Appl Intell (2012) 36:61–74
DOI 10.1007/s10489-010-0244-1

A robust missing value imputation method for noisy data

Bing Zhu · Changzheng He · Panos Liatsis

Published online: 28 July 2010

Abstract Missing data imputation is an important research topic in data mining. The impact of noise is seldom considered in previous works while real-world data often contain much noise. In this paper, we systematically investigate the impact of noise on imputation methods and propose a new imputation approach by introducing the mechanism of Group Method of Data Handling (GMDH) to deal with incomplete data with noise. The performance of four commonly used imputation methods is compared with ours, called RIBG (robust imputation based on GMDH), on nine benchmark datasets. The experimental result demonstrates that noise has a great impact on the effectiveness of imputation techniques and our method RIBG is more robust to noise than the other four imputation methods used as benchmark.

Keywords Missing data imputation · Noise · Group method of data handling (GMDH)

This work is supported by National Natural Science Foundation of China (Grant No. 70771067) and the NSFC/RS (Royal Society of the UK) International Joint Project (Grant No. 70911130133).

B. Zhu · C. He (✉)
Business School, Sichuan University, Chengdu 610064, P.R. China
e-mail: hechangzheng@scu.edu.cn

B. Zhu
e-mail: zhubing1866@hotmail.com

P. Liatsis
School of Engineering and Mathematical Sciences, City University, London EC1V 0HB, UK
e-mail: P.Liatsis@city.ac.uk

1 Introduction

Data in business are often corrupted by missing values, especially the data collected from surveys. For example, consumer data obtained from questionnaires usually contain missing values because the consumers refuse to answer some sensitive questions (e.g., income level, age) or they simply have no opinions about them and so on. Industrial databases are another data source which contains a lot of missing data. The databases maintained by Honeywell company, for instance, have more than 50% of its items (or values) missing, despite great efforts taken in data collection [23]. Such nonresponses complicate the data mining process because most data mining algorithms cannot be immediately and straightforwardly applied to incomplete data. The simplest method to deal with missing data is data reduction which deletes the instances with missing values. But such method will lead to great information loss since in many cases the datasets contain a large amount of missing data [27]. In order to solve this problem, two categories of techniques have been developed. First, there are missing data toleration techniques which integrate the techniques of missing values handling in specific data mining algorithms such as in classification [39, 48], clustering [16] and feature selection [4]. Second, there are missing data imputation techniques which fill in missing values before using complete-data methods. One advantage of imputation is that the treatment of missing data is independent of the succeeding learning algorithm, and people can select a suitable learning algorithm after imputation [36]. Therefore, imputation has received considerable attention and many methods have been proposed in recent years [25, 43].

However, an important issue has been neglected by previous research: real-world data often contain much noise in addition to the missing values while the impact of noise is

seldom considered in current literature. The noise comes from the process of data collection, data entry, data transformation, etc. The presence of noise may introduce some negative effects. For example, most classification algorithms are sensitive to noise [28, 50]. Consequently, noise will affect the performance of imputation methods that are based on such classification algorithms. Therefore, it is necessary to investigate the performance of current imputation methods in the presence of noise.

Group Method of Data Handling (GMDH) proposed by Ivakhnenko [20] is a heuristic self-organizing data mining technique for complex system modeling and identification. One of the main advantages of the GMDH method is its noise immunity. As Aksenova and Yurachkovsky have stated in [2], GMDH will produce a so-called non-physical model when the dataset includes noise. The non-physical model has simpler structure and better generalization than a complete physical model. Therefore, GMDH has been applied to many real-world data mining applications in recent years [31]. In this paper, we try to design a new imputation method using the mechanism of GMDH and expect that the missing values can be predicted accurately in a noisy environment. Extensive experiments and comparisons are done on 9 benchmark datasets from different domains. Experimental results show that noise has a remarkable impact on the effectiveness of imputation techniques and our new method RIBG is more robust to noise than the other imputation methods used as benchmark.

The rest of the paper is organized as follows. In Sect. 2, we discuss related works on imputation method. In Sect. 3, we introduce some definitions to be used in this paper and give a brief description of GMDH. In Sect. 4, we describe in detail the proposed new approach RIBG. Section 5 explores the impact of noise on imputation methods and evaluate the effectiveness of RIBG, particularly its robustness in the presence of noise, through experiments on 9 bechmark datasets. Finally in Sect. 6, we conclude the whole paper.

2 Related work

Methods to deal with missing values are not something new. In 1976, Rubin developed a framework of inference from incomplete data that is still in use today [38]. After that many researchers have run into this area and proposed a great number of methods. All the imputation methods can be roughly classified into the following six categories:

- *Mean substitution*: It is the simplest imputation method. It replaces the missing values by the mean of all the observed values or a subgroup at the same variable. It is fast, simple and easily implemented.
- *Hot-deck imputation* [14]: For Hot-deck imputation, missing values are recovered from similar cases drawn from the same dataset. It is often used to handle missing data of survey.
- *Regression imputation* [7]: Regression imputation uses regression models to predict missing values. Many forms of regression models can be used for regression imputation such as linear regression, logistic regression and semi-parametric regression [36].
- *EM imputation*: The EM imputation is based on the Expectation-Maximization (EM) algorithm proposed by Dempster, Laird and Rubin [10]. It uses the iterative procedure of the EM algorithm to calculate the sufficient statistics and estimate the parameters. The missing values will be produced in the process.
- *Multiple imputation* [15]: Multiple imputation was first proposed by Rubin [38] and now is an increasingly popular way to handle missing data. It produces m complete datasets, and then each of the datasets is analyzed by complete-data method. At last, the results derived from these m datasets are combined. Multiple imputation reflects the uncertainty of the missing values.
- *Machine-learning-based approach*: Most methods described above come from statistics. Recently, some machine learning techniques have been introduced to estimate missing values. For example, a decision tree approach was suggested by Quinlan [37]. Lakshminarayan, Harp and Samad [23] used Autoclass clustering to handle missing data. Batista and Monard [6] suggested a k-nearest-neighbor approach to fill in the missing data. Huang and Lee [18] employed a grey-based nearest neighbor method to impute the missing data. Hruschka Jr., Hruschka and Ebecken [17] used Bayesian network to substitute missing values. Also there is another work by Chen and Huang [8, 9], which used the weighted fuzzy rules to estimate null values in relational database. These methods mainly use the predictive power of machine learning algorithms to estimate missing values.

To examined the performance of aforementioned imputation methods, several works have been done. Myrtveit, Stensrud and Olsson [32] presented an empirical evaluation of imputation methods in the context of software cost modeling. Olinsky, Chen and Harlow [34] compared the efficacy five imputation methods in structural equation modeling. Farhangfar, Kurgan and Dy [13] studied how the choice of different imputation methods affects the performance of various classifiers. Twala [44] investigated the impact of popular imputation approaches on tree-based model.

Although there have been a large amount of work on imputation, relatively little attention has been given to the impact of noise. As far as we know, the only work that considered the influence of noise on imputation methods was [47]. In [47], Van Hulse and Khoshgoftaar analyzed the performance of five popular imputation techniques on noisy software measurement data, namely mean imputation, regression imputation, REPTree imputation, Bayesian multiple

imputation, and k-nearest-neighbor imputation. It concluded that noise has a dramatic negative impact on the effectiveness of these imputation techniques. But there are some limitations in that paper just as the authors have pointed out. First, the conclusion was drawn from experiments on single software measurement dataset. Second, it assumed the missing values and noise appear only on one variable (dependent variable).

