Impact of Dimension and Sample Size on the Performance of Imputation Methods

Yanjun Cui[1] and Junhu Wang[1,2(✉)]

[1] Hebei Academy of Sciences, Institute of Applied Mathematics, Shijiazhuang, China
[2] Griffith University, Gold Coast Campus, Southport, Australia
j.wang@griffith.edu.au

Abstract. Real-world data collections often contain missing values, which can bring serious problems for data analysis. Simply discarding records with missing values tend to create bias in analysis. Missing data imputation methods try to fill in the missing values with estimated values. While numerous imputations methods have been proposed, these methods are mostly judged by their imputation accuracy, and little attention has been paid to their efficiency. With the increasing size of data collections, the imputation efficiency becomes an important issue. In this work we conduct an experimental comparison of several popular imputation methods, focusing on their time efficiency and scalability in terms of sample size and record dimension (number of attributes). We believe these results can provide a guide to data analysts when choosing imputation methods.

Keywords: Imputation · RMSE · MissForest · MICE · Matrix Completion

1 Introduction

With the rapid development of Internet of Things and wireless networks, massive amounts of data are being collected daily. Such data are a valuable resource from which new knowledge can be discovered and new models can be built, e.g., using data mining, machine learning or statistical methods. However, raw data collected in the real world often contain missing values. Missing values are especially common in some areas. For example, in industrial databases, the ratio of missing data can be up to 50% [1]; and in bioinformatics, if we discard samples with missing data, some databases will lose about 90% of its data [2]. Even if there may be many *complete* records (i.e., records with no missing values) in the data set, simply discarding incomplete records tend to cause bias in analysis when the data is not missing completely at random. Therefore, missing data *imputation* has been widely used by data analysts to fill in the missing values with estimates.

Supported by The Excellent Going Abroad Experts' Training Program in Hebei Province.

J. He et al. (Eds.): ICDS 2019, CCIS 1179, pp. 538–549, 2020.
https://doi.org/10.1007/978-981-15-2810-1_51

Over the last decades many imputation methods have been proposed. Generally, different methods suit different data analysis tasks, different causes of missing values, and different types of data (e.g., categorical and numerical). In the literature, a variety of imputation methods have been compared for their effectiveness, i.e., accuracy. However, the efficiency of imputation algorithms has not been adequately addressed. Yet efficiency is an important problem when the data size is large. In our experiments, some imputation methods takes several days to complete on a modern PC over moderate record size. In this paper, we provide an experimental comparison of four popular imputation methods in terms of efficiency and scalability as well as accuracy, using real industrial data sets. We hope the results will be able to guide the choice of imputation methods for data analysis practitioners.

The remainder of this paper is organized as follows. Section 2 provides the preliminaries. Section 3 presents our experimental results. Section 4 discusses related work, and Sect. 5 draws the conclusion.

2 Preliminaries

2.1 Type of Missing Values

Missing values can be categorized into three main types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [6]. In MCAR, the probability of a variable to have missing values is independent of other variables. In MAR, the probability of a variable to have missing values depends only on the variables whose values are observed, not on variables whose values are missing; In NMAR, the probability may depend on values that are not missing as well as values that are missing. Complete case analysis (i.e., discarding records with missing values) does not lead to bias only for MCAR, but can create bias for MAR and NMAR.

2.2 Imputation Methods

Imputation methods can be categorized into traditional statistical methods and modern machine learning(ML) methods. They can also be hot-deck or cold-deck, the former uses a randomly selected similar record to impute the missing value, and the latter selects donors from another dataset. Imputation methods can be simple such as mean/median value substitution and linear interpolation. Most state-of-art imputation methods are based on machine learning techniques and can also be divided into two categories: local based and global based methods [2]. Local based methods include kNN (k-Nearest Neighbors), K-means, Maximum Likelihood [17], linear regression [50], LSimpute [46] and missForest [47]. These methods are based on the hypothesis that the data that are close in distance have the similar distribution of values. The disadvantage of local methods is that the missing values need to be imputed one by one, hence is generally more time-consuming. Global based methods include MC (Matrix Completion) [18],

SVT (Singular Value Thresholding) [19], bPCA (Bayesian Principle component analysis) [20] and so on. The advantage of these methods is that they can impute all the missing values simultaneously. The disadvantage is that the accuracy of the imputation is lower than the local-based ones. Imputation methods can be single or multiple, the former uses a single estimated value, and the latter uses multiple estimated values to add a degree of randomness. The most popular multiple imputation method is multiple imputation by chained equations (MICE) [48].

