

# Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering

Jing Tian · Bing Yu · Dan Yu · Shilong Ma

Published online: 5 September 2013  
© Springer Science+Business Media New York 2013

**Abstract** Researchers and practitioners who use databases usually feel that it is cumbersome in knowledge discovery or application development due to the issue of missing data. Though some approaches can work with a certain rate of incomplete data, a large portion of them demands high data quality with completeness. Therefore, a great number of strategies have been designed to process missingness particularly in the way of imputation. Single imputation methods initially succeeded in predicting the missing values for specific types of distributions. Yet, the multiple imputation algorithms have maintained prevalent because of the further promotion of validity by minimizing the bias iteratively and less requirement on prior knowledge to the distributions.

This article carefully reviews the state of the art and proposes a hybrid missing data completion method named *Multiple Imputation using Gray-system-theory and Entropy based on Clustering (MIGEC)*. Firstly, the non-missing data instances are separated into several clusters. Then, the imputed value is obtained after multiple calculations by utilizing the information entropy of the proximal category for each incomplete instance in terms of the similarity metric based on *Gray System Theory (GST)*.

Experimental results on *University of California Irvine (UCI)* datasets illustrate the superiority of *MIGEC* to other

current achievements on accuracy for either numeric or categorical attributes under different missing mechanisms. Further discussion on real aerospace datasets states *MIGEC* is also applicable for the specific area with both more precise inference and faster convergence than other multiple imputation methods in general.

**Keywords** Missing data · Multiple imputation · Gray System Theory · Entropy · Clustering

## 1 Introduction

Machine learning and data mining algorithms are frequently used for knowledge discovery in databases since they are non-trivial processes of exploring the new facts and identifying helpful relationships or patterns in data [5, 21, 26]. However, analysts using real-world databases or datasets constantly encounter data imperfection in the form of incompleteness [3, 8]. Thus, plenty of resolutions have been devised to tackle the unfavorable phenomenon. Even though missing data might not cause any serious trouble especially when the missing ratio is not significantly high, it would not be an ideal case to ensure the data quality. Additionally, some opinions argue that fragmentary data should be directly discarded without any further considerations. Nevertheless, the opinion has obvious shortcomings which are articulated that the other observed factual values of the same instance may simultaneously be absent [11, 18, 35]. In some high-missing-rate environment, this strategy is presumed unreasonable and infeasible. Consequently the handling for substitution or replacement draws increasing attentions, termed as imputation. In broad outline, the methods available can be separated into two categories: single imputation and multiple imputation methods. Single imputation,

---

J. Tian (✉) · B. Yu · D. Yu · S. Ma  
State Key Laboratory of Software Development Environment,  
Beihang University, Beijing, China  
e-mail: tianjing@nlsde.buaa.edu.cn

B. Yu  
e-mail: yubing@nlsde.buaa.edu.cn

D. Yu  
e-mail: yudan@nlsde.buaa.edu.cn

S. Ma  
e-mail: slma@nlsde.buaa.edu.cn

*i.e.* filling in precisely one value for each missing one, intuitively has many appealing features, *e.g.* standard complete-data methods can be applied directly and the substantial effort required to create imputations needs to be carried out only once [23, 47]. At the same time, it cannot provide valid standard errors and confidence intervals, because it ignores the uncertainty implicit in the fact that the imputed values are not the actual values. Oppositely, multiple imputation generates a quantity of simulated values for each missing item, in order to reflect properly the uncertainty attached to missing data [42, 58]. More importantly, multiple imputation accounts for missing data by restoring not only the natural variability in the missing-data, but also by also incorporating the uncertainty caused by the interpolation process. This type of data imputation has been advocated as a statistically sound approach, but so far its use has been limited mainly to the social and medical sciences.

Multiple imputation recently has emerged as an interesting and quite visible direction in missing data analyses. The versions of the sophisticated approach are advantageous to these conventional techniques because they require less stringent assumptions and mitigate the pitfalls of traditional ones [19, 36, 49]. Nevertheless, for most of the existing solutions, the following facets retain defective: (a) The clustering strategy combining complete instances with incomplete instances violates the formation of good clusters. In other words, the entire-instances-involved clustering generates biased values due to the imperfection [57, 59]. (b) The common distance metrics such as *Minkowski's*  $L_p$  ( $p = 1, 2, \infty$ ) and *Cosine Correlation (CC)* [34, 41] is imprecise to scale the dissimilarity among instances; (c) The current methodologies are hardly applicable to handling the missing aerospace data due to the their underperformance on validity.

In this paper, data imputation is formulated as a problem of estimation of missing values by multiple operations based on clustering. Furthermore, the prime contribution of this paper could be described as: (a) Dividing non-missing items into a finite number of well-partitioned clusters contributes to make the completion in the optimal tailored area. (b) *GST*, which signifies the situational variation of the curve, could characterize the relative discrepancy more precisely; (c) *MIGEC* performs accurately on the real aerospace datasets than other multiple imputation strategies.

The rest of this paper is organized as: Sect. 2 first briefly introduces both the missingness mechanisms and the emblematic patterns of missing data treatment then reviews the diverse related literatures about imputation. In Sect. 3, the detailed process of the *MIGEC* algorithm is illustrated in three primary procedural sub-items. Section 4 demonstrates a series of experimental results on both *UCI* datasets and empirical aerospace datasets to compare the performance with other state of the art including both single imputation

and multiple imputation. Finally conclusions are given in Sect. 5.

## 2 Related work

### 2.1 Missingness mechanisms

Before the discussions on different handling options, it is necessary to have a solid comprehension of missingness mechanism. That is because the performance of methods does not depend only on the amount of absent data but on the characteristics of the missing data patterns. According to *Rubin* and his colleagues' taxonomy, the mechanisms are categorized as follows [5, 19, 31, 48]:

1. *Missing Completely At Random (MCAR)*. The *MCAR* refers to the case that the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data. The probability that units provide data on a particular variable, thus, does not depend on the value of that variable or the value of any other variable. An example of the mechanism is that a laboratory sample is dropped, so the resulting observation is missing.
2. *Missing At Random (MAR)*. Once the distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data, the mechanism is *MAR*. As the probability of a value being missing will generally depend on observed values, it does not correspond to the intuitive notion of 'random'. For example, if income is more likely to be missing for the more educated and education is fully observed.
3. *Not Missing At Random (NMAR)*. It implies that the pattern of data missingness is non-random and it is unpredictable from other variables in the database. If missing data are *NMAR*, even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. In other words, the missing entry relies on the observed data as well as on the value of the data which is missing.

These terms own precise probabilistic and mathematical implications since they explain why the data are missing. Moreover, the conceptual descriptions state the relationships between observed variables and the probability of missing data. So they have to be involved into missing data analysis.

