Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Comparative Study of Missing Value Imputation Methods for Education Data

Phimmarin KEERIN
Faculty of Science and Technology, Pibulsongkram Rajabhat University, Thailand
k.song@psru.ac.th

Abstract: Missing data are often inevitable in real-world problems and indeed affect the overall result of research. Similar to other domains, missing values occurring in education data require a solid imputation to arrive at valid findings. As such, the objective of this research paper is to provide better understanding of the aforementioned issue as well as imputation methods, and to assess performance of benchmark alternatives on actual data. In particular, it aims to provide a comparative study, using various techniques of mean imputation, K-nearest neighbor (KNN) Imputation, Cluster-K-nearest neighbor (CKNN) Imputation, Local Least Square (LLS) Imputation, Cluster-base Local Least Square (CLLS) imputation, Iterated Local Least Square (ILLS) imputation and Bayesian Principal Component Analysis (BPCA) Imputation. The comparison is conducted on five real datasets of same sizes, under a missing completely at random (MCAR) assumption, and based on the evaluation metric of normalized root mean square error (NRMSE). The corresponding result suggests that BPCA and ILLS are two most effective imputation methods for these small-size datasets.

Keywords: Missing values, imputation methods, comparative study, educational data

1. Introduction

Datasets obtain from surveys, experiments, and administrative recodes usually contain missing data. Quality of data is main concern of researchers working in the filed of data science and data analytics. Although quality of result of the machine learning algorithm depends on several factors such as selection of algorithm, missing rate, feature selection and validation datasets. In various domains such as microarray experiments in cancer studies, survey responses in social science and marketing research, the problem of missing data has led to sub-optimal performance of machine learning models, most of which are designed to work only with a complete data (N. Iam-on, 2019). One approach to overcome this is to amend an algorithm with the ability to handle missing values on the fly, while the other relies on an imputation technique that is exploited to clean data prior analysis (F. Ridzuan, W. Zainon, 2019). Among these, a simple wayout is to get rid of samples with missing values, which is recommended only when a large size of data is available (T. Aittokallio, 2010). Besides, a heuristic of zero as well as other statistical representatives like means, maximum and minimum have been introduced to provide estimates of missing data entries (M. Pattanodom, N. Iam-On, T. Boongoen, 2016). Apart from the initial imputation methods mentioned thus far, a rich collection of learning-based models is proposed in the literature (e.g., P. Keerin, W. Kurutach, and T. Boongoen, 2016; Z. Zhu, J. Wang, and B. Sun, 2021). Appropriately dealing with missing values is important and challenging task because it requires careful examination of all instance of dataset to identify of missingness in the data.

Objective of this research these focuses on the concept of imputation that exploits a set of techniques to produce different estimates for a missing value and to assess performance of most common and widely used data imputation techniques namely mean imputation, K-nearest neighbor (KNN) Imputation, Cluster K-nearest neighbor (CKNN) Imputation, Local Least Square (LLS) Imputation, Cluster-based Local Least Square (CLLS) imputation, Iterated Local Least Square (ILLS) imputation and Bayesian Principal Component Analysis (BPCA) Imputation. This will help practitioners and data scientists to select appropriate data imputation method while carrying data mining task. Focus of this study is to analyze and compare performance of imputation methods for numeric dataset.

The rest of this paper is organized as follows. To start with, Section 2 presents the methodology and data collection. After that, Section 3 describes research methodology for comparison of imputation methods. Section 4 describes the experiment design including compared techniques and parameter settings, followed by the report of experimental results and discussion. The work is concluded in Section 5, along with the future-research direction.