2.3 Root Mean Square Error (MISE)

The most frequently used measurements for evaluating imputation accuracy is the Root Mean Square Error (RMSE). Let M denote the number of missing values and $y, \hat{y}$ be the i-th imputed and observed value respectively. Then RMSE is defined as [10]:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{N}(y - \hat{y})^2}$$

RMSE measures the difference between imputed value and the observed value, the less the better.

In this work, we choose two simple imputation method (mean value substitution), hot-dec, two local based imputation methods (kNN and missForest), one global based method (Matrix Completion), and one multiple imputation method (MICE), in our comparison. Mean substitution is the easiest way used in data imputation. MICE and kNN are the most popular methods that are used in many research fields. MissForest [47] can be used to impute missing values particularly in the case of mixed-type data, and MC is the most popular global based method.

Next we present a brief description of each of these methods.

Mean Substitution. Here we use mean to replace the missing values. It is a highly efficient imputation method that barely needs computing capability and can be implemented easily. In R environment, they can be done by one command.

Hot Deck. Hot deck is a simple imputation method too. The function we used imputes the missing values in any variable by replicating the most recently observed value in that variable. This is by far the fastest imputation method. Only one pass of the data is needed.

MICE is a multivariate imputation method [23], it can infer more than one data sets at the same time, and provide a tool for the user to choose which one is better. Theoretically, MICE can reflect the uncertainty of the missing values, and should have better results in machine learning algorithms. MICE draws

imputation from their conditional distributions by Markov chain Monte Carlo (MCMC) techniques.

$$\begin{aligned}
\theta_1^{*(t)} &\sim P\left(\theta_1 | Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{t-1}\right) \\
Y_1^{*(t)} &\sim P\left(Y_1 | Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}\right) \\
&\vdots \\
\theta_p^{*(t)} &\sim P\left(\theta_p | Y_p^{obs}, Y_1^{(t)}, \ldots, Y_{p-1}^{(t)}\right) \\
Y_p^{*(t)} &\sim P\left(Y_p | Y_p^{obs}, Y_1^{(t)}, \ldots, Y_p^{(t)}, \theta_p^{*(t)}\right)
\end{aligned}$$

where θ is a vector of multivariate distribution that is used to impute missing data Y with p-variate multivariate distribution $P(Y|\theta)$. Starting from a simple draw from observed marginal distributions, the tth iteration of chained equations is a Gibbs sampler that successively draws. $Y_j^{(t)} = (Y_j^{obs}, Y_j^{*(t)})$ is the jth imputed variable at iteration t.

From the above we can see that each time the MICE tries to impute a missing value, it uses all the other attributes excluding the missing one to construct a regression model. Where the missing one is the dependent variable in a regression model and all the other variables are independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing linear, logistic, or Poison regression outside of the context of imputing missing data [48]. Finally, it uses the predicted value to replace the missing one. This step will repeat several times to gain better result. Because the regression model include all the attributes in the dataset, the larger the number of attributes, the more complex the regression model. That makes MICE more time-consuming.

kNN imputation is a local imputation method. It first finds the nearest neighbors of the record with missing value, and then calculates the missing values from that of its neighbors [24,25].

MissForest is based on random forest algorithm. Missforest turns data imputation into data prediction problems. First, the observed variables are used to regress the missing variables, and then the random forest is used to classify the data, so that the dependent variables can be used to predict the missing values [47].

Matrix Completion is a global imputation method. For a low rank matrix, the missing values can be inferred by the observed ones if we figure out the rank of the matrix. The calculation of the rank of a matrix is a NP-hard problem, nuclear norm can provide an approximate result [18].

3 Experimental Evaluation

In this section we present our experimental results of six imputation methods: mean substitution, hotdec, KNN, missForrest, MICE, and MC. We focus on the time-cost and scalability in terms of record dimension and sample size. We also use RMSE to compare their imputation accuracy. Mean substitution and hot deck method are simple imputation methods, they are very fast. Therefore we only test the time cost of MICE, kNN, missForest and MC.