### 2.2 Methods for missing data analysis

Current administrations of processing missing data can be approximately divided into three categories: tolerance, ignoring and imputation-based procedures.

### 2.2.1 Tolerance

The straightforward method aims to maintain the source entries in the incomplete fashion. It may be a practical and computationally low cost solution, whereas it requires the techniques to work robustly even if the data quality stays low [20, 48].

### 2.2.2 Ignoring

Missing data ignorance often refers to “*Case Deletion*”. It is the most frequently applied procedure, but it is prone to diminish the data quality. The strength lies in the ease of application: deleting the elements with missing values in two manners [15, 55]:

- (a) List-wise/Case-wise Deletion: it performs indiscriminately deleting from the database any elements with missing data for any of the attributes being examined.
- (b) Pairwise Deletion: incomplete cases are removed on an analysis-by-analysis basis, such that any given case may contribute to some analyses but not to others.

### 2.2.3 Imputation

*Mean/Mode Substitution (MMS)*: It replaces the missing values by the mean (the arithmetic average value) or mode (the highest frequency of occurrence) of all the observations or a subgroup at the same variable. It consists of replacing the unknown value for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute. Replacing all missing records with a single value distorts the input data distribution [9, 22].

*Hot-deck/Cold-deck Imputation* [9, 21]: Given an incomplete pattern, *Hot-Deck Imputation (HDI)* replaces the missing data with the values from the input vector that is closest in terms of the attributes that are known in both patterns. This method attempts to preserve the distribution by substituting different observed values for each missing item. Another possibility is the *Cold-Deck Imputation (CDI)* method, which is similar to hot deck but the data source must be other than the current dataset. For example, in a survey context, the external source can be a previous realization of the same survey.

*Regression imputation*: This method uses multiple linear regression to obtain estimates of the missing values. It is applied by estimating a regression equation for each variable, using the others as predictors. This solves the problems concerning variance and covariance raised by the previous method but leads to polarization of all the variables if they are not linked in a linear fashion. Possible errors are due to the insertion of highly correlated predictors to estimate the variables. The advantage of this method is that existing relationships between the variables can be used to calculate

missing data, but it is rarely used as it amplifies the correlation between variables [18, 44, 56].

*Expectation Maximization Estimation (EME)*: The algorithm can handle parameter estimation in the presence of missing data, based on *Expectation-Maximization (EM)* algorithm proposed by *Dempster, Laird and Rubin*. These methods are generally superior to case deletion methods, because they utilize all the observed data. However, they suffer from a strict assumption of a model distribution for the variables, such as a multivariate normal model, which has a high sensitivity to outliers [17, 24].

*Machine-learning-based imputation*: It acquires the features of interested unknown data by behavior evolution after sample data processed. The essence is to automatically learn sample for complicated pattern cognition and intelligently predict the missing values. The methods mainly include decision tree based imputation, association rules based imputation and clustering-based imputation [39, 43, 55].

*Multiple imputation*: Several, usually likelihood, ordered choices for imputing the missing value are computed. Each of the two or more resulting complete data sets is then analyzed using standard complete-data methods. All the analysis becomes combined to reflect both the inter-imputation variability and intra-imputation variability [30, 40, 46].

### 2.2.4 State of the art for missing data imputation

Historically, people have relied on diverse ad hoc techniques to deal with missing data. These related methodologies have accomplished substantial and rapid developments during the last decades. *Baraldi A. and Enders C.* [5] made a comparison of the conventional literatures with the most recent methodological researches. They pointed out that the traditional techniques could only work in some limited circumstances with strict assumptions proposed by *Little and Rubin* [35]. *Magnani* [38] also investigated the main missing data techniques, including conventional methods, global imputation, local imputation, parameter estimation and direct management of missing data. They tried to highlight the advantages and disadvantages for all kinds of missing data mechanisms. Clearly, the major problem of these techniques is under strong model assumptions. *Chen and Chen* [11] developed an estimating null value method, where a fuzzy similarity matrix is used to represent fuzzy relations, and the method is used to deal with one missing value in an attribute. The *K Nearest Neighbors (KNN)* [8, 22] is another prevailing means to explore missing data completion, such as *Sequential KNN (SKNN)* and *Iterative KNN (IKNN)*. *Li D. et al.* [33] represented the missing data as intervals which were clustered by a nearest-neighbor-intervals-based *Fuzzy C-Means (FCM)* resulting in interval cluster prototypes that reflect both uncertainty and the shape of clusters. *García-Laencina, P.J. et al.* [22] established a new *KNN*

variant that selects the  $k$  nearest cases considering the relevance between the input and target attribute to classify and impute missing data using mutual-information-based distance metric. Bose S. et al. [8] invented a local interpolation based imputation method which each time generates a similarity sub-matrix about a missing position in a target gene and interpolate the missing data rather than using all genes and samples. Soft computing techniques were also introduced to seek for the solutions. García, J.C.F. et al. [23] presented a genetic algorithm taking advantages of the flexibility and non-linear capability and got successful results even without conditional information. Huang and Lee [27] employed a gray-based nearest neighbor method to handle the missing data problem. The gray association analysis is employed to determine the nearest neighbors of an instance with missing values. And those unknown values are inferred by the known attribute values from these nearest neighbors. Chen and Huang [12] used the weighted fuzzy rules to estimate null values in relational database. Di Nuovo, A.G. [15] made the comparisons among four solutions of FCM in the psychological research environment. The result revealed that the FCM based on *Optimal Completion Strategy (FCMOCS)* lead to effective data imputation instead of deleting elements with missing values. The theoretical underpinnings of multi-imputation are Bayesian [52]. Hruschka Jr. et al. [26] used Bayesian networks to fulfill missing values in a hybrid model, which applies the clustering genetic algorithm in objects without missing values and generates Bayesian networks to substitute the missing values. *Predictive Mean Matching (PMM)* [10, 16] is to fill in the blanks based on the combined residuals, with the residual value forecasts to reflect the uncertainty. The distribution of residuals can be either normal or non-normal. Yet the random error term is often difficult to determine. *Propensity Score (PS)* [37, 45] is a particular processing of conditional probability as the observed covariates in a provided time-scale. Each parametric value with missing scores owns a tendency to imply the probability of lost observations. The observations were categorized into groups according to the trends scores and consequently applied to each group to predict the approximate Bayesian bootstrap. *Markov Chain Monte Carlo (MCMC)* [2, 7] is another kind of Bayesian inference that it explores the posterior distribution. Zhang C. et al. [53] generalized the random regression estimation with a method named “*Clustering-based Random Imputation (CRI)*”, which fills the unknown values with those plausible ones generated from the same cluster using a kernel based random method after splitting the raw data into complete and incomplete sets. *Clustering-based Multiple Imputation (CMI)* [58] was designed to utilize the kernel function nonparametric random imputation to make inference for the missing data after k-means clustering. Zhang S. et al. [56] utilized the information within the incomplete instances

since the second imputation iteration. The *Non-Parametric Iterative Imputation (NIIA)* is an improvement of the classic multiple imputation, which is based on kernel function. The experimental results on UCI datasets unfolded that the NIIA could easily capture the distribution of a dataset even when there is no prior knowledge of the datasets.

