Empirical Performance Evaluation of Imputation Techniques using Medical Dataset

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012055 doi:10.1088/1757-899X/551/1/012055

Empirical Performance Evaluation of Imputation Techniques using Medical Dataset

O A Alade[1,2] R.Sallehuddin[1] and A.Selamat[1]

[1]School of Computing, Universiti Teknologi Malaysia, Skudai, Johor, 81310, Malaysia
[2]Computer Science Department, Federal Polytechnic, Bida, 912101, Nigeria

E-mail : kaleabel@gmail.com , roselina@utm.my and aselamat@utm.my

Abstract. This paper evaluates the error measures of missing value imputations in medical research. Several imputation techniques have been designed and implemented, however, the evaluation of the degree of deviation of the imputed values from the original values have not been given adequate attention. Predictive Mean Matching Imputation (PMMI) and K-Nearest Neighbour Imputation (KNNI) techniques were implemented on imputation of fertility dataset. The implementation was on three mechanisms of missing values: Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not At Random (MNAR). The results were evaluated by mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE). PMMI performed better than KNNI in all the results. MSE for example, has the ratio of 0.0260/2.8555 (PMMI/KNNI) for 1-10% MAR – 99.09% reduced error rate; 0.1108/3.0120 (PMMI/KNNI) for 30-40% MCAR – 96.32 reduced error rate; and 0.0642/3.7187 (PMMI/KNNI) for 40-50% MNAR – 98.27% reduced error rate. MCAR was the most consistent missingness mechanism for the evaluations. Density distributions of the imputed dataset were compared with the original dataset. The distribution plots of the imputed missing data followed the curve of the original dataset.

1. Introduction

Data is the major operational facts of any organization including medical sector. However, medical datasets are blighted by missingness that characterize other data driven sectors. This makes discretization of medical data difficult for classifiers. Missing data comes from error of procedure, equipment error, measurement error, or respondents declining answer to queries of personal information. Most classifiers lack inbuilt routine to handle datasets with missing values, therefore there is need for explicit software for missing data imputation.

Several imputation techniques have been used to explicitly impute missing data. Zero substitution [1] has been proposed, but it is unacceptable by research community because it leads to wrong conclusion. Mean imputation is another approach [4], this approach is simple but it also undermines the variance and standard deviation of the imputed data. Support vector regression performed effectively well, but it was solely based on regression. Gaussian mixture model and extreme learning machines (GMM-ELM) was proposed as a reliable approximation technique for imputing missing data by Sovilj *et al.* in [5]. The result of their work improved imputation of missing values over mean imputation technique, however the evaluation of imputation result was not sufficiently investigated.

Pohar-Perme method [6] was used for the imputation of missing data in colorectal cancer dataset. They recommended sensitivity analysis of the results, but they did not evaluate the error measure of their technique. Hybrid of fuzzy c-means and GA was used in [7], but it is computationally complex. Monte Carlo simulation method was used to examine the performance of Bayesian imputation technique [8]. They concluded that the performance of Bayesian imputation technique depends on risk factors and the

mechanism of missingness. This work of [8] motivated part of this work. KNNI is a state-of-the-art imputation technique. It is simple and works well with Gaussian distribution [9]. Also, PMMI imputes values that are like the original [10]. It is used to construct metrics matching of missing data instances with the observed instances. Despite all the advantages of KNNI and PMMI, there are less emphases on their evaluation of their performances [10].

The major missing patterns fall in three basic characteristics of missingness namely, Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR). Most researchers concentrate on MAR and MCAR, assuming that it is proper to find, and discard MNAR values in advance. During the analysis of incomplete datasets, it is necessary to consider the process that generates the missing values, and perform valid analyses based on that assumptions.

This work therefore posits to evaluate performance imputation of PMMI and KNNI based on the MAR, MCAR and MCAR.

2. Materials and Methods

Fertility medical dataset from UCI machine learning repository was used in this work. The dataset describes the quality of semen sample of 100 volunteers using artificial intelligence. The dataset consists of nine (9) continuous predictor variables and a binary class decision variable. The predictor variables are childish disease (C_disease), frequency of alcohol (Alcohol_Freq) and smoking habit (Smok_hab). The simulations were repeated for all the three mechanisms of missing values, that is MAR, MCAR and MNAR. All methods in this study were implemented in RStudio version 1.1.456 2018 on a HP computer with Core i3-4030U CPU @ 1.9Hz X64-based processor, RAM 6.00GB, TOSHIBA MQ01AB075 HDD and 64-bit Windows OS.