3.1 Experimental Setup

Hardware and Software Packages. The experiments are conducted on a desktop computer with Intel Core i5-7200U 2.71 GHz CPU, 8 GB memory, and Samsung MZCLW256HEHP-000L7 Flash disk, running Windows 10 (64 bit) Enterprise Edition. We used R x64 3.51 as the programming environment. The packages we used include HotDeckImputation [41], caret [40], missForest [47], MICE [44], RSNNS [41], DMwR [43] and VIM [45].

Dataset. We used two real-world data sets in our experiments: Turbine and Spectral. Turbine is a real operational data set collected from the National Wind Turbine Grid of China. It has about 37000 samples, each sample has 720 attributes. These data were collected from 10 points independently, each point has 72 attributes, all of them are continuous numerical variables. There is also a label column to indicate whether there was a function failure. The Spectra data set is related to bacterial identification using MALDI-TOF mass-spectrometry data which has 571 samples and 1300 attributes. We use SMOTE method to expand the data set to 2160 samples for our test.

3.2 Impact of Number of Attributes

For Turbine, we first divide the dataset into 10 subsets based on their collection points, then we concatenate the records in i ($i \in [1, 10]$) subsets to generate 10 datasets with 72, 144, ..., 720 attributes respectively. Each subset is given 10% of missing values randomly. Then, we invoke the 4 imputation methods to impute each subset and record the time cost. The result is given in Fig. 1.

Notice that we only use 6 subsets in our experiment, because the time cost of MICE grows too fast to finish all the test. As we can see from the figure, the performance of MICE is heavily influenced by the number of attributes.

For the spectra data set, we fixed the number of samples to 360 and randomly chose 100, 150, 200, and 250 attributes. The results are shown in Fig. 2.

It can be seen that for both data sets, the time cost of MICE increases exponentially with the number of attributes, while it increases moderately with the other methods. Note that MC is extremely fast compared with other methods.

3.3 Impact of Number of Samples

We divide the Turbine dataset into 6 subsets, each has 10%, 30%, 50%, 70% and 90% and 100% samples of the original dataset. Each of the dataset is given 10% missing values. Then we invoke the imputation methods to impute the missing data. The results are shown in Fig. 3.

We also take 6 random subsets of the Spectra data of 360, 720, up to 2160 records, and fixed the number of attributes to 100. The experimental results are shown in Fig. 4.

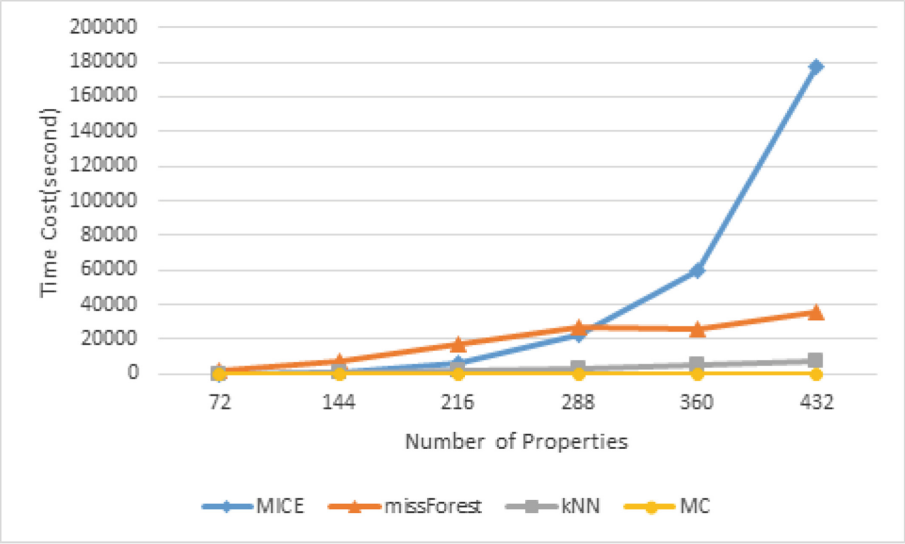

Fig. 1. Time cost with different number of attributes over Turbine data set

As can be seen, the time cost of missForest increases much more dramatically than all other methods.