In this paper, we examine the impact of noise on imputation methods in a more general situation. Specifically, experiments are done on 9 benchmark datasets from several different domains. Meanwhile, we assume noise and missingness distribute throughout the dataset which means both dependent and independent variables in a dataset contain noise and missing values. This is a more realistic situation, because for real-world data, the number of independent variables is large and they are much dirtier than dependent variables [50]. Experimental results demonstrate that noise has considerable negative effects on imputation methods and our proposed method RIBG outperforms other competing methods in the presence of noise.

3 Preliminaries

3.1 Basic notions

In this section, we define some notions about missing values and then introduce missing data mechanism which is considered in our experiments. Finally, a brief description of the GMDH algorithm is given.

Let D denote an incomplete dataset with r variables $D = \{A_1, A_2, \ldots, A_r\}$ and n instances. For each variable A_j, $j = 1, 2, \ldots, r$, it contains two parts: $A_j = \{A_j^{obs}, A_j^{mis}\}$, where A_j^{obs} is the set of observed elements and A_j^{mis} is the set of missing elements. Similarly, the entire dataset D also consists of two components, $D = \{D^{obs}, D^{mis}\}$, where D^{obs} is the set of observed values and D^{mis} is the set of missing values.

We can also introduce a response indicator matrix R which is the same size as D to describe the missingness. Each element of R is defined as follows:

$$R_{ij} = \begin{cases} 0 & \text{if } v_{ij} \text{ is missing} \\ 1 & \text{if } v_{ij} \text{ is observed} \end{cases} \quad (1)$$

where v_{ij} is the value of the i-th instance at variable A_j, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, r$.

The aim of imputation is to fill in all the blanks of incomplete dataset D, so that the estimated complete dataset $\widehat{D}$ can be used for succeeding data mining algorithm.

3.2 Missingness mechanism

The missingness mechanism determines how the missing data are generated and it is a potential factor that will affect the imputation results. Thus, a comprehensive study of the noise impact on imputation methods must take different missingness mechanisms into account. There are three types of missing data mechanisms according to Little and Rubin [25]:

- *Missing Completely At Random (MCAR)*
  If $\Pr(R|D^{mis}, D^{obs}) = \Pr(R)$, then the missing mechanism is defined as MCAR, where Pr presents the probability. MCAR implies that the missingness is unrelated to both the missing and observed values in the dataset.
- *Missing at Random (MAR)*
  If $\Pr(R|D^{mis}, D^{obs}) = \Pr(R|D^{obs})$, then the missingness mechanism is called MAR. MAR means the missingness depends on observed values but not on missing values.
- *Not Missing At Random (NMAR)*
  If $\Pr(R|D^{mis}, D^{obs})$ is not equal to $\Pr(R|D^{obs})$ and it depends on D^{mis}, then the missing data is NMAR.

3.3 The GMDH algorithm

GMDH is an inductive modeling method that constructs a hierarchical (multi-layered) network structure to identify complex input-output functional relationships from data. It was first developed by Ivakhnenko as a multivariate analysis method for complex systems modeling and identification in the 1960s [19]. During the 1980s its theoretical background was formulated [21, 41] and later considerable improvements were introduced by versions of the polynomial network training algorithms (PNETTR) by Barron [5] and the algorithm for synthesis of polynomial networks (ASPN) by Elder and Brown [11]. In the 1990s, Mueller and Lemke further developed the GMDH algorithm into the self-organizing data mining algorithm [24]. After entering the 2000s, many researchers have attempt to use computational intelligence technology such as genetic algorithm [33] to optimize the network structure of the GMDH model. Now GMDH has become a set of several algorithms for different problem solutions. It consists of parametric, clusterization, analogs complexing, and probability algorithms. It has now been successfully applied to many domains such as economics [29], ecology [45], engineering [35], medical science [1], etc.

The process of GMDH is analogous to the natural evolution of wheat. To obtain wheat with a certain property, a large number of wheat are sown which may have this property. From the harvest of the first generation, wheat which better satisfy the requirements as compared to others are chosen. The seeds of these wheat are sown again. From the second harvest certain seeds are once again selected and

sown. After several generations, some wheat will be obtained in which the desired property are more predominant than in others. Similarly, the process of GMDH is a self-organizing process based on sorting-out of gradually complicated models and selection of the best solution by external criterion. GMDH first produces some simple elementary models by reference functions and uses them as initial input models at the start of modeling process. After generating a large number of competition models by these initial input models (inheritance), the algorithm selects certain more optimal intermediate models (selection), so that it regenerates a large number of new competition models by these intermediate models. Such procedure of inheritance and selection is repeated until an optimal complex model has been created. According to the theory of optimal complexity, as the complexity of the model increases, the value of external criterion usually decreases first, then reaches a minimum and later starts to increase again. The GMDH algorithm will stop when an eternal criterion reaches its minimum which corresponds to the optimal complex model [26]. In this way, the algorithm can determine the input variables and structure of the final model automatically, and has accomplished it by the process of self-organizing modeling [31].

One desirable characteristic of GMDH is its noise immunity. As we all know, when data contain noise, the most dangerous thing is overfitting [42], which implies that models tend to be excessively complex, and have poor generalization. But for GMDH, this problem can be avoided. As Ivakhnenko and Stepashko have stated in [22] that when the dataset contains noise the minimal value of the external criterion of GMDH usually indicates a non-physical model. Aksenova and Yurachkovsky [2] have proved the non-physical model has simpler structure and better generalization than a complete physical model.

4 Algorithm of RIBG

The main idea of RIBG is using the mechanism GMDH to impute missing data in the hope it will give more accurate imputation results than traditional imputation approaches even when data contain noise. Let us consider an incomplete dataset D with r variables $D = \{A_1, A_2, \ldots, A_r\}$. RIBG will fill in the original incomplete dataset D by simple mean imputation to get an initial complete dataset. We use mean imputation to initially impute the missing values because it has been proved to be an efficient pre-imputation method [12]. Then we use the mechanism of GMDH to predict and update these initial estimated missing values with an iterative process. When using the mechanism of GMDH, we present a new combined criterion RM which integrates the systematic regularity criterion (SR) and minimum bias criterion (MB) criterion together:

$$\begin{aligned} RM &= SR + MB \\ &= \left\{ \left(\sum_{i \in B} (y_i - \hat{y}_i^C)^2 + \sum_{i \in C} (y_i - \hat{y}_i^B)^2 \right) \right\} \\ &\quad + \sum_{i \in B \cup C} (\hat{y}_i^B - \hat{y}_i^C)^2 \end{aligned} \tag{2}$$

where B and C are two disjoint subset of the entire datasets D ($B \cup C = D$), y_i is the actual output, $\hat{y}_i^B$ and $\hat{y}_i^C$ are the estimated outputs of the model constructed on dataset B and dataset C, respectively. Algorithm RIBG shows the whole learning process step by step.

