Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09709-4

Missing value imputation: a review and analysis of the literature (2006–2017)

Wei-Chao Lin[1,2,3] · Chih-Fong Tsai[4]

Abstract

Missing value imputation (MVI) has been studied for several decades being the basic solution method for incomplete dataset problems, specifically those where some data samples contain one or more missing attribute values. This paper aims at reviewing and analyzing related studies carried out in recent decades, from the experimental design perspective. Altogether, 111 journal papers published from 2006 to 2017 are reviewed and analyzed. In addition, several technical issues encountered during the MVI process are addressed, such as the choice of datasets, missing rates and missingness mechanisms, and the MVI techniques and evaluation metrics employed, are discussed. The results of analysis of these issues allow limitations in the existing body of literature to be identified based upon which some directions for future research can be gleaned.

Keywords Missing values · Imputation · Supervised learning · Incomplete dataset · Data mining

1 Introduction

Data mining or big data analysis has been recognized as an important and challenging task for many problems in daily life. To perform big data analysis or data mining, a specific dataset for a chosen target problem is collected. However, in practice, the collected dataset usually contains some proportion of incomplete data that have one or more missing attribute values. There are many reasons for incompleteness in datasets, arising from a variety of sources, the database system per se, the network, improper, mistaken data entries, and so on.

According to Strike et al. (2001) and Raymond and Roberts (1987), when the dataset contains a very small amount of missing data, e.g. the missing rate is less than 10% or 15%

✉ Chih-Fong Tsai
cftsai@mgt.ncu.edu.tw

[1] Department of Information Management, Chang Gung University, Taoyuan, Taiwan

[2] Healthy Aging Research Center, Chang Gung University, Taoyuan, Taiwan

[3] Department of Thoracic Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan

[4] Department of Information Management, National Central University, Zhongli, Taoyuan, Taiwan

Published online: 05 April 2019

for the whole dataset, the missing data can simply be removed from the dataset without having a significant effect on the final mining or analysis result. However, when the missing rate exceeds 15%, careful consideration needs to be given to dealing with these missing data (Acuna and Rodriguez 2004). It should be noted that this does not mean that every domain problem dataset follows this kind of rule. Often small amounts of missing data may contain important information that cannot be ignored, such as the records containing very high amount of money spent by some consumers but some of their personal information is missing, e.g. age, income, education, etc.

Unlike the case deletion strategy, missing value imputation (MVI) is the solution method most commonly used to deal with the incomplete dataset problem. In general, MVI is a process in which some statistical or machine learning techniques are used to replace the missing data with substituted values. Statistical techniques, such as mean/mode and regression, have been applied for this purpose, for several decades (Little and Rubin 1987), with machine learning techniques, such as the k nearest neighbor, artificial neural network, and support vector machine techniques being employed in the last 10 years (Garcia-Laencina et al. 2010).

There is variety of MVI techniques suitable for application to different domain problems. A large number of surveys of MVI from different perspectives have already appeared in the literature, such as for operation management (Tsikriktsis 2005), medical problems (Aittokallio 2009; Donders et al. 2006; Harel and Zhou 2007; Liew et al. 2011), pattern classification (Garcia-Laencina et al. 2010), and questionnaires and surveys (Baraldi and Enders 2010; De Leeuw 2001).

Most of these surveys have mainly focused on describing the basic concepts of the relevant MVI techniques. However, from the experimental design procedure viewpoint, there are many technical issues that have not been adequately reviewed and analyzed. For example, it is not known which technique(s) are the most widely used, what kinds of domain problem datasets have been studied, how many dataset missing rates are considered in the simulations, and so on.

Therefore, unlike previous surveys, this survey provides statistical analyses of technical questions related the experimental design procedure. Specifically, 111 journal papers published over the past decade, from 2006 to 2017, are reviewed and analyzed. In addition, some limitations of related works are also discussed for indication of possible future research directions.

The rest of this paper is organized as follows. Section 2 describes the commonly used experimental design procedure for MVI. The related literature for each of the major components of the procedure is analyzed, with Sects. 3, 4, and 5 focused on the datasets used, as well as the related information including missing rates and missingness mechanisms, MVI techniques, and evaluation metrics, respectively. Section 6 discusses the limitations of related work and Sect. 7 offers some conclusions.

2 The experimental design procedure for missing value imputation

There are three technical issues that need to be considered in the experimental design procedure for MVI outlined in Fig. 1. The first one is the chosen datasets for related experiments. The experimental dataset may contain a number of missing data or it may be a complete dataset. For complete datasets, a missing value simulation is performed. That is, the chosen dataset is simulated with different missing rates (e.g., 10% or 20%) using three

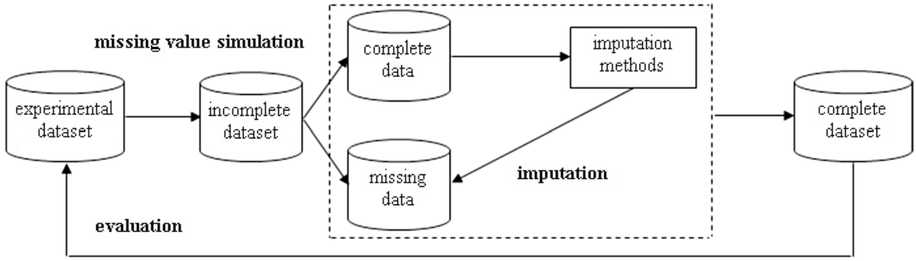

Fig. 1 The experimental design procedure for MVI

different missingness mechanisms: missing completely at random (MCAR), missing at random (MAN), and missing not at random (MNAR). This results in different incomplete datasets having different proportions of missing data.

The second technical issue is the techniques used for missing value imputation. During the imputation process, each incomplete dataset can be divided into a set of complete data and a set of missing data. The former is used for the 'estimation' of suitable values by different imputation methods to replace the missing values in the set of missing data. Consequently, this produces a 'pseudo' complete dataset for later data mining or analysis tasks, if any.

The third technical issue is performance evaluation of imputation results. The most straightforward method to evaluate the performance of the imputation method is to assess the differences between the real values in the original dataset and the estimated values in the 'pseudo' dataset. Another method involves using the 'pseudo' dataset to perform some data mining task, such as classification or clustering, then observing the final mining performance.

According to the three technical issues, related literatures are reviewed and analyzed by their experimental datasets collected, MVI techniques used, and performance evaluation methods considered, which are discussed in Sects. 3, 4, and 5, respectively.

3 Analysis of experimental datasets

3.1 Dataset domains

Table 1 shows the number of works using each type of dataset from the 111 related studies. As can be seen, most researchers use the UCI (University of California at Irvine) datasets[1] for their experiments. Often, several UCI datasets are used in each study, which can include a variety of domain problems. In contrast, medical related datasets, such as microarray or gene datasets, are the most widely considered domain problem in MVI. Other domain problems considered less often include image data, software measurement and project, financial data, questionnaire based data.

The results indicate that in most MVI studies, more than one specific domain problem is considered. The advantage of doing this is to prove the domain scalability of the MVI

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

Table 1 The numbers of works using different domain datasets

	UCI	Medical	Images	Software	Financial	Questionnaire	KDD[a]	Other
2017	5	3	0	1	0	0	0	1
2016	3	2	0	0	0	0	1	0
2015	7	2	0	0	1	0	0	0
2014	3	3	0	1	0	1	0	3
2013	5	3	1	0	0	0	0	0
2012	5	2	1	0	1	0	0	2
2011	6	3	0	0	0	1	0	0
2010	3	4	0	0	1	0	0	1
2009	3	4	0	0	0	1	0	1
2008	3	4	0	2	0	1	1	1
2007	4	1	0	0	0	1	0	3
2006	2	4	0	0	0	0	0	0
Total	48	34	2	4	3	5	1	12

[a] http://www.kdd.org/kdd-cup

method used. However, further analysis of the dataset characteristics, including the number of features (i.e., attributes) and data samples, allows some limitations of past work to be identified. One major limitation is the problem of dataset size. That is, most studies use small scale UCI datasets that contain small numbers of features and/or data samples, with the number of features ranging from 4 to 89 and the usual number of data samples ranging from several hundred to thousands. Some exceptions are Folino and Pisani (2016) and Farhangfar et al. (2007), who used very large scale datasets containing a very high number of feature dimensions, i.e., 216, and a very large amount of data samples, i.e., 581,012 and 256,000, respectively.