2. Methodology

In this section, the proposed method to estimate the missing values of an observation based on valid values of other variables is called as data imputation (R. Little and D.B. Rubin, 1987). In general, there are three major families of these techniques: knowledge-assisted, global and local categories, respectively. Firstly, the knowledge-assisted approach integrates domain knowledge or external information into the imputation process. This has been reported to be useful for a data set with small number of samples or with a high missing level, where both global and local data driven counterparts would become largely ineffective. Examples belonging to this family are Projection Onto Convex Set (X. Gan, A. Liew, and H. Yan, 2006), which is introduced to solve the problem in gene expression datasets. In cases with sufficient amount of data, the global approach can estimate missing values using a global correlation measurement extracted from the entire data matrix (Z. Zhu, J. Wang, and B. Sun, 2021). Well-known methods in this group include Singular Value Decomposition (SVD) (O. Troyanskaya et al., 2001) and Bayesian Principal Component Analysis (BPCA) (S. Oba et al., 2003).

In contrast, algorithms in the local category exploit only local similarity structure in a dataset for missing value imputation. Only a subset of samples that exhibits high correlation with the one containing missing values is used to approximate estimates. Examples of these include k-nearest neighbor imputation or KNNimpute (O. Troyanskaya et al., 2001), and local least square imputation or LLSimpute (H. Kim, G. Golub, and H. Park, 2005). Unlike those discussed above, a new trend to make a good use of data structure or clustering result to determine the scope of neighbor selection has recently emerged. In addition to this attempt, the selection of nearest neighbors can be constrained by clusters of data under examination, i.e., neighbors of any sample must belong to the same cluster. With such a concept, CKNNimpute (P. Keerin, W. Kurutach, and T. Boongoen, 2012) applies the k-means clustering technique to obtain those clusters that provide basis of improved KNNimpute. In fact, based on experiments with published gene expression datasets, it usually outperforms global methods like BPCA as well as those local techniques such as KNNimpute and its weighted modification. Despite this success, there is a gap for further improvement that is the organic combination of cluster-based nearest neighbor selection and multi-stage refinement of IKNNimpute. Besides this innovative method called Iterative-CKNN, another novel model focuses on the similar extension to LLSimpute, i.e., CLLSimpute (P. Keerin, W. Kurutach, and T. Boongoen, 2013).

2.1 Missing Data Mechanisms

Missing values are certain values in datasets that are not observed. It occurs in the phase of data collecting for various reasons, such as administrative error, defective technique, or technology failure. Most compelling evidence are an intended replication may be omitted, a feature of the robotic apparatus may fail, a scanner may have insufficient resolution, or an image may be corrupted. It is beneficial to classify missing values on the basis of the mechanism that produces them. Roughly all the causes of missing values can be classified by the following classification system based on the relationship between the missing values and data points that have been observed.

Missing completely at random (MCAR): Missing completely at random (MCAR) is the only missing data mechanism that can actually be verified. Missing data are MCAR when the probability of missing data on a variable is unrelated to any other measured variable and to the variable with missing values itself. The assumption of MCAR is that probability of the missingness depends neither on observed values in any variable of data nor on unobserved part of dataset.

Missing at random (MAR): Missing data are missing at random (MAR) when the probability of missing data on a variable related to some other measured variable in the model, but not to the value of the variable with missing values itself. Missing value of any of the variable in the dataset depends

on observed values of other variables in the dataset because some correlation exists between attribute containing missing value and other attributes in the dataset. The pattern of missing data may be traceable from the observed values in the dataset.

Missing Not at Random (MNAR): Data are missing not at random (MNAR) when the missing values on a variable are related to the values of that variable itself even after other variables controlling. In the same way, the problem with the MNAR mechanism is that it is impossible to verify the scores of MNAR without knowing the missing values. The pattern of missing data is not random and is non-predictable from observed values of the other variables in the dataset.

2.2 Imputation Method

In this paper, present a comprehensive evaluation on the performance of seven imputation algorithms on a wide variety of types and sizes of datasets, which assessed the performance of different algorithms on each dataset. Algorithms used can be divided into two categories: local imputation algorithms and global imputation algorithms. Local imputation algorithms select a group of data with the highest relevance to the target data to impute missing values. For local imputation algorithms, used mean imputation, k-Nearest-Neighbors (KNN), Cluster base k-Nearest-Neighbors (CKNN), local least squares (LLS), iterative LLS (ILLS) and Cluster-based LLS (CLLS). For global imputation algorithms, used Bayesian principal components analysis (BPCA). The KNN impute were run with the parameter k = 10, the CKNN algorithm was run with the parameter k = 5 for non-time series data. The parameter estimator was used for LLS and CLLS were run with the parameter k = 5. The BPCA and ILLS methods do not contain any free parameters. A brief information of these algorithms being used is presented in Table 1.

