

Received March 27, 2020, accepted April 26, 2020, date of publication May 6, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991669

Takagi-Sugeno Modeling of Incomplete Data for Missing Value Imputation With the Use of Alternate Learning

XIAOCHEN LAI^{1,3}, LIYONG ZHANG², AND XIN LIU⁴

¹School of Software, Dalian University of Technology, Dalian 116600, China

²School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China

³Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China

⁴School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Xiaochen Lai (laixiaochen@dlut.edu.cn)


This work was supported by the National Key R&D Program of China under Grant 2018YFB1700200.

ABSTRACT Missing values often occur in real-world datasets, which undermines the data integrity and reduces the reliability of data mining. In this paper, a method of Takagi-Sugeno (TS) fuzzy modeling for incomplete data is proposed and utilized to estimate missing values. Considering the difference of attribute relationship within different clusters, this method performs regression analysis on the subsets obtained by fuzzy clustering and constructs the global model with the weighted sum of regression models, which describes the relationship between attributes more precisely on the basis of traditional regression imputation. Meanwhile, focusing on the problem of incomplete model input caused by missing values, we propose an alternate learning strategy to train model parameters and imputations, which treats missing values as variables to drive the advance of incomplete data modeling and updates imputations with the adjustment of model parameters. Through the alternate learning strategy, not only the problem of incomplete model input is well solved, but also the accuracy of the model and the performance of imputation are improved together in a collaborative way. Experimental results on several UCI datasets with different missing ratios and missing data mechanisms demonstrate the effectiveness of the proposed method and strategy.

INDEX TERMS Incomplete data, missing value imputation, clustering-based modeling, alternate learning.

I. INTRODUCTION

In real-world datasets, the problem of missing values is almost inevitable, which is often caused by many factors such as equipment failure, limitation of data collection and human fault in storage [1], [2]. These missing values undermine data integrity and have become a major obstacle in data mining. Therefore, how to deal with missing values is a crucial issue in the analysis of incomplete data. Generally speaking, the simplest way is to discard incomplete records directly and analyze with the remaining complete records, but it only works for datasets with a small number of incomplete records [3]. In practice, the incomplete records usually cannot be overlooked, because if they are discarded directly, it may result in misleading conclusions due to the loss of information. By contrast, missing value imputation is an effective way, which can further improve data quality.

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Jahan Khan .

Missing value imputation aims to replace missing values with reasonable ones derived from present data, and many imputation methods have been proposed to make the results of data mining more effective and valuable [4], [5]. In the past few decades, commonly utilized methods include mean imputation, median imputation, hot-deck imputation, k nearest neighbor (kNN) imputation, class center-based imputation, exception maximization imputation (EMI) and regression-based imputation. The first two imputation methods take mean values and median values of present data in each incomplete attribute as the replacements. Hot-deck imputation method selects a complete record nearest to the current incomplete one and regards its corresponding attribute values as expected imputations [6]. Similar to the hot-deck imputation, the kNN imputation method performs replacement with mean values of corresponding attributes in k nearest neighbors [7], [8]. To improve the imputation accuracy, Song *et al.* took similarity neighbors into consideration when searching for the nearest neighbors [9]. The class

center-based imputation method defines a threshold by the distances between each class center and present values for missing value imputation [10]. EMI is a parametric model-based imputation method, which is realized through the iteration of E-step and M-step. E-step estimates conditional expectations of missing values and takes them as imputations. M-step calculates parameters that maximize the expectations of log-likelihood function based on the imputed dataset [11]. Considering that instances in the same cluster are very similar to each other, Rahman *et al.* first made a fuzzy clustering of the dataset for finding similar records, and then applied a fuzzy EM algorithm to impute the missing values [12]. These methods mentioned above are widely adopted, but sometimes with limited imputation performance due to ignoring relationships between attributes [13].

Taking attribute relationships into consideration, the regression-based imputation method establishes several regression models with each missing attribute as the output variable, which has received wide attention [14]–[17]. For example, Kim *et al.* proposed local least squares imputation to estimate missing values in the gene expression data, where the target gene with missing values is represented as a linear combination of similar genes chosen based on similarity measures [18]. Cheng *et al.* incorporated the clustering idea into the framework of local least squares imputation for characterizing the gene similarity [19]. Inspired by the Bayesian inference method, Shah *et al.* developed a Bayesian regression model called BayesMetab which systematically estimates missing values based on a Markov chain Monte Carlo (MCMC) algorithm [20]. Stein *et al.* first pre-imputed missing values of continuous attributes by the mean values and replaced missing values of discrete variables with the most frequent values, followed by predicting missing values through a series of regression models with the class label of each sample as an extra predictor variable [21]. Zhang *et al.* combined Bayesian regression and EM algorithm to construct the predictive model. Moreover, the experiment results showed that training the model with present values and missing values has a better prediction performance compared with ignoring the missing values [22]. Aydilek *et al.* combined fuzzy c-means clustering with support vector regression and a genetic algorithm to estimate missing values. In this method, support vector regression model is trained by complete records before being utilized for imputation, which attempts to make the output values more approximate to their corresponding inputs [23]. Sefidian *et al.* imputed missing values by a novel grey-based fuzzy c-means, mutual information-based feature selection, and regression model, which achieved good performance of imputation through the construction of a specialized regression model for each cluster [24].