3.4 Root Mean Square Error

We used the Turbine data set and RMSE to test the imputation accuracy. We randomly give from 10% up to 50% missing values to the dataset to verify how the RMSE change with missing ratios. Figure 5 shows that as the missing ratio grows, all the imputation results became worse, but the missForest method still has the best accuracy.

4 Related Work

Schmitt et al. [14] compared Mean, kNN (k-Nearest Neighbors), FKM (fuzzy K-means), SVD (Singular Value Decomposition), BPCA (Bayesian Principle Component Analysis) and MICE with Iris and Breast cancer data sets, and use RMSE, UCE (Unsupervised Classification Error), SCE (Supervised Classification Error) and Time cost as criterion. Their experimental results show that FKM and bPCA are more robust and more accurate than other methods. Without considering the time cost, FKM outperforms all the other methods. Pan et al. [2] compared KNN, bPCA, MC (Matrix Completion), LSimple (Least Square adaptive) and EM (Expectation Maximization) imputation methods on 5 data sets. Their results indicate that none of them can be better than others in all 5 data sets in terms of RMSE. Liu [36] uses classification accuracy and covariance as the criterion to compare five imputation methods: GIP (general iterative principal component imputation), SVD, r-EM (regularized EM with multiple ridge regression), t-EM (regularized EM with truncated total least squares), and MICE. The results show that covariance criterion does not always correlate with classification results. The r-EM imputation has better performance when the missing proportion is under 20%. Johnston et al. [38] compared five imputation program: AlphaImpute, BEAGLE, FImpute, findhap, PHASEBOOK with two data sets of genotypes. All the missing values are categorical data, so hitting rate instead of RMSE was used to evaluate the methods. The results shown that each of them had certain strengths and weaknesses and the author suggested that using a combination of 2 programs to improve imputation results. Musil et al. [37] used the CES-D (Center for Epidemiological Studies–Depression) to evaluate imputation results of data set on the stress and health of older adults. It compared EM imputation with simple regression imputation, regression with error term imputation and mean substitution. The results shown that although some methods of imputation may be better than others for recovering essential parameters such as the mean or standard deviations, all have some limitations in approximating the original data. Waljee, Mukherjee, Singal et al. [39] used the accuracy of MAAA (Multianalyte Assays with Algorithmic Analyses) model as measurement to evaluate imputation methods. The results shown that on small laboratory values, missForest is more robust and accurate. Muchlinski et al. [8] Compared random forest with logistic regression on civil war data. They found that random forests offers superior predictive power compared to several forms of logistic regression in an important applied domain–the quantitative analysis of civil war. Huang et al. [15] compared reconstruction method and MICE on social network data imputation. Their results indicate that the two methods have small bias, but MICE has smaller RMSE than reconstruction method. To the best of our knowledge, there has been no previous comparison of MICE, misForrest and MC in terms of scalability based on sample size and attribute size, nor comparsions of accuracy between these methods.