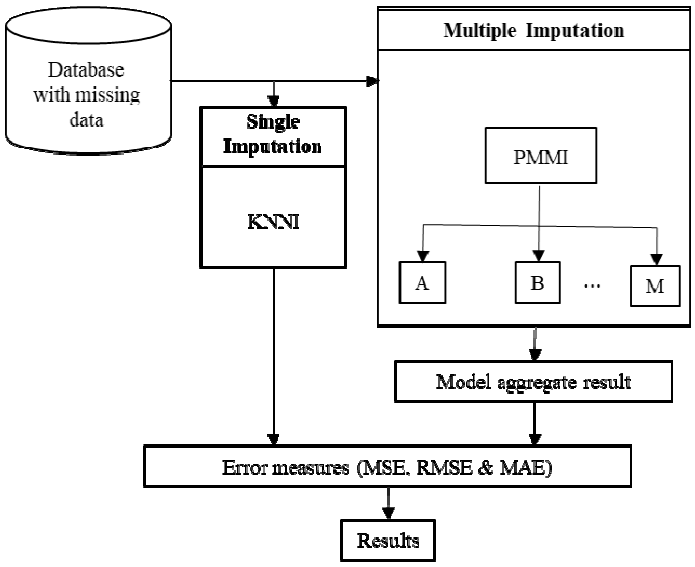


Figure 1: Framework of the proposed evaluation models for KNNI and PMMI imputation techniques.

2.1 Proposed evaluation models for KNNI and PMMI imputation techniques

Error! Reference source not found. shows the proposed framework of this study. Two methods of missing value imputations were considered – single imputation and multiple imputation. KNNI was used for single imputation, while PMMI was used for multiple imputation. The imputed datasets are labelled A, B, …, M. Multiple imputation algorithm generated five (5) imputed datasets for each run of 1-10%, 10-20%, 20-30%, 30-40%, and 40-50% of MAR, MCAR and MNAR. A sum of seventy five (75) imputed datasets were generated. The five imputed datasets for each run were aggregated; then, the error measures were used to evaluate the performance of the imputation models. The error performance for

PMMI and KNNI are compared and evaluated using MSE, MAE and RMSE. The method that produce small error indicates the better performance.

3. Result of the Evaluation

The evaluation results of the predictive mean matching imputation (PMMI) and k-nearest neighbour (KNNI) techniques are shown in Table 1. It shows different percentages of missing values ranging from 1 to 50% with and interval of 10%. The evaluations were based on the MAR, MCAR and MNAR using three error evaluation metrics that is MSE, RMSE and MAE. From the table, it shows that PMMI perform better than KNNI in all missing data mechanisms and in all missing values ratio. From the results in figures 2 – 4, the most obvious trend is the great disparity of error measures between PMMI and KNNI. PMMI has a sharp reduced error measures in the order of 10^{-1} - 10^{-2} than its counterpart KNNI. MSE for example, has the ratio of 0.0260/2.8555 (PMMI/KNNI) for 1-10% MAR – 99.09% reduced error rate; and 0.0642/3.7187 (PMMI/KNNI) for 40-50% MNAR – 98.27% reduced error rate. This high rate is observed for all the characterics of missing data under study. This shows that imputation accuracy of PMMI is higher than KNNI in all the experiments. This better performance of PMMI over KNNI may also be a proof of superiority of multiple imputation over single imputation model.

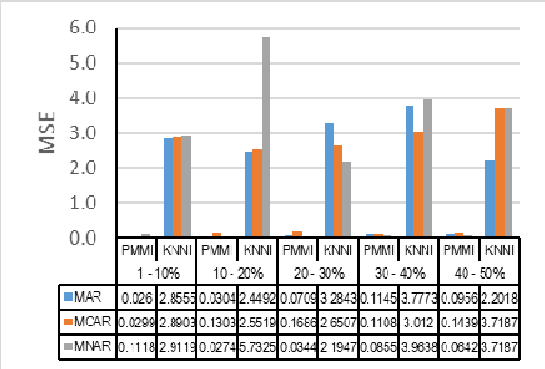
The error measures apparently increase for MAR, MCAR and MNAR as the percentage error increases for both PMMI and KNNI. In fact, looking at the figures 2-4, MCAR is relatively more consistent among the three mechanisms for every error measure, while MNAR is unstable for the two imputation algorithms with changes in the percentages of missing values. This may account for why MAR and MCAR are the choices of many researchers in the study of imputation of missing data. The slight fluctuations observed in MAR and MCAR may be due to the randomness of seed[3] which is subject to further study.