Another limitation of past studies is that although there are three different types of features, categorical, numerical, and mixed types, very few have analyzed differences in performance of MVI methods between different feature types. There are, however, two exceptions to this, namely, Tsai and Chang (2016) and Stekhoven and Buhlmann (2012).

3.2 Missing rates

In general, most studies have examined imputation performance by performing different missing value simulations over a chosen dataset using different missing rates. Some have considered very small missing rates, e.g., less than 30%, while others have focused on large ranges of missing rates, such as from 5 to 80%. Figure 2 shows the number of works that consider missing rates that are less than 30%, between 30 and 50%, and greater than 50%.

As we can see from the figure, most studies discuss missing rates that are less than 30% with only twelve works considering missing rates that are larger than 50%. Of the studies using very large missing rates, seven of them used the UCI datasets (Eirola et al. 2013; Kapelner and Bleich 2015; Kiasari et al. 2017; Mesquite et al. 2017; Purwar and Singh 2015; Qin et al. 2009; Zhu et al. 2011), one the Digital Bibliographic Library Browser

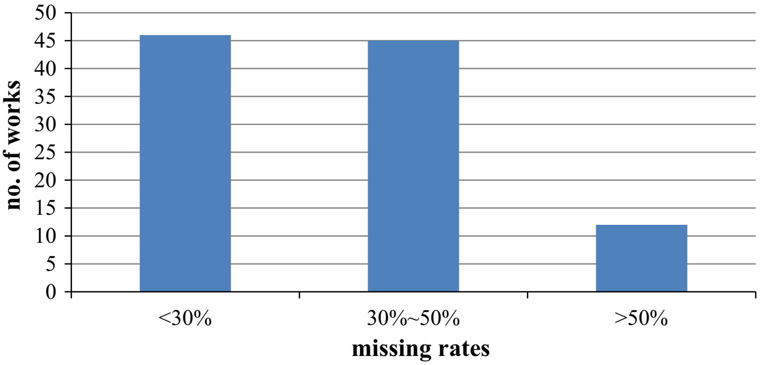

Fig. 2 Number of works using missing rates that are less than 30%, between 30 and 50%, and greater than 50%

(DBLP) dataset[2] (Li et al. 2014), one the wireless sensor network dataset (Li and Parker 2014), one the medical dataset (Janssen et al. 2010), one the traffic flow dataset (Chen et al. 2017), and one a synthetic dataset (Graham et al. 2007).

In short, most of the datasets which have been used are relatively small, containing several hundreds to thousands of data samples, in contrast to the DBLP and wireless sensor networks datasets which contain 10,000 and 12,000 data samples, respectively.

3.3 Missingness mechanisms

According to Little and Rubin (1987), there are three types of missingness mechanisms that can cause an incomplete dataset. They are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data.

On the other hand, MAR occurs when the probability of an instance having a missing value for an attribute may depend on the value of that attribute. In other words, when the distribution of an instance having missing values for an attribute depends on the observed data, but does not depend on the missing data. NMAR occurs when the probability of an instance having a missing value for an attribute may depend on the value of that attribute.

Therefore, there are three different ways to artificially simulate a collected dataset as an incomplete dataset containing a controlled missing rate. Figure 3 shows the number of works detailing simulations of the three types of missingness mechanisms.

The results show that most researchers have only used the MCAR mechanism for their experiments. Very few (i.e., 15 works) have considered all three mechanisms for each chosen dataset. Specifically, seven have used the UCI datasets (Garciarena and Santana 2017; Kapelner and Bleich 2015; Pan et al. 2015; Tian et al. 2014; Twala 2009; Twala et al. 2008; Valdiviezo and Van Aelst 2015; Xia et al. 2017; Zhu et al. 2012), three the software measurement/project datasets (Khoshgoftaar and Van Hulse 2008; Song et al. 2008; Van Hulse

[2] http://dblp.uni-trier.de/db/index.html.

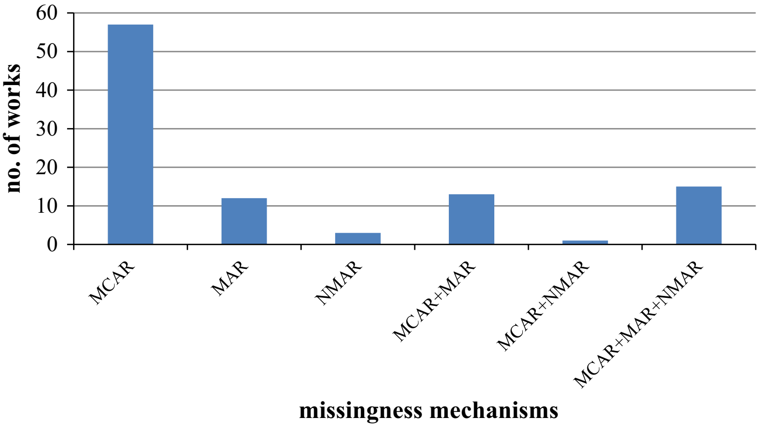

Fig. 3 Number of works discussing the different missingness mechanisms

and Khoshgoftaar 2014), one for the medical dataset (Armitage et al. 2015), and two synthetic datasets (Ding and Simonoff 2010; Hapfelmeier and Ulm 2014).

4 Missing value imputation techniques

In general, missing value imputation techniques can be classified into two types, namely, statistical and machine learning based techniques (Aittokallio 2009; Garcia-Laencina et al. 2010). Related studies have considered some of these techniques as the basis for relevant experiments regardless of whether the work focuses on proposing a novel imputation technique or on comparing some chosen imputation techniques for specific domain problems.

The following subsections describe the analysis related to the questions as to what kinds of baseline techniques have been used for MVI and which is the most popular. Note that describing the concepts of these statistical and machine learning based techniques is not the main focus of this paper.

4.1 Statistical techniques

Table 2 lists the statistical techniques that have been used in studies published from 2006 to 2017. As we can see, expectation management (EM), linear regression (LR), least squares (LS), and mean/mode are the top four most widely used statistical techniques, having been applied in 23, 15, 33, and 28 works respectively.

Among these top four statistical techniques, the mean and mode methods are the simplest imputation methods for imputing numerical and categorical attribute values, respectively. In the mean approach, missing attributes are filled in by the average value of that attribute in all the observed data. On the other hand, the mode approach uses the attribute value in all the observed data that appears most often to fill in the missing attribute values.

In the EM algorithm, it consists of two steps where the E-step calculates the expectation of the complete data sufficient statistics given the observed data and current parameter estimates, and the M-step updates the parameter estimates through the maximum likelihood approach based on the current values of the complete sufficient statistics. The EM

