Efficient Utilization of Missing Data in Cost-Sensitive Learning

Xiaofeng Zhu, Jianye Yang, Chengyuan Zhang, and Shichao Zhang, *Senior Member, IEEE*

Abstract—Different from previous imputation methods which impute missing values in the incomplete samples by using the information in the complete samples, this paper proposes a Date-drive Incremental imputation Model, DIM for short, which uses all available information in the data set to impute missing values economically, effectively, orderly, and iteratively. To this end, we propose a scoring rule to rank the missing features by taking into account both the economical criterion and the effective imputation information. The economical criterion takes both the imputation cost and the discriminative ability of the feature into account, while the effective imputation information enables to use all observed information in the data set including the imputed missing values to impute the left missing values. During the imputation process, our DIM first detects the neednot-impute samples for reducing the imputation cost and noise, and then selects the missing features with the top rank to impute first. The imputation process orderly imputes the missing features until all missing values are imputed or the imputation cost is exhausted. Experimental results on UCI data sets demonstrated the advantages of our proposed DIM, compared to the comparison methods, in terms of prediction accuracy and classification accuracy.

Index Terms—Missing data imputation, cost-sensitive learning, decision tree, classification, imputation order, C4.5 algorithm, imputation cost

1 Introduction

Incomplete samples whose part of elements or entries are missed can often be found in the real-world data sets [1]. It usually is an challenging task to construct effective classification models on the incomplete data set which includes incomplete samples and complete samples (where all the elements are observed), as many machine learning algorithms were designed for dealing with complete data sets where all the samples are observed [2], [3]. To conduct knowledge discovery from the incomplete data set, previous methods usually discard the incomplete samples and only use the complete samples to construct learning models [4], [5]. As a result, the limited complete samples may be unavailable to guarantee the effectiveness of classification models. Hence, it attracts a number of research interests to focus on the study of dealing with incomplete data sets in real applications [6], [7], [8], [9], such as industrial applications [1], [10], web mining applications [4], and text clustering analysis [11], [12].

Traditional methods for dealing with incomplete data include Expectation Maximization (EM) algorithm [13], C4.5 algorithm [14], etc. Recently, cost-sensitive methods are designed to handle incomplete data sets, such as the parimputation method [12] and time-sensitive decision tree [15]. To deal with the missing values in the incomplete data sets, it is very popular to first impute the missing values (i.e., missing value imputation) followed by using all the samples for the construction of the classification tasks. Specifically, in the imputation process, discriminative models guess possible values for the missing values using classification models, such as decision tree and maximum mean discrepancy, while generative models estimate the missing values by statistical models [13], [16], [17], such as probabilistic principal component analysis [6] and nonparametric models [16]. After the imputation process, all information in the incomplete samples become available. As a result, both the incomplete samples and the complete samples are used to construct the classification models. However, previous methods of missing value imputation still have limitations to be addressed.

First, conventional imputation methods impute every missing feature equivalently to achieve high imputation performance (e.g., the maximal prediction accuracy) or classification accuracy. However, these methods do not take the diversity of missing features into account [15]. For example, some missing features are more important than others for knowledge discovery. On the other hand, cost-sensitive methods are designed to impute missing values for achieving either the minimal mis-classification cost or the minimal total cost (including the mis-classification cost, test cost, and so on [18]). However, previous cost-sensitive methods still equivalently consider every missing feature by replacing the classification accuracy with either the mis-classification cost or the total cost [19]. Hence, it will be interesting to regard both the classification accuracy and the imputation cost as the goal of missing value imputation by taking into account of the diversity of missing features.

- *X. Zhu is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China. E-mail: xfzhu0011@hotmail.com.*
- *J. Yang is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China. E-mail: jyyang@hnu.edu.cn.*
- *C. Zhang and S. Zhang are with the School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China. E-mail: {216070, zhangsc}@csu.edu.cn.*

Manuscript received 20 Nov. 2018; revised 17 Nov. 2019; accepted 17 Nov. 2019. Date of publication 28 Nov. 2019; date of current version 29 Apr. 2021. (Corresponding author: Shichao Zhang.)
Recommended for acceptance by Muhammad Aamir Cheema.
Digital Object Identifier no. 10.1109/TKDE.2019.2956530

Second, previous methods of missing value imputation do not use the observed information in the incomplete samples for the imputation process. That is, they use all the complete samples to construct the imputation models by ignoring the observed information in the incomplete samples. As a result, the observed information in the incomplete sample was wasted. Recently, a few literature have paid attention to this issue. For example, both EM algorithm [13] and C4.5 algorithm [14] could use the observed information in the incomplete samples as well as the complete samples to impute missing values. However, not all observed information are useful or effective for the imputation process. Identifying effective observed information should be important for improving the robustness of the imputation models.

In this paper, to address aforementioned issues, we investigate a Domain-driven Incremental imputation Model, DIM for short, by extending the pioneering work [20] to impute missing values. To do this, we assume that the quality of missing value imputation models may be related to the factors, such as imputation cost, discriminative ability of the feature, effective imputation information (EII), and imputation order. Specifically, (i) imputing missing value should suffer from cost and we call all costs in the imputation process as imputation cost; (ii) the feature is diverse and its importance (or discriminative ability) is measured by the mutual information between it and the decision feature, i.e., the class label; (iii) EII indicates the effective information in the incomplete data set that is available for the imputation process; and (iv) the imputation order is essential while an incremental imputation method is employed.

Based on the above assumptions, the goal of our DIM is to use available EII to orderly impute missing values, aiming at achieving the minimal imputation cost and the maximal imputation accuracy. More specifically, our DIM detects neednot-impute samples including absent samples and predictable samples to reduce imputation cost and noise. This is because that the missing values in the absent samples will not be imputed if they do not change the effectiveness of the imputation model, while the missing values in the predictable samples could be correctly predicted by current complete samples in the incomplete data set. Second, our DIM automatically learns a scoring rule to rank all missing features whose elements contain missing information to achieve a trade-off between the economical criterion (the combination of the imputation cost and the discriminative ability for every missing feature) and EII. Last, our DIM imputes the top-ranked missing feature first, and then combines this missing feature and exiting complete samples to select the next missing feature to impute until the imputation process is finished. In this way, our DIM is able to impute missing values economically, effectively, orderly, and iteratively.

Compared to previous methods, our proposed DIM has the following contributions.

- Our DIM is a cost-sensitive imputation method, which is essential and practical in real applications as missing value imputation always suffers from cost. However, previous methods only focused on achieving classification accuracy while other existing methods took misclassification cost into account. Moreover, our proposed DIM integrates the imputation cost (taking the diversity of missing data into account) with the discriminative ability (i.e., mutual information) in a unified framework.
- Our DIM uses the observed information in the incomplete samples to add observed information for the imputation process. Moreover, our DIM orderly selects the most economical feature to be imputed first and then imputes the left missing features with all available information including original complete samples and new complete samples (i.e., the missing features have been imputed before). Thus our DIM may output reliable imputation models.
- Our DIM simultaneously takes economical criterion, imputation cost reduce, and making the best use of observation information, into account. Experimental results on 12 UCI data sets demonstrated that our proposed DIM outperformed the comparison methods and individual part in our DIM is feasible, in terms of imputation performance.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 presents the preliminaries and the problem definition. The proposed DIM is presented in Section 4. Experimental analysis is conducted in Section 5. Finally, the conclusion of this work is listed in Section 6.

2 Related Work

In this section, we briefly review state-of-the-art literature on missing data imputation techniques and cost-sensitive learning.

2.1 Missing Value Imputation

Four strategies for handling missing values have widely been applied in real applications, e.g., complete case analysis, inference from the data set with missing values, missing data imputation, and parimputation. The complete case analysis uses only the complete samples to impute the missing values in the incomplete samples [2], [21], [22], [23]. The method of inference from the data set with missing values does not deal with the missing values, e.g., [24], [25], as it directly conducts knowledge discovery using generative models from incomplete data sets. Missing value imputation guesses possible values for missing information based on the complete samples [3], [26]. The parimputation strategy advocates that missing data is imputed if and only if its neighborhood includes enough complete samples. Otherwise, the missing values will not be imputed [12], [23].