Algorithm RIBG

Input:

D—$n \times r$ incomplete dataset

Output:

$\widehat{D}$—$n \times r$ complete dataset

Step 1: Divide dataset D into two disjoint subsets: $D = B \cup C$, where B is the training set and C is the validation set;

Step 2: For each variable in D, replace missing elements by mean (if the variable is a numeric variable) or mode (if the variable is nominal) of the observed elements to get an initial complete dataset;

Step 3: Select variable A_j that needs to be imputed as output variable ($y = A_j$) and all the remaining variables as input variables ($x = \{A_s | s = 1, \ldots, r, s \neq j\}$) to enter the first layer of the GMDH network;

Step 4: Combine input variables in pairs (x_i, x_j), $1 \leq i, j \leq m$ and generate model candidates from each combination using the following quadratic polynomial:

$$y = c_0 + c_1 x_i + c_2 x_j + c_3 x_i \cdot x_j + c_4 x_i^2 + c_5 x_j^2 \tag{3}$$

where $c_0, c_1, \ldots, c_5$ are parameters to be estimated by the ordinary least square (OLS) method. For example, if we use $x_1, x_2, \ldots, x_5$ as the input variables to estimate the output variable y, then 10 model candidates are produced in Fig. 1 and input variables x_1 and x_2 will be combined to produce the model candidate Z_{11} as follows:

$$Z_{11} = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_1 \cdot x_2 + c_4 x_1^2 + c_5 x_2^2 \tag{4}$$

Step 5: Evaluate the external criterion of each model using the combined criterion RM. Record the minimum of the external criterion R_i from the current layer. Select F_i best models with lower criterion values and

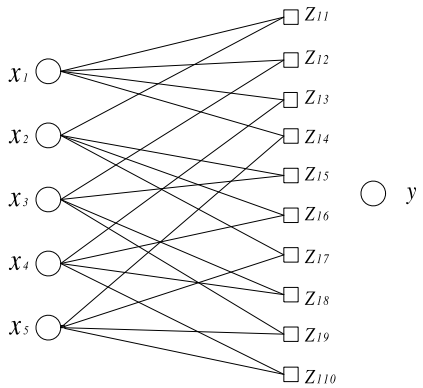


Fig. 1 Generation of candidate models in the first layer

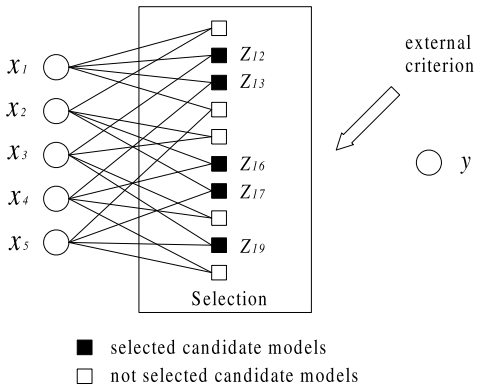


Fig. 2 Selection of candidate models

their outputs z_{ti} are employed as new input variables for the second layer of the GMDH network ($x_{it} = z_{it}$, $t = 1, \ldots, F_i$). For instance, model candidates which have lower external criterion values z_{1t} ($t = 2, 3, 6, 7, 9$) are selected and used as input variables for the second layer in Fig. 2;

Step 6: Repeat Steps 4–5 to produce model candidates of the second layer, the third layer, ... until the lowest value of external criterion at the current layer R_i is greater than that in previous layer. The model with the minimum external criterion at the i-1 layer is then selected as the final optimal complex model. Figure 3 gives an example that the optimal complex model z_{32} is obtained at the third layer;

Step 7: Use the corresponding estimates of the optimal complex model to update the missing elements A_j^{mis} of variable A_j;

Step 8: Repeat Steps 3–7 until the change of missing element estimates $\widehat{A}_j^{mis}$ becomes smaller than a threshold or maximum number of iterations is reached. Then assign current values of A_j to the corresponding elements in $\widehat{D}$;

Step 9: Follow Steps 3–8 to update the missing values of remaining variables.

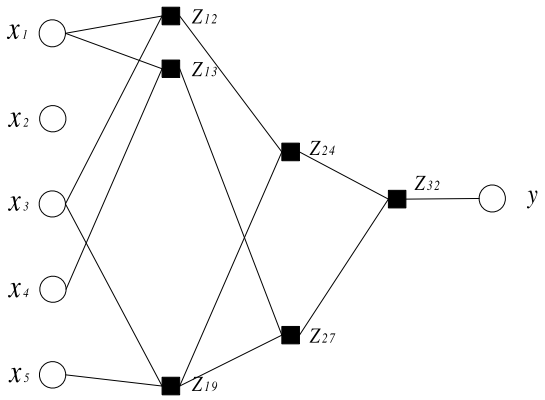


Fig. 3 Generation of optimal complex model

5 Experiments

In this section, we evaluate the impact of noise on imputation methods and verify the robustness of RIBG in a noisy environment through experiments.

5.1 Experimental design

5.1.1 Datasets

Nine datasets from the UCI (University of California at Irvine) machine learning repository [3] were used in the experiments. The basic information of these datasets is listed in Table 1. The 9 datasets come from several different domains such as economics (Housing), medical science (Breast), life science (Bupa, Cmc, Iris), social science (Balance) and physics (Glass2, Ionosphere and Wine). They were chosen because they have no missing data (for the Breast dataset 16 instances with missing values were removed). Consequently, we can have total control over the generation of missing data in the datasets to produce missing data with specified patterns and evaluate the performance of imputation methods by comparing the imputed values with original ones.

5.1.2 Simulation of missingness and noise

To introduce artificial missingness, we considered two important factors which may affect the imputation results: missing rate and missing data mechanism. Three different levels of missing rate were considered, i.e., 5%, 10% and 20%. Meanwhile, three missing mechanisms were taken into consideration, namely MCAR, MAR and NMAR. We used the similar approach to produce artificial missingness as Twala has done in [44]:

Table 1 Datasets used in the experiments

Dataset	Abbr.	Size	#Variable
Balance Scale	Balance	625	5
Breast Cancer Wisconsin	Breast	683	10
Liver Disorders	Bupa	345	7
Contraceptive Method Choice	Cmc	1473	10
Glass Identification (2 class)	Glass2	163	10
Housing	Housing	506	14
Ionosphere	Ionosphere	351	35
Iris Plant	Iris	150	5
Wine	Wine	178	14

Table 2 An example dataset for the simulation of missingness

Instance number	A_s		A_j	
	Numerical	Nominal	Numerical	Nominal
1	24	S1	48	J4
2	30	S2	75	J4
3	31	S2	83	J3
4	35	S3	58	J3
5	60	S1	83	J2
6	76	S2	32	J2
7	81	S3	45	J1
8	82	S2	50	J1
9	88	S1	86	J1

For MCAR, we let every data in the dataset have the same probability α to be missing, where α was the specified missing rate.

Simulating MAR was more complex and it worked as follow: we first randomly separated the variables into pairs (A_j, A_s), $1 \leq j, s \leq r$, where A_j was the variable into which missing values were introduced, and A_s was the variable that affected the missingness of A_j. Given a pair of variables (A_j, A_s) and missing rate α, we first split the instances into two equal-sized subsets according to their values at A_s. If the variable A_s is numerical, we would find the median a_s^{med} of A_s and then assigned all the instances into two subsets according to weather the instances have bigger values than a_s^{med} at A_s. If A_s is nominal, we randomly divided the categorical values of A_s into two parts, and then split instances according to which part their corresponding categorical values at A_s belong to. After the splitting of instances, we randomly selected one subset of instances and let their values at A_j to be missing with the probability of 4α. The probability of 4α will result in a missing rate of 2α on the whole variable A_j which is equivalent to having a missing rate of α on the two variables (A_j, A_s), since we did not introduce any missing values into A_s. We can describe the simulation process by a concrete example. Suppose there is a complete dataset with two variables and nine instances as in Table 2. The two variables can be numerical or nominal. If A_s is numerical, we may let the instances whose values at A_s are lower than the median 60 (instance number 1–5) to be missing with the probability of 4α, that is to say, $\Pr(A_j = missing | A_s \leq 60) = 4\alpha$. If A_s is nominal, we can divide the instances into two subsets. On one subset the instances have categorical value *S2* at A_s and on the other subset the instances have values *S1* or *S3*. We may generate missing values at variable A_j on those instances coming from the first subset (instance number 2, 3, 6 and 8) with a probability of 4α, i.e., $\Pr(A_j = missing | A_s = S2) = 4\alpha$. In addition, there is one thing that needs to be mentioned. If the number of variables in the dataset is odd, there would be a variable A_k $(1 \leq k \leq r)$ that does not fall into any pair of variables. For this variable we randomly selected a pair of variables for it to form a variable triple (A_j, A_k, A_s). In this triple, the missingness of A_j and A_k all depends on A_s and we used the same way as described above to produce missingness on A_j and A_k, but we made one subset of instances to be missing at A_j and A_k only with the probability of 3α, as missingness was only introduced into two of the three variables in the triple.