Table 2 The statistical techniques used in the literature

Techniques	Works
Expectation maximization (EM)	Aussem and de Morais (2010), De Souto et al. (2015), Di Nuovo (2011), Ding and Ross (2012), Doquire and Verleysen (2012), Folino and Pisani (2016), Garcia et al. (2011), Garciarena and Santana (2017), Ghanad-Rezaie et al. (2010), Ghorbani and Desmarais (2017), Graham et al. (2007), Hron et al. (2010), Hruschka et al. (2007), Jerez et al. (2010), Kang (2013), Li and Parker (2014), Lin et al. (2006), Luengo et al. (2012), Merlin et al. (2010), Peng and Zhu (2008), Polikar et al. (2010), Stekhoven and Buhlmann (2012), Twala (2009), Twala et al. (2008), Zhang and Liu (2009) and Zhu et al. (2012)
Gaussian mixture model (GMM)	Di Zio et al. (2007), Garcia-Laencina et al. (2013) and Kang (2013)
Hot deck (HD)	Di Zio et al. (2007), Farhangfar et al. (2007, 2008), Jerez et al. (2010), Ghorbani and Desmarais (2017), Kang (2013), Nishanth and Ravi (2016), Silva-Ramirez et al. (2011) and Tian et al. (2014)
Linear discriminant analysis (LDA)	Farhangfar et al. (2007)
Linear/logistic regression (LR)	De Souto et al. (2015), Di Nuovo (2011), Farhangfar et al. (2007, 2008), Eirola et al. (2013), He et al. (2009), Iacus and Porro (2007), Jerez et al. (2010), Peng and Zhu (2008), Qin et al. (2007, 2009), Saar-Tsechansky and Provost (2007), Shao et al. (2017), Silva-Ramirez et al. (2011, 2015) and Zhu et al. (2012)
Least squares (LS)	*Iterative local LS (ILLS)*: Cheng et al. (2012), Chiu et al. (2013), Hron et al. (2010), Hu et al. (2006), Pati and Das (2017), Rahman and Islam (2013), Sun et al. (2009) and Tuikkala et al. (2008) *Least trimmed squares (LTS)*: Hron et al. (2010) *LS adaptive (LSA)*: Chiu et al. (2013) and Rao et al. (2013) *Local LS (LLS)*: Brock et al. (2008), Celton et al. (2010), Cheng et al. (2012), Chiu et al. (2013), Gan et al. (2006), Hu et al. (2006), Liu et al. (2010), Luengo et al. (2012), Oh et al. (2011), Pati and Das (2017), Rahman and Islam (2013), Rao et al. (2013), Sehgal et al. (2008, 2009), Tuikkala et al. (2008), Wang et al. (2006), Yu et al. (2011) and Zhang et al. (2008) *Noniterative partial least squares (NIPALS)*: Rao et al. (2013) *Ordinary LS (OLS)*: Brock et al. (2008) and Oh et al. (2011) *Partial LS (PLS)*: Brock et al. (2008) and Oh et al. (2011) *Sequential local LS (SLLS)*: Chiu et al. (2013)
Low rank matrix decomposition/tensor completion	Liu et al. (2013)
Markov chain Monte Carlo (MCMC)	Ding and Ross (2012) and Saha et al. (2017)

Table 2 (continued)

Techniques	Works
Mean/mode	Armitage et al. (2015), Chen et al. (2017), De Souto et al. (2015), Di Zio et al. (2007), Ding and Ross (2012), Farhangfar et al. (2007, 2008), Folino and Pisani (2016), Hron et al. (2010), Hruschka et al. (2007), Huang et al. (2017), Jerez et al. (2010), Kang (2013), Khoshgoftaar and Van Hulse (2008), Kiasari et al. (2017), Lin et al. (2006), Luengo et al. (2012), Liao et al. (2014), Mesquite et al. (2017), Moons et al. (2006), Munoz and Rueda (2009), Nishanth and Ravi (2016), Pati and Das (2017), Polikar et al. (2010), Rao et al. (2013), Silva-Ramirez et al. (2011, 2015), Sun et al. (2009), Tian et al. (2014), Tuikkala et al. (2008), Twala (2009), Twala et al. (2008), Valdiviezo and Van Aelst (2015) and Xia et al. (2017)
Multiple imputation by chained equations (MICE)	Burgette and Reiter (2014), Ding and Ross (2012), Garciarena and Santana (2017), Jerez et al. (2010), Liao et al. (2014), Janssen et al. (2010), Moons et al. (2006), Stekhoven and Buhlmann (2012) and Valdiviezo and Van Aelst (2015)
Multiple imputation by sequential regression trees (MIST)	Valdiviezo and Van Aelst (2015)
Naïve Bayes (NB)	Farhangfar et al. (2007, 2008), Graham et al. (2007), Khoshgoftaar, and Van Hulse (2008), Leung and Leung (2013), Nishanth and Ravi (2016) and Subasi et al. (2011)
Principal component analysis (PCA)	*Bayesian PCA*: Brock et al. (2008), Celton et al. (2010), Cheng et al. (2012), Chiu et al. (2013), Liu et al. (2010), Luengo et al. (2012), Oh et al. (2011), Pati and Das (2017), Rahman and Islam (2013), Rao et al. (2013), Sehgal et al. (2008, 2009), Sun et al. (2009), Tuikkala et al. (2008), Wang et al. (2006), Xia et al. (2017), Yu et al. (2011) and Zhang et al. (2008) *PCA*: Chen et al. (2017), Saha et al. (2017), Van Ginkel and Kroonenberg (2014) and Zuccolotto (2012)
Sampling	*Data augmentation (DA)*: Hruschka et al. (2007), Van Ginkel et al. (2007) and Zhang and Liu (2009) *Random sampling (RS)*: Farhangfar et al. (2007) and Lin et al. (2006)
Singular value decomposition (SVD)	Brock et al. (2008), Gan et al. (2006), Liu et al. (2010), Luengo et al. (2012), Oh et al. (2011), Paul et al. (2017), Rao et al. (2013), Saha et al. (2017) and Yu et al. (2011)

algorithm then proceeds in an iterative manner until the difference between the last two consecutive parameter estimates converges to a specified criterion. According to the final parameter estimates and the observed data, the expectation of each missing value can be calculated, which will be used as the imputation value.

For regression based imputation methods, the relationships among attributes are estimated, and then the regression coefficients are used to estimate the missing attribute values.

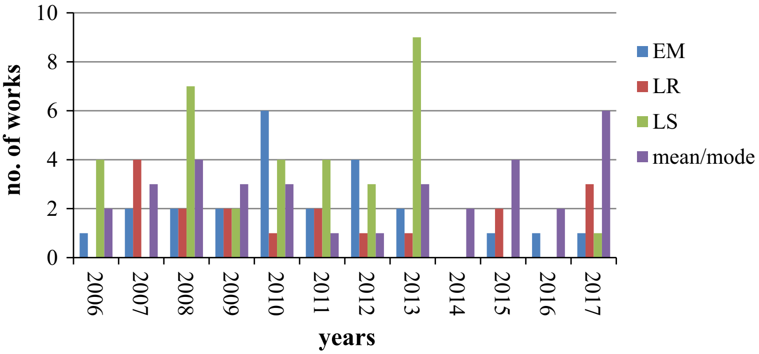

Fig. 4 The year-wise distribution of the number of works using the EM, LR, LS, and mean/mode techniques

Particularly, linear regression and logistic regression are used for the prediction of numerical and categorical attribute values, respectively. In general, the method of least squares (LS) (or ordinary least squares) is used in linear regression to produce the final estimation by minimizing the measured and predicted values of the attributes.

More specifically, for the category of least squares (LS), various estimation techniques can be used to replace ordinary least squares (OLS) to produce the final prediction result, such as ILLS, LTS, LSA, LLS, NIPALS, OLS, PLS, and SLLS.

Figure 4 shows the year-wise distribution of works applying the top four statistical techniques. Although LS is the most widely used MVI technique, there has had been no related work considering this technique within the last 3 years, except for Pati and Das (2017). Instead, researchers have tended to prefer the EM, LR, and mean/mode techniques. Of these, the mean/mode technique is the second mostly widely used, having been used in publications appearing each year from 2006 and 2017. This survey indicates that the mean/mode technique should be regarded as one of the representative baseline statistical MVI techniques.

Regarding the researches using one of the top four statistical MVI techniques, we can further analyze the relationships between these techniques and their experimental datasets, which are shown in Table 3.

As we can see that the major statistical MVI technique for medical domain datasets is LS. This indicates that most medical domain datasets contain the numerical data type of missing values. Moreover, these medical domain datasets are usually simulated with the missing rates that are smaller than 30%, and related studies using the LS techniques only consider the MCAR and MAR missingness mechanism. On other hand, EM, LR, and mean/mode are the widely used statistical MVI techniques for UCI datasets where the simulated missing rates mostly range from 30 to 50%.

4.2 Machine learning based techniques

Table 4 lists the machine learning based techniques that have been applied in the literature from 2006 to 2017. The top four techniques are clustering, DT, KNN and RF, which have been used in 14, 17, 52, and 9 related works, respectively.