### 3 The MIGEC algorithm

The global procedure of the MIGEC algorithm is schematized in Fig. 1. And each of the key components is detailed in the following subsections.

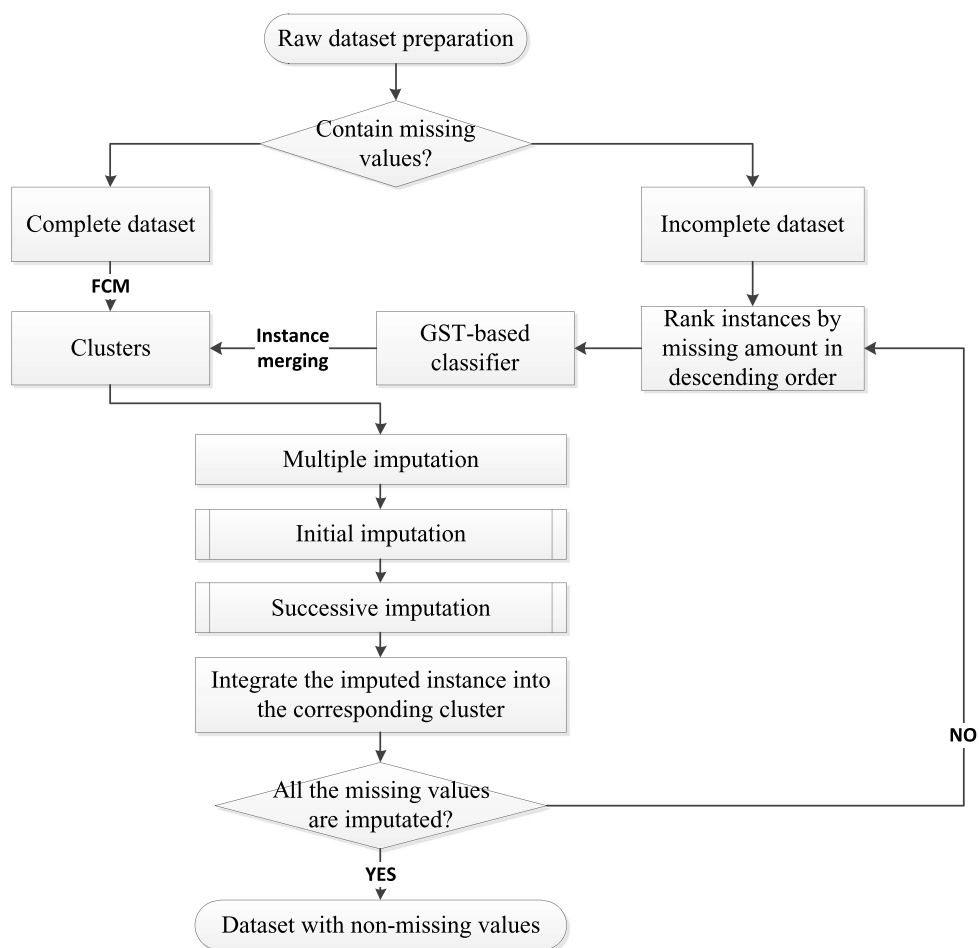
Our method pursues to make full use of the uncorrupt information at instance level. For this reason, the items from the raw data set are divided into two disjoint subsets, namely the complete dataset and the incomplete dataset. It is expected to minimize the negative impact due to the information loss of missing values by the way of separation. On one hand, the objects of the complete set constitute a number of clusters via FCM. On the other hand, the items in the incomplete set are reordered according to the missing severity from high to low. That is, the specific record with the least missing parametric values is firstly allocated to the closest group quantified by the GST-based distance metric. Next, each missing attributive value of the record is estimated by the proposed multiple imputation (including the initial and successive stages). Then the imputed item is included into the complete set along with excluding the original copy from the incomplete set. And the next element in the rearranged incomplete data set repeats the similar solution until no more elements exist in that set. The strategy ensures that all the missing-valued instances could be processed through utilizing the known information as much as possible in the most similar region.

#### 3.1 The clustering strategy

Clustering divides target elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. This is important because it helps to categorize the raw samples into groups with high level of intra-cohesion and inter-separation. In hard clustering (such as  $k$ -means [1]), data is identified into distinct clusters, where each item precisely belongs to one cluster. Soft clustering ensures data elements, which gain a set of membership levels, belong to more than one cluster, though. Moreover, the soft clustering has been successfully applied to a variety of domains with variants.

The specific clustering schema utilizes the standard FCM [6, 25], which aims to minimize the following objective function with respect to fuzzy memberships  $U^{(r)} = [u_{ij}^{(r)}]$  and cluster centroids  $C^{(r)} = c_j^{(r)} : J =$

**Fig. 1** The flowchart of the MIGEC algorithm



$\sum_{j=1}^G \sum_{i=1}^M (u_{ij}^{(r)})^s d(x_i, c_j^{(r)})$ . In these expressions,  $r$  is the ordinal number of the iterations with  $x_k$  and  $c_j^{(r)}$  respectively denoting the  $k$ th complete data instance and the  $j$ th cluster, while  $d(\cdot, \cdot)$  is the distance metric between two instances and  $u_{ij}^{(r)}$  is the degree of membership that the  $i$ th instance is subordinate to the  $j$ th cluster under the “fuzzier”  $s$ , as  $G$  defines the total number of clusters and  $M$  represents the number of data instances.

The *FCM* can be summarized in 4 steps:

- Step 1. Randomly initialize the matrix  $U^{(0)}$ , which satisfies  $\sum_{j=1}^G u_{ij}^{(0)} = 1; i = 1, 2, \dots, M$
- Step 2. From the  $r$ th iteration ( $r > 0$ ), calculate the centroids  $C^{(r)}$ :

$$c_j^{(r)} = \frac{\sum_{i=1}^N u_{ij}^{(r)} x_i}{\sum_{i=1}^N u_{ij}^{(r)}} \tag{1}$$

- Step 3. Update the membership matrix  $U^{(r)}$ :

$$u_{ij}^{(r)} = \sum_{g=1}^G \left[ \frac{d(x_i, c_j^{(r)})}{d(x_i, c_g^{(r)})} \right]^{-\frac{2}{s-1}} \tag{2}$$

- Step 4. Check whether the following *STOP* conditions are satisfied. If not, it returns to Step 2, otherwise the iterative procedure immediately ends with formed clusters.

$$\|U^{(r)} - U^{(r-1)}\| < \varepsilon; \quad \varepsilon > 0$$

or  $r$  accumulatively reaches the predefined number  $R$ .

