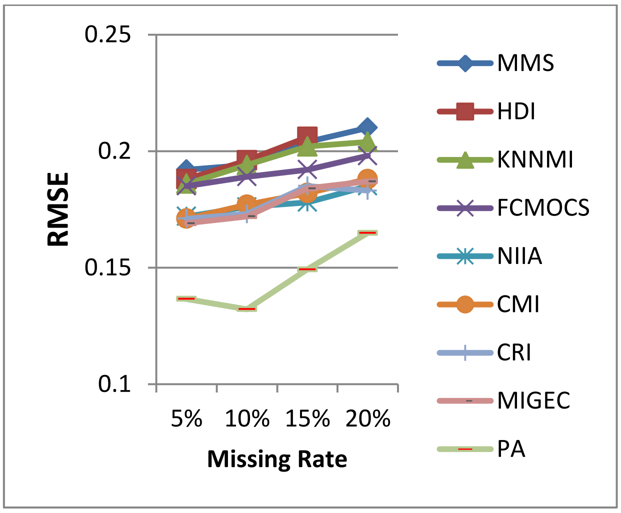


Figure (9): Comparison between proposed algorithm (PA) and other methods for imputation (MAR)

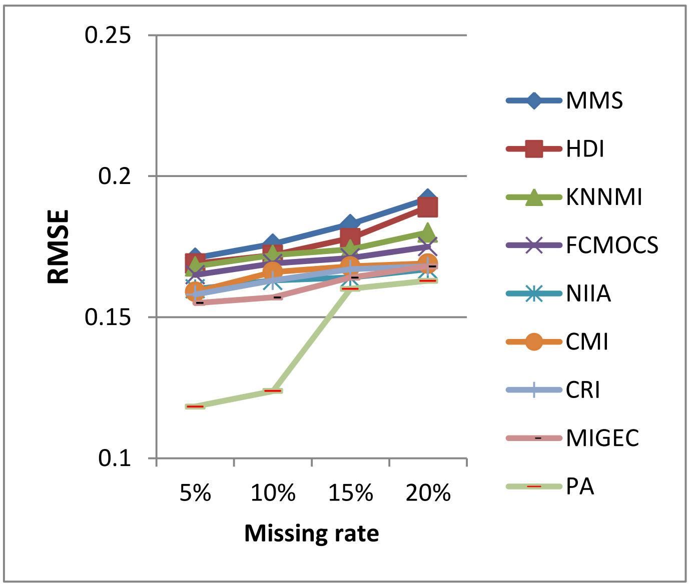


Figure (10): Comparison between proposed algorithm (PA) and other methods for imputation (NMAR)

Figures (8), (9) and (10) show some results that we would like to discuss as follows:

i- Outcome results demonstrate that the proposed algorithm performs better than the other eight approaches under all missingness mechanisms at varying missing rates.

ii- Different missing rate have different impacts on imputation accuracy. The RMSE increases with increasing missing proportion for all the methods approximately. This is understandable because with more missing rate introduced into the datasets, more information of data will be loss. However sometimes the nature of data, outlier and noise also effect on accuracy of imputation.

iii-The worse RMSE achieved by methods are for MCAR mechanism, followed by MAR and MCAR mechanism..

6: Conclusions and Future works

6-1: Conclusions

The problem of incomplete data is one which researchers must handle it. Many researchers fail to consider missing values of varying natures in their analyses, treating them as a singular type or not considering the impact of the missing values at all. In this paper an extension algorithm based on MIGEC for dealing with incomplete data has been proposed.

The experimental results show

1- Experimental results on wine dataset from University of California Irvine (UCI) repository illustrate the superiority of proposed algorithm to other imputation methods on accuracy of imputing missing data on three different missing types MCAR, MAR and NMAR.

2- The RMSE shows that our proposed algorithm has better results (namely, the minimal value of RMSE) than MIGEC algorithm, with average absolute difference beyond (0.025108).