Among the top four machine learning based techniques, cluster analysis is only the unsupervised learning technique whose task is to group a set of similar objects into the

Table 3 The number of works for the relationships between the top four statistical MVI techniques and their experimental datasets

Techniques	Missingness mechanisms	Dataset domains	Missing rates
EM	MCAR: 19 MAR: 11 NMAR: 7	UCI: 12 Medical: 4 Images: 1 Finance: 1 Questionnaire: 2 Others: 4	< 30%: 9 30–50%: 12 > 50%: 2
LR	MCAR: 12 MAR: 7 NMAR: 3	UCI: 9 Medical: 4 Questionnaire: 1 Others: 3	< 30%: 4 30–50%: 7 > 50%: 2
LS	MCAR: 28 MAR: 3 NMAR: 0	UCI: 3 Medical: 30 Others: 4	< 30%: 30 30–50%: 2 > 50%: 0
Mean/mode	MCAR: 20 MAR: 11 NMAR: 3	UCI: 19 Medical: 8 Images: 1 Software: 2 Questionnaire: 2 Others: 1	< 30%: 11 30–50%: 16 > 50%: 3

same clusters. Specifically, each cluster center (or centroid) is the mean of the objects in the same cluster. To impute the missing values, the distance between incomplete data and the identified cluster centroids is calculated where the closest centroid's values are used to fill in missing values.

On the other hand, KNN is one representative supervised learning (classification) technique where missing values are imputed using the values calculated from the *k* nearest observed data. In particular, the nearest neighbors can be identified by some specific distance function, usually the Euclidean distance. For missing value imputation, the missing data is used as the testing case, in which the complete and missing attributes represent the input features and output class label (or prediction), respectively. Next, its *k* nearest observed data using the complete attributes can be identified whose class label is used to impute the missing attribute.

In DT, it is a tree-like model that each internal node denotes a test of an attribute and each branch represents an outcome of the test. The leaf nodes represent classes or class distributions. The upper-most node in a tree is the root node with the highest entropy. In the tree-growing process, the attribute having the highest information gain is chosen to split the node into child nodes. In related literatures, C4.5/5.0 and CART are used for imputing categorical and numerical attribute values, respectively.

About RF, multiple decision trees are constructed based on the bootstrapping procedure and the final predictions are given by averaged values or majority votes of each tree's prediction. The imputation process by DT and RF is similar to KNN where the internal and leaf nodes represent the input features and output class label, respectively.

Figure 5 shows the year-wise distribution for the number of works that have used the top four machine learning based techniques. KNN is no doubt the most popular MVI technique in this category, and can be regarded as the most representative baseline machine learning based MVI technique. On the other hand, the clustering and RF

Table 4 The machine learning based techniques used in the literature

Techniques	Works
Artificial neural networks (ANN)	Aydilek and Arslan (2012), Gautam and Ravi (2015), Nishanth and Ravi (2016) and Shao et al. (2017)
Association rule (AR)	Li et al. (2014)
Collateral missing value estimation (CMVE)	Sehgal et al. (2008, 2009)
Clustering	*Fuzzy c-means (FCM)*: Aydilek and Arslan (2013), Di Nuovo (2011), Li et al. (2010), Somasundaram and Nedunchezhian (2011), Tian et al. (2014) and Zhang et al. (2015)
	Fuzzy k-means (FKM): Luengo et al. (2012) and Pan et al. (2015)
	Hierarchical clustering (HC): Celton et al. (2010)
	K-means (KM): Eirola et al. (2013), Kang (2013), Luengo et al. (2012) and Somasundaram and Nedunchezhian (2011)
	Self-organizing map (SOM): Garcia-Laencina et al. (2013), Jerez et al. (2010), Merlin et al. (2010) and Somasundaram and Nedunchezhian (2011)
Decision tree (DT)	*Classification and regression tree (CART)*: Burgette and Reiter (2014), Ding and Simonoff (2010), Doove et al. (2014), Ghanad-Rezaie et al. (2010), Hapfelmeier et al. (2012), Iacus and Porro (2007) and Purwar and Singh (2015)
	C4.5: Ding and Simonoff (2010), Fortes et al. (2006), Hruschka et al. (2007), Nishanth and Ravi (2016), Saar-Tsechansky and Provost (2007), Twala (2009) and Twala et al. (2008)
Extreme learning machine	Shao et al. (2017)
Genetic algorithm (GA)	Garcia et al. (2011)
K-nearest neighbor (KNN)	*KNN*: Armitage et al. (2015), Bras and Menezes (2007), Brock et al. (2008), Celton et al. (2010), Chiu et al. (2013), De Souto et al. (2015), Ding and Ross (2012), Doquire and Verleysen (2012), Farhangfar et al. (2008), Eirola et al. (2014), Gan et al. (2006), Garcia-Laencina et al. (2009, 2013), Hron et al. (2010), Hu et al. (2006), Huang et al. (2016, 2017), Iacus and Porro (2007), Jerez et al. (2010), Kang (2013), Khoshgoftaar, and Van Hulse (2008), Li and Parker (2014), Liu et al. (2010), Luengo et al. (2012), Nishanth and Ravi (2016), Oh et al. (2011), Pan et al. (2015), Pati and Das (2017), Paul et al. (2017), Rao et al. (2013), Sehgal et al. (2008, 2009), Stekhoven and Buhlmann (2012), Song et al. (2008), Sun et al. (2009), Tian et al. (2014), Tsai and Chang (2016), Tuikkala et al. (2008), Valdiviezo and Van Aelst (2015), Van Hulse and Khoshgoftaar (2014), Wang et al. (2006), Xia et al. (2017), Yu et al. (2011), Zhang (2008, 2011, 2012), Zhang et al. (2008, 2011, 2015) and Zhu et al. (2011)
	Grey based KNN: Zhang (2012) and Zhu et al. (2012)
	Iterative KNN: Bras and Menezes (2007), Chiu et al. (2013), Hu et al. (2006) and Pati and Das (2017)
	Sequential KNN: Bras and Menezes (2007), Chiu et al. (2013) and Pati and Das (2017)
	Weighted KNN: Luengo et al. (2012) and Shao et al. (2017)
Kernel-based imputation	Zhu et al. (2011)
Multilayer perceptron (MLP)	Garcia-Laencina et al. (2013), Gautam and Ravi (2015), Jerez et al. (2010), Nishanth et al. (2012) and Silva-Ramirez et al. (2011, 2015)

Table 4 (continued)

Techniques	Works
Random forest (RF)	Doove et al. (2014), Hapfelmeier and Ulm (2014), Hapfelmeier et al. (2012), Iacus and Porro (2007), Kapelner and Bleich (2015), Nishanth and Ravi (2016), Purwar and Singh (2015), Shah et al. (2014), Valdiviezo and Van Aelst (2015) and Xia et al. (2017)
Rough set theory (RST)	Clark et al. (2014)
Support vector machine/regression (SVM/SVR)	Aydilek and Arslan (2013), Iacus and Porro (2007), Luengo et al. (2012), Tuikkala et al. (2008) and Wang et al. (2006)

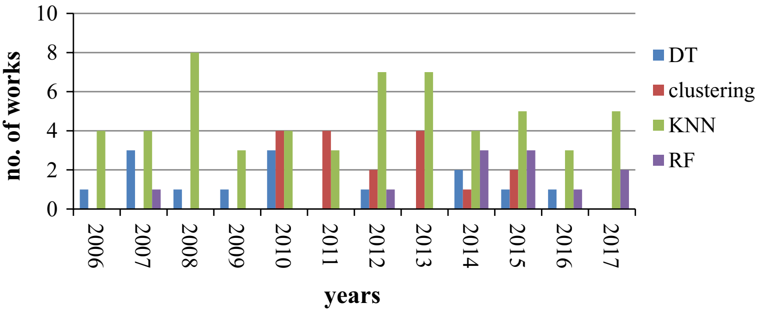

Fig. 5 The year-wise distribution for the number of works that have used the DT, clustering, KNN, and RF techniques

techniques have recently been utilized in a number of studies, whereas DT has been consistently used each year from 2006 and 2016.

According to the researches using one of the top four machine learning based MVI techniques, Table 5 shows the relationships between these techniques and their experimental datasets.

Regarding Table 5, KNN is the most widely used machine learning based MVI technique, especially for UCI and medical datasets. However, for the questionnaire datasets, related studies have never considered machine learning based MVI techniques before. On other hand, for the missing rates, if we count the number of works by DT and RF (an ensemble of DTs) together, very few studies simulated with the missing rates that are smaller than 30% (i.e. 3 out of 22) whereas 13 and 6 works consider the 30–50% and larger than 50% missing rates, respectively. This is different from clustering and KNN that most studies simulated with the missing rates that are smaller than 30% (i.e. 10 out of 16 and 34 out of 60, respectively).