Research on missing value imputation is very popular in both statistics and machine learning [4], [9], [23], [27]. Statistical imputation methods often pay attention on the data sets with continuous features, or the data sets with the majority of continuous features and a few discrete features [12]. Lobato *et al.* [28] proposed a multi-objective genetic algorithm to process mixed-attribute data sets which include both categorical and continuous features. Recently, a number of literature investigated the data imputation problem on non-quantitative string data [7], [29]. Missing data imputation has also widely been studied in machine learning, such as associate-rule based imputation [30], rough set method [31], C4.5 algorithm [14], and other methods [1], [3]. More recently, Zhao *et al.* [32]

employed deep neural network to impute the imcomplete data in cyber-physical systems.

Based on the imputation times, existing literature of missing data imputation can be categorized into the following subgroups, including single imputation [33], multiple imputation [34], [35], fractional imputation [25], and iterative imputation [36]. In particular, single imputation provides a single estimation for every missing data, such as C4.5 algorithm and kNN method [12]. Multiple imputation generates multiple values for every missing data [37], [38]. The iterative imputation approach can utilize all useful information, e.g., [1], [13], [39], [40] and this work. Fractional imputation combines single imputation with multiple imputation to impute missing values [41], [42], [43]. That is, this method imputes missing values once, but each missing value contains multiple imputation results. By comparing to existing methods for missing data imputation, Peng and Zhu suggested that multiple imputation methods outperform EM algorithm [31], while Kang *et al.* thought that fractional imputation is more efficient than either single imputation or multiple imputation [41].

2.2 Cost-Sensitive Learning From Incomplete Data

Cost-sensitive learning is a hot research topic in machine learning. Turney discussed many types of cost in machine learning, such as misclassification costs (i.e., cost incurred by misclassification errors), test cost (i.e., cost incurred for obtaining feature values [19]) [18]. Recently, the literature investigated active feature value acquisition by the principle of the maximal cost [44], [45].

Imputation order, which orderly imputes missing values one by one with a predefined criterion, has been considered to improve the classification accuracy [20], [46], [47]. For example, Estevam *et al.* designed an exhaustive search method to demonstrate that the use of an order process could improve the classification efficiency of imputation [47]. Conversano and Siciliano designed a lexicographic order to impute multiple missing values without considering either classification accuracy or other principles [46]. Lobo and Numao proposed to first impute the missing features with the maximal mutual information to the decision feature [48]. However, it did not pay any attention to the imputation cost and the effectiveness.

2.3 Efficient Imputation Information for Classification

Zhu *et al.* indicated that both imputation order and efficient imputation information should be considered for missing value imputation [20]. In real applications, data sets usually contain a number of incomplete samples leading to that imputing missing values with a few complete samples will easily deteriorate the model performance. For example, Lakshminarayan, *et al.* [10] and Pearson [49] analyzed an industrial data set which has 4,383 incomplete samples and zero complete sample. In this case, it is impossible to impute missing values in the data set with traditional methods [25], [41]. Moreover, in the incomplete data sets, the number of the incomplete samples is often small, but the percentages of incomplete samples is very high. For example, the percentages of missing values, i.e., missing rate, on six UCI data sets Water-Treatment, Hepatitis, Bridge, Echocardiogram, Soybean, and House-Voting, are 2.95, 5.67, 5.56, 7.69, 6.63, and 4.13 percent, respectively, while the percentages of their incomplete samples are 26.56, 48.38, 35.18, 53.79, 13.36, and 46.68 percent, respectively.

3 Preliminaries and Problems

In this section, we first introduce basic concepts and definitions related to missing data imputation, and then list the problem of missing value imputation.

3.1 Incomplete Data Sets

Let $\mathcal{D}$ be a data set with n samples and m features, where some features contain missing values. In this paper, we call the samples containing missing values as incomplete samples, and the samples with all observed values as complete samples. For example, Fig. 1a shows a data set with missing values, where observed values are represented by white cells and missing values are marked as black cells. Based on the principle of the lexicographic order [46], [48], the rearranged data set of the original data set (i.e., Fig. 1b) satisfies the following conditions:

$$c_1 \leq c_2 \leq \cdots \leq c_m; \text{ and } r_1 \leq r_2 \leq \cdots \leq r_n, \tag{1}$$

where c_i is the ith feature and r_j is the jth sample. In this way, the features and the samples are both sorted in terms of the number of missing values. The rearranged data matrix, denoted by M, of the original data set $\mathcal{D}$, can be viewed as four parts of entries as follows:

$$M_{n,m} = \begin{bmatrix} A_{p,d} & C_{p,m-d} \\ B_{n-p,d} & D_{n-p,m-d} \end{bmatrix}. \tag{2}$$

The data set M includes four blocks, i.e., A, B, C, and D, where $A_{p,d}$ denotes that Block A has p rows and d columns. Moreover, the yellow part, i.e., the first p rows in Fig. 1c, includes Block A (i.e., d columns from the left) and Block C (i.e., (m - d) columns from the right), while the non-yellow part, i.e., the last (n - p) rows in Fig. 1c, includes Block B (i.e., d columns from the left) and Block D (i.e., (m - d) columns from the right).

3.2 Economical Criterion

By analyzing previous literatures [3], [4], we have the following observations. First, feature has diversity. That is, different features have different importance (such as discriminative efficiency) for the classification task. Therefore, building an imputation model should pay attention to the feature importance. Second, the imputation process should suffer from cost. Ignoring the imputation cost may possibly use uneconomical features, and therefore result in more cost compared to their contribution to the imputation process. Third, under the assumption that features have different imputation cost, it is useful to pay attention to the features with the lowest cost. Last, the classification ability must be taken into account for the construction of the imputation models as the classification performance is the goal of missing value imputation.

Based on above analysis, we propose to define an economical criterion to meet the above four observations.

Definition 1 (Economical Criterion). *Let $\mathcal{D}$ be a data set, C be the decision features with respect to the ith feature, MI_i be*

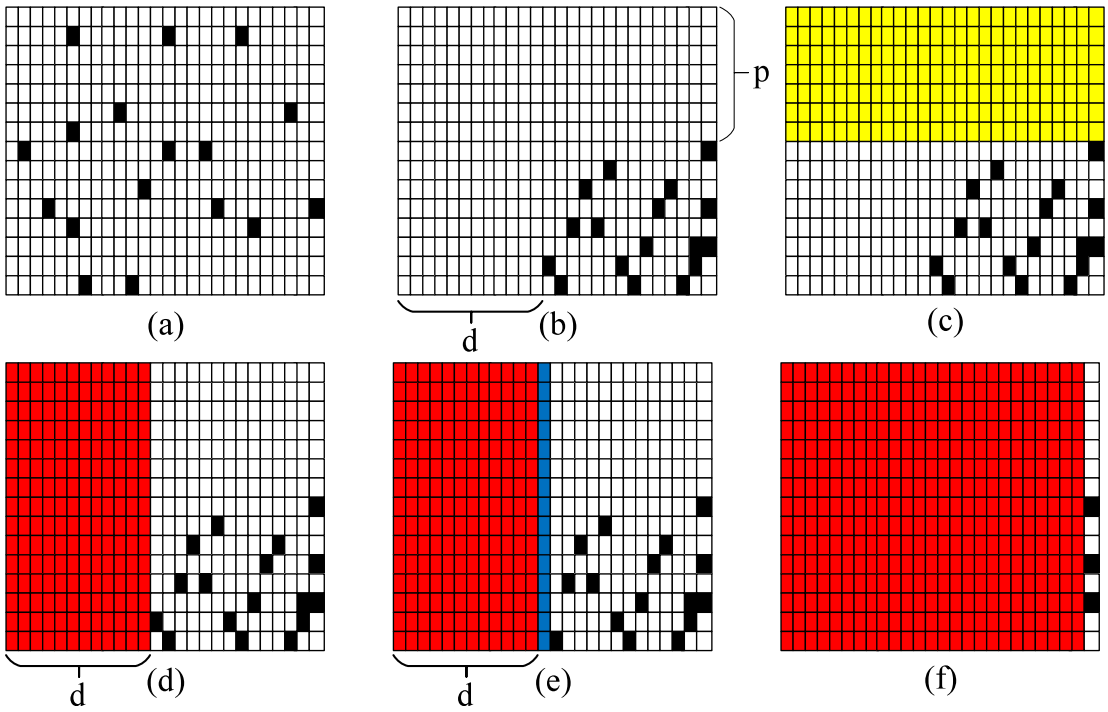


Fig. 1. The flowchart of the proposed Data-drive Incremental imputation Model (DIM) where every white cell and every black cell, respectively, represent an observed element and a missing element. (a) Original data set without the decision feature (or the class labels); (b) The rearranged data set of original data set based on the lexicographic order has *d* complete columns and *p* complete rows. (c) EII (i.e., *p* yellow rows) results in a row imputation order; (d) EII (i.e., *d* red columns) results in a column imputation order; (e) Imputation results (i.e., the blue cells) by using the EII in the red cells. After the imputation, EII in the next imputation includes the red cells and the blue cells; (f) EII (i.e., the red part) for imputing the last missing feature with the column imputation order.