Table 1. *Missing Value Imputation Methods Used in This Study*

Methods	***Author***	***Year***	***Type***
Mean imputation	-	-	local
K-Nearest Neighbor (KNN) imputation	O. Troyanskaya et al.	2001	local
Cluster *K*-Nearest Neighbor (CKNN) imputation	P. keerin et al.	2012	local
Local Least Square (LLS) Imputation	H. Kim et al.	2005	local
Cluster Local Least Square (CLLS) imputation	P. keerin et al.	2013	local
Iterated Local Least Square (ILLS) imputation	Z. Cai et al.	2006	local
Bayesian Principal Component Analysis (BPCA) Imputation	S. Oba et al.	2003	global

2.3 Dataset

Considering that dataset from different species and type of dataset may have different effects on the quality of imputation algorithm. Before analyses, generate missing values entries with different missing rate (1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%) were randomly introduced into these complete data.

Table 2. *Missing Data Used for Imputation Methods Comparison*

Dataset	***Description***	***No. of instances***	***No. of attributes***
Edu	These data are the academic results of educational information at Mae Fah Luang University	151	32
Wine	These data are the result of a chemical analysis of wines grown	178	13
Ecoli	This data contains protein localization sites	336	7
Seed	Measurements of geometrical properties of kernels belonging to three different varieties of wheat.	210	7
Iris	The data are iris flowers contains 3 classes of 50 instances each, where each class refers to a type of iris plant.	150	4

2.4 Evaluation criteria

Similar to many previous studies, normalized root means square error or NRMSE (T. Aittokallio, 2010) is used to determine a goodness of imputation. It is based on the difference between values estimated by an imputation technique and their true values. Intuitively, the lower such a difference is the better the performance is. Formally, NRMSE is defined by the following. Note that x_{truth} is the actual value in the original data matrix, $x_{estimate}$ is the corresponding estimated value, $var(x_{truth})$ is the variance of the actual values. The lower NRMSE is, the better the value estimated by a computerized method becomes.

$$NRMSE = \sqrt{\frac{mean(x_{estimate} - x_{truth})^2}{var(x_{truth})}}$$

Each experiment setting (imputation method, dataset and missing rate) is repeated for 20 trials to generalize the results and comparison. Lower is value of NRMSE; better is estimate of missing values.

3. Results and Discussion

3.1 Experimental Result

This section describes evaluation of performance of seven different imputation methods namely mean imputation, k-Nearest-Neighbors (KNN), Cluster base k-Nearest-Neighbors (CKNN), local least squares (LLS), iterative LLS (ILLS), Cluster-based LLS (CLLS) and Bayesian principal components analysis (BPCA). According to Tables 3 to 7 that represent the NRMSE measures of the proposed against seven compared techniques on five datasets, BPCA usually delivers better estimates for missing values than other single-pass imputation methods. This improvement is obvious when the missing ratio is ≥ 5%, as shown in Tables 3, 5 and 7 that correspond to the experiments on edu, ecoli and iris. In particular to Table 4 and 6, NRMSE of ILLS at all missing rates have been decreased with BPCA from all missing rate, respectively. Note that the presented NRMSE values in these tables are the average across 20 trials of each experimental setting, with the corresponding standard deviations being shown in parentheses.

The NRMSE for each dataset for different percentage of imputed data using different imputation methods is calculated and given in the Tables 3–7. Each column in the table indicates percentage of imputed data and each row indicates method used for imputation of data. The value in bold indicates lowest NRMSE. It means that bold value indicates the imputation method that gives better imputation result when applied on the given dataset.