Table 1: MSE, RMSE, and MAE evaluations of missing data for MAR, MCAR and MNAR characteristics with 1 - 50% range of missingness at an interval of 10.

		MSE			RMSE			MAE		
% MISSING		MAR	MCAR	MNAR	MAR	MCAR	MNAR	MAR	MCAR	MNAR
1 - 10%	PMMI	**0.0260**	0.0299	0.1118	0.1613	0.1729	0.3343	0.1613	0.1729	0.2554
	KNNI	**2.8555**	2.8903	2.9119	1.6898	1.7001	1.7064	2.7164	2.2792	2.0889
10 - 20%	PMMI	0.0304	0.1303	0.0274	0.1744	0.3610	0.1655	0.1744	0.2684	0.1655
	KNNI	2.4492	2.5519	5.7325	1.5650	1.5975	2.3943	2.1981	2.1812	4.1111
20 - 30%	PMMI	0.0709	0.1686	0.0344	0.2663	0.4107	0.1854	0.2053	0.2943	0.1854
	KNNI	3.2843	2.6507	2.1947	1.8123	1.6281	1.4815	2.6447	2.5188	1.8922
30 - 40%	PMMI	0.1145	**0.1108**	0.0855	0.3384	0.3329	0.2925	0.2514	0.2417	0.2325
	KNNI	3.7773	**3.0120**	3.9638	1.9435	1.7355	1.9909	2.9120	2.4720	2.8062
40 - 50%	PMMI	0.0956	0.1439	**0.0642**	0.3092	0.3794	0.2534	0.2044	0.2772	0.2179
	KNNI	2.2018	3.7187	**3.7187**	1.4839	1.9284	1.9284	2.0802	2.8706	2.8706

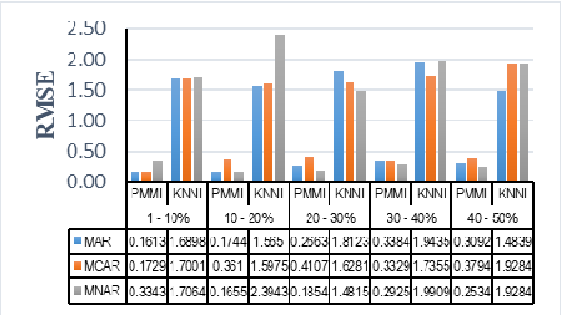
NB: C_Disease – Childish disease, Alcohol_Freq – Alcohol frequency, and Smok_Hab – Smoking habit

MCAR was chosen to compare the distribution of multiple imputation with the original dataset for the percentage ranges of the missing values (1—10%, 10-20%, …, 40-50%). This is because it was relatively consistency of as equally observed. Figures 5 represents the distribution charts used to compare imputed and the original dataset. The distributions follow the curve of the original data; i.e. the resultant imputation was very close to the original data, leading to a high degree of accuracy.



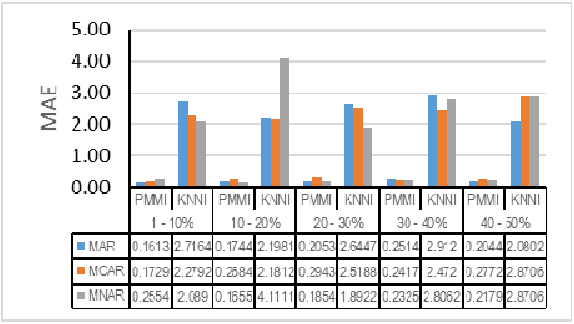
	PMMI	KNNI	PMM	KNNI	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI
	1 - 10%		10 - 20%		20 - 30%		30 - 40%		40 - 50%	
MAR	0.026	2.8555	0.0304	2.4492	0.0709	3.2843	0.1145	3.7773	0.0956	2.2018
MCAR	0.0299	2.8903	0.1303	2.5519	0.1686	2.6507	0.1108	3.012	0.1439	3.7187
MNAR	0.1118	2.9119	0.0274	5.7325	0.0344	2.1947	0.0855	3.9638	0.0642	3.7187

Figure 2: Comparative mean square error (MSE) of predictive mean matching imputation (PMMI) model and k-nearest neighbour imputation (KNNI) model for MAR, MCAR and MNAR with increase in percentage missingness.