the classification ability to $\mathcal{D}$, and $Cost_i$ be the imputation cost caused by imputing the missing values in the ith feature in $\mathcal{D}$, the economical criterion of ith feature EC_i is defined as follows:

$$EC_i = \frac{MI_i}{Cost_i}, \tag{3}$$

where

$$MI_i = \sum_{x \in F_i} P(x) log \frac{1}{P(x)} - \sum_{y \in C} P(y) \sum_{x \in F_i} P(x|y) log \frac{1}{P(x|y)}. \tag{4}$$

It is noteworthy that $P(x)$ indicates the probability of the condition feature x and $P(y)$ indicates the probability of the decision feature y. Based on Eq. (1), features with high mutual information with respect to the class labels have a high chance of being incorporated into the final decision, and vice versa. Hence, the larger value of EC_i, the more economical the ith feature is. Moreover, the definition of the economical criterion allows the user to take a trade-off between the discriminative ability measured by mutual information and the imputation cost, i.e., economically using the resources to conduct missing value imputation.

In our imputation model, we normalized every MI_i and $Cost_i$ to be zero mean and unit variance for avoiding the bias of different magnitudes.

Definition 2 (Imputation Cost). *The $Cost_i$, i.e., the imputation cost of the ith missing feature, is the sum of computing cost (i.e., cost for running the imputation process for the ith feature) and mis-imputation cost, i.e., the cost of wrong imputation for imputing the missing values in the ith feature.*

Note that, the imputation cost can be measured in a monetary unit, as well as can be quantified in many aspects, such as distance, time, labor, and danger [50]. In this paper, a general form of the imputation cost is considered, i.e., all the cost are assigned in the same unit, e.g., dollar, as in [19]. Apparently, the larger the value of EC_i, the more economical the imputation cost of the ith feature is. The economical criterion achieves little imputation cost as well as much classification ability.

3.3 Cost-Sensitive Learning on Incomplete Data Sets

In this work, we expect to impute missing values to minimize the imputation cost and maximize the discriminative ability. In the imputation process, we regard the feature which will be imputed as the decision feature, as well as regard the other features (including other condition features and the class labels) as the condition features, as in the literature [46], [51], [52], [53]. As a result, the imputation process is transferred to the task of predicting the class labels, i.e., a classification task. Based on previous literature [46], [51], [52], our proposed DIM only deals with discrete features where the continuous values must be discretized before hand. Finally, the problem of missing data imputation is formally defined as follows.

Definition 3 (Missing Data Imputation). *Given a data set D containing a set of incomplete samples, the issue of missing*

value imputation is solved by a classification problem, such that the total imputation cost is minimized while the classification accuracy is maximized.

4 Approach

As discussed in Section 2, a good imputation model should take into account the aspects, including imputation cost, discriminative ability, imputation order, and all the observed information. Even though the economical criterion is one of the possible candidates for orderly imputing the missing feature, it may not be the best; that is, the missing feature with the top value of economical criterion may not be the first missing feature to be imputed. This is because 1) the missing values in this feature might have inefficient observed information to build good imputation models, and 2) this feature may contain more missing values than other features.

Hence, utilizing observed information to iteratively or incrementally impute missing features should be considered. Recall the example in Fig. 1. In traditional imputation methods, such as C4.5 algorithm, the complete samples, i.e., the yellow cells in Fig. 1c, and the class labels are employed to impute every missing value. In some iterative methods, e.g., the EM algorithm, all the observed information in the data set, i.e., all the white cells in Fig. 1b, will be employed to impute the missing values. Different from these two methods, we develop an incremental imputation method to select the missing feature with the top rank to be imputed in each imputation round.

4.1 Scoring Missing Features

In this section, we first propose a new scoring rule to rank all missing features, by integrating economical criterion and efficient imputation information (EII). Compared to traditional methods, e.g., C4.5 algorithm [14] and the methods in [1], [4], [12], our proposed DIM can use more information to impute missing values.

4.1.1 Efficient Imputation Information

In Fig. 1, traditional imputation methods (e.g., C4.5 algorithm [14] and the methods in [3], [4]) may select the yellow cells and the class labels to impute all missing values, as shown in Fig. 1c. On the contrary, the proposed DIM prefers to select the red cells in Fig. 1d and the class labels to first impute the missing values in the $(d+1)$th column if its EII is larger than that in the yellow cells in Fig. 1c. After the selected missing values are imputed, we recalculate EII to determine the next missing feature to impute.

Based on above analysis, EII is important for the determination of the imputation order, i.e., either the row imputation order which imputes all the missing values row by row (e.g., Fig. 1c) or the column imputation order which imputes all missing values column by column (e.g., Fig. 1c). We present the definition of EII as follows.

Definition 4 (Efficient Imputation Information). *Efficient Imputation Information (EII) of a missing feature is defined as the percentage of available samples (including the complete samples and the imputed samples) that can be used for constructing imputation models, i.e.,*

$$EII_i = \max(info_{obs}, info_{eff}), \tag{5}$$

where $info_{obs}$ and $info_{eff}$ are the information used in traditional methods (the yellow cells in Fig. 1c) and the information for the missing values in the ith feature (the red cell in Fig. 1d), respectively. We further have

$$info_{obs} = \frac{row_{obs} \times col_{obs}}{row_{data} \times col_{data}},$$

and

$$info_{eff} = \frac{row_{eff} \times col_{eff}}{row_{data} \times col_{data}}.$$

It is noteworthy that the case in Fig. 1c will not be considered in this paper if its EII does not obviously increase by comparing with that in Fig. 1b based on Definition 4.

Given Definition 4, the EII of the $(d+1)$th feature in Fig. 1e is defined as

$$\begin{aligned} EII_{d+1} &= \max\left\{\frac{n \times (d+1)}{n \times (m+1)}, \frac{p \times (m+1)}{n \times (m+1)}\right\} \\ &= \max\left\{\frac{d+1}{m+1}, \frac{p}{n}\right\}, \end{aligned} \tag{6}$$

where n and m, respectively, are the numbers of the samples and the features in the data set, and d and p, respectively, are the numbers of the complete columns and the complete rows in the data set. From Definition 4, we can obtain three corollaries as follows.

Corollary 1. *The EII in traditional algorithms (such as C4.5 algorithm and EM algorithm) is a constant*

$$EII_{trad} = \frac{p \times (m+1)}{n \times (m+1)} = \frac{p}{n}. \tag{7}$$

Proof. It is obvious that the value of EII_{trad} equals to the values of $info_{obs}$ based on Definition 4. □

Corollary 2. *After imputing t missing features, the EII in Eq. (6) becomes*

$$\begin{aligned} EII_{d+1} &= \max\left\{\frac{n \times (d+t+1)}{n \times (m+1)}, \frac{p \times (m+1)}{n \times (m+1)}\right\} \\ &= \max\left\{\frac{d+t+1}{m+1}, \frac{p}{n}\right\}. \end{aligned} \tag{8}$$

Proof. Because our DIM is an incremental imputation method. That is, after finishing to impute all missing values in one missing feature, DIM obtains n new observed values because the imputed samples are regarded as the complete samples for the next imputation. Therefore, after imputing t missing features, the observed information will increase to $t \times n$, i.e., Eq. (8) holds.