5 Evaluation methods

5.1 Direct evaluation

The final step after imputation of the missing values is to evaluate the imputation results. The most commonly used method is to directly assess the difference between the original

Table 5 The number of works for the relationships between the top four machine learning based MVI techniques and their experimental datasets

Techniques	Missingness mechanisms	Dataset domains	Missing rates
Clustering	MCAR: 14 MAR: 5 NMAR: 3	UCI: 13 Medical: 6 Finance: 1 Others: 1	< 30%: 10 30–50%: 5 > 50%: 1
DT	MCAR: 11 MAR: 7 NMAR: 5	UCI: 9 Medical: 2 Others: 6	< 30%: 1 30–50%: 7 > 50%: 4
KNN	MCAR: 45 MAR: 20 NMAR: 10	UCI: 23 Medical: 30 Images: 1 Software: 4 Others: 8	< 30%: 34 30–50%: 24 > 50%: 2
RF	MCAR: 7 MAR: 6 NMAR: 4	UCI: 6 Medical: 2 Others: 4	< 30%: 2 30–50%: 6 > 50%: 2

values in the collected dataset and the estimated or predicted values in the simulated incomplete dataset. There are two types of attribute values, namely, discrete and continuous. For evaluation of results of the imputation of discrete values, the percentage of values that have been predicted correctly (or incorrectly) for the missing values is usually used, e.g. Nishanth and Ravi (2016) and Valdiviezo and Van Aelst (2015). The percentage of correct predictions (PCP) can be obtained by

$$PCP = 100 \times \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \tag{1}$$

On the other hand, for the imputation of continuous values, the mean absolute percentage error (MAPE) and/or root-mean-square error (RMSE) related measures are used. Gautam and Ravi (2015) and Silva-Ramirez et al. (2015). MAPE and RMSE can be computed by Eqs. (2) and (3), respectively.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2} \tag{3}$$

where x_i is the actual value, $\hat{x}_i$ is the predicted value and n is the total number of missing value.

5.2 Classification accuracy

Another strategy for assessing the imputation quality is to examine the classification performance of some chosen classifiers trained by the imputed datasets. Different from the direct evaluation strategy, after the imputation process is completed, the imputed dataset

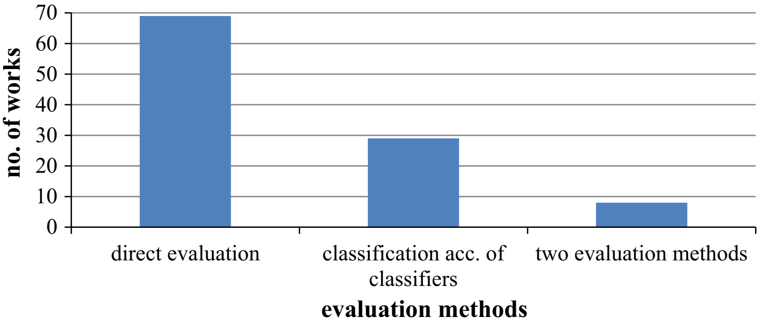

Fig. 6 The number of works using the different evaluation methods

Table 6 Evaluation strategies of related works

Year	Classification accuracy of classifiers	Two evaluation methods
2017	Garciarena and Santana (2017), Ghorbani and Desmarais (2017), Kiasari et al. (2017) and Xia et al. (2017)	
2016	Folino and Pisani (2016), Huang et al. (2016) and Tsai and Chang (2016)	
2015	De Souto et al. (2015), Pan et al. (2015) and Purwar and Singh (2015)	Pan et al. (2015)
2014	Clark et al. (2014) and Li and Parker (2014)	Li and Parker (2014)
2013	Garcia-Laencina et al. (2013), Leung and Leung (2013) and Kang (2013)	Kang (2013)
2012	Doquire and Verleysen (2012), Luengo et al. (2012), Nishanth et al. (2012) and Zhang (2012)	Zhang (2012)
2011	Di Nuovo (2011) and Oh et al. (2011)	Oh et al. (2011)
2010	Ghanad-Rezaie et al. (2010) and Polikar et al. (2010)	
2009	Garcia-Laencina et al. (2009) and Sun et al. (2009)	Sun et al. (2009)
2008	Farhangfar et al. (2008), Song et al. (2008) and Twala et al. (2008)	
2007	Hruschka et al. (2007) and Saar-Tsechansky and Provost (2007)	Hruschka et al. (2007)
2006	Lin et al. (2006)	Lin et al. (2006)

without missing values is used to train some specific classifier(s), and another testing set is chosen to test their classification performance. Since using different imputation methods for the same incomplete datasets is likely to produce different imputation results, the classifier with higher classification accuracy is indicated by the better imputation quality of its training and datasets. Consequently, the better imputation methods can be identified.

The proportion of related works that utilize this type of evaluation strategy is much smaller than the number utilizing the direct evaluation strategy. Moreover, studies which consider both evaluation strategies at the same time are even rarer. Figure 6 shows the number of works using these evaluation methods and Table 6 lists the related works that consider the classification accuracy of classifiers and two evaluation methods at the same time. As we can see that very few studies consider both evaluation methods at the same time.

Ten different classification techniques have been employed in related works for analysis of the classification accuracy of the classifiers, as shown in Fig. 7. As can be seen, KNN,

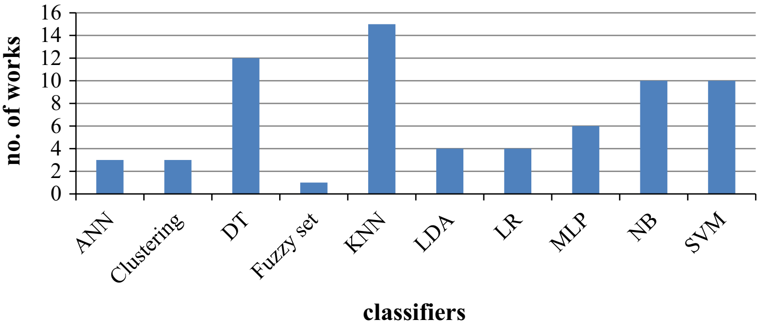

Fig. 7 The number of works using the ten different classifiers

DT, NB, SVM, and MLP are the top five classifiers constructed for evaluating the imputation results.

5.3 Computational time

In addition to the two afore-mentioned evaluation strategies, it is also important to take into consideration the computational time for each MVI method. This issue is especially critical for machine learning based MVI techniques which require some time for the model training step. Moreover, when the size of the dataset as well as the missing rate is very large, the imputation process is likely to take a lot of time. Among the articles reviewed, only 16 examined the computation time, these are Huang et al. (2017), Kiasari et al. (2017), Saha et al. (2017), Valdiviezo and Van Aelst (2015), Li and Parker (2014), Shah et al. (2014), Tian et al. (2014), Aydilek and Arslan (2013), Liu et al. (2013), Rahman and Islam (2013), Stekhoven and Buhlmann (2012), Zhu et al. (2012), Zhang et al. (2011), Tuikkala et al. (2008), Farhangfar et al. (2007), and Lin et al. (2006).

5.4 Missing data simulation strategies for missing value imputation

The simulation of an incomplete dataset with a specific missing rate is usually performed several times in order to avoid producing biased imputation results. This is because the missing data in the incomplete dataset can be different for each simulation even with the same missing rate. In general, there are two strategies used to perform a missing data simulation. The first one is to directly use the whole of the chosen dataset, making it become an incomplete dataset based on a specified missing rate.

The second strategy is to first divide the chosen dataset into training and testing subsets by some method such as *n*-fold cross validation (CV) (Kohavi 1995), or by adopting fixed proportions for the training and testing subsets, e.g. 70% and 30%, respectively. Then, either the training or testing subset is used to perform the missing data simulation for a specific missing rate.