### 3.2 The classification of incomplete instances

*GST* is established by Deng [13, 14], combined with Gray Control Theory. Both of the two branches adapt to the situation where partial information is difficult to acquire as well as information stays extensional explicit or intensional implicit. For *GST*, the concept of *Gray Relational Analysis (GRA)* remains crucial, as it includes *Gray Relational Coefficient (GRC)* and *Gray Relational Grade (GRG)*. *GRA* is used to scale the influence of a compared series on the reference series in the gray space without prior assumption about the distribution type. Furthermore, it could generate the satisfactory outcome among target objects even if the amount is small or with great variability. Thus, the correlation can



be regarded as a new distance metric, for smaller distance indicates strong influence.

Two steps for distance calculation are needed in a missing environment:

Step 1. The processing of initialization is indispensable in order to map the original data into a particular interval and eliminate the outliers. Then *GRC* is formulated by (3):

$$GRC(x_k^{mis}, c_i) = \frac{\min_i \min_p |x_{kp}^{mis} - c_{ip}| + \zeta \max_i \max_p |x_{kp}^{mis} - c_{ip}|}{|x_{kp}^{mis} - c_{ip}| + \zeta \max_i \max_p |x_{kp}^{mis} - c_{ip}|}, \quad p = 1, 2, \dots, N, \quad 0 \leq \zeta \leq 1 \quad (3)$$

In (3),  $x_k^{mis}$  is the  $k$ th incomplete instance and  $p$  is the  $p$ th attribute with non-missing values, while  $c_i$  denotes the centroid of the  $i$ th cluster. In other words, the calculation only happens when the  $p$ th attributive value of  $x_k^{mis}$  exists. Another important parameter is  $\zeta \in (0, 1]$ , which is used to control the level of differences with respect to the relational coefficient. When  $\zeta = 0$ , the comparison environment does not occur any more. On the contrary,  $\zeta = 1$  shows that the comparison environment remains the unchanged status. A proper value of  $\zeta$  could favorably manage the impact of the maximum value in the matrix. Nevertheless, no methods have been convinced about the optimum value selection so far. Instead, researchers usually choose to empirically set it as 0.5 or learn the optimized one from experimental results [28].

Step 2. Integrating each parameter's *GRC* between an incomplete instance and the reference, the *GRG* is calculated in (4).

$$GRG(x_k^{mis}, c_i) = \frac{1}{N} \sum_{p=1}^N GRC(x_k^{mis}, c_i), \quad i = 1, 2, \dots, M \quad (4)$$

In terms of the maximal value of *GRG*, each incomplete instance is individually incorporated into the closest cluster [6, 54].

From the above descriptions, it should be pointed out that the rationale of the *GRA* is to quantify the similarity and the degree of compactness for the different items based on their geometric relationship.

### 3.3 The entropy-based multiple imputation

Each time that one instance has been assigned to the most proximate group, an internal multiple imputation strategic approach starts as follows:

#### 3.3.1 First imputation

The *MMS* is employed to initialize missing values in the first imputation. The simple technique could perform well only when the data is normally distributed. Yet, it is believed that it could produce excellent performance provided that the missing ones are initialized by *MMS* before the multiple imputation, even without any prior knowledge about the pattern of distribution [56].

#### 3.3.2 Successive imputation

In the authors' opinion, the imputation of one specific missing instance could benefit from all the other instances within the particular cluster. As a result, a method involving all the instances via entropy is proposed. In this context, the term "entropy" refers to the *Shannon's* entropy [29, 50], which quantifies the expected value of the information contained in a message. It states that a broad distribution represents more uncertainty than does a sharply peaked one. And it is used to determine the relative importance of each criterion in the matrix as follows.

$R = (r_{ij})_{m \times n}$  associates with the data matrix of the cluster, into which  $x_i^{mis} \in X_{ic}$  is attached. That is, it includes  $m - 1$  complete elements and one initialized element.

Step 1. Calculate the entropy value of the  $f$ th data instance [29]:

$$I_f = -k * h_f * \ln p_f, \quad k = \frac{1}{\ln m}, \quad h_f = \frac{\|r_{fl} - r_{il}\|}{\sum_{i=1}^m \|r_{fl} - r_{il}\|}, \quad (l \neq j) \quad (5)$$

$I_f$  measures the decision information that the  $f$ th parameter contains.

Step 2. Compute the coefficient of difference for the  $f$ th instance:

$$t_f = 1 - I_f, \quad f = 1, 2, \dots, n \quad (6)$$

$t_f$  represents the inherent contrast intensity of the  $f$ th parameter. The greater value of  $t_f$  signifies the more significance of that parameter.

Step 3. Elicit the coefficient of weight for the  $f$ th copy:

$$w_f = \frac{t_f}{\sum_{f=1}^n t_f} \quad (7)$$

Step 4. Estimate the  $j$ th attributive missing value of  $x_i^{mis}$ :

$$x_{ij}^{mis} = \sum_{q=1, q \neq j}^n w_q x_{iq}^{mis} \quad (8)$$

If the estimated values of the individual instance vary beyond a tolerable interval compared with the calculated value

of the last iteration or the number of iteration times does not reach the threshold, the operations from (5) to (8) continue iteratively. Otherwise, the iterative process mentioned above terminates as the assessed value is considered as the imputed one. Consequently, the imputed instance is aggregated into the corresponding cluster afterwards with updated centroids.

### 3.4 The framework of the MIGEC algorithm

**Input:**  $X_{raw}$ , the  $n \times m$  dimensional dataset with missing values

$G$ , the number of clusters

**Output:**  $X_{full}$ , the  $n \times m$  dimensional complete dataset with imputed values

$X_{raw} \rightarrow X_{obs}, X_{mis}$ ,

where  $X_{raw} = X_{obs} \cup X_{mis}$  and  $\emptyset = X_{obs} \cap X_{mis}$

$FCM(X_{obs}, G) \rightarrow C = \{C_1, C_2, \dots, C_G\}$  according to Sect. 3.1

For each element  $x_k$  in  $X_{mis}$

Allocate  $x_k$  to the closest cluster  $c_q$  according to Sect. 3.2

Complete the missing values of  $x_k$  and according to Sect. 3.3

Integrate the  $x_k$  into corresponding cluster and update  $c_q$  according to Sect. 3.3

$$X_{full} \leftarrow \bigcup_{i=1}^G c_i$$

## 4 Experimental evaluation

In this section, the assessment criteria are primarily explained in terms of the types of the attributes in Sect. 4.1. Then the general effectiveness of our algorithmic approach is presented by comparative experiments on two UCI datasets [4], remaining superior to other methods in Sect. 4.2.1. Section 4.2.2 shows that the technique also outperforms these aforementioned approaches by applying MIGEC to a real case analysis in two aerospace datasets.