3- When calculating GRA on attributes instead of instances we work with more homogeneous values in comparison with calculating GRA based on instances and as a result the attribute belong to proper cluster.

4- Increasing rate of missing records suffer the precision of the fulfillment in RMSE. It states that incomplete values negatively impact on the completion.

5- Proposed algorithm can handle missing values and perform better either with small or huge amount of the raw data, we can conclude that proposed algorithm remain stable with increasing the size of dataset which means our proposed algorithm is suitable for large data repositories.

6- Proposed algorithm reach results with less imputed iterations in comparison with other algorithms which means less run time needed in case of huge amount of data in data repositories.

7- The drawback of our proposed algorithm on MIGEC is appeared in cases when there is large amount of heterogeneity inside the attributes since GRA in our proposed algorithm depends on attribute values instead of instance values. This conclusion appears when we run the algorithm on difference simulated data.

6.2: Future Works

1- Working with data mining techniques need powerful computer to implement our work speedily and not restrict us. In this paper because of computer limitation we cannot increase the size of simulated dataset because it needs days to get results of proposed algorithm with vast size of dataset.

2- Hybrid proposed algorithm with another data mining or statistical techniques like (Neural Network, Nearest Neighbor, …).

3- Extending proposed algorithm to work with categorical attributes.

4- Distortion and noise has a great effect on the imputation techniques, while real-world data often contain much noise, therefore, another preprocessing algorithm can be implemented to clean the data before implementing PA.

5- Implementing different data mining algorithms such as association rule mining on PA and compare the results with other existing algorisms.

Biography

Nzar Ali: is an Associate Professor of Computer Science at the Cihan University, and his Ph.D. degree from the Sulaimaniya University. His research interests in data mining, spatial database and database indexing. He has published over 9 research papers and supervise one Phd and 6 Msc Students.

.

Zhyan Omer: is an Associate Lecture of Statistics and Informatic at the Sulaimaniya University and her MSc. degree from the Sulaimaniya University Her research interests include statistical data mining and missing value estimation .

REFERENCE

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth ,(1996) *"From data mining to knowledge discovery"*, American Association for Artificial Intelligence, San Francisco, Vol. 17, No. 3.

[2] R. Nisbet , J. Elder, G. Miner , (2009) *"Handbook of Statistical Analysis and Data Mining Applications"* . Academic Press, Boston.

[3] J. Han and M. Kamber, (2011) *"Data Mining: Concepts and Techniques",* Morgan Kaufmann,San Francisco .

[4] J. Tian , B. Yu , D. Yu , Sh. Ma , (2014) *"A hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering"* Appl Intell 40:376–388, DOI 10.1007/s10489-013-0469-x, Springer Science+Business Media New York.

[5] X.Y. Zhou , J. S. Lim , (2014) *"Replace Missing Values with EM algorithm based on GMM and Naïve Bayesian"* International Journal of Software Engineering and Its Applications Vol.8, No.5, pp.177-188.

[6] O. B. Shukur, M.H. Lee , (2015) *"Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method ",* Published by Canadian Center of Science and Education ,Modern, ISSN 1913-1844 E-ISSN 1913-1852, Applied Science; Vol. 9, No. 11 .

[7] J.Barnard, X.Meng, (1999) *"Applications of multiple imputation in medical studies: from aids to nhanes",* Stat. Methods Med. Res. 8(1), 17–36 .

[8] J.A. Boyko, (2013) *"Handling Data with Three Types of Missing Values" ,* Ph.D. Thesis , University of Connecticut .

[9] R.J.A. Little, D.B. Rubin, (1987) *"Statistical Analysis with Missing Data"* , 1st edn. Wiley Series in Probability and Statistics, New York.

[10] S .Zhang , (2011) *"Shell-neighbor method and its application in missing data imputation ",* Appl Intell 35(1):123–133.