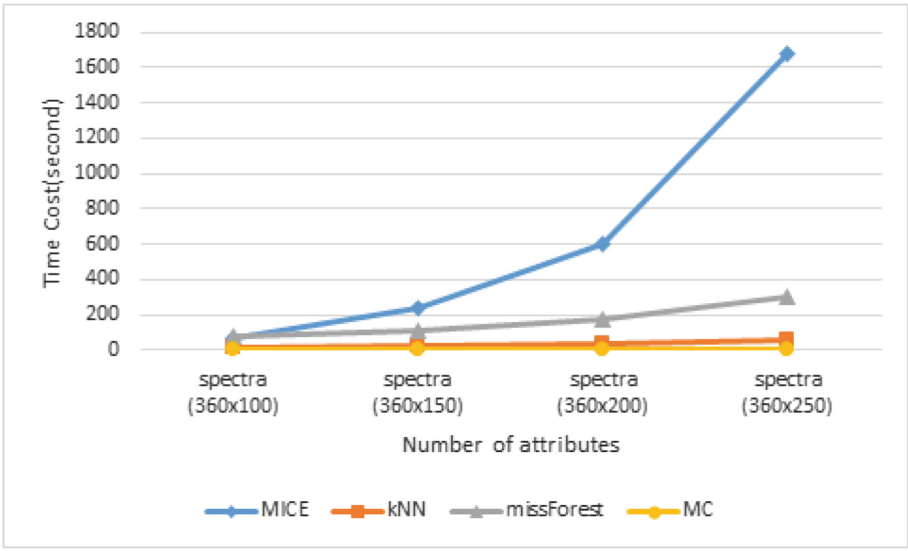

Fig. 2. Time cost with different number of attributes over Spectra data set

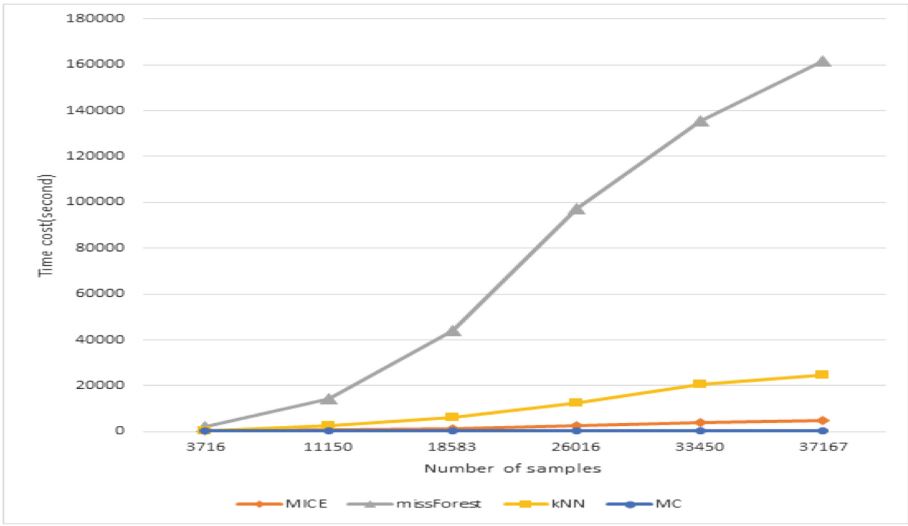

Fig. 3. Time cost with different number of samples with the Turbine data set

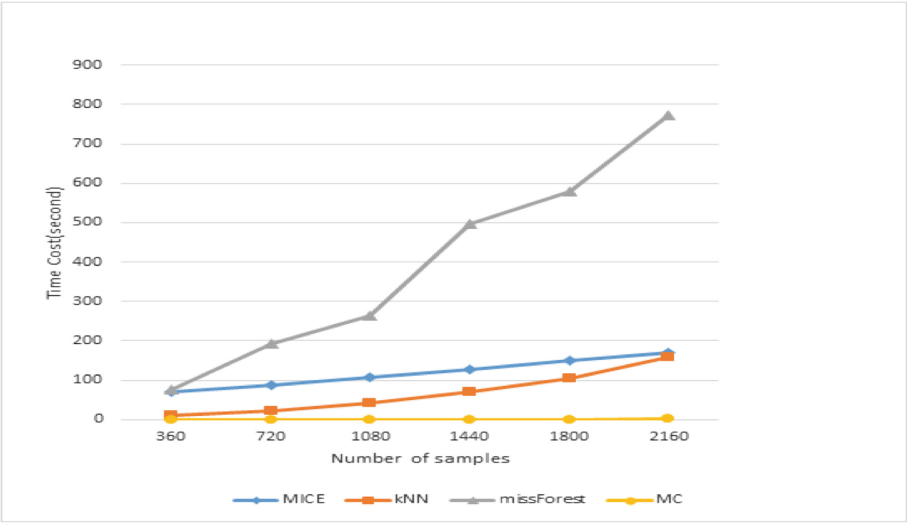

Fig. 4. Time cost with different number of samples with the Spectral data set

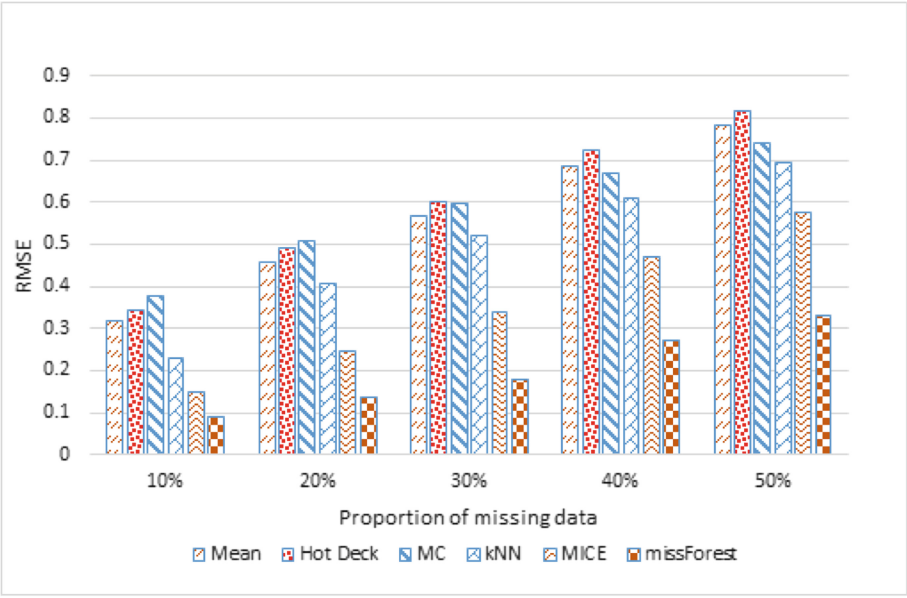

Fig. 5. Comparison of RMSE with Wind Turbine data

5 Conclusion

Missing values are almost inevitable in the real world, especially with sensor networks, social networks, bioinformatics and so on. Our experiments show that, For MICE, the imputation cost grows exponentially with the number of attributes. The time cost of missForest, on the other hand, grows drastically with the number of samples. For datasets with hundreds of attributes, we can divide the whole data set into several subsets, each time we do imputation on one subset to overcome the problem.

References

1. Lakshminarayan, K., Harp, S.A., Samad, T.: Imputation of missing data in industrial databases. Appl. Intell. **11**, 259–275 (1999)
2. Pan, X.-Y., Tian, Y., Huang, Y., Chen, H.-B.: Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. Genomics **97**, 257–264 (2011)
3. Pooler, P.S.: Handling missing data: applications to environmental analysis. J. Am. Stat. Assoc. **101**, 400–401 (2006)
4. Schneider, T.: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. **14**, 853–871 (2001)
5. Sun, Y., Braga-Neto, U., Dougherty, E.R.: Impact of missing value imputation on classification for DNA microarray gene expression data: a model-based study. EURASIP J. Bioinform. Syst. (2009)
6. Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)
7. Yu, L.-M., Burton, A., Rivero-Arias, O.: Evaluation of software for multiple imputation of semi-continuous data. Stat. Methods Med. Res. **16**, 243–258 (2007)
8. Muchlinski, D., Siroky, D., He, J., Kocher, M.: Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Polit. Anal. **24**, 87–103 (2016)
9. Montgomery, J.M., Olivella, S., Potter, J.D., Crisp, B.F.: An informed forensics approach to detecting vote irregularities. Polit. Anal. **23**, 488–505 (2015)
10. Chen, X., Xiao, Y.: A novel method for air quality data imputation by nuclear norm minimization. J. Sens. (2018)
11. White, I.R., Daniel, R., Royston, P.: Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Comput. Stat. Data Anal. **54**, 2267–2275 (2010)
12. Shao, J., Meng, W., Sun, G.: Evaluation of missing value imputation methods for wireless soil datasets. Pers. Ubiquit. Comput. **21**, 113–123 (2017)
13. Kornelsen, K., Coulibaly, P.: Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. J. Hydrol. Eng. **19**, 26–43 (2017)
14. Schmitt, P., Mandel, J., Guedj, M.: A comparison of six methods for missing data imputation. Biometrics Biostatistics **6**, 1 (2015)
15. Huang, H., Huang, F.: A comparison study of reconstruction and multiple imputation in social network analysis. Adv. Psychol. **8**, 642–648 (2018)
16. Van Buuren, S., Boshuizen, H.C., Knook, D.L.: Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. **18**, 681–694 (1999)