### 4.1 The evaluation criterion

#### 4.1.1 Missing data on numeric attributes

The *Root Mean Square Error (RMSE)* is used to evaluate the predictive ability of the various data imputation algorithms within which the attributes are quantitative:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (e_i - \tilde{e}_i)^2} \tag{9}$$

where  $e_i$  is the original value,  $\tilde{e}_i$  is the predicted plausible value and  $m$  is the total number of estimations. The larger value of *RMSE* suggests the less accuracy that the algorithm holds [58].

#### 4.1.2 Missing data on nominal attributes

The performances of the algorithms for categorical attributes are appraised by the *Classification Accuracy (CA)*:

$$CA = \frac{1}{n} \sum_{i=1}^n l(EC_i, TC_i) \tag{10}$$

where  $EC_i$  and  $TC_i$  are the estimated and true class label for the  $i$ th missing value respectively with  $n$  indicating the total number of the missing values. The function  $l(x, y) = 1$  if  $x = y$ , otherwise  $l(x, y) = 0$ . For this reason, the larger value of function  $l$  indicates the more correct imputed value [56].

### 4.2 Empirical result analysis

#### 4.2.1 UCI datasets

Two UCI datasets, *i.e.* *Wine* and *Thyroid Disease*, are selected to test the validity of the algorithms. *Wine* contains 178 instances and 13 attributes. The variable values are either real or integer. *Thyroid Disease* includes 7200 instances and 21 attributes. The multivariate factual data are either categorical or real.

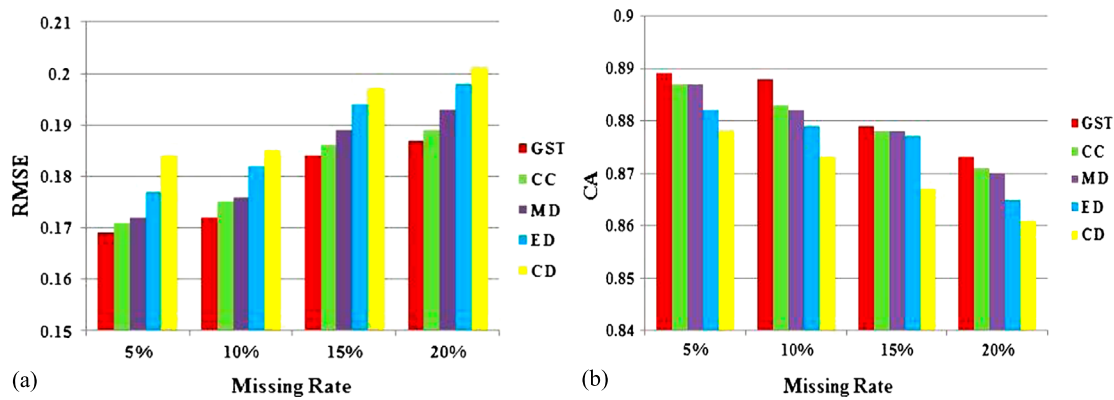
*Missing data generation* To intrinsically examine the effectiveness and validity and ensure the systematic nature of the research, we artificially generated a lack of data at four distinct missing ratios, *i.e.* 5 %, 10 %, 15 % and 20 % under three different modalities, namely *MCAR*, *MAR*, *NMAR* in the complete datasets via the means that *Twala* did [51, 59].

*Parametric values determination* Before the comparative demonstrations, it is requisite to select the optimum values for fuzzier  $s$  and  $\zeta$ . In the section, they are both determined by practical testing in a specific interval ( $s \in (1, 2), \zeta \in (0, 1)$ ). Here, we choose a typical scenario “at *MAR* with 10 % missing rate and three clusters” and the variables processes are displayed in Table 1. These parametric values under the other circumstances could be found in a similar way.

From Table 1, it can be seen that the selection of the values for  $s$  and  $\zeta$  could influence the precision of the imputation in a manner. The worst situation, which yields the maximum *RMSE*, usually occurs when approaching to the upper and (or) lower boundaries of the intervals. And when  $s = 1.3$  and  $\zeta = 0.5$ , the *RMSE* declines to the least. So they are the optimum parametric values respectively.

**Table 1** The impact on RMSE in Wine at MAR with 10 % missing rate by  $s$  and  $\zeta$

$\zeta$	$s$									
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	
0.1	0.190	0.188	0.183	0.186	0.185	0.183	0.191	0.185	0.191	
0.2	0.186	0.185	0.181	0.188	0.192	0.193	0.186	0.189	0.181	
0.3	0.179	0.181	0.178	0.179	0.182	0.186	0.183	0.177	0.178	
0.4	0.182	0.175	0.173	0.175	0.185	0.187	0.192	0.188	0.189	
0.5	0.187	0.174	0.172	0.177	0.184	0.183	0.196	0.201	0.199	
0.6	0.188	0.178	0.174	0.176	0.179	0.181	0.196	0.220	0.213	
0.7	0.195	0.181	0.177	0.179	0.187	0.185	0.199	0.213	0.220	
0.8	0.197	0.192	0.188	0.195	0.198	0.208	0.205	0.202	0.211	
0.9	0.201	0.196	0.198	0.203	0.199	0.207	0.213	0.218	0.222	



**Fig. 2** The performances of distance metrics on (a) Wine and (b) Thyroid Disease

*Distance metric selection* To clarify the different distance metrics that influence the accuracy of the results, the test is supposed to happen under MAR. Then the GST-based metric is practically compared with both the Minkowski Distance family (referring to Manhattan Distance (MD), Euclidean Distance (ED) and Chebyshev Distance (CD)) and CC.

As seen in Fig. 2, we could assume that GST-based distance metric generates the least bias at different missing rates comparing with both the Minkowski’s  $L_p$  and CC. Particularly, the discrimination is even more significant when GST is contrasted with CD according to either RMSE or CA.

*Comparative experiments* In consideration of making comparisons as extensively as possible, we thoughtfully select seven other approaches, which are MMS, HDI, Garcia’s KNN Imputation with Mutual Information [22] (denoted as KNNMI), FCMOCS [15], CRI [53], CMI [58] and NIIA [56]. These methods involve both single methods (i.e. MMS, HDI, KNNMI and FCMOCS) and multiple methods (i.e. CRI, CMI and NIIA).