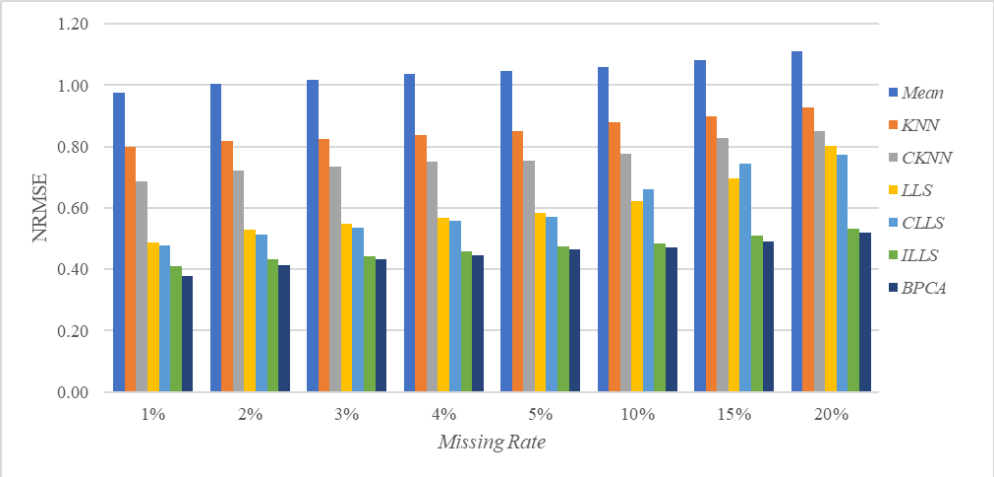


Figure 1. Method-specific NRMSE scores as averages across datasets and multiple trials, categorized by different rates of missing values: 1%, 2%, 3%, 4%, 5%, 10%, 15% and 20%, respectively.

Table 3. *NRMSE for Edu Dataset*

Method for missing imputation	***Percentage of missing Rate***							
	1%	2%	3%	4%	5%	10%	15%	20%
Mean	*1.0316*	*1.0321*	*1.0332*	*1.0342*	*1.0346*	*1.0374*	*1.0394*	*1.0430*
K-Nearest Neighbour (KNN)	*0.8030*	*0.8350*	*0.8549*	*0.8552*	*0.8656*	*0.9894*	*1.0507*	*1.0825*
Cluster K-Nearest Neighbour (CKNN)	*0.8306*	*0.8331*	*0.8483*	*0.8430*	*0.8458*	*0.8584*	*0.8624*	*0.8805*
Local Least Square (LLS)	*0.8307*	*0.8606*	*0.8622*	*0.8901*	*0.9201*	*0.9461*	*0.9579*	*0.9858*
Cluster Local Least Square (CLLS)	*0.7340*	*0.7566*	*0.7808*	*0.7518*	*0.7475*	*0.7940*	*0.8214*	*0.8355*
Iterated Local Least Square (ILLS)	*0.7604*	*0.7632*	*0.7756*	*0.7960*	*0.8040*	*0.8126*	***0.3530***	*0.8529*
Bayesian Principal Component Analysis (BPCA)	***0.5293***	***0.5356***	***0.5571***	***0.5871***	***0.6269***	***0.6343***	*0.6570*	***0.6863***