For example, by setting $t = 1$, EII in Fig. 1e is

$$EII_{d+1} = \max\left\{\frac{d+2}{m+1}, \frac{p}{n}\right\}. \tag{9}$$

□

Corollary 3. *The value of EII in our proposed DIM is not less than EII_{trad}.*

The proof of Corollary 3 can directly be obtained by Eq. (6) and Corollary 2.

4.1.2 Scoring Rule

We investigate two rules to score missing features so that we use the obtained score to yield the imputation order. The first rule is that the feature with fewer missing values is imputed prior to the feature with more missing values. For example, in Fig. 1d, the $(d+1)$th feature is imputed prior to the last feature based on this rule. The reason is that imputing missing values might introduce noise, and the feature with more missing values could easily result in generating noise, compared to the feature with less missing values. Therefore, this rule can reduce noise.

The second rule is the preference between EC_i and EII_i. Based on Sections 3.2 and 4, our DIM can result in cheap cost and much information, while combining EC_i with EII_i for conducting missing value imputation. In this paper, the weighted harmonic mean [54] is employed to make a trade-off between EC_i and EII_i through a nonnegative coefficient $\alpha \in (0, +\infty)$. To do this, we have the following definition.

Definition 5 (Rank Score). *Assuming that EII_i has a unit weight and EC_i has a weight of α ($\alpha \in (0, +\infty)$), the rank score $Rank_i$ of the ith missing feature is defined as the weighted harmonic mean between EC_i and EII_i as follows:*

$$Rank_i = \frac{(\alpha+1)EII_i \times EC_i}{EII_i + \alpha EC_i}. \tag{10}$$

In Eq. (10), missing features obtain the scores according to the rank as follows: compute $Rank_i$ for each missing feature and sort them with the descending order. It is important to take into account the magnitude gap between EC_i ($EC_i \in (0, +\infty)$) and EII_i ($0 < EII_i < 1$). Generally, the result is usually prone to the data with big magnitude. Therefore, EC_i need to be normalized, i.e., $0 < EC_i < 1$, because the value of EII_i is between 0 and 1.

In Eq. (10), on one hand, the coefficient α is regarded as a preference between EC_i and EII_i. On the other hand, it also demonstrates the principle that features with the least missing values are first chosen in the imputation process, i.e., the first rule for the scoring rule.

In this paper, we denote the missing features as $I(i)(i = 1, \ldots, t)$. Moreover, the array $I(1), \ldots, I(t)$ is sorted in descending order. More specifically, the value of $I(1)$ is the maximal value while the value $I(t)$ is the minimal value. In particular, $I(t)$ is always 1. We also assume that $I(i)$ and $I(j)$ belong to the same class if their numbers of missing values are the same. Based on such an assumption, we have an array with a descending order, i.e., $C(1), \ldots, C(k)$, where $k = I(1)$ and $k \leq t$. We need to obtain the value of α to meet above two scoring rules. To do this, we first denote the the maximal value and the minimal value of all the rank scores of the kth class, respectively, as $\max\{Rank_{c(k)}\}$ and $\min\{Rank_{c(k)}\}$, and then obtain the value of α by satisfying

$$\underset{k=1,2,\ldots,I(1)}{\arg} \{\max\{Rank_{c(k)}\} \leq \min\{Rank_{c(k+1)}\}\}. \tag{11}$$

The solution of Eq. (11) may suffer from at least two issues as follows. i) $(I(1)-1)$ inequalities in Eq. (11) need to be solved. The monotonicity of the variable α in Eq. (10) is related to the value of $EII_i - EC_i$, but its sign (i.e., minus or plus) is different for other features. In this case, it is difficult to solve Eq. (11); ii) Setting the initial value of α is challenging. The reason is that the value of α should simultaneously satisfy both Eqs. (11) and (10).

In this paper, we propose a heuristic method to address above two issues for the setting of α. Specially, by randomly setting the value of α as any positive number,[1] Eq. (11) can be improved as follows:

$$\arg\{\max\{Rank_{c(k)}\} \leq \min\{Rank_{c(k+1)}\}\}. \tag{12}$$

According to Eq. (12) and the initial value of α for $C(t+1)$, the value of α for $C(t)$ can be solved by setting $k = t$. After this, Eq. (12) may sequentially output $C(t-2), \ldots, C(1)$. As a result, solving $(I(t)-1)$ inequalities in Eq. (11) is transformed to only one inequality in Eq. (12) when giving an initial value for α. More specifically, our DIM uses the following steps to generate the value of α by considering Eq. (12): First, it computes $\max\{Rank_{c(t)}\}$ which can satisfy Eq. (12) based on the initial value of α for $\min\{Rank_{c(t+1)}\}$; Second, the value of α will be updated by Eq. (12) to output the new value of α , i.e., the value of α for $\{Rank_{c(t-1)}\}$. After this, we could finally select an appropriate α to meet the two rules aforementioned. Note that, in the DIM method, EII is varied with the increase of the imputed missing features. Therefore, it is necessary to adjust α so as to satisfy Eq. (12) when constructing next imputation model.

4.2 Identifying Neednot-Impute Samples

In this section, we discuss two types of neednot-impute samples which are incomplete samples and do not need to be imputed for saving the imputation cost, namely absent sample and predictable sample [55], which are defined as follows.

Definition 6 (Absent Sample). *An incomplete sample is regarded as an absent sample if it already contains enough information such that a classifier can correctly classify it.*

Definition 7 (Predictable Sample). *An incomplete sample is regarded as a predictable sample if its missing values are predictable from the complete samples of the data set.*

Imputing either an absent sample or a predictable sample is unnecessary due to 1) It may generate bias and noise, as well as increases the imputation cost; and 2) The information for the imputation process is not changed. As a result, these two types of samples are identified and left unimputed in our DIM method.

Identifying Absent Samples. To identify absent samples, a classifier T_{abs} (e.g., C4.5 algorithm in this paper) is constructed with all complete samples in a given data set (i.e., the union of blocks A and C in Eq. (2), denoted by S). After that, T_{abs} is employed to evaluate each incomplete sample in blocks B and D in Eq. (2), where the union of B and D is denoted by MS. If an incomplete sample in MS can be correctly classified by T_{abs}, the sample is regarded as an absent sample. Once the absent samples have been identified, they are added to S to form a new set S_a.

1. It is noteworthy that the subsequent α will be very small if the initial value of α is small in our heuristic method.

Identifying Predictable Samples. To detect whether a missing value in MS is predictable, a classifier T'_j (e.g., C4.5 algorithm in this paper) is constructed for the jth feature in the new set S_a based on [52]. For each missing feature in the original data set, the EII of this feature is used to generate a classifier T'_j for the jth feature, where the jth feature is taken as the decision feature, and other complete features and the decision feature are taken as the conditional features. Accordingly, each missing value $MV_{i,j}$ in the jth feature is replaced by $M\hat{V}_{i,j}$ which is generated by T'_j, and then this classifier is applied to test whether all the imputed samples (i.e., their missing values are replaced by $M\hat{V}_{i,j}$) can be correctly classified. If the response is positive, the missing values are predictable; otherwise, they are unpredictable. Incomplete samples with only predictable missing values are added to form a new set S_p.