The experimental data provided in Table 2 and Table 3 illustrate some phenomena that we would like to discuss as follows:

- (a) Increasing proportion of missing instances deteriorates the accuracy of the interpolation in either RMSE or CA. It states that incomplete values negatively impact on the completion, in other words, more available information could promote the precision of the final predictions.
- (b) For each individual method, the best result (namely, the minimal value of RMSE or the maximal value of CA) at the same missing ratio always appears when data are NMAR distributed, whilst MCAR yields the opposite occasions.
- (c) Concerning imputation types, the performance of single imputation techniques (MMS, HDI, KNNMI and FCMOCS) stays inferior to the multiple imputation ones (NIIA, CMI, CRI and MIGEC). There are clear improvements between the two categories of methods. Therefore, imputing the absent value by multiple times can significantly alleviate the biased effect of single imputation.
- (d) It could be obviously observed that MMS does worst while MIGEC does best, which has the average absolute difference beyond 0.020 (measured by RMSE in Table 2) or 0.030 (measured by CA in Table 3). Nonetheless, MIGEC amalgamates MMS into itself as the first step. Hence, it is a feasible and proper option to take



**Table 2** RMSE on Wine under different mechanism with varying missing rates

Missing Mechanism	Missing Rate	Methods							
		MMS	HDI	KNNMI	FCMOCS	NIIA	CMI	CRI	MIGEC
MCAR	5 %	0.201	0.197	0.191	0.188	0.179	0.176	0.172	0.174
	10 %	0.203	0.202	0.195	0.189	0.182	0.180	0.179	0.179
	15 %	0.205	0.203	0.196	0.192	0.186	0.181	0.181	0.182
	20 %	0.213	0.205	0.198	0.195	0.188	0.184	0.188	0.189
MAR	5 %	0.192	0.188	0.186	0.185	0.172	0.171	0.171	0.169
	10 %	0.194	0.196	0.194	0.189	0.176	0.177	0.173	0.172
	15 %	0.204	0.206	0.202	0.192	0.178	0.182	0.185	0.184
	20 %	0.21	0.208	0.204	0.198	0.185	0.188	0.183	0.187
NMAR	5 %	0.171	0.169	0.168	0.165	0.160	0.159	0.158	0.155
	10 %	0.176	0.172	0.172	0.169	0.163	0.166	0.163	0.157
	15 %	0.183	0.178	0.174	0.171	0.164	0.168	0.167	0.164
	20 %	0.192	0.189	0.180	0.175	0.167	0.169	0.168	0.168

**Table 3** CA on Thyroid Disease under different mechanism with varying missing rates

Missing Mechanism	Missing Rate	Methods							
		MMS	HDI	KNNMI	FCMOCS	NIIA	CMI	CRI	MIGEC
MCAR	5 %	0.852	0.856	0.864	0.868	0.875	0.876	0.879	0.881
	10 %	0.849	0.852	0.863	0.866	0.873	0.874	0.873	0.879
	15 %	0.846	0.848	0.862	0.863	0.872	0.871	0.868	0.875
	20 %	0.841	0.842	0.860	0.861	0.870	0.870	0.863	0.865
MAR	5 %	0.877	0.883	0.889	0.892	0.887	0.889	0.895	0.901
	10 %	0.873	0.882	0.887	0.887	0.884	0.886	0.892	0.889
	15 %	0.867	0.880	0.886	0.885	0.881	0.883	0.883	0.885
	20 %	0.861	0.873	0.882	0.883	0.872	0.880	0.872	0.874
NMAR	5 %	0.894	0.891	0.896	0.897	0.897	0.901	0.905	0.908
	10 %	0.89	0.886	0.893	0.895	0.895	0.897	0.894	0.903
	15 %	0.886	0.883	0.887	0.894	0.891	0.894	0.892	0.899
	20 %	0.878	0.872	0.882	0.886	0.888	0.891	0.890	0.886

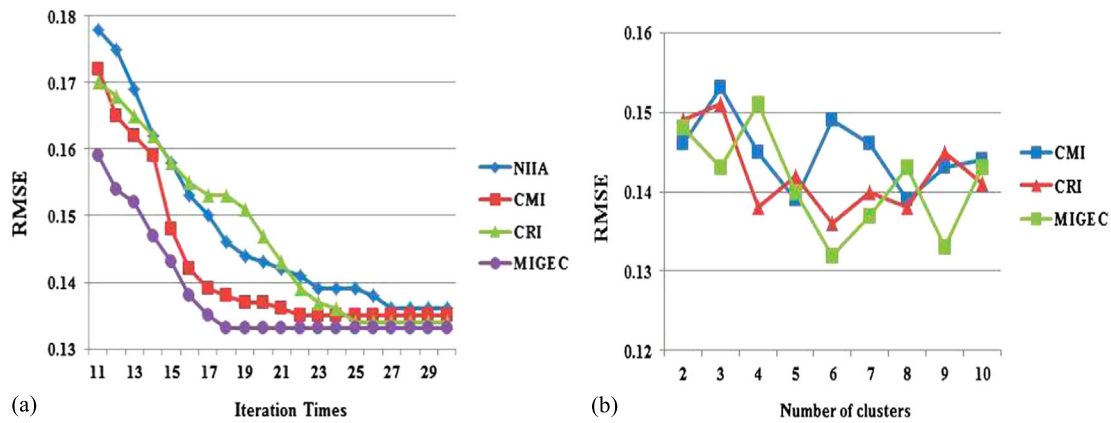
*MMS* to initialize the unknown data before the subsequent multiple completion.

- (e) For the multiple imputation methodologies, though the results generated from *NIIA*, *CMI*, *CRI* and *MIGEC* overlaps slightly at some points, the three clustering-based methods (*CMI*, *CRI* and *MIGEC*) outperform the non-clustered one (*NIIA*). Hence, clustering could actually help ameliorate the accuracy of the prediction through narrowing the potential space of the target value.
- (f) For the clustering-based solutions, *CMI* works slightly worse than those strictly split missing and non-missing values before dividing them into a finite set of groups. So it could be presumed that information loss can disturb the formation of clustering.

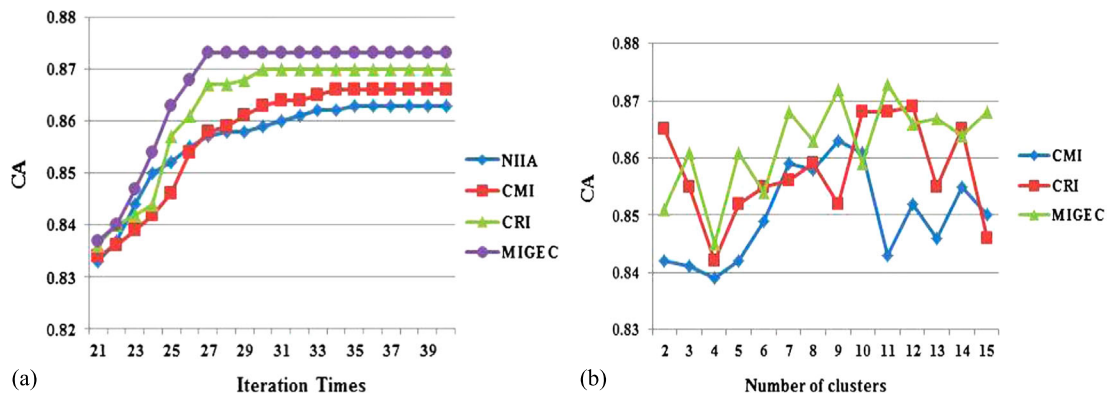
According to factual values and the above analyzing, *MIGEC* performs better than the other seven approaches under any missingness mechanism no matter the data is numeric or categorical, in general.