Table 4. *NRMSE for Wine Dataset*

Method for missing imputation	***Percentage of missing Rate***							
	1%	2%	3%	4%	5%	10%	15%	20%
Mean	*1.1134*	*1.1213*	*1.1391*	*1.1456*	*1.1611*	*1.1797*	*1.2435*	*1.3637*
K-Nearest Neighbour (KNN)	*0.9404*	*0.9703*	*0.9781*	*0.9893*	*0.9951*	*0.9973*	*0.9971*	*0.9986*
Cluster K-Nearest Neighbour (CKNN)	*0.8313*	*0.8569*	*0.8703*	*0.9032*	*0.9081*	*0.9489*	*0.9707*	*0.9948*
Local Least Square (LLS)	***0.1791***	***0.2308***	*0.3196*	*0.3338*	*0.3412*	*0.3497*	*0.4573*	*0.5328*
Cluster Local Least Square (CLLS)	*0.3880*	*0.4611*	*0.4689*	*0.5331*	*0.5847*	*0.7902*	*0.9625*	*0.8945*
Iterated Local Least Square (ILLS)	*0.1876*	*0.2427*	***0.2533***	***0.2608***	***0.2749***	***0.2803***	***0.3043***	***0.3106***
Bayesian Principal Component Analysis (BPCA)	*0.2151*	*0.3591*	*0.3868*	*0.4003*	*0.4216*	*0.4375*	*0.4441*	*0.4731*

Table 5. *NRMSE for Ecoli Dataset*

Method for missing imputation	***Percentage of missing Rate***							
	1%	2%	3%	4%	5%	10%	15%	20%
Mean	*0.8164*	*0.8417*	*0.8672*	*0.8969*	*0.9150*	*0.9454*	*0.9658*	*0.9709*
K-Nearest Neighbour (KNN)	*0.9046*	*0.9117*	*0.9288*	*0.9435*	*0.9348*	*0.9407*	*0.9416*	*0.9495*
Cluster K-Nearest Neighbour (CKNN)	*0.7541*	*0.7887*	*0.7964*	*0.8090*	*0.8177*	*0.8229*	*0.8343*	*0.8684*
Local Least Square (LLS)	***0.8284***	***0.8836***	*0.8864*	*0.9046*	*0.9164*	*0.9328*	*0.9819*	*0.9996*
Cluster Local Least Square (CLLS)	*0.7016*	*0.7247*	*0.7350*	*0.7485*	*0.7570*	*0.7610*	*0.7872*	*0.8116*
Iterated Local Least Square (ILLS)	*0.7098*	*0.7501*	***0.7612***	***0.7657***	***0.7765***	***0.7820***	***0.8468***	***0.8904***
Bayesian Principal Component Analysis (BPCA)	*0.7247*	*0.7255*	*0.7383*	*0.7436*	*0.7545*	*0.7635*	*0.7777*	*0.7997*