The process of generating missing values by NMAR was similar to MAR. The only difference was that there was no need to split variables into pairs, NMAR produced missingness on every variable directly. For a given variable A_j and specified missing rate α, if A_j is numerical, we first calculated the median a_j^{med} of A_j and then randomly let the values that are lower (or higher) than a_j^{med} to be missing with probability of 2α. Take the dataset in Table 2 for instance once more, we may let the values at A_j that are smaller than the median 58 to be missing with probability of 2α. If A_j is nominal, categorical values of A_j were randomly divided into two parts. We then randomly selected one part of values and made them to be missing with the probability of 2α, For example, we may split the categorical values of A_j in Table 2 into two parts: *J1* and *J2* on one part, while *J3* and *J4* on the other part. Then categorical values on the first part (*J3* and *J4*) are selected to be missing with the probability of 2α.

The UCI datasets have been carefully examined by the domain experts and they do not contain much noise [49]. So we exploited a manual mechanism to add injected noise to each variable in the datasets. Three noise levels were considered, i.e., 0%, 10% and 20%. We used the mechanism adopted by Zhu and Wu [50] to corrupt the data. To corrupt variable A_j with a given noise level δ, we let every value at the variable A_j have a δ chance to be changed to any other random value. For discrete variable, the random value is another possible value at this variable. For continuous variable, the random value is between the maximal and minimal value.

5.1.3 Imputation methods

To verify the effectiveness of RIBG, four popular imputation methods were used in our experiments as base line. They are regression imputation (RI), EM imputation (EI), grey-based nearest neighbor imputation (GBNN), multiple imputation based on fully conditional specification (MI).

RI uses a multivariate linear regression model to impute the missing values. In order to reduce the number of regressors, forward stepwise selection was used in building the multivariate linear regression model.

EM imputation in our experiments assumes a distribution for the data, and uses the iterative procedure of EM algorithm to impute the missing values.

GBNN is a machine-learning-based approach with high accuracy proposed by Huang and Lee [18]. It first determines the nearest neighbors of an instance with missing values using grey-relation analysis and then uses the known values of nearest neighbors to impute the missing values. To determine the best choice of k (number of nearest neighbors), we let k vary from 1 to 50 and chose the best one according to imputation accuracy as Huang and Lee has done in [18].

There are generally two strategies for multiple imputation: joint modeling and fully conditional specification [46]. We selected fully conditional specification in our experiments because it is more flexible than joint modeling. It allows user to specify imputation model for each variable separately. Typically, 5–10 imputations are enough for most problems according to Schafer [40]. To get better results, we utilized MI to obtain 15 complete datasets in our experiments, and then integrated the 15 complete datasets into one dataset by taking the average as Van Hulse and Khoshgoftaar has done in [47].

EM imputation, regression imputation, multiple imputation were implemented in SPSS with default settings. Matlab codes were developed for GBNN and RIBG. In Algorithm RIBG, we assigned a fixed value to F_i (number of candidate models selected in each layer) in each layer and tried three different numbers (10, 20 and 30) and observed the results was not very sensitive to F_i. As a consequence, we set F_i to be 10 in every layer.

5.1.4 Experimental design summary

In summary, our experiments considered the following three aspects:

- *Missing rate*: 5%, 10%, and 20%
- *Noise level*: 0%, 10%, and 20%
- *Missingness mechanisms*: MCAR, MAR, and NMAR

In total, $3 \times 3 \times 3 = 27$ scenarios (combinations of different missing rates, noise levels, and missingness mechanisms) were considered for every dataset, we assumed that all the variables had the same level of noise and all the variables with missing values had the same missing rate and missing mechanism. To avoid bias, five independent experiments (runs) were implemented for each scenario of every dataset.

In general, the experiments were done as follows. Original complete datasets were first corrupted by artificial missingness and noise. Next, RIBG and other benchmark methods were used to fill in the missing values. Finally, the performance of each method was evaluated by comparing the imputed values with original ones as Farhangfar, Kurgan and Pedrycz have done in [12].

5.1.5 Performance measure

To evaluate the precision of imputation, the normalized mean absolute error (NMAE) is used and its value at variable A_j is calculated as follows:

$$NMAE_j = \begin{cases} \frac{1}{n_j^{mis}} \sum_{i=1}^{n_j^{mis}} \left(\frac{\hat{v}_{ij} - v_{ij}}{v_j^{\max} - v_j^{\min}} \right) & \text{if } A_j \text{ is numerical} \\ 1 - \frac{n_j^{cor}}{n_j^{mis}} & \text{if } A_j \text{ is nominal} \end{cases} \quad (5)$$

where n_j^{mis} is the number of missing values at A_j, v_{ij} and $\hat{v}_{ij}$ denote the true value and imputed value of the missing data respectively, $v_j^{\max}$ and $v_j^{\min}$ are the maximum and minimum value at A_j, n_j^{cor} is the number of missing nominal values that are correctly predicted. The value of NMAE on the whole dataset takes the average over all the variables. The NMAE of one imputation method at a scenario is calculated as the average over all the five runs of that scenario.

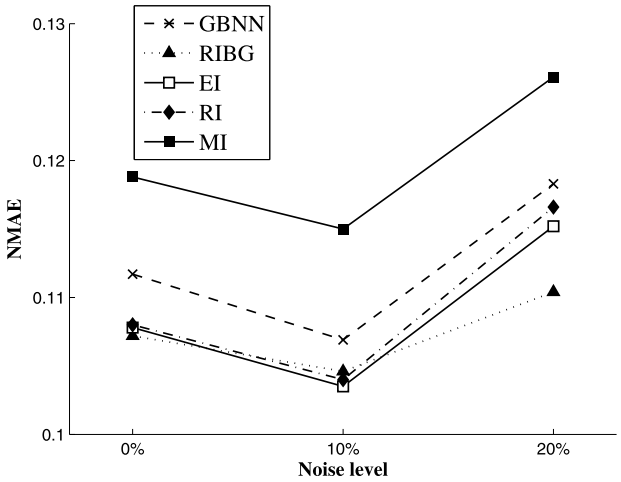
5.2 Experimental results and analysis

5.2.1 Illustration of results

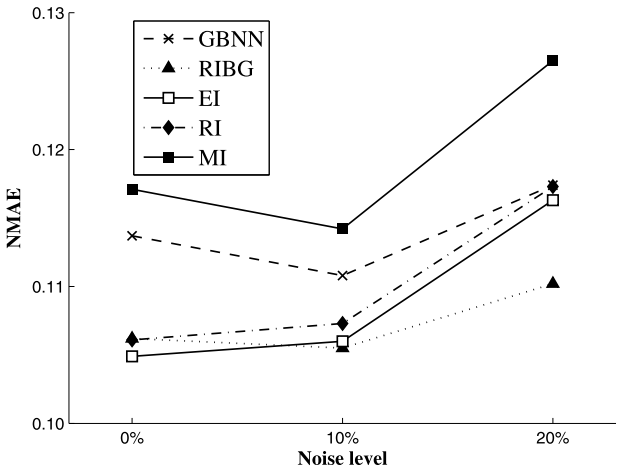
We have carried out the experiments with all the 9 datasets. Figure 4 only demonstrates one set of representative results (different scenarios of the Breast dataset), where the x-axis indicates the noise level and the y-axis represents the imputation error in terms of NMAE. Each curve in the figure represents the results from one imputation method. Take Fig. 4(a) for example, it presents the results of the five imputation methods at three different noise levels (0%, 10% and 20%) in Breast dataset when the missing rate is 5% and the missing data mechanism is MCAR. Figures 4(b)–4(i) illustrate the results of other scenarios.