Algorithm 1. The Pseudo of the Proposed DIM

Input: A data set with missing values;
Output: A score list with the descending order for all missing features;
S: the set containing all complete samples;
MS: the set containing all incomplete samples;
S_a: the set containing all absent samples;
S_p: the set containing all predictable samples;
T_{abs}: C4.5 classifier for finding absent samples;
T'_j: C4.5 classifier for the jth missing feature;
$Flag_1(i,j) = 1$: the missing value in ith row and jth column of the set MS;
$Flag_2(i,j) = 1$: the missing value in ith row and jth column of the set MS is unpredictable;
$Flag_3(i,j) = 1$: the missing value in ith row and jth column of the set MS must be imputed;

```
1  Build classifier T_abs in S;
2  for each sample I_i in MS do
3    if CorrectClassify(I_i, T_abs) then
4      S = S ∪ I_i;
5      MS = MS − I_i;
6  for each MV_{i,j} in the jth feature in MS do
7    Flag_1(i,j) = 1;        /* is missing */
8    Build Classifier T'_j;
9  if !CorrectClassify(MV_{i,j}, T'_j) then
10   Flag_2(i,j) = 1;    /* is unpredictable */
11 for each sample I_i in MS do
12   if each Flag_2(i,j) == 1 then
13     MS = MS − I_i;
14     S = S ∪ I_i;
15   else
16     S_0 = s_0 + I_i;
17 MS' = MS − S_a − S_p;
18 while ∃ i,j, Flag_1(i,j) == 1 do
19   for each MV_{i,j} in jth feature in MS ∪ S_0 do
20     if Flag_3(j) == 1 then
21       Calculate Rank(j);
22   Sort Rank(j) in descending order;
23   CMA(j) = MV_{i,j} with maximal Rank(j);
24   CMA(i,j) = M̂V_{i,j};
25   Flag_1(i,j) = 0;
```

Consequently, samples in original MS can be classified into three subsets: a set of absent samples S_a, a set of samples with only predictable missing values S_p, and a set MS' including the samples with unpredictable missing values or the samples with both predictable and unpredictable missing values. Therefore, we have $MS = MS' \bigcup S_a \bigcup S_p$. It is noteworthy that our proposed DIM only imputes the set MS' and previous methods (e.g., C4.5 algorithm, EM algorithm, and the methods in [1], [4]) impute the set MS. Obviously, our proposed DIM needs less imputation cost than these previous methods.

4.3 Data-Driven Incremental Imputation Model (DIM)

Algorithm 1 illustrates the details of our proposed DIM. It is noteworthy that any of the well-established imputation methods can be used in the DIM. Without loss of generality, C4.5 algorithm is adopted in our DIM. Specifically, DIM identifies absent samples and predictable samples in Lines 1-16, which is an offline process. The missing features are then determined in Line 17, followed by the score rules and the imputation process in Lines 18-25.

The key part of the DIM is the construction of decision trees for imputing missing values, i.e., Lines 18-25. Based on the construction of decision tree in [3], [23], pruning techniques could be designed for the imputation process. For example, if the ith feature is independent of the class label, i.e., the mutual information between the ith feature and the decision feature, i.e., $MI_i = 0$, this feature will be pruned away and its missing values are not imputed. Such a pruning process can also be used in the imputation steps. For example, neither absent samples nor predictable samples are imputed for constructing imputation models. On the other hand, features that are independent of the current target feature are not taken into account while constructing an imputation model for both the decision feature and the identification of both absent samples as well as predictable samples. These two pruning techniques assist in efficiently selecting the top rank missing features and an appropriate α in each loop. As a result, this can reduce the imputation cost, improve the imputation performance, and decrease the algorithm complexity. Therefore, the proposed DIM is different from previous imputation methods. That is, our proposed DIM can orderly and incrementally impute missing values. Moreover, the imputation process is cost-sensitive.

5 Experimental Analysis

In this section, we empirically evaluated the effectiveness of the proposed DIM, compared to several existing methods, on 12 UCI data sets in terms of prediction accuracy and classification accuracy.

5.1 Experimental Settings

5.1.1 Set-Up

In our experiments, first, we conducted experiments to test the feasibility of the imputation order, the economical criterion, and the EII, in terms of unlimited imputation cost and limited imputation cost. Second, we evaluated the necessity of the detection of the neednot-impute samples as well as the selection of the parameter α in Eq. (10).

The continuous features in the selected data sets were discretized with WEKA software, because the algorithm is based

TABLE 1
Prediction Accuracy (%) of All Methods on Six UCI Data Sets

Data sets	MR	RI	IIA	ME	LE	MI	DIM
Auto-mpg	0.1	78.98	79.51	80.18	80.03	80.25	**82.21**
	0.3	77.12	77.63	79.90	79.41	79.20	**80.92**
	0.5	75.59	75.32	78.01	78.02	77.28	**79.82**
Wine	0.1	65.40	66.08	66.88	66.63	67.25	**68.94**
	0.3	64.17	65.34	66.03	65.89	66.87	**68.14**
	0.5	63.73	64.29	65.13	65.16	65.83	**67.59**
Tie-tac-toe	0.1	77.52	79.76	80.29	80.32	80.98	**82.23**
	0.3	73.32	74.98	76.58	76.75	77.01	**79.92**
	0.5	71.23	73.52	74.55	74.44	74.82	**78.29**
Vote	0.1	83.86	85.22	87.28	87.05	87.24	**90.79**
	0.3	82.08	84.73	86.43	86.02	86.10	**88.75**
	0.5	76.67	80.27	83.86	83.67	84.06	**86.74**
Balance scale	0.1	82.13	85.33	87.97	88.05	88.59	**89.35**
	0.3	80.29	82.93	84.75	84.33	84.46	**86.17**
	0.5	76.37	79.72	81.34	80.87	81.03	**83.45**
Breast cancer	0.1	88.14	91.07	92.45	91.85	91.41	**93.05**
	0.3	85.76	88.79	89.91	89.96	90.36	**91.79**
	0.5	77.76	80.53	82.36	81.94	82.71	**85.56**

MR: missing rates.

TABLE 2
Classification Accuracy (%) of All Methods on Six UCI Data Sets

Data sets	MR	RI	IIA	ME	LE	MI	DIM
Auto-mpg	0.1	66.82	68.02	69.71	69.78	70.03	**71.18**
	0.3	66.12	67.23	69.11	68.58	68.94	**70.25**
	0.5	64.93	66.38	68.01	68.04	68.39	**69.66**
Wine	0.1	80.83	81.52	84.78	85.21	84.67	**86.81**
	0.3	77.32	78.58	81.72	81.27	82.71	**84.08**
	0.5	74.30	75.62	77.66	77.51	79.32	**80.69**
Tie-tac-toe	0.1	83.65	84.96	85.68	85.43	85.92	**86.64**
	0.3	75.34	77.67	78.23	78.39	78.59	**83.83**
	0.5	72.15	73.95	75.79	76.34	77.83	**80.32**
Vote	0.1	90.55	92.34	94.25	94.67	95.74	**96.57**
	0.3	86.01	89.59	91.48	91.42	92.28	**95.00**
	0.5	81.65	83.98	86.85	86.43	87.73	**90.21**
Balance scale	0.1	89.07	90.75	92.94	92.68	92.24	**93.25**
	0.3	83.33	87.14	88.84	87.49	89.85	**92.11**
	0.5	81.37	83.69	86.33	85.96	87.34	**89.84**
Breast cancer	0.1	81.42	84.64	88.45	89.94	91.03	**91.48**
	0.3	80.24	83.03	86.14	87.18	87.38	**90.02**
	0.5	78.83	82.37	84.36	84.66	85.73	**88.65**

MR: missing rates.

on C4.5 algorithm that can only deal with discrete features. Imputation cost was generated with a random mechanism and was obtain at the beginning of building each imputation model.

5.1.2 Comparison Algorithms

The comparison methods include two imputation algorithms and three variants of our proposed DIM.

- Random Imputation (*RI*) imputes missing values without taking the imputation order into account. The imputation process does not stop until the imputation cost exceeds the fixed maximal cost or all missing values have been imputed.
- Incremental Imputation Algorithm (*IIA*) [46] does not consider either the economical criterion or EII for the construction of the imputation model. However, it uses the lexicographic order to impute missing values by taking the imputation order into account.
- Maximal Economics (*ME*), a variant of the proposed DIM, does not consider the discriminative ability of every feature, i.e., $MI_i = 1$ for every feature. Moreover, it prefers to first impute the missing feature with the maximal imputation cost.
- Least Economics (*LE*) prefers to first impute the missing features with the minimal imputation cost, while setting $MI_i = 1$ for every feature in the DIM.
- Mutual Information (*MI*) only takes the discriminative ability by setting $Cost_i = 1$ in our DIM.

It is noteworthy that both RI and IIA do not consider the efficient imputation information.