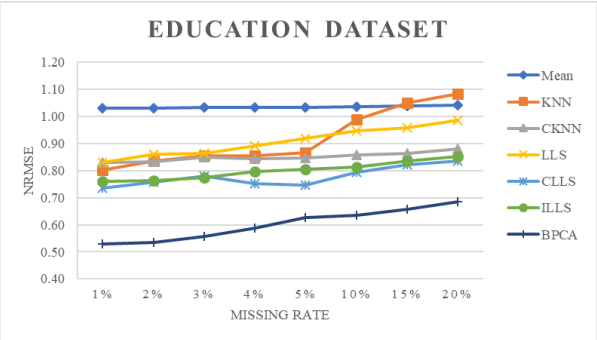
Table 6. *NRMSE for Seed Dataset*

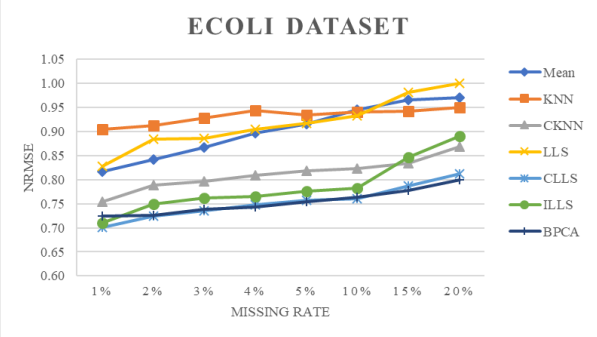
Method for missing imputation	***Percentage of missing Rate***							
	1%	2%	3%	4%	5%	10%	15%	20%
Mean	*1.1610*	*1.1703*	*1.1746*	*1.1748*	*1.1811*	*1.1848*	*1.1871*	*1.1909*
K-Nearest Neighbour (KNN)	*0.5480*	*0.5659*	*0.5457*	*0.5596*	*0.5676*	*0.5714*	*0.5820*	*0.6554*
Cluster K-Nearest Neighbour (CKNN)	*0.4063*	*0.4135*	*0.4153*	*0.4341*	*0.4361*	*0.4689*	*0.6755*	*0.6770*
Local Least Square (LLS)	*0.3926*	*0.4169*	*0.4246*	*0.4403*	*0.4667*	*0.5811*	*0.6040*	*0.8966*
Cluster Local Least Square (CLLS)	*0.3428*	*0.3827*	*0.4544*	*0.4676*	*0.4743*	*0.6540*	*0.8425*	*0.9974*
Iterated Local Least Square (ILLS)	***0.1477***	***0.1513***	***0.1544***	***0.1965***	***0.2069***	***0.2183***	***0.2301***	***0.2613***
Bayesian Principal Component Analysis (BPCA)	*0.1973*	*0.2095*	*0.2425*	*0.2544*	*0.2612*	*0.2619*	*0.2683*	*0.3055*

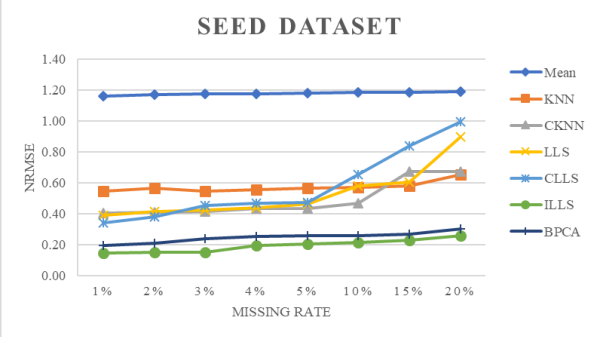
To generalize these findings, Figure 2 illustrates NRMSE measures that are the averages across all missing rates, with five single-pass imputation methods. According to this figure, ILLS and BPCA are of higher quality than their baseline counterparts. This reinforces the aforementioned suggestion, directly drawn from Tables 3–7. In addition, BPCA is often the best technique, with comparative performance to ILLS in wine and seed datasets. This may due to the fact that LLS exploits and differentiates the significance of multiple attributes more effectively than the basic KNN method. Of course, not all attributes equally correlate to the target one, which is in line with the practice of feature weighting or feature selection used for data classification and data clustering (Boongoen, T. and Shen, Q., 2010)

Table 7. *NRMSE for Iris Dataset*

Method for missing imputation	***Percentage of missing Rate***							
	1%	2%	3%	4%	5%	10%	15%	20%
Mean	*0.7590*	*0.8531*	*0.8771*	*0.9365*	*0.9390*	*0.9525*	*0.9690*	*0.9795*
K-Nearest Neighbour (KNN)	*0.7973*	*0.8044*	*0.8173*	*0.8350*	*0.8831*	*0.8993*	*0.9210*	*0.9552*
Cluster K-Nearest Neighbour (CKNN)	*0.6148*	*0.7126*	*0.7396*	*0.7609*	*0.7633*	*0.7800*	*0.7906*	*0.8301*
Local Least Square (LLS)	***0.2049***	***0.2530***	*0.2540*	*0.2665*	*0.2784*	*0.3078*	*0.4769*	*0.5894*
Cluster Local Least Square (CLLS)	*0.2242*	*0.2338*	*0.2430*	*0.2824*	*0.2925*	*0.3037*	*0.3136*	*0.3284*
Iterated Local Least Square (ILLS)	*0.2529*	*0.2530*	***0.2662***	***0.2785***	***0.3107***	***0.3277***	***0.3365***	***0.3454***
Bayesian Principal Component Analysis (BPCA)	*0.2311*	*0.2374*	*0.2376*	*0.2445*	*0.2578*	*0.2637*	*0.3064*	*0.3390*