There are two interesting observations from these figures, which also apply to the other 8 datasets:

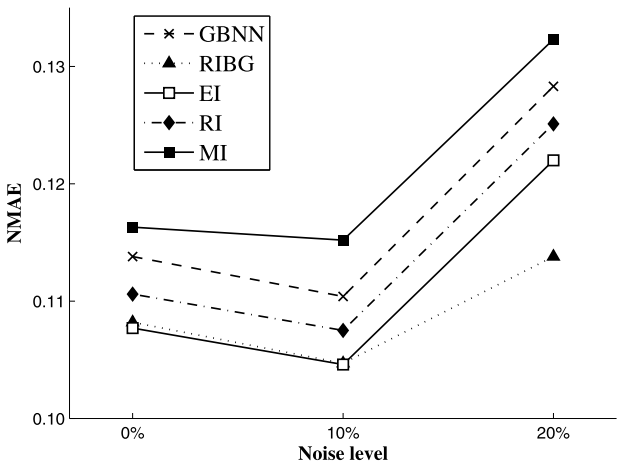
- First, different noise levels have different impacts on imputation accuracy. Generally speaking, the imputation error increases with the level of noise for all the methods. This is understandable because with more noise introduced into the datasets, more negative effects will be



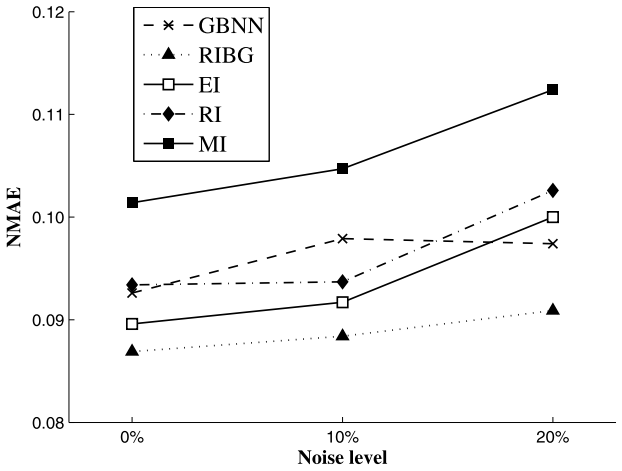
(a) Mechanism=MCAR, Missing rate=5%



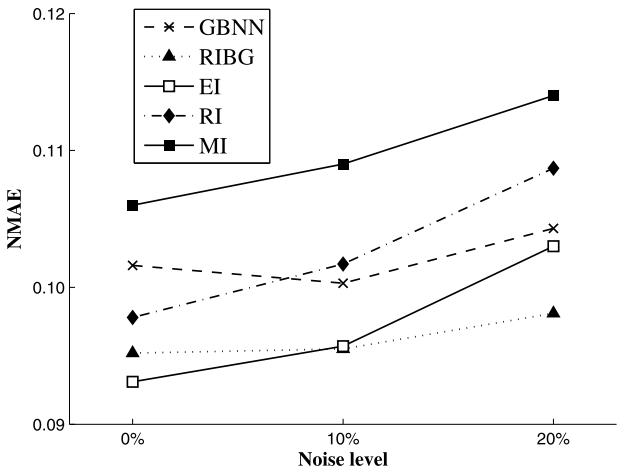
(b) Mechanism=MCAR, Missing rate=10%



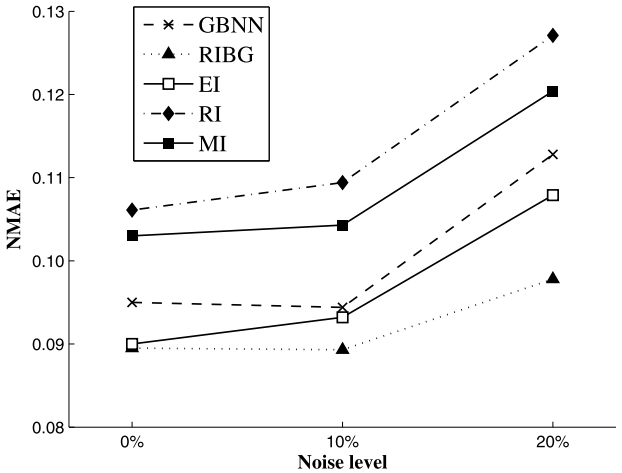
(c) Mechanism=MCAR, Missing rate=20%



(d) Mechanism=MAR, Missing rate=5%

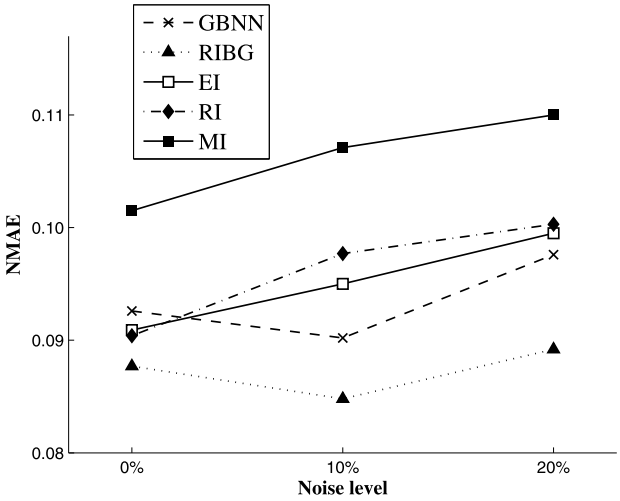


(e) Mechanism=MAR, Missing rate=10%

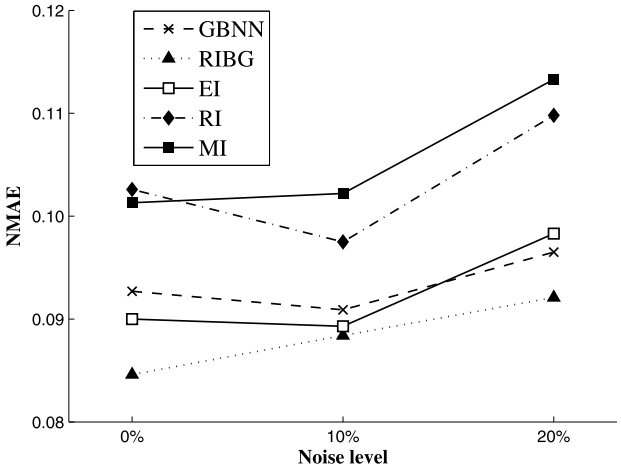


(f) Mechanism=MAR, Missing rate=20%

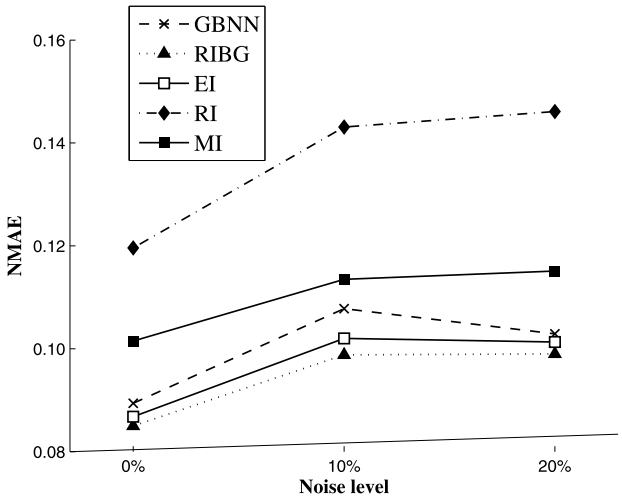
Fig. 4 Experimental results on Breast dataset