5.1.3 Evaluation Metrics

We employed two kinds of evaluation metrics, i.e., prediction accuracy (PA) and classification accuracy (CA), to evaluate the imputation performance of all methods

$$\begin{aligned} PA &= \frac{1}{t}\sum\nolimits_{i=1}^{t} l(IV_i, RV_i) \\ CA &= \frac{1}{n}\sum\nolimits_{i=1}^{n} l(IC_i, RC_i), \end{aligned} \tag{13}$$

where t is the number of the imputed missing values and n is the number of samples in the data set. $l(x, y)$ is an indicator function. Specifically, $l(x, y) = 1$ if $x = y$, otherwise 0. IV_i and RV_i are the imputed value and the real value for ith missing value, respectively. IC_i and RC_i, respectively, are the prediction result and the real class label for the ith sample. After imputing all missing values, all samples in the data set were used to construct a classifier, which is then used to classify each sample in the data set. Obviously, the higher either the prediction accuracy or the classification accuracy, the more effective the imputation algorithm is.

5.2 Experimental Result Analysis

5.2.1 Experiments With Various Missing Rates

In this section, we evaluated the performances of all algorithms on 6 complete UCI data sets. We randomly generated three incomplete data sets for each of them. The missing rates of thee incomplete data sets are 10, 30, and 50 percent, respectively. Moreover, the imputation cost was not considered, i.e., each missing value in a data set is directly imputed with unlimited imputation cost. Tables 1 and 2, respectively, report the results of prediction accuracy and classification accuracy. From the experimental results, we have the following observations.

With the increase of the missing rate, all methods decrease in terms of both prediction accuracy and classification accuracy. Moreover, when the missing rate is low, e.g., 10 percent, both DIM and MI perform very well. However, their

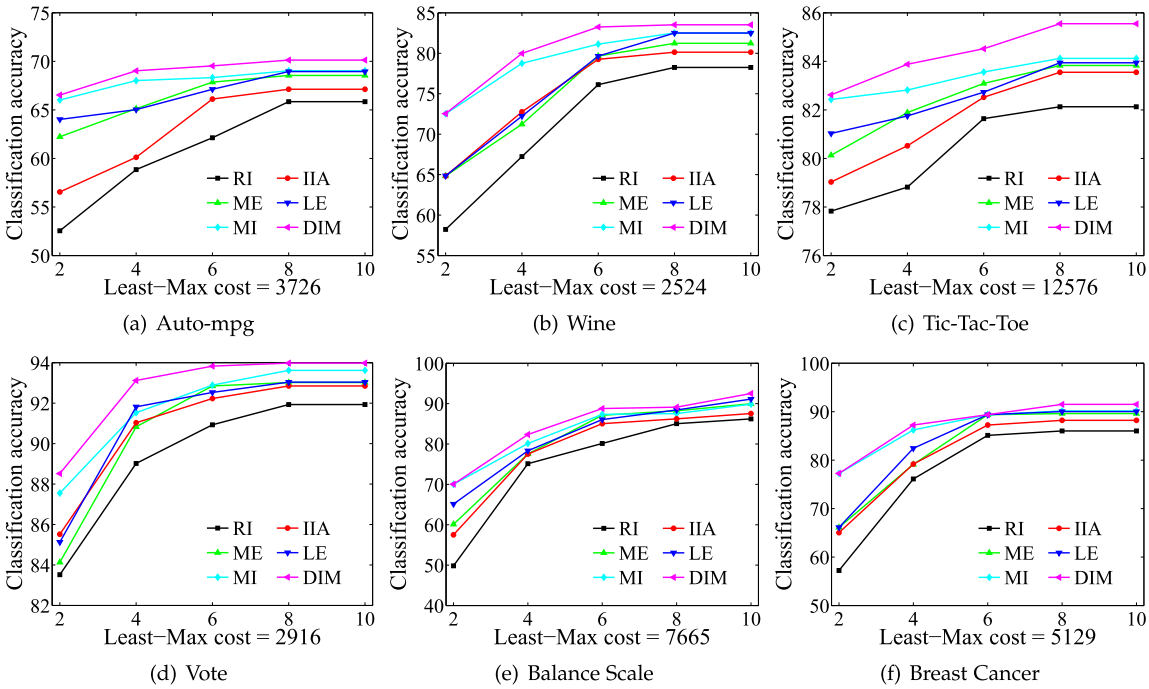


Fig. 2. Classification accuracy (%) of all methods with limited imputation cost (i.e., 1,000 unit for the imputation cost) on six data sets.

performance begins to decrease with the increase of the missing rate e.g., from 30 to 50 percent. Furthermore, the algorithms (e.g., ME, LE, MI, and DIM) output the best imputation results while the missing rate is moderate, e.g., 30 percent. However, our proposed DIM always achieves the best performance in terms of both the prediction and classification accuracy, followed by MI, ME, LE, IIA, and RI, at different missing rates. The reasons are in two folds: 1) it is difficult to construction a good imputation model with a large number of missing values, which concequently deteriorates the performance of both the classifier and the imputation model degrades as the the missing rate increases; 2) When the missing rate is high, i.e., the number of the complete samples is small, the absent samples and the predictable samples are difficult to detect.

Compared to RI which does not take the imputation order into account, the imputation methods considering the imputation order (e.g., IIA, ME, LE, MI, and our DIM) significantly outperform RI in terms of both prediction accuracy and classification accuracy. This indicates that it is necessary to consider the imputation order for imputing missing data. Among the algorithms taking the imputation order into account, e.g., IIA, ME, LE, and DIM, IIA is the worst as it only takes the imputation order into account by overlooking the economical criterion and the EII. This implies that the proposed scoring rule in this paper is feasible as our method achieves the best performance.

In a nutshell, each part of the imputation order, the economical criterion, and the EII, is feasible, while the imputation cost is enough. Moreover, all methods have similar trend in terms of prediction accuracy and classification accuracy.

5.2.2 Experiments With Limited Imputation Costs

In Section 5.2.1, the imputation cost was set as unlimited and every missing value was imputed. As a result, the prediction accuracy was found to have similar trend to the classification accuracy for all the methods. However, in real-world applications, it is difficult to conduct missing value imputation with unlimited imputation cost. Therefore, it is interesting to demonstrate whether our proposed DIM outperforms the comparison methods in terms of classification accuracy while each data set has a limited imputation cost. To do this, we assigned the imputation cost randomly drawn from [1,10] for every missing feature on each data set, as well as fixed the missing rate as 30 percent. Note that, if all missing values are imputed, the sum of imputation cost is called the Least-Max cost. The Least-Max cost for the data sets Auto-mpg, Wine, Vote, Tic-tac-toe, Balance scale, and Breast cancer, respectively, is 3726, 2524, 12576, 2916, 7665, and 5129. In our experiments, we set the imputation cost as 1,000 so that all the missing values cannot be imputed for every incomplete data set.

In our experiments, we set five different imputation cost, i.e., 2, 4, 6, 8, and 10, for every missing feature, to test the classification accuracy of all the algorithms, and reported the result in Fig. 2.[2] We list our observations as follows.

First, it is obvious that all algorithms achieve good classification accuracy with the increase of the imputation cost. In particular, our DIM achieves the best performance. The reason is that more incomplete samples are imputed means that more information is provided to impute subsequent missing features.

Second, our proposed DIM achieves the best performance, regardless of the limitation of imputation cost. In particular, when all imputation methods reach their maximal classification accuracy, DIM is still the best. The possible reasons includes 1) the scoring rule is able to balance the weight between EC and EII with limited imputation cost for missing

2. In Section 5.2.1, we have demonstrated that the prediction accuracy has similar trend to the classification accuracy for all the methods on six data sets, so did the experiments in this section. For simplicity, we only reported the result of classification accuracy in Section 5.2.1.

TABLE 3
Illustration the Effectiveness of our DIM for the Detection of Neednot-Impute Samples on Six Data Sets, in Terms of Classification Accuracy (%)

Data sets	MR	CA-DIM	CA-no
Water-Treatment	2.95%	**92.85**	91.56
Hepatitis	5.67%	**90.37**	88.46
Bridge	5.56%	**92.92**	91.81
Echocardiogram	7.69%	**88.33**	86.50
Soybean	6.63%	**88.47**	87.97
House-Voting	4.13%	**92.21**	90.12

MR: missing rate, CA: classification accuracy, and CA-no: classification accuracy for our DIM without detecting neednot-impute samples.

value imputation; and 2) the economical criterion guarantees the advantages of our proposed DIM, compared to the comparison methods.