Figure 8 shows the number of works that have used the different strategies, including the whole dataset (denoted as whole), training set obtained by cross validation (CV-training),

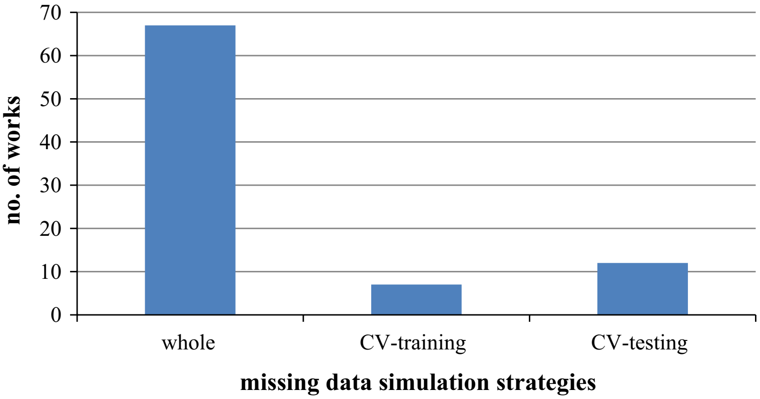

Fig. 8 The number of works using the whole, CV-training, and CV-testing simulation strategies

and the testing set obtained by cross validation (CV-testing). Note that studies that do not clearly describe their simulation strategies are not counted here.

This result shows that most studies use the whole dataset strategy, with the imputation results usually evaluated by the direction evaluation method. Only a small proportion of related studies use the cross validation method for missing data simulation. It should be noted that there is no study which has considered both training and testing subsets when performing missing data simulation for a specific missing rate. This simulation method is much closer to real world problems based on historical data where the collected data contain some missing values and the new unknown testing data may also contain some missing values.

6 Discussion

The above results of analysis of the related literature show the existence of some limitations related to technical issues in the experimental procedure, which can be regarded as future research directions for MVI. They are discussed in greater detail below.

6.1 The chosen datasets

The domain problem datasets for MVI can be broadly classified into two categories. The first type is based on a number of UCI datasets that cover various domain problems. The second is based on specific (real world) domain datasets. Although a survey of related works in the first category study show variety of different domain datasets, they are not as large in scale as real world datasets, which often contain a very large number of feature dimensions (e.g. over 100) and/or data samples (e.g. over 100,000).

In addition, in terms of attribute values, the chosen dataset can contain categorical, numerical, or both categorical and numerical types of data. Using different types of data may affect the imputation performances of different MVI methods.

We now discuss the missing rate used for the simulation for a chosen dataset. It is very hard to define the missing rates as practical, say less than 30%. To deal with real world problems, however, it would be better to be able to perform simulations with larger missing rates (e.g. 70%) or a wide range of missing rates (e.g. from 10 to 90%). The findings from this kind of simulation would be more practical.

About the missing mechanisms, different domain problem datasets with missing data may be occurred based on the MCAR, MAR, and NMAR scenarios. Consideration of only one type of scenario in the incomplete dataset simulation is not enough to fully understand the imputation performance of the MVI methods.

Another issue that could affect the imputation results is whether to perform feature and/or instance selection before or after MVI. Feature and instance selection are aimed at filtering out unrepresentative features and data samples from a given dataset, respectively. Performing one or both of these tasks before MVI could make the complete dataset 'cleaner', which might lead to produce better imputation results.

Alternatively, performing feature and/or instance selection over an imputed dataset after MVI could make the classifier perform better than one based on the imputed dataset without feature and/or instance selection. There have been very few studies which have considered the effect of feature/instance selection on MVI (Aussem and de Morais 2010; Doquire and Verleysen 2012; Hapfelmeier and Ulm 2014; Huang et al. 2016; Sun et al. 2009; Tsai and Chang 2016).

6.2 The MVI techniques

The various baseline MVI techniques discussed in related works can be classified into two types, statistical or machine learning based techniques. The analytical results detailed in Sect. 4 clearly identify the most popular MVI techniques. However, there has been no comprehensive study comparing these well-known MVI techniques in terms of different domain datasets containing a wide range of missing rates based on different missing mechanisms. The findings of this study allow us to understand which technique(s) are more suitable for which kind of incomplete dataset. The results can be regarded as guidelines for the choice of the most representative MVI technique(s) in future work.

Several novel approaches have been proposed that do not require performing the MVI process to tackle the incomplete datasets; see for example, Conroy et al. (2016), Polikar et al. (2010), and Yan et al. (2017). It would very useful to examine the final classification accuracy of these constructed classifiers based on these approaches as well as the imputed datasets by the representative baseline MVI technique(s). This study can answer the question: When should we perform missing value imputation?

Furthermore, most of the proposed novel (hybrid) approaches are either statistical techniques (such as the studies on dynamic Fisher's linear discrimination by Leung and Leung 2013; iterative bi-cluster based least squares by Cheng et al. 2012; interval imputation by PCA by Zuccolotto 2012, etc.), or machine learning based techniques (such as the studies by Folino and Pisani (2016) who combined genetic programming and ensemble learning models; Zhang et al. (2015) who combined particle swarm optimization and fuzzy c-means; Silva-Ramirez et al. (2015) who combined MLP and KNN, and so on). The results show that there have been very few studies where a combination of both types of MVI techniques is considered.

6.3 The evaluation methods

Evaluation is very critical for validation of the performance of the MVI technique and reaching the final conclusions. As noted in Sect. 5, the direct evaluation method, the classification accuracy of the classifiers, and consideration of the computational time are the three main ways to evaluate the MVI technique. Using all three of these evaluation metrics would allow us to fully understand the performance as well as provide suggestions for the development of better technique(s). However, this has not been the case, all three evaluation metrics are not used together in most related studies, which is one of the main limitations of current work in the literature and should be considered in future research.

It is suggested that when the chosen dataset is divided into training and testing subsets for missing data simulation, it would be more practical to make both subsets become incomplete rather than focus on only one of them. For instance, start with a collected historical dataset which is incomplete. After performing the MVI process the imputed dataset is used to train a classifier, after which a new testing dataset can be collected, ready for the classification task. However, it could happen that this testing dataset is also incomplete, so MVI is required. After performing MVI over the incomplete testing dataset, it can then be fed into the constructed classifier.

It this case, the missing rates of the training and testing datasets can significantly affect the final classification accuracy of the classifiers and what is the best combination of MVI techniques and classifiers should be answered.

7 Conclusion

Missing value imputation (MVI) for incomplete datasets is a very important problem in data mining and big data analysis. If the incomplete datasets are not well imputed, the final mining or analysis result could be affected. This paper discusses a literature review and analysis of 111 related journal articles published from 2006 to 2017.

The review and analysis focus on the issues encountered during the MVI process. They include (1) the chosen datasets as well as their domain problems, missing rates, and missingness mechanisms in the simulation, (2) the MVI techniques employed, and (3) the evaluation methods considered.

The analysis results show the existence of many limitations encountered in the current literature, which can be improved upon in future. In summary, these include the scalability of datasets, the wide range of missing rates with the MCAR, MAR, and NMAR missingness mechanisms, the representative MVI baseline techniques, the development of novel hybrid approaches by combining statistical and machine learning based techniques, the consideration of three evaluation metrics together, and missing data simulation for both training and testing datasets.

Acknowledgements The work of the first author was supported in part in part by the Healthy Aging Research Center, Chang Gung University from the Featured Areas Research Center Program within the Framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan under Grants EMRPD1I0481 and EMRPD1I0501, and in part by Chang Gung Memorial Hospital, Linkou under Grant CMRPD3I0031. This research of the second author was supported by the Ministry of Science and Technology of Taiwan (MOST 105-2410-H-008-043-MY3).