(g) Mechanism=NMAR, Missing rate=5%



(h) Mechanism=NMAR, Missing rate=10%



(i) Mechanism=NMAR, Missing rate=20%

Fig. 4 (*Continued*)

brought to the imputation results. Nevertheless, although the noise will deteriorate the imputation accuracy, when comparing the results from three noise levels, the patterns of deterioration are different. When the noise level is low, the impact of noise is limited, and the increasing of imputation errors is not significant. Sometimes a small amount of noise even seems to improve the results. However, when level noise is relatively high, the introducing of more noise will deteriorate the imputation results dramatically. Take Fig. 4(f) in which the mechanism is MAR and the missing rate equals to 20% as an example, when noise level increases from 0% to 10%, the error of RI increases slightly from 0.1061 to 0.1094. But when noise level reaches to 20%, the error degrades seriously to 0.1271. All the other curves in the same figure demonstrate the same trend. This contrast indicates that there may be a threshold in each scenario. When the noise level is below this threshold, the imputation methods are insensitive to noise. The rising noise level has limited effects on imputation results. But once the noise level is beyond this point, the errors usually begin to increase dramatically.

- Another observation is that different methods have different reactions to noise. As we can see from these figures, when there is no noise or the noise level is low, i.e.,10%, the imputation accuracy is relatively good. When comparing different methods, EI and RIBG achieve better accuracy than the other three methods and their accuracy difference is indiscernible. GBNN and RI appear to be the second best methods and MI is worst of all the five methods. When the noise level goes higher, i.e., 20%, the accuracy difference between RIBG and EI turns to be significant and RIBG becomes the best method in all the

nine scenarios. EI sets the second best. It takes the second place six times and the third place three times in term of imputation accuracy in the nine scenarios. GBNN and RI set the third best and MI is still worst of all. All these reveal although no methods can be absolutely robust to noise, but RIBG is really more robust than others at high noise level.

5.2.2 *The impact of noise on imputation methods*

To justify the first observation that noise has a great impact on imputation methods is statistically significant, analysis of variance (ANOVA) is used to analyze the experimental results on each dataset as Van Hulse and Khoshgoftaar have done in [47]. ANOVA is a powerful statistical model that can be used to test the hypothesis that different levels of factor have equal means when there are many factors influencing the experimental results simultaneously [30]. We use the ANOVA model to test the null hypothesis that imputation results are similar at different noise levels. If the null hypothesis is rejected, we then use a post-hoc test to find which noise levels are significantly different from the others. Four factors are considered in our ANOVA model: noise level, missing rate, missingness mechanism and imputation method.

The analysis of the main effect *noise level* on the 9 datasets is tabulated in Table 3, where the second column gives the degrees of freedom (DF) (since there are three different noise levels, the DF for that factor is 2), the third column lists the Type III Sum of Squares (SS), the fourth and fifth column report the F-values and p-values, and the sixth column tells will the null hypothesis be rejected at 5% significance level. Since our main purpose is to explore the influence of noise, we do not present the analysis of other main effects due to lack of space.

We can see from this table that 7 datasets have p-values much less than 0.001, one dataset (Wine) has the p-value 0.01 and one dataset (Ionosphere) has a p-value 0.036. The results of Table 3 indicate the main effect *noise level* is significant at the 5% level on all the 9 datasets. So we reject the null hypothesis and conclude that of all the three noise levels there is at least one that is different from others on the 9 datasets.

To determine which noise levels are significantly different from others, two popular post hoc tests are used for multiple comparison: Fisher's Least Significant Difference test (LSD) and Tukey's Honestly Significant Difference test (HSD) [30]. Table 4 and Table 5 give marginal means and the grouping of the three noise levels on each dataset based on these two post hoc tests at 5% significance level. Given a noise level, the marginal mean is calculated as the average NMAE across all the scenarios (combinations of different missingness mechanisms, missing rates and imputation methods) at that noise level. A, B and C in the Table 5 denote which group the noise level belongs to. Group A is the one that has the minimum NMAE and group C has the maximum NMAE. If two noise levels have significantly different means, then they fall into different groups; otherwise, they stay in the same group.

Table 3 Main effect *noise level*

Dataset	DF	SS	F-value	p-value	Hypothesis
Balance	2	0.0083	45.59	<0.001	Reject
Breast	2	0.0025	30.53	<0.001	Reject
Bupa	2	0.0056	18.90	<0.001	Reject
Cmc	2	0.0069	30.53	<0.001	Reject
Glass2	2	0.0107	49.66	<0.001	Reject
Housing	2	0.0023	31.90	<0.001	Reject
Ionosphere	2	0.0125	3.398	0.036	Reject
Iris	2	0.0207	37.38	<0.001	Reject
Wine	2	0.0011	8.36	0.010	Reject

Table 4 Mean of different noise levels

Dataset	Mean		
	0%	10%	20%
Balance	0.2357	0.2458	0.2548
Breast	0.1003	0.1022	0.1103
Bupa	0.1299	0.1381	0.1457
Cmc	0.2213	0.2293	0.2387
Glass2	0.1042	0.1145	0.1260
Housing	0.0870	0.0938	0.0972
Ionosphere	0.1382	0.1496	0.1618
Iris	0.0905	0.0952	0.1188
Wine	0.1100	0.1115	0.1167

Table 5 Grouping of different noise levels

Dataset	LSD			HSD		
	0%	10%	20%	0%	10%	20%
Balance	A	B	C	A	B	C
Breast	A	A	B	A	A	B
Bupa	A	B	C	A	B	C
Cmc	A	B	C	A	B	C
Glass2	A	B	C	A	B	C
Housing	A	B	C	A	B	C
Ionosphere	A	AB	B	A	AB	B
Iris	A	A	B	A	A	B
Wine	A	A	B	A	A	B

As Table 4 has shown, the imputation errors on all the 9 datasets always increase with the noise level. But the increase demonstrates different patterns at different noise levels. The increase from the noise level 10% to 20% is significantly larger than that from 0% to 20% on most datasets. Take the Iris dataset for example, when raising noise level from 0% to 10%, the error increase is 0.0047 (from 0.0905 to 0.0952), but when noise level moves from 10% to 20% the increase reaches 0.0236 (from 0.0952 to 0.1188) which is almost as five times big as the former increase.

LSD and HSD have the same grouping of the three noise levels on all the 9 datasets in Table 5. On all the 9 datasets the noise level 20% belongs to a different group from noise level 0%, which means it has significantly higher means than noise level 0%. On 5 datasets (Balance, Bupa, Cmc, Glass2 and Housing), the difference of mean between noise level 10% and 0% is significant, they fall into different groups. On the remaining 4 datasets (Breast, Ionosphere, Iris and Wine), the noise level 0% and 10% perform similarly and they belong to the same group.

The results of Tables 3–5 indicate that noise has a great negative impact on imputation methods. Even though sometimes low noise levels have limited impact on imputation, high noise levels are doomed to degrade the imputation results drastically.

5.2.3 *Robustness of RIBG*

To validate the second observation in Sect. 5.2.1, i.e., the robustness of RIBG at high noise level, the ANOVA model is used again to test the null hypothesis that all imputation methods perform the same at higher noise level (20% noise level). In the ANOVA model, three factors are considered: missing rate, missingness mechanism and imputation method.

Table 6 tabulates the analysis of the main effect *imputation method* on the 9 datasets, where the second column presents the degrees of freedom (DF) (since there are five imputation methods, the DF for that factor is 4), the third column reports the Type III Sum of Squares (SS), the fourth and fifth column give the F-values and p-values, and the sixth column reports will the null hypothesis be rejected at 5% significance level. We can see from this table that the main effect *imputation method* is significant on all the 8 datasets except Bupa at 5% level. Therefore, we can reject the null hypothesis and infer that at least one imputation method has significantly lower errors than other methods one the 8 datasets.