Third, all methods considering the imputation order (i.e., IIA, ME, LE, MI, and DIM) outperform RI at different levels of imputation cost. Moreover, the methods (i.e., ME, LE, MI, and DIM) outperform IIA as IIA only considers the lexicographic order. Besides, LE performs better than ME when the budget is low. However, there is no significant difference between them when the budgets increase. The reason may be that the number of the available information for LE is more than that for ME. Compared both ME and LE with MI, if the budget is low, MI outperforms either ME or LE. Moreover, the higher the budget is, the smaller the difference between them is. The reason is that MI could first impute the missing values with high classification ability, so it quickly improves the classification performance, compared to either ME or LE. However, such difference decreases with the increase of the budgets.

To sum up, our DIM achieves the best result and the methods (e.g., our DIM, MI, LE, ME, and IIA) outperform RI, in terms of limited imputation cost. This indicates again that each part of the imputation order, economical criterion, and EII is effective while imputing missing values with limited imputation cost.

5.2.3 Experiments for Detecting Neednot-Impute Samples

In this section, we investigated the ability of our DIM for detecting the neednot-impute samples, i.e., the absent samples and the predictable samples, on six UCI data sets, which contain missing values. In Table 3, we reported the missing rate of each data set in the second column and the results of classification accuracy in the last two columns. We did not report the prediction accuracy in this section as we have no ground truth for the missing data.

Table 3 shows that our DIM with detecting neednot-impute samples significantly outperforms the same method without detecting the neednot-impute samples, on different data sets. This implies that detecting the neednot-impute samples could help reducing imputation cost and noise, and thus improving the classification performance.

5.2.4 Experiments for the Sensitivity of α

The tuning parameter α in Eq. (10) should meet two criteria in Section 4.1.2, i.e., the feature with less missing values is imputed prior to the feature with more missing values, and the trade-off between economical criterion and EII. More specifically, based on Eq. (10), our DIM prefers to economical criterion if we set $\alpha \geq 1$. Otherwise, our DIM prefers to EII. Moreover, the selection of the value of α should also meet the first scoring rule, i.e., Eq. (11). Furthermore, our DIM is an incremental model, so the value of α will change for imputing the next missing feature. To do this, a heuristic solution was proposed in Eq. (12). In this section, we set different initial values for the parameter α (i.e., $\alpha \in \{0.1, 0.3, 0.6, 1, 2, 5, 50, 500\}$) to test the sensitivity of our proposed DIM on six UCI data sets which contain different levels of missing values.

From Table 4, our proposed DIM achieves the best result while the value of α was set between 0.6 and 2 in our experiments. In particular, our DIM always obtain good classification accuracy while setting $\alpha = 1$, where the weight of economical criterion is equivalent to the weight of efficient imputation information. This indicates that both economical criterion and efficient imputation information are important for missing value imputation. In real applications, we suggest to select the values of the α under the assumption of our proposed heuristics method by the cross-validation scheme.

6 Conclusion

We have proposed a new imputation method by taking the imputation order, economical criterion, efficient imputation information, and detecting neednot-impute samples into account for missing value imputation. Experimental results on 12 UCI data sets demonstrated the advantages of our proposed method, compared to the comparison methods, in terms of prediction accuracy and classification accuracy. Moreover, experimental result indicated the effectiveness of the individual part (i.e., imputation order, economical criterion, efficient imputation information, and detecting neednot-impute samples) of our proposed DIM.

In our future work, we would like to extend our proposed framework to other tasks of data mining and machine learning, such as feature selection [22], [56], [57] and clustering [5],

TABLE 4
Classification Accuracy (%) of our Proposed DIM With Different Values of the Parameter α on Six Data Sets

Data sets	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 1$	$\alpha = 2$	$\alpha = 5$	$\alpha = 50$	$\alpha = 500$
Water-Treatment	86.15	91.70	91.76	**92.85**	92.01	91.67	89.52	87.23
Hepatitis	82.35	88.75	89.21	**90.37**	89.37	89.05	86.02	84.21
Bridge	85.30	90.95	**92.96**	92.92	92.01	91.98	86.25	85.32
Echocardiogram	80.12	86.75	88.17	88.33	**88.34**	86.93	82.64	81.94
Soybean	83.51	88.40	88.19	**88.47**	88.15	87.80	85.01	84.76
House-Voting	84.78	90.36	91.65	92.21	**92.26**	90.51	86.38	85.67

[9], [58], [59], instead of only the classification tasks [60], [61] in this work.

ACKNOWLEDGMENTS

This work is partially supported by the China Key Research Program (Grant No: 2016YFB1000905), the Natural Science Foundation of China (Grants No: 61836016, 61876046, 61573270, and 61672177), the Project of Guangxi Science and Technology (GuiKeAD17195062), the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing, the Guangxi High Institutions Program of Introducing 100 High-Level Overseas Talents, and the Research Fund of Guangxi Key Lab of Multisource Information Mining and Security (18-A-01-01).

REFERENCES

[1] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 110–121, Jan. 2011.

[2] S. Shan *et al.*, "WebPut: A web-aided data imputation system for the general type of missing string attribute values," in *Proc. IEEE Int. Conf. Data Eng.*, 2019, pp. 1952–1955.

[3] S. Zhang, "Cost-sensitive KNN classification," *Neurocomput.*, 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.11.101

[4] Z. Li, M. A. Sharaf, L. Sitbon, S. Sadiq, M. Indulska, and X. Zhou, "A web-based approach to data imputation," *World Wide Web*, vol. 17, no. 5, pp. 873–897, 2014.

[5] X. Zhu, J. Gan, G. Lu, J. Li, and S. Zhang, "Spectral clustering via half-quadratic optimization," *World Wide Web*, 2019. [Online]. Available: https://doi.org/10.1007/s11 280–019–00 731–8

[6] A. Sportisse, C. Boyer, and J. Josse, "Estimation and imputation in probabilistic principal component analysis with missing not at random data," 2019.

[7] Z. Li, L. Qin, H. Cheng, X. Zhang, and X. Zhou, "TRIP: An interactive retrieving-inferring data imputation approach," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2550–2563, Sep. 2015.

[8] X. Zhu, S. Zhang, Y. Li, J. Zhang, L. Yang, and Y. Fang, "Low-rank sparse subspace for spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 8, pp. 1532–1543, Aug. 2019.

[9] X. Zhu, S. Zhang, W. He, R. Hu, C. Lei, and P. Zhu, "One-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 2022–2034, Oct. 2019.

[10] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Appl. Intell.*, vol. 11, no. 3, pp. 259–275, 1999.

[11] M. S. Aktaş, S. Kaplan, H. Abac1, O. Kalipsiz, U. Ketenci, and U. O. Turgut, "Data imputation methods for missing values in the context of clustering," in *Big Data and Knowledge Sharing in Virtual Organizations*. Hershey, PA, USA: IGI Global, 2019, pp. 240–274.

[12] S. Zhang, "Parimputation: From imputation and null-imputation to partially imputation," *IEEE Intell. Informat. Bulletin*, vol. 9, no. 1, pp. 32–38, 2008.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[14] J. R. Quinlan, *C4.5:Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[15] S. Zhang, Z. Qin, C. X. Ling, and S. Sheng, "Missing is useful: Missing values in cost-sensitive decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1689–1693, Dec. 2005.

[16] P. Mozharovskyi, J. Josse, and F. Husson, "Nonparametric imputation by data depth," *J. Amer. Statist. Assoc.*, pp. 1–24, 2019.

[17] S. Zhang, "Shell-neighbor method and its application in missing data imputation," *Appl. Intell.*, vol. 35, no. 1, pp. 123–133, 2011.

[18] P. D. Turney, "Types of cost in inductive concept learning," 2002, *arXiv preprint cs/0212034*.

[19] C. X. Ling, Q. Yang, J. Wang, and S. Zhang, "Decision trees with minimal costs," in *Proc. Int. Conf. Mach. Learn.*, 2004, Art. no. 69.