References

Acuna E, Rodriguez C (2004) The treatment of missing values and its effect in the classifier accuracy. In: Banks D et al (eds) Classification, clustering and data mining applications. Springer, Berlin, pp 639–648

Aittokallio T (2009) Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief Bioinform 11(2):253–264

Armitage EG, Godzien J, Alonso-Herranz V, Lopez-Gonzalvez A, Barbas C (2015) Missing value imputation strategies for metabolomics data. Electrophoresis 36:3050–3060

Aussem A, de Morais SR (2010) A conservative feature subset selection algorithm with missing data. Neurocomputing 73:585–590

Aydilek IB, Arslan A (2012) A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. Int J Innov Comput Inf Control 8(7):4705–4717

Aydilek IB, Arslan A (2013) A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci 233:25–35

Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48:5–37

Bras LP, Menezes JC (2007) Improving cluster-based missing value estimation of DNA microarray data. Biomol Eng 24:273–282

Brock GN, Shaffer JR, Blakesley RE, Lotz MJ, Tseng GC (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinform 9:12–23

Burgette LF, Reiter JP (2014) Multiple imputation for missing data via sequential regression trees. Am J Epidemiol 172(9):1070–1076

Celton M, Malpertuy A, Lelandais G, de Brevern AG (2010) Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genom 11:15–30

Chen X, Wei Z, Li Z, Liang J, Cai Y, Zhang B (2017) Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation. Knowl Based Syst 132:249–262

Cheng KO, Law NF, Siu WC (2012) Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. Pattern Recogn 45:1281–1289

Chiu C-C, Chan S-Y, Wang C-C, Wu W-S (2013) Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Syst Biol 7:S12

Clark PG, Grzymala-Busse JW, Rzasa W (2014) Mining incomplete data with singleton, subset and concept probabilistic approximations. Inf Sci 280:368–384

Conroy B, Eshelman L, Potes C, Xu-Wilson M (2016) A dynamic ensemble approach to robust classification in the presence of missing data. Mach Learn 102:443–463

De Leeuw ED (2001) Reducing missing data in surveys: an overview of methods. Qual Quant 35:147–160

De Souto MCP, Jaskowiak PA, Costa IG (2015) Impact of missing data imputation methods on gene expression clustering and classification. Bioinformatics 16:64–72

Di Nuovo AG (2011) Missing data analysis with fuzzy c-means: a study of its application in a psychological scenario. Expert Syst Appl 38:6793–6797

Di Zio M, Guarnera U, Luzi O (2007) Imputation through finite Gaussian mixture models. Comput Stat Data Anal 51:5305–5316

Ding Y, Ross A (2012) A comparison of imputation methods for handling missing scores in biometric fusion. Pattern Recogn 45:919–933

Ding Y, Simonoff JS (2010) An investigation of missing data methods for classification trees applied to binary response data. J Mach Learn Res 11:131–170

Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59:1087–1091

Doove LL, Van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal 72:92–104

Doquire G, Verleysen M (2012) Feature selection with missing data using mutual information estimators. Neurocomputing 90:3–11

Eirola E, Doquire G, Verleysen M, Lendasse A (2013) Distance estimation in numerical data sets with missing values. Inf Sci 240:115–128

Eirola E, Lendasse A, Vandewalle V, Biernacki C (2014) Mixture of Gaussians for distance estimation with missing data. Neurocomputing 131:32–42

Farhangfar A, Kurgan LA, Pedrycz W (2007) A novel framework for imputation of missing values in databases. IEEE Trans Syst Man Cybern A Syst Humans 37(5):692–709

Farhangfar A, Kurgan LA, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recogn 41:3692–3705

Folino G, Pisani FS (2016) Evolving meta-ensemble of classifiers for handling incomplete and unbalanced datasets in the cyber security domain. Appl Soft Comput 47:179–190

Fortes I, Mora-Lopez L, Morales R, Triguero F (2006) Inductive learning models with missing values. Math Comput Model 44:790–806

Gan X, Liew AW-C, Yan H (2006) Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucleic Acids Res 34(5):1608–1619

Garcia JCF, Kalenatic D, Bello CAL (2011) Missing data imputation in multivariate data by evolutionary algorithms. Comput Hum Behav 27:1468–1474

Garcia-Laencina PJ, Sancho-Gomez J-L, Figueiras-Vidal AR, Verleysen M (2009) *K* nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72:1483–1493

Garcia-Laencina PJ, Sancho-Gomez J-L, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. Neural Comput Appl 19:263–282

Garcia-Laencina PJ, Sancho-Gomez J-L, Figueiras-Vidal AR (2013) Classifying patterns with missing values using multi-task learning perceptrons. Expert Syst Appl 40:1333–1341

Garciarena U, Santana R (2017) An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst Appl 89:52–65

Gautam C, Ravi V (2015) Data imputation via evolutionary computation, clustering and a neural network. Neurocomputing 156:134–142

Ghanad-Rezaie M, Soltanian-Zadeh H, Ying H, Dong M (2010) Selection-fusion approach for classification of datasets with missing values. Pattern Recogn 43:2340–2350

Ghorbani S, Desmarais MC (2017) Performance comparison of recent imputation methods for classification tasks over binary data. Appl Artif Intell 31(1):1–22

Graham JW, Olchowski AE, Gilreath TD (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 8:206–213

Hapfelmeier A, Ulm K (2014) Variable selection by random forests using data with missing values. Comput Stat Data Anal 80:129–139

Hapfelmeier A, Hothorn T, Ulm K (2012) Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Comput Stat Data Anal 56:1552–1565

Harel O, Zhou X-H (2007) Multiple imputation: review of theory, implementation and software. Stat Med 26:3057–3077

He Y, Zaslavsky AM, Harrington DP, Catalano HP, Landrum MB (2009) Multiple imputation in a large-scale complex survey: a practical guide. Stat Methods Med Res 19(6):653–670

Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Comput Stat Data Anal 54:3095–3107

Hruschka ER Jr, Hruschka ER, Ebecken NFF (2007) Bayesian networks for imputation in classification problems. J Intell Inf Syst 29:231–252

Hu J, Li H, Waterman MS, Zhou XJ (2006) Integrative missing value estimation for microarray data. BMC Bioinform 7:449–462

Huang MW, Lin W-C, Chen C-W, Ke S-W, Tsai C-F, Eberle W (2016) Data preprocessing issues for incomplete medical datasets. Expert Syst 33(5):432–438

Huang J, Keung JW, Sarro F, Li Y-F, Yu YT, Chan WK, Sun H (2017) Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. J Syst Softw 132:226–252

Iacus SM, Porro G (2007) Missing data imputation, matching and other applications of random recursive partitioning. Comput Stat Data Anal 52:773–789

Janssen KJM, Donders ART, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, Moons KGM (2010) Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 63:721–727

Jerez JM, Molina I, Garcia-Laencina PJ, Alba E, Ribelles N, Martin M, Franco L (2010) Missing data imputation using statistical and machine learning methods in real breast cancer problem. Artif Intell Med 50:105–115

Kang P (2013) Locally linear reconstruction based missing value imputation for supervised learning. Neurocomputing 118:65–78

Kapelner A, Bleich J (2015) Prediction with missing data via Bayesian additive regression trees. Can J Stat 43(2):224–239

Khoshgoftaar TM, Van Hulse J (2008) Imputation techniques for multivariate missingness in software measurement data. Softw Qual J 16:563–600

Kiasari MA, Jang G-J, Lee M (2017) Novel iterative approach using generative ad discriminative models for classification with missing features. Neurocomputing 225:23–30

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Conf Artif Intell 2:1137–1143

Leung KC, Leung CH (2013) Dynamic discriminant functions with missing feature values. Pattern Recogn Lett 34:1548–1556

Li YY, Parker LE (2014) Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks. Inf Fusion 15:64–79

Li D, Gu H, Zhang L (2010) A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Expert Syst Appl 37:6942–6947

Li Z, Sharaf MA, Sitbon L, Sadiq S, Indulska M, Zhou X (2014) A web-based approach to data imputation. World Wide Web 17:873–897

Liao S, Lin Y, Kang DD, Chandra D, Bon J, Kaminski N, Sciurba FC, Tseng GC (2014) Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinform 15:346–357

Liew AW-C, Law N-F, Yan H (2011) Missing value imputation for gene expression data: computation techniques to recover missing data from available information. Brief Bioinform 12(5):498–513

Lin T, Lee JC, Ho HJ (2006) On fast supervised learning for normal mixture models with missing information. Pattern Recogn 39:1177–1187

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, Hoboken

Liu C-C, Dai D-Q, Yan H (2010) The theoretic framework of local weighted approximation for microarray missing value estimation. Pattern Recogn 43:2993–3002