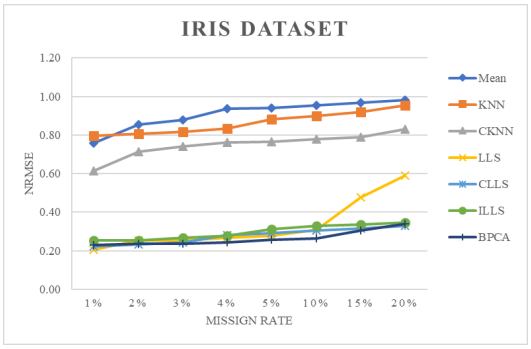


Figure 2. Average NRMSE values for different percentages of missing values in different datasets

Figure 2 specifically, BPCA outperforms all other methods when applied to the 3 datasets (i.e., edu, ecoil and iris). The improved imputation model that is denoted as ILLS, provides lower NRMSE measures than the basic LLS and cluster LLS in all examined datasets. Note that BPCA appears to perform significantly better than ILLS, except for the wine and seed dataset where their performance is comparative. Given these observations, the use of other conventional methods in addition to mean, LLS and KNN for the initial data preparation may enhance the accuracy of the resulting cluster framework. This view triggers the possible development of multi-level structure in which existing methods can be reused for the matrix generation at the low level and for the actual imputation at the top.

3.2 Discussion

Tables 8 provide rank of each imputation method for given dataset for different percentage of missing values. Each table indicates performance of different imputation method for different percentage of missing data for a given dataset. The reason for doing this is to assess consistency in performance of each imputation method for different percentage of missing data for a given dataset. The last two columns in each table indicate the average rank and rank obtained using mode method.

Table 8. *Rank Of Imputation Method for All Dataset for Different Percentage of Estimate Data*

Imputation method	Rank of imputation method for different missing rate in all Dataset								Rank By Mean	Rank By Mode
	1%	*2%*	*3%*	*4%*	*5%*	*10%*	*15%*	*20%*		
Mean	7	7	7	7	7	7	7	7	7	7
KNN	6	6	6	6	6	6	6	6	6	6
CKNN	5	5	5	5	5	5	5	5	5	5
LLS	4	4	4	4	4	3	3	4	3.75	4
CLLS	3	3	3	3	3	4	4	3	3.25	3
ILLS	2	2	2	2	2	2	2	2	2	2
BPCA	1	1	1	1	1	1	1	1	1	1

Therefore, can conclude that there is an agreement among rankings of different imputation methods and the rank of imputation method is independent of dataset and percentage of missing data in the dataset. In other words, conclude that the ranking or performance of the imputation method is consistent across five different datasets used in the study and with different percentages of missing data. It means ranking or performance of the imputation method neither changes with percentage of missing data nor with the different datasets. We also found that average NRMSE is lowest for BPCA imputation method and hence can conclude that BPCA imputation method outperforms the other methods. But these results are applicable only to datasets and one must always consider that there is no universal method always performing best in every situation.

4. Conclusion

This paper has investigated a neutral comparison of seven imputation methods based on five real datasets of various sizes, under an MCAR assumption. To obtain a realiable conclusion, Normalized Root mean squared error or NRMSE is exploited as an evaluation metric herein. In particular, the overall results across different experimental settings are summarized in Table 8. At first, these suggest that the most popular methods (Mean, KNN and LLS) are not necessarily the most efficient, a conclusion also shared by Celton et al., 2010. This is not a surprise for Mean that has an edge for simplicity, as it does not make use of the underlying correlation structure of the examined data and thus performing poorly. Both KNN and CKNN represent an improvement of Mean because they make use of the observed data structure. CLLS and LLS are based on a much more complex algorithmic procedure and their behaviors appear to be related to the size of dataset. In other words, they can be fast and efficient on small datasets, while their performance decreases and become time-intensive when applied to larger ones.