To identify which methods have lower imputation errors, Tables 7–11 report marginal means and the grouping of the five imputation methods according to LSD and HSD at 5% significance level on each dataset. A, B and C in the tables denote the first, second and third group, respectively.

Table 6 Main effect *imputation method*

Dataset	DF	SS	F-value	p-value	Hypothesis
Balance	4	0.0037	14.02	<0.001	Reject
Breast	4	0.0021	13.70	<0.001	Reject
Bupa	4	6.1E-1	1.27	0.301	Accept
Cmc	4	0.0025	3.73	0.012	Reject
Glass2	4	0.0014	4.40	0.005	Reject
Housing	4	0.0016	9.56	<0.001	Reject
Ionosphere	4	0.0938	28.12	<0.001	Reject
Iris	4	0.0180	16.27	<0.001	Reject
Wine	4	9.8E-4	3.02	0.030	Reject

Table 7 Means and grouping of imputation methods (Balance and Breast)

Method	Balance			Method	Breast		
	Mean	LSD	HSD		Mean	LSD	HSD
RIBG	0.2403	A	A	RIBG	0.1002	A	A
EI	0.2516	B	B	EI	0.1071	B	B
RI	0.2541	B	B	GBNN	0.1084	B	B
MI	0.2612	BC	B	RI	0.1171	C	C
GBNN	0.2670	C	C	MI	0.1189	C	C

Table 8 Means and grouping of imputation methods (Bupa and Cmc)

Method	Bupa			Method	Cmc		
	Mean	LSD	HSD		Mean	LSD	HSD
RIBG	0.1410	A	A	RIBG	0.2314	A	A
GBNN	0.1428	A	A	EI	0.2350	A	A
EI	0.1449	A	A	RI	0.2364	A	A
RI	0.1486	A	A	GBNN	0.2376	A	A
MI	0.1509	A	A	MI	0.2532	B	B

Table 9 Means and grouping of imputation methods (Glass2 and Housing)

Method	Glass2			Method	Housing		
	Mean	LSD	HSD		Mean	LSD	HSD
RIBG	0.1196	A	A	RIBG	0.0896	A	A
GBNN	0.1225	A	A	GBNN	0.0918	AB	A
MI	0.1249	A	A	EI	0.0961	B	B
EI	0.1268	AB	AB	MI	0.1036	C	BC
RI	0.1363	B	B	RI	0.1049	C	C

The grouping is obtained in the same way as described in Sect. 5.2.2. All the five imputation methods are sorted by imputation accuracy in each table in descending order.

Table 10 Means and grouping of imputation methods (Ionosphere and Iris

Method	Ionosphere			Method	Iris		
	Mean	LSD	HSD		Mean	LSD	HSD
RIBG	0.1318	A	A	RIBG	0.0963	A	A
GBNN	0.1327	A	A	GBNN	0.1003	A	A
EI	0.1448	A	A	MI	0.1168	AB	
MI	0.1477	A	A	EI	0.1296	BC	B
RI	0.2523	B	B	RI	0.1510	C	C

Table 11 Means and grouping of imputation methods (Wine)

Method	Wine		
	Mean	LSD	HSD
RIBG	0.1110	A	A
EI	0.1128	AB	AB
GBNN	0.1162	AB	AB
MI	0.1198	BC	AB
RI	0.1239	C	B

As can be seen from Tables 7–11, in terms of imputation error, RIBG gives the lowest error on all the 9 datasets. The next is GBNN and it takes the second place on 5 datasets and either the third or fourth place on 3 datasets. EI is worse than the above two methods. It ranks second on 4 datasets, ranks third on 3 datasets and ranks either fourth or fifth on 2 datasets. RI and MI are the worst methods, and they never take the first or second place. Meanwhile, the last position is taken by one of them on 8 datasets.

According to LSD and HSD test, RIBG belongs to the first group (group A) on all the 9 datasets and takes this position alone on 2 datasets. For GBNN, it belongs to the first group (group A) on 7 datasets but never takes it by itself. It also falls into the second group (group B) on 3 datasets and the third group (group C) on 1 dataset. For EI, RI and MI, they sometimes stays in the first group, but never take it by themselves. It is worth noting that on dataset Bupa the difference between methods is not statistically significant and all the methods belong to the same group.

All the above results from Tables 6–11 demonstrate that RIBG achieves higher imputation accuracy at high noise level (20%) when comparing with other benchmark methods.

5.3 Time complexity

We compare the runtime performance of RIBG with multiple imputation, which is the most popular imputation method. All the experiments are conducted on a same PC with a 1.73 GHz Intel Core 2 Duo processor and 2 GB memory running on Windows Vista operating system. Table 12 presents the average time (in seconds) required for producing the imputation result on one dataset. According to Table 12, the time complexity RIBG is acceptable in comparison with multiple imputation. In view of the higher imputation accuracy, our method is worth paying attention to.

Table 12 Execution times (in seconds) to impute missing values

Dataset	MI	RIBG
Balance	19.12	30.15
Breast	32.74	44.69
Bupa	13.24	16.02
Cmc	104.15	98.91
Glass2	11.55	18.69
Housing	36.26	53.97
Ionosphere	66.93	287.31
Iris	7.58	12.81
Wine	15.36	44.12

5.4 Discussion of the results

In the previous subsections, we presented numerous experiments and comparative studies. Two important conclusions can be drawn as follows:

1. Noise has a great impact on imputation methods and the patterns of influence are different for low noise level and high noise level. The former has a limited impact on imputation accuracy and the latter deteriorates the results dramatically.
2. RIBG outperforms other benchmark methods at high noise level. This does not surprise us, because most commonly used imputation methods are not intentionally designed for noisy environment, whereas RIBG inherits the noise-immunity of GMDH, so its robustness is expectable. But we do not claim RIBG outperforms other benchmark methods in all situations. Actually, as we have noticed in the Breast dataset, when the noise level is low the performance of EI is comparable to RIBG. However, as the noise level goes higher, the merit of our method becomes obvious. Given a new incomplete dataset, we usually do not have any prior knowledge about whether the noise level is low or high, so it is preferable to use RIBG.

Another interesting finding in our experiments is that MI and RI which were considered to be effective techniques in noisy environment [47] perform poorly. One possible reason for this is that the experiments in [47] assume only the dependent variable contains noise while our experiments make the assumption that all the variables in the dataset include

noise. Intuitively, the imputation results using noise-free independent variables are different from those using noisy ones. The second explanation is that the software measurement dataset is utilized in [47] while the UCI datasets from several different domains are used in our experiments. The special features of software measurement data make itself fit MI and RI, whereas the characteristics of UCI datasets may make MI and RI unsuitable. Which factors determine this fitness remains to be answered.

6 Conclusions

In this paper, we systematically studied the impact of noise on missing value imputation methods when noise and missing values distributed throughout the dataset. By observing the behavior of the different imputation methods at different noise levels, we drew the conclusion that noise has great negative effects on imputation methods, especially when the noise level is high. Meanwhile, we designed a robust method RIBG based on GMDH to impute missing values in noisy environment. Comparative studies have shown that RIBG performs quite well in comparison with other four popular imputation methods in the presence of noise. Given the frequent occurrence of missing values and noise, RIBG is a good choice in imputing incomplete data and has great potential in real-world data mining applications.