	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI
	1 - 10%		10 - 20%		20 - 30%		30 - 40%		40 - 50%	
MAR	0.1613	1.6898	0.1744	1.555	0.2563	1.8123	0.3384	1.9436	0.3092	1.4839
MCAR	0.1729	1.7001	0.351	1.5975	0.4107	1.6281	0.3329	1.7355	0.3794	1.9284
MNAR	0.3343	1.7064	0.1655	2.3943	0.1854	1.4815	0.2925	1.9904	0.2534	1.9284

Figure 3: Comparative root mean square error (RMSE) of predictive mean matching imputation (PMMI) model and k-nearest neighbour imputation (KNNI) model for MAR, MCAR and MNAR with increase in percentage missingness.



	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI	PMMI	KNNI
	1 - 10%		10 - 20%		20 - 30%		30 - 40%		40 - 50%	
MAR	0.1613	2.7164	0.1744	2.1981	0.2053	2.6447	0.2514	2.9172	0.2044	2.0802
MCAR	0.1729	2.2792	0.2584	2.1812	0.2943	2.5188	0.2417	2.472	0.2772	2.8706
MNAR	0.2554	2.089	0.1655	4.1111	0.1854	1.8922	0.2325	2.8062	0.2179	2.8706

Figure 4: Comparative mean average error (MAE) of predictive mean matching imputation (PMMI) model and k-nearest neighbour imputation (KNNI) model for MAR, MCAR and MNAR with increase in percentage missingness.

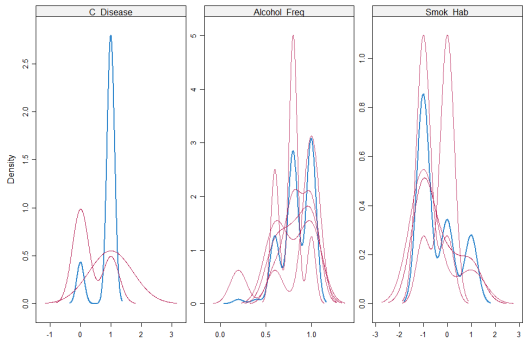


Figure 5: Density plot represents the missing value distributions for missingness in Fertility dataset. The blue line showing distribution of the original data plot, while the magenta lines shows the imputation

4. Conclusion

This work considered missing values as a serious problem in medical research that needs to be handled before classification, analysis and inferences for diagnosis. It focused on evaluation of error measure to determine the performance of imputations techniques in medical classification. The underlining causes of missingness were discussed, two state-of-the-art imputation techniques (PMMI and KNNI) were used to impute missing values in a dataset. The results of the imputations were empirically evaluated with three error metrics – MSE, RMSE and MAE – to measure and compare the degree of error in the imputed datasets under MAR, MCAR, and MNAR. The results showed that PMMI is more promising under all the mechanism of missing pattern than KNNI, and MCAR provide a better optimal missing pattern, while MNAR was the least. The density plots also showed that the distribution of the imputed datasets followed the distribution curve of the original dataset. However, the error measures of the imputation techniques were evaluated in their natural forms, not minding the correlation of the dataset. In future work, the effect of correlation on the error measures will be considered, this would be in a view to improve the algorithms' performances.

References

[1] Newgard CD *et al* 2015 *Jama* **314** 940.
[2] Garciarena U *et al* 2017 *Expert Syst Appl* **89** 52.
[3] Zeng D *et al IEEE 19th Int Conf e-Health Networking, Appl Serv* **1** 4.
[4] Liu Y *et al* 2017 *Data* **2** 8.

[5] Sovilj D *et al* 2015 *Neurocomputing* **174** 220.
[6] Falcaro M *et al* 2017 *Int J Cancer Epidemiol Detect Prev* **48** 16.
[7] Tang J *et al* 2015 *Transp Res Part C Emerg Technol* **51** 29.
[8] Austin PC *et al* 2005 *Comput Stat Data Anal* **49** 821.
[9] Wasito I *et al* 2006 *Comput Stat Data Anal* **50** 926.
[10] Gil D *et al* 2012 *Expert Syst Appl* **39** 12564.