#### 4.2.2 Aerospace datasets

Aerospace is a special but critical field that associates with both nations' military and residents' daily life. In this paper, it refers to the diverse procedures of spacecrafts. Like a great quantity of other businesses or industries, aerospace analysis confronts the data quality problem as well. The specific issue derives from numerous facets such as difficulties in sensing the target objects restricted by the physical environment or in data acquisition because of autonomous units' regulations. The loss of information negatively blocks further handling or analysis, for instance, curve fitting and rules mining. Nonetheless, open-access researches in aerospace data imputation remain hard to achieve for its confidentiality and particularity. Thereupon, we try our best to explore some of the known parts by using the current accomplishments in other industries *e.g.* medicine or biology for reference together with our aforementioned *MIGEC* [32].



**Fig. 3** The RMSE influenced by (a) imputation times and (b) number of clusters on RCSF dataset



**Fig. 4** The CA influenced by (a) imputation times and (b) number of clusters on SOMD dataset

To capture the result of the data imputation accurately, we chose the complete data entries and artificially simulated the missing situation according with [Missing data generation](#). In this part, the missingness is established under *MAR* at missing rate 15 % on two datasets, as three imputation methods in [Parametric values determination](#) (*NIIA*, *CMI* and *CRI*) are selected as the competitors of *MIGEC*.

The *Remote Controlling for Spacecraft Flying (RCSF)* dataset is comprised of the data produced by one particular unmanned spaceship in real time condition when flying in the outer space with remote controlling by the experts on the ground. Due to the huge amount of the raw data, we just extract the data produced within one minute. Subsequently, the experiment is designed on the 953 records of 20 continuous attributes.

When *MIGEC* is applied to *RCSF* dataset, the maximum times of the iteration in all the clusters is 18 loops, which is faster than *NIIA*'s 27 times, *CMI*'s 22 times and *CRI*'s 25 times of iterations respectively in Fig. 3(a). What is more, the *RMSE* is slightly lower than in the other counterparts.

As versions of clustering principles, interrelationship between *RMSE* and the number of clusters in these techniques should be discussed except the non-clustering *NIIA*.

In Fig. 3(b), it appears that when the whole data is agglomerated into 6 groups, the *RMSE* of *MIGEC* declines to the minimum. Differently, *CMI* performs best with 5 clusters while *CRI* requires 6 partitions.

The *Spacecraft Overall Mechanical Design (SOMD)* dataset comprises the data related to the assembling and fabrication of one specific model of manned spaceship. Both the numeric and categorical values are mixed in the dataset. The total number of instances is beyond 300,000. 1,221 elements with the 30 variables belonging to a certain step of the entire manufacturing process are chosen.

It is easy to perceive that the three algorithms advance *CA* as the number of iteration aggrandizes until the convergence emerges in Fig. 4(a). Concurrently, *MIGEC* attains the best *CA* in the minimum time of the repetitions comparing with the other opponents on *SOMD*.

The *CA* fluctuates irregularly in the interval [0.83, 0.88], when the amount of clusters rises. And *MIGEC* approaches to the optimized *CA* when 11 clusters exist in Fig. 4(b). Generally, *CMI* and *CRI* undulate in an inferior range of *CA* to *MIGEC*, which demands the different optimal number of clusters respectively.

## 5 Conclusion

Revisiting various missing data analysis techniques, this study advocates the clustering-based imputation via partitioning original data into two non-overlapped subsets, *i.e.* the missing-valued subsets and the complete-valued subsets. Then the iterative imputation is combined within the categorized groups after each missing-valued entry has been merged into the most homogeneous cluster through *GST*-based distance metric. The experiments demonstrate that *MIGEC* operates better than the existing methods, like *MMS*, *HDI*, *KNNMI*, *FCMOCS*, *CMI* and *CRI*, in terms of the *RMSE* (for continuous missing attributes), and the *CA* (for discrete missing attributes) at different missing ratios in two canonical *UCI* datasets. In particular, *MIGEC* has been successfully applied into the aerospace datasets. The *RMSE* and *CA* affected by the iteration times indicate that *MIGEC* converges more rapidly than the other iterative imputation techniques with better accuracy in the real application environment. The ongoing research focuses on how to infer and impute missing values more effectively when the dimensionality becomes high.

**Acknowledgements** This work is supported by Project of the State Key Laboratory of Software Development Environment, Beihang University (SKLSDE-2011ZX-09) and National Natural Science Foundation of China (61003016).