A second major conclusion is that BPCA and ILLS appeared to be the most robust imputation methods in the conditions tested here, with a significant advantage for ILLS when applied to small datasets. Based on the quality metric of NRMSE, the proposed BPCA and ILLS outperform their baselines and other compared methods, across different missing rates from 1% to 20%. Specific to the case of high missing level, i.e., 20%, they are able to sustain the quality of estimates a lot better than other alternatives from the family of local approach. In fact, the global algorithm like BPCA becomes more effective as the missing rate grows, but still sub-optimal against the new techniques. Despite this success, it is recommended to make use of other alternatives to resolve the problem of missing values when the rate of missing values grows higher than the level of 20-40%. Perhaps, a re-run of imaging process might be a better choice than analyzing uncertain data.

Our study has several limitations. The treatment of missing data is a very widespread broad statistical problem and one should consider that there is no universal imputation method performing best in every situations. Our results are limited to data matrices of numerical values, and we did not consider the case of longitudinal or nominal data which would merit to be considered with careful attention (Horton, NJ., Kleinman KP., 2007). In addition, our intention is also to provide general conclusions independent from the domain of application, and one could certainly further improve the accuracy of imputation methods by integrating specific domain knowledge into the imputation process (Liew, AWC., Law, NF., and Yan, H., 2011). Despite these limitations, this study provides a set of coherent observations across different settings. In conclusion, BPCA and ILLS are two imputation methods of interest. They outperform more popular approaches such as Mean, KNN, CKNN, LLS or CLLS, and hence deserve further consideration in practice.

Acknowledgements

This research work is partly supported by Pibulsongkram Rajabhat University.

References

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond, *Briefings in Bioinformatics,* 11:253-264.

Boongoen, T. & Shen, Q. (2010). Nearest-neighbor guided evaluation of data reliability and its Applications. *IEEE Transactions on Systems, Man and Cybernetics, Part B.* 40(6):1622–1633.

Cai, Z., Heydari, M., & Lin, G. (2006). Iterated local least squares microarray missing value imputation. *Journal of Bioinformatics and Computational Biology*. 4(5):935–957.

Gan, X., Liew, A. & Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, 34 (5), 1608-1619.

Horton, NJ., Kleinman, KP. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61:79-90.

Iam-On, N. (2019). Improving the consensus clustering of data with missing values using the link-based approach. *Data-Enabled Discovery and Applications,* 3 (7).

Keerin, P., Kurutach, W., & Boongoen, T. (2012). Cluster-based KNN missing value imputation for DNA microarray data. In *Proceedings of IEEE International Conference on System, Man and Cybernetics*, 445–450.

Keerin, P., Kurutach, W., & Boongoen, T. (2013). An improvement of missing value imputation in DNA microarray data using cluster-based LLS method. In *Proceedings of International Symposium on Communications and Information Technologies*, 559–564.

Keerin, P., Kurutach, W., & Boongoen, T. (2016). A cluster-directed framework for neighbour based imputation of missing value in microarray data, *International Journal of Data Mining and Bioinformatics*, 15, 165-193.

Kim, H., Golub, G., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation, *Bioinformatics,* 21:187-198.

Little, R. J., & D. B. Rubin. (1987). Statistical analysis with missing data. Second ed. Hoboken, *NJ: John Wiley & Sons.*

Liew, AWC., Law, NF., Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics,*12: 498-513.

Oba S., Sato Ma., Takemasa I., Monden M., Matsubara Ki., et al. (2003). A bayesian missing value estimation method for gene expression profile data. *Bioinformatics,* 19:2088-2096.

Pattanodom, M., Iam-On, N., & Boongoen, T. (2016). Clustering data with the presence of missing values by ensemble approach, *Proceedings of Asian Conference on Defence Technology*, 114-119.

Ridzuan, F. & Zainon, W. (2019). A review on data cleansing methods for big data, *Procedia Computer Science,* 161:731-738.

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., et al. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics 17:520-525.

Zhu, Z., Wang, J., & Sun, B. (2021). An efficient ensemble method for missing value imputation in microarray gene expression data, *BMC Bioinformatics*, 22 (188).