References

1. Abdel-Aal RE (2005) GMDH-based feature ranking and selection for improved classification of medical data. J Biomed Inf 38(6):456–468
2. Aksenova TI, Yurachkovsky YP (1988) A characterisation at unbiased structure and conditions of their J-optimality. Sov J Autom Inf Sci 21(4):36–42
3. Asuncion A, Newman DJ (2007) UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. http://www.ics.uci.edu/mlearn/MLRepository.html
4. Aussem A, de Morais SR (2008) A conservative feature subset selection algorithm with missing data. In: Kellenberger P (ed) Proc eighth IEEE int conf on data mining, ICDM'08, Pisa, Italy, pp 725–730
5. Barron AR, Barron RL (1988) Statistical learning networks: A unifying view. In: Wegman E (ed) Proc the 20th symposium on the interface: computing science and statistics. American Statistical Association, Washington, pp 192–203
6. Batista G, Monard MC (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5–6):519–533
7. Beaumont JF (2000) On regression imputation in the presence of nonignorable nonresponse. In: Proc of the survey research methods section, ASA, pp 580–585
8. Chen S, Huang C (2003) Generating weighted fuzzy rules from relational database systems for estimating null values using genetic algorithms. IEEE Trans Fuzzy Syst 11(4):495–506
9. Chen S, Huang C (2008) A new approach to generate weighted fuzzy rules using genetic algorithms for estimating null values. Expert Syst Appl 35(3):905–917
10. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38
11. Elder JF, Brown DE (2000) Induction and polynomial networks. In: Fraser MD (ed) Proc network models for control and processing, induction and polynomial networks. Intellect Books, Exeter, pp 143–198
12. Farhangfar A, Kurgan L, Pedrycz W (2007) A novel framework for imputation of missing values in databases. IEEE Trans Syst Man Cybern Part A, Syst Humans 37(5):692–709
13. Farhangfar A, Kurgan L, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recogn 41(12):3692–3705
14. Ford BL (1983) An overview of hot-deck procedures. In: Madow WG, OIkin I, Rubin DB (eds) Incomplete data in sample surveys, vol II: theory and bibliographies. Academic Press, New York, pp 85–207
15. Harel O, Zhou XH (2007) Multiple imputation: Review of theory, implementation and software. Stat Med 26(16):3057–3077
16. Hathaway RJ, Bezdek JC (2002) Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. Pattern Recogn Lett 23(1-3):151–160
17. Hruschka ER Jr, Hruschka ER, Ebecken N (2007) Bayesian networks for imputation in classification problems. J Intell Inf Syst 29(3):231–252
18. Huang CC, Lee HM (2004) A grey-based nearest neighbor approach for missing attribute value prediction. Appl Intell 20:239–252
19. Ivakhnenko AG (1968) The group method of data handling-a rival of the method of stochastic approximation. Sov Autom Control 1–3:43–55
20. Ivakhnenko AG (1971) Polynomial theory of complex systems. IEEE Trans Syst Man Cybern 1(4):364–378
21. Ivakhnenko AG, Kocherga YL (1983) Theory of two-level GMDH algorithms for long-range quantitative prediction. Sov Autom Control 16(6):7–12
22. Ivakhnenko AG, Stepashko VS (1985) Noise stability of modeling. Naukova Dumka, Kiev
23. Lakshminarayan K, Harp SA, Samad T (1999) Imputation of missing data in industrial databases. Appl Intell 11(3):259–275
24. Lemke F, Mueller J (2003) Self-organising data mining. Syst Anal Model Simul 43(2):231–240
25. Little R, Rubin D (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
26. Madala HR, Ivakhnenko AG (1994) Inductive learning algorithms for complex systems modeling. CRC Press, Boca Raton
27. Mani S, Valtorta M, McDermott S (2005) Building Bayesian network models in medicine: The MENTOR experience. Appl Intell 22(2):93–108
28. Mannino M, Yang Y, Ryu Y (2009) Classification algorithm sensitivity to training data with non representative attribute noise. Decis Support Syst 46(3):743–751
29. Mehrara M et al (2009) Investigating the efficiency in oil futures market based on GMDH approach. Expert Syst Appl 36(4):7479–7483
30. Miller RG (1997) Beyond ANOVA: basics of applied statistics. Chapman & Hall, Boca Raton
31. Mueller JA, Lemke F (2000) Self-organizing data mining: an intelligent approach to extract knowledge from data. Libri Books, Berlin

32. Myrtveit I, Stensrud E, Olsson U (2001) Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Trans Softw Eng 27(11):999–1013
33. Oh S, Pedrycz W (2002) The design of self-organizing polynomial neural networks. Inf Sci 141(3–4):237–258
34. Olinsky A, Chen S, Harlow L (2003) The comparative efficacy of imputation methods for missing data in structural equation modeling. Eur J Oper Res 151(1):53–79
35. Puig V et al (2007) A GMDH neural network-based approach to passive robust fault detection using a constraint satisfaction backward test. Eng Appl Artif Intell 20(7):886–897
36. Qin Y et al (2007) Semi-parametric optimization for missing data imputation. Appl Intell 27(1):79–88
37. Quinlan JR (1993) C4. 5: Programs for machine learning. Morgan Kauffman, Los Altos
38. Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592
39. Saar-Tsechansky M, Provost F (2007) Handling missing values when applying classification models. J Mach Learn Res 8:1625–1657
40. Schafer JL (1999) Multiple imputation: A primer. Stat Methods Med Res 8(1):3–15
41. Stepashko VS, Yurachkovskiy YP (1986) The present state of the theory of the group method of data handling. Sov J Autom Inf Sci c/c of Avtomatika 19(4):36–46
42. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Addison-Wesley, Boston
43. Tsikriktsis N (2005) A review of techniques for treating missing data in OM survey research. J Oper Manag 24(1):53–62
44. Twala B (2009) An empirical comparison of techniques for handling incomplete data when using decision trees. Appl Artif Intell 23(5):373–405
45. Ungaro F, Calzolari C, Busoni E (2005) Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano-Veneta region of North Italy: Water retention properties. Geoderma 124(3–4):293–317
46. Van Buuren S et al (2006) Fully conditional specification in multivariate imputation. J Stat Comput Simul 76(12):1049–1064
47. Van Hulse J, Khoshgoftaar TM (2008) A comprehensive empirical evaluation of missing value imputation in noisy software measurement data. J Syst Softw 81(5):691–708
48. Williams D et al (2007) On classification with incomplete data. IEEE Trans Pattern Anal Mach Intell 29(3):427–436
49. Wu X, Zhu X (2008) Mining with noise knowledge: Error-aware data mining. IEEE Trans Syst Man Cybern Part A 38(4):917–932
50. Zhu X, Wu X (2004) Class noise vs. attribute noise: A quantitative study. Artif Intell Rev 22(3):177–210

Bing Zhu is currently a Ph.D. student at Sichuan University. He received his B.Sc. and M.Sc. in 2005 and 2008, respectively, all from the Sichuan University, Chengdu, China. His research interests include knowledge discovery, data mining, fuzzy modeling, and forecasting.

Changzheng He is a professor at Business school of Sichuan University. He received his M.Sc. degree in mathematics from Southwest China Normal University, Chongqing, China. His research interests include data mining, forecasting and knowledge discovery in customer relationship management. He has published extensively in these areas in various journals and conferences, including Knowledge-based Systems, Journal of Forecasting, International Journal of Systems Science, etc. He has been a member of numerous conference program committees in the area of inductive modeling.

Panos Liatsis is a senior lecture at School of Engineering and Mathematical Sciences, City University, UK. He is also the director of the Information and Biomedical Engineering Centre of City University. He graduated with the Diploma degree in electrical engineering from the Department of Electrical and Computer Engineering at the Democritus University of Thrace, Greece. He received the Ph.D. degree in electrical engineering and electronics from the Control Systems Center at the University of Manchester (UMIST). His research interests include pattern recognition, neural and evolutionary systems. He has published over 100 research contributions in high-impact factor journals, books and international conference proceeding.