## References

- Al-Harbi SH, Rayward-Smith VJ (2006) Adapting *k*-means for supervised clustering. *Appl Intell* 24(3):219–226
- Ahn KW, Chan K-S (2010) Efficient Markov chain Monte Carlo with incomplete multinomial data. *Stat Comput* 20(4):447–456
- Allison PD (2001) Missing data. Sage university papers series on quantitative applications in the social sciences. Sage, Thousand Oaks
- Bache K, Lichman M (2013) UCI machine learning repository. <http://archive.ics.uci.edu/ml>. Irvine, CA: University of California, School of Information and Computer Science
- Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. *J Sch Psychol* 48(1):5–37
- Bezdek JC, Keller J, Krishnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing. In: Dubois D, Prade H (eds) *The handbooks of fuzzy sets series*. Kluwer Academic, Boston/London/Dordrecht
- Biba M, Ferilli S, Esposito F (2011) Boosting learning and inference in Markov logic through metaheuristics. *Appl Intell* 34(2):279–298
- Bose S, Das C, Dutta S, Chattopadhyay S (2012) A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data. In: Proceedings of 2012 international conference on communications, devices and intelligent systems (CODIS), pp 318–321
- Bras LP, Menezes JC (2007) Improving cluster-based missing value estimation of DNA microarray data. *Biomol Eng* 24:273–282
- Calle J, Castaño L, Castro E, Cuadra D (2013) Statistical user model supported by R-tree structure. *Appl Intell*. doi:10.1007/s10489-013-0432-x
- Chen SM, Chen HH (2000) Estimating null values in the distributed relational databases environments. *Cybern Syst* 31(8):851–871
- Chen SM, Huang CM (2003) Generating weighted fuzzy rules from relational database systems for estimating null values using genetic algorithms. *IEEE Trans Fuzzy Syst* 11(4):495–506
- Deng JL (1982) Control problems of grey system. *Syst Control Lett* 1:288–294
- Deng JL (1988) Properties of relational space for grey system. In: Deng JL (ed) *Essential topics on grey system theory and applications*. China Ocean, Beijing, pp 1–13
- Di Nuovo AG (2011) Missing data analysis with fuzzy C-means: a study of its application in a psychological scenario. *Expert Syst Appl* 38(6):6793–6797
- Di Zio M, Guarnera U (2009) Semiparametric predictive mean matching. *ASTA Adv Stat Anal* 93(2):175–186
- Di Zio M, Guarnera U, Luzi O (2007) Imputation through finite Gaussian mixture models. *Comput Stat Data Anal* 51(11):5305–5316
- Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 59(10):1087–1091
- Enders CK (2010) *Applied missing data analysis*. Guilford Press, New York
- Enders C, Dietz S, Montague M, Dixon J (2006) Modern alternatives for dealing with missing data in special education research. *Adv Learn Behav Disabil* 19:101–129
- Farhangfar A, Kurgan L, Pedrycz W (2004) Experimental analysis of methods for imputation of missing values in databases. In: *Intelligent computing: theory and applications II*, Orlando, Florida, 12 April 2004. Proceedings of SPIE, vol 5421. SPIE Press, Bellingham, pp 172–182
- García-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(7–9):1483–1493
- García JCF, Kalenatic D, Bello CAL (2011) Missing data imputation in multivariate data by evolutionary algorithms. *Comput Hum Behav* 27:1468–1474
- González S, Rueda M, Arcos A (2008) An improved estimator to analyse missing data. *Stat Pap* 49(4):791–796
- Hathaway R, Bezdek J (2001) Fuzzy C-means clustering of incomplete data. *IEEE Trans Syst Man Cybern, Part B, Cybern* 31(5):735–744
- Hruschka ER Jr., Hruschka ER, Ebecken NFF (2011) A Bayesian imputation method for a clustering genetic algorithm. *J Comput Methods Sci Eng* 11(4):173–183
- Huang CC, Lee HM (2004) A grey-based nearest neighbor approach for missing attribute value prediction. *Appl Intell* 20(3):239–252
- Huang CC, Lee HM (2006) An instance-based learning approach based on grey relational structure. *Appl Intell* 25(3):243–251
- Jaynes ET (1957) *Information theory and statistical mechanics*. *Phys Rev* 106(4):620–630
- Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. *Atmos Environ* 38(18):2895–2907
- Kim KY, Kim BJ, Yi GS (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinform*. doi:10.1186/1471-2105-5-160
- Lakshminarayanan K, Harp SA, Samad T (1999) Imputation of missing data in industrial databases. *Appl Intell* 11(3):259–275
- 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

34. Li D, Deogun J, Spaulding W, Shuart B (2004) Towards missing data imputation: a study of fuzzy  $k$ -means clustering method. In: Rough sets and current trends in computing. Lecture notes in computer science, vol 3066. Springer, Berlin, pp 573–579
35. Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
36. Liu XH (1999) Progress in intelligent data analysis. Appl Intell 11(3):235–240
37. Lubinsky D (1994) Classification trees with bivariate splits. Appl Intell 4(3):283–296
38. Magnani M (2004) Techniques for dealing with missing data in knowledge discovery tasks. <http://magnanim.web.cs.unibo.it/index.html>
39. McLachlan GJ, Do KA, Ambrose C (2004) Analyzing microarray gene expression data. Wiley, New York
40. Muñoz JF, Rueda M (2009) New imputation methods for missing data using quantiles. J Comput Appl Math 232(2):305–317
41. On BW, Lee I (2011) Meta similarity. Appl Intell 35(3):359–374
42. Pan M (2011) Based on kernel function and non-parametric multiple imputation algorithm to solve the problem of missing data. In: Proceedings of international conference on management science and industrial engineering (MSIE), pp 905–909
43. Parveen S, Green P (2004) Speech enhancement with missing data techniques using recurrent neural networks. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP '04), vol 1, pp 733–738
44. Peng CJ, Zhu J (2008) Comparison of two approaches for handling missing covariates in logistic regression. Educ Psychol Meas 68(1):58–77
45. Posner MA, Ash AS, Freund KM, Moskowitz MA, Shwartz M (2001) Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. Health Serv Outcomes Res Methodol 2(3–4):279–290
46. 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(2):2794–2804
47. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
48. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
49. Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall/CRC Press, London
50. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
51. Twala B (2009) An empirical comparison of techniques for handling incomplete data when using decision trees. Appl Artif Intell 23(5):373–405
52. Yap GE, Tan AH, Pang HH (2008) Explaining inferences in Bayesian networks. Appl Intell 29(3):263–278
53. Zhang C, Qin Y, Zhu X, Zhang J, Zhang S (2006) Clustering-based missing value imputation for data preprocessing. In: Proceedings of IEEE international conference on industrial informatics, Singapore, 16–18 Aug 2006, pp 1081–1086
54. Zhang ML, Zhou ZH (2009) Multi-instance clustering with applications to multi-instance prediction. Appl Intell 31(1):47–68
55. Zhang S (2011) Shell-neighbor method and its application in missing data imputation. Appl Intell 35(1):123–133
56. Zhang S, Jin Z, Zhu X (2011) Missing data imputation by utilizing information within incomplete instances. J Syst Softw 84(3):452–459
57. Zhang S, Jin Z, Zhu X, Zhang J (2009) Missing data analysis: a kernel-based multi-imputation approach. In: Transactions on computational science III. Lecture notes in computer science, vol 5300. Springer, Berlin, pp 122–142
58. Zhang S, Zhang J, Zhu X, Qin Y, Zhang C (2008) Missing value imputation based on data clustering. In: Transactions on computational science I. Lecture notes in computer science, vol 4750, pp 128–138
59. Zhu B, He C, Liatsis P (2012) A robust missing value imputation method for noisy data. Appl Intell 36(1):61–74



**Jing Tian** received his B.S. degree in Software Engineering from University of Electronic Science and Technology of China in July 2008. Since September 2008, he has been a member of State Key Laboratory of Software Development Environment, Beihang University. And he has also been a doctor candidate via a successive postgraduate and doctoral program since 2009. His current research interests include data integration and data quality.



**Bing Yu** received her B.S. degree in Computer Science and Technology from Beihang University in June 2007. She has been a doctor candidate in State Key Laboratory of Software Development Environment, Beihang University since September 2007. Her research interests include wireless sensor networks, data quality and grid computing.



**Dan Yu** received her Ph.D. degree from Wuhan University in 2007 and obtained her Post-Doctor at Beihang University in 2011. Her research interests are spacecraft test language and safety critical system.



**Shilong Ma** is a professor and a doctoral advisor of School of Computer Science and Engineering at Beihang University. His current research interests include software formal theory, software automatic testing, computation model in network and the Internet of things.