[20] X. Zhu, S. Zhang, J. Zhang, and C. Zhang, "Cost-sensitive imputing missing values with ordering," in *Proc. AAAI Conf. Artif. Intell.*, 2007, pp. 1922–1923.

[21] C. Lei and X. Zhu, "Unsupervised feature selection via local structure learning and sparse learning," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29 605–29 622, 2018.

[22] X. Zhu, X. Li, S. Zhang, Z. Xu, L. Yu, and C. Wang, "Graph PCA hashing for similarity search," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2033–2044, Sep. 2017.

[23] S. Zhang, "Multiple-scale cost sensitive decision tree learning," *World Wide Web*, vol. 21, no. 6, pp. 1787–1800, 2018.

[24] U. Dick, P. Haider, and T. Scheffer, "Learning from incomplete data with infinite imputations," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 232–239.

[25] M. Huisman, "Missing data in social network," in *Sunbelt,XXVII*, 2007.

[26] W. Zheng, X. Zhu, G. Wen, Y. Zhu, H. Yu, and J. Gan, "Unsupervised feature selection by self-paced learning regularization," *Pattern Recognit. Lett.*, 2018. [Online]. Available: https://doi.org/10.1016/j.patrec.2018.06.029

[27] W. Zheng, X. Zhu, Y. Zhu, R. Hu, and C. Lei, "Dynamic graph learning for spectral feature selection," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29 739–29 755, 2018.

[28] F. Lobato *et al.*, "Multi-objective genetic algorithm for missing data imputation," *Pattern Recognit. Lett.*, vol. 68, pp. 126–131, 2015.

[29] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri, "InfoGather: Entity augmentation and attribute discovery by holistic matching with web tables," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 97–108.

[30] W. Zhang, "Association-based multiple imputation in multivariate datasets: A summary," in *Proc. IEEE Int. Conf. Data Eng.*, 2000, pp. 310–310.

[31] C.-Y. J. Peng and J. Zhu, "Comparison of two approaches for handling missing covariates in logistic regression," *Educ. Psychol. Meas.*, vol. 68, no. 1, pp. 58–77, 2008.

[32] L. Zhao, Z. Chen, Z. Yang, Y. Hu, and M. S. Obaidat, "Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1610–1620, Jun. 2018.

[33] J. Junior, M. D. C. Nicoletti, and L. Zhao, "An embedded imputation method via attribute-based decision graphs," *Expert Syst. Appl.*, vol. 57, pp. 159–177, 2016.

[34] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: Wiley, 2002.

[35] J. Tian, B. Yu, D. Yu, and S. Ma, "Missing data analyses: A hybrid multiple imputation algorithm using gray system theory and entropy based on clustering," *Appl. Intell.*, vol. 40, no. 2, pp. 376–388, 2014.

[36] X. Liao, H. Li, and L. Carin, "Quadratically gated mixture of experts for incomplete data classification," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 553–560.

[37] T. M. Pham, J. R. Carpenter, T. P. Morris, A. M. Wood, and I. Petersen, "Population-calibrated multiple imputation for a binary/categorical covariate in categorical regression models," *Statist. Med.*, vol. 38, no. 5, pp. 792–808, 2019.

[38] N. Metternich, F. Hollenbach, I. Bojinov, S. Minhas, M. Ward, and A. Volfovsky, "Multiple imputation using Gaussian copulas," *Sociol. Methods Res.*, 2019.

[39] S. Rafsunjani, R. S. Safa, A. Al Imran, M. S. Rahim, and D. Nandi, "An empirical comparison of missing value imputation techniques on APS failure prediction," 2019.

[40] V. Nassiri, G. Molenberghs, G. Verbeke, and J. Barbosa-Breda, "Iterative multiple imputation: A framework to determine the number of imputed datasets," *Amer. Statistician*, pp. 1–17, 2019.

[41] S. Kang, K. Koehler, and M. Larsen, "Partial FEFI for incomplete tables with covariates Iowa State University," *JSM*, pp. 1038–1047, 2007.

[42] I. Song, Y. Yang, J. Im, T. Tong, H. Ceylan, and I.-H. Cho, "Impacts of fractional hot-deck imputation on learning and prediction of engineering data," *IEEE Trans. Knowl. Data Eng.*, 2019.

[43] X. She and C. Wu, "Fully efficient joint fractional imputation for incomplete bivariate ordinal responses," *Statistica Sinica*, vol. 29, no. 1, pp. 409–430, 2019.

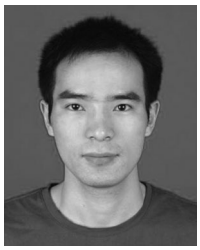
[44] M. Saar-Tsechansky, P. Melville, and F. Provost, "Active feature-value acquisition," *Manage. Sci.*, vol. 55, no. 4, pp. 664–684, 2009.

[45] S. Zhang, "Decision tree classifiers sensitive to heterogeneous costs," *J. Syst. Softw.*, vol. 85, no. 4, pp. 771–779, 2012.

[46] C. Conversano and R. Siciliano, "Incremental tree-based missing data imputation with lexicographic ordering," *J. Classification*, vol. 26, no. 3, pp. 361–379, 2009.

[47] E. R. Hruschka, E. R. Hruschka, and N. F. Ebecken, "Bayesian networks for imputation in classification problems," *J. Intell. Inf. Syst.*, vol. 29, no. 3, pp. 231–252, 2007.
[48] O. O. Lobo and M. Numao, "Ordered estimation of missing values," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 1999, pp. 499–503.
[49] R. K. Pearson, "The problem of disguised missing data," *ACM SIGKDD Explorations Newslett.*, vol. 8, no. 1, pp. 83–92, 2006.
[50] M. Núñez, "The use of background knowledge in decision tree induction," *Mach. Learn.*, vol. 6, no. 3, pp. 231–250, 1991.
[51] M. M. Hassan, A. F. Atiya, N. El-Gayar, and R. El-Fouly, "Regression in the presence missing data using ensemble methods," in *Proc. Int. Joint Conf. Artif. Intell.*
[52] C.-M. Teng, "Correcting noisy data," in *Proc. Int. Conf. Mach. Learn.*, 1999, pp. 239–248.
[53] B. Zadrozny, "One-benefit learning: Cost-sensitive learning with restricted cost information," in *Proc. 1st Int. Workshop Utility-Based Data Mining*, 2005, pp. 53–58.
[54] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2100–2110, Oct. 2017.
[55] X. Zhu and X. Wu, "Cost-constrained data acquisition for intelligent data preparation," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1542–1556, Nov. 2005.
[56] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1263–1275, Jun. 2017.
[57] X. Zhou *et al.*, "Graph convolutional network hashing," 2018, doi: 10.1109/TCYB.2018.2883970.
[58] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2894139.
[59] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.
[60] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, 2017, Art. no. 43.
[61] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

Xiaofeng Zhu received the PhD degree in computer science from the University of Queensland, Australia. His research interests include machine learning and image analysis. Specifically, he is focusing on mining useful knowledge or information from big multimedia data and medical imaging data.



Jianye Yang received the BSc and MSc degrees in computer science from Xidian University, and the PhD degree in computer science from the University of New South Wales. He is an associate professor with the College of Computer Science and Electronic Engineering, Hunan University. His research interests include large scale similarity search and graph data analysis.

Chengyuan Zhang received the BS degree from Sun-Yat Sen University, in 2008, and the master's and PhD degrees in computer science from the University of New South Wales, in 2011 and 2015, respectively. Currently, he is a lecturer with the School of Computer Science and Engineering, Central South University, China. His main research interests include information retrieval, query processing on spatial data, and multimedia data.

Shichao Zhang received the PhD degree in computer science from Deakin University, Australia. He is currently a China National distinguished professor with the School of Computer Science and Technology, Central South University, China. His research interests include information quality and pattern discovery. He has published about 80 international journal papers and more than 80 international conference papers. He has won 16 national-class Grants. He served/is serving as an associate editor of the *ACM Transactions on Knowledge Discovery from Data*, the *IEEE Transactions on Knowledge and Data Engineering*, *Knowledge and Information Systems*, and the *IEEE Intelligent Informatics Bulletin*. He is a senior member of the IEEE and IEEE Computer Society and a member of the ACM.