Liu J, Musialski P, Wonka P, Ye J (2013) Tensor completion for estimating missing values in visual data. IEEE Trans Pattern Anal Mach Intell 35(1):208–220

Luengo J, Garcia S, Herrera F (2012) On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst 32:77–108

Merlin P, Sorjamaa A, Maillet B, Lendasse A (2010) X-SOM and L-SOM: a double classification approach for missing value imputation. Neurocomputing 73:1103–1108

Mesquite DPP, Gomes JPP, Junior AHS, Nobre JS (2017) Euclidean distance estimation in incomplete datasets. Neurocomputing 248:11–18

Moons KGM, Donders RART, Stijnen T, Harrell FE Jr (2006) Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 59:1092–1101

Munoz JF, Rueda M (2009) New imputation methods for missing data using quantiles. J Comput Appl Math 232:305–317

Nishanth KJ, Ravi V (2016) Probabilistic neural network based categorical data imputation. Neurocomputing 218:17–25

Nishanth KJ, Ravi V, Ankaiah N, Bose I (2012) Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts. Expert Syst Appl 39:10583–10589

Oh S, Kang DD, Brock GN, Tseng GC (2011) Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics 27(1):78–86

Pan R, Yang T, Cao J, Lu K, Zhang Z (2015) Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. Appl Intell 43:614–632

Pati SK, Das AK (2017) Missing value estimation for microarray data through cluster analysis. Knowl Inf Syst 52(3):709–750

Paul A, Sil J, Mukhopadhyay CD (2017) Gene selection for designing optimal fuzzy rule base classifier by estimating missing value. Appl Soft Comput 55:276–288

Peng C-Y, Zhu J (2008) Comparison of two approaches for handling missing covariates in logistic regression. Educ Psychol Measur 68:58–77

Polikar R, DePasquale J, Mohammed HS (2010) Learn^{++}.MF: a random subspace approach for the missing feature problem. Pattern Recogn 43:3817–3832

Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 42:5621–5631

Qin Y, Zhang S, Zhu X, Zhang J, Zhang C (2007) Semi-parametric optimization for missing data imputation. Appl Intell 27(1):79–88

Qin Y, Zhang S, Zhu X, Zhang J, Zhang C (2009) POP algorithm: kernel-based imputation to treat missing values in knowledge discovery from databases. Expert Syst Appl 36:2794–2804

Rahman MdG, Islam MdZ (2013) Missing value imputation using decision trees and decision forests by splittling and merging records: two novel techniques. Knowl Based Syst 53:51–65

Rao SSS, Shepherd LA, Bruno AE, Liu S, Miecznikowski JC (2013) Comparing imputation procedures for affymetrix gene expression datasets using MAQC datasets. Adv Bioinform 2013:790567

Raymond M, Roberts D (1987) A comparison of methods for treating incomplete data in selection research. Educ Psychol Meas 47:13–26

Saar-Tsechansky M, Provost F (2007) Handling missing values when applying classification models. J Mach Learn Res 8:1625–1657

Saha B, Gupta S, Phung D, Venkatesh S (2017) Effective sparse imputation of patient conditions in electronic medical records for emergency risk predictions. Knowl Inf Syst 53(1):179–206

Sehgal MSB, Gondal I, Dooley LS, Coppel R (2008) Ameliorative missing value imputation for robust biological knowledge inference. J Biomed Inform 41:499–514

Sehgal MSB, Gondal I, Dooley LS, Coppel R (2009) How to improve postgenomic knowledge discovery using imputation. EURASIP J Bioinform Syst Biol 2009:717136

Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H (2014) Comparison of random forest and parametric imputation models for imputing missing data using MICE: a caliber study. Am J Epidemiol 179(6):764–774

Shao J, Meng W, Sun G (2017) Evaluation of missing value imputation methods for wireless soil datasets. Pers Ubiquit Comput 21(1):113–123

Silva-Ramirez E-L, Pino-Mejias R, Lopez-Coello M, Cubiles-de-la-Vega M-D (2011) Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw 24:121–129

Silva-Ramirez E-L, Pino-Mejias R, Lopez-Coello M (2015) Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. Appl Soft Comput 29:65–74

Somasundaram RS, Nedunchezhian R (2011) Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. Int J Comput Appl 12(10):14–19

Song Q, Shepperd M, Chen X, Liu J (2008) Can k-NN imputation improve the performance of C4.5 with small software project datasets? A comparative evaluation. J Syst Softw 81:2361–2370

Stekhoven DJ, Buhlmann P (2012) MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118

Strike K, Emam KE, Madhavji N (2001) Software cost estimation with incomplete data. IEEE Trans Softw Eng 27(10):890–908

Subasi MM, Subasi E, Anthony M, Hammer PL (2011) A new imputation method for incomplete binary data. Discrete Appl Math 159:1040–1047

Sun Y, Braga-Neto U, Dougherty ER (2009) Impact of missing value imputation on classification for DNA microarray gene expression data—a model-based study. EURASIP J Bioinform Syst Biol 2009:504069

Tian J, Yu B, Yu D, Ma S (2014) Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering. Appl Intell 40:376–388

Tsai C-F, Chang F-Y (2016) Combining instance selection for better missing value imputation. J Syst Softw 122:63–71

Tsikriktsis N (2005) A review of techniques for treating missing data in OM survey research. J Oper Manag 24:53–62

Tuikkala J, Elo LL, Nevalainen OS, Aittokallio T (2008) Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC Bioinform 9:202–215

Twala B (2009) An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell 23(5):373–405

Twala BETH, Jones MC, Hand DJ (2008) Good methods for coping with missing data in decision trees. Pattern Recogn Lett 29:950–956

Valdiviezo HC, Van Aelst S (2015) Tree-based prediction on incomplete data using imputation or surrogate decision. Inf Sci 311:163–181

Van Ginkel JR, Kroonenberg PM (2014) Using generalized procrustes analysis for multiple imputation in principal component analysis. J Classif 31:242–269

Van Ginkel JR, Van der Ark LA, Sijtsma K, Vermunt JK (2007) Two-way imputation: a Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. Comput Stat Data Anal 51:4013–4027

Van Hulse J, Khoshgoftaar TM (2014) Incomplete-case nearest neighbor imputation in software measurement data. Inf Sci 259:596–610

Wang X, Li A, Jiang Z, Feng H (2006) Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. BMC Bioinform 7:32–41

Xia J, Zhang S, Cai G, Li L, Pan Q, Yan J, Ning G (2017) Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recogn 69:52–60

Yan Y-T, Zhang Y-P, Zhang Y-W, Du X-Q (2017) A selective neural network ensemble classification for incomplete data. Int J Mach Learn Cybern 8(5):1513–1524

Yu T, Peng H, Sun W (2011) Incorporating nonlinear relationships in microarray missing value imputation. IEEE/ACM Trans Comput Biol Bioinf 8(3):723–731

Zhang S (2008) Parimputation: from imputation and null-imputation to partially imputation. IEEE Intell Inform Bull 9(1):32–38

Zhang S (2011) Shell-neighbor method and its application in missing data imputation. Appl Intell 35:123–133

Zhang S (2012) Nearest neighbor selection for iteratively kNN imputation. J Syst Softw 85:2541–2552

Zhang Y, Liu Y (2009) Data imputation using least squares support vector machines in urban arterial streets. IEEE Signal Process Lett 16(5):414–417

Zhang X, Song X, Wang H, Zhang H (2008) Sequential local least squares imputation estimating missing value of microarray data. Comput Biol Med 38:1112–1120

Zhang S, Jin Z, Zhu X (2011) Missing data imputation by utilizing information within incomplete instances. J Syst Softw 84:452–459

Zhang L, Bing Z, Zhang L (2015) A hybrid clustering algorithm based on missing attribute interval estimation for incomplete data. Pattern Anal Appl 18:377–384

Zhu X, Zhang S, Jin Z, Zhang Z, Xu Z (2011) Missing value estimation for mixed-attribute data sets. IEEE Trans Knowl Data Eng 23(1):110–121

Zhu B, He C, Liatsis P (2012) A robust missing value imputation method for noisy data. Appl Intell 36:61–74

Zuccolotto P (2012) Principal component analysis with interval imputed missing values. AStA Adv Stat Anal 96:1–23

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.