17. Troyanskaya, O., et al.: Missing value estimation for DNA microarray. Bioinformatics **17**, 520–525 (2001)
18. Lei, C., Song-Can, C.: Survey on matrix completion models and algorithms. J. Softw. **28**, 1547–1564 (2017)
19. Cai, J.-F., Candes, E.J., Shen, Z.: A singular value Thresholding Algorithm for matrix completion. Soc. Ind. Appl. Math. **20**, 1956–1982 (2010)
20. Oba, S., Sato, M.-A., et al.: Bayesian missing value estimation method for gene expression profile data. Bioinformatics **19**, 2088–2096 (2003)
21. Vach, W.: Missing values: statistical theory and computational practice. Comput. Stat., 345–354 (1994)
22. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (2002)
23. White, I.R., Royston, P., Wood, A.M.: Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. **30**, 377–399 (2010)
24. Finley, A.O., McRoberts, R.E., Ek, A.R.: Applying an efficient k-Nearest Neighbor search to forest attribute imputation. For. Sci. **52**, 130–135 (2006)
25. Crookston, N.L., Finley, A.O.: yaImpute: an R Package for kNN Imputation. J. Stat. Softw. **23**, 16 (2008)
26. Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. Oper. Res. **43**, 570–577 (1995)
27. SuykensJ, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**, 293–300 (1999)
28. Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**, 18–22 (2002)
29. Ho, T.K.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278–282 (1995)
30. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
31. Zhou, Z.: Machine Learning. Tsinghua University Press, Beijing (2016)
32. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton (2004)
33. Luengo, J., Garca, S., Herrera, F.: On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl. Inf. Syst. **32**, 77–108 (2012)
34. Brock, G., Shaffer, J., Blakesley, R., Lotz, M., Tseng, G.: Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinf. **9**, 1–12 (2004)
35. Deb, R., Liew, A.W.-C.: Missing value imputation for the analysis of incomplete traffic accident data. Inf. Sci. **339**, 274–289 (2016)
36. Liu, Y., Brown, S.D.: Comparison of five iterative imputation methods for multivariate classification. Chemometr. Intell. Lab. Syst. **120**, 106–115 (2013)
37. Musil, C.M., Warner, C.B., et al.: A comparison of imputation techniques for handling missing data. West. J. Nurs. Res. **24**, 815–829 (2002)
38. Johnston, J., Kistemaker, G., Sullivan, P.G.: Comparison of different imputation methods. Interbull Bull. **44**, 26–29 (2011)
39. Waljee, A.K., Mukherjee, A., et al.: Comparison of imputation methods for missing laboratory data in medicine. BMJ Open **3** (2013)
40. Kuhn, M.: e classification and regression training (2018). https://cran.r-project.org/package=caret

41. Bergmeir, C.: Neural networks using the stuttgart neural network simulator (SNNS) (2018). https://cran.r-project.org/package=RSNNS
42. Joenssen, D.W.: Hot deck imputation methods for missing data (2015). https://cran.r-project.org/package=HotDeckImputation
43. Torgo, L.: Functions and data for data mining with R (2015). https://cran.r-project.org/package=DMwR
44. van Buuren, S.: Multivariate imputation by chained equations (2018). https://cran.r-project.org/package=mice
45. Templ, M., Alfons, A., Kowarik, A., Prantner, B.: Visualization and imputation of missing values (2017). https://cran.r-project.org/package=VIM
46. Bø, T.H., Dysvik, B., Jonassen, I.: LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. **32** (2004)
47. Stekhoven, D.J.: Nonparametric missing value imputation using random forest (2013). http://www.r-project.org. https://github.com/stekhoven/missForest
48. Azur, M.J., Stuart, E.A., et al.: Multiple imputation by chained equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. **20**, 40–49 (2011)
49. Zhang, S., Li, X., et al.: Efficient kNN classification with different numbers of nearest neighbors. IEEE Trans. Neural Netw. Learn. Syst. **5**, 1774–1784 (2018)
50. Chen, Y., Li, Y., et al.: Data envelopment analysis with missing data: a multiple linear regression analysis approach. Int. J. Inf. Tech. Decis. Making **13**, 137–153 (2015)