From the above, we can conclude that analysis aiming at each subset rather than the whole dataset is more capable in describing the relationships between attributes during the regression modeling of incomplete data, thus obtaining a better performance of imputation. Hence, the partition-based

models have been widely employed in data mining, especially rule-based fuzzy models. As the most concerned rule-based fuzzy model, Takagi-Sugeno (TS) model performs regression analysis on the premise of fuzzy partition and establishes linear regression models aiming at the relationships in each subset [25]–[27]. Due to this feature, TS model can be utilized as a universal approximator to handle nonlinear problems, and has an outstanding performance in working out the relationships than traditional regression model [28], [29]. On the other hand, pre-imputation of missing values is commonly adopted to deal with the problem of incomplete input during modeling. Whereas, the perturbation caused by different pre-imputed values can easily lead to the variations of model parameters, which has a great influence on the model accuracy. Hence, in the process of incomplete data modeling, the way to handle the incomplete model input deserves great attention.

In this paper, we take TS model for modeling incomplete data and realize missing value imputation in collaboration with the dynamic modeling. Aiming to describe the relationship between attributes more precisely, the method first divides the dataset into several subsets and performs regression analysis for each subset, followed by constructing the global model with the weighted sum of regression models, which improves the model accuracy on the basis of traditional regression modeling. Meanwhile, owing to the problem of incomplete model input caused by missing values, an alternative learning strategy used for training the parameters of incomplete data-based model together with imputations is presented. In this strategy, missing values are treated as the variable to drive the advance of incomplete data modeling, and imputations are promoted to update dynamically together with the adjustment of model parameters. Therefore, a collaborative improvement of model accuracy and imputation performance can be realized additionally as the problem of incomplete model input is resolved.

The rest of this paper is organized as follows. Section II introduces the basic structure of TS model. Section III describes the proposed method of incomplete data modeling and missing value imputation. Meanwhile, the alternate learning strategy that treats missing values as variables is carried out to train model parameters together with imputations. Section IV demonstrates the effectiveness of the proposed method by several UCI datasets with different missing ratios and missing data mechanisms. Section V concludes the paper.

II. TAKAGI-SUGENO FUZZY MODEL

The Takagi-Sugeno fuzzy model was proposed by Takagi and Sugeno in 1985, whose basic idea is to divide the nonlinear problem into several linear sub-problems and describe them separately with “IF-THEN” rules [30]. It obtains the premise parts by fuzzy partition, and then linear regression models are established as corresponding consequence parts to describe the relationships between input-output variables. Given a dataset composed of N records in s -dimensional real space, i.e. $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and $\mathbf{x}_k = [x_{k1}, \dots, x_{ks}]$ for

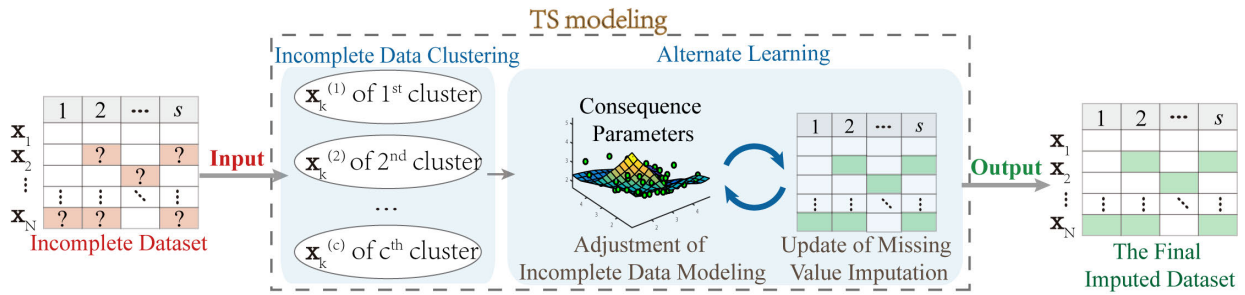


FIGURE 1. Alternate learning-based TS modeling and missing value imputation.

$k = 1, \dots, N$, the fuzzy rule of TS model can be described in (1):

$$\begin{aligned}
 &R^{(i)} : \\
 &\text{IF } x_{k1} \text{ is } A_1^{(i)}, \text{ and } \dots, \\
 &\quad \text{and } x_{k(j-1)} \text{ is } A_{j-1}^{(i)}, \text{ and } x_{k(j+1)} \text{ is } A_{j+1}^{(i)}, \text{ and } \dots, \\
 &\quad \text{and } x_{ks} \text{ is } A_s^{(i)}, \\
 &\text{THEN} \\
 &\quad \hat{x}_{kj}^{(i)} = \theta_0^{(i)} + \theta_1^{(i)}x_{k1} + \dots + \theta_{j-1}^{(i)}x_{k(j-1)} \\
 &\quad + \theta_{j+1}^{(i)}x_{k(j+1)} + \dots + \theta_s^{(i)}x_{ks}, \tag{1}
 \end{aligned}$$

where $R^{(i)}$ is the i th fuzzy rule for $i = 1, \dots, c$ and c is the number of fuzzy rules; $x_{k1}, \dots, x_{k(j-1)}, x_{k(j+1)}, \dots, x_{ks}$ are the input variables of $R^{(i)}$, $A_1^{(i)}, \dots, A_{j-1}^{(i)}, A_{j+1}^{(i)}, \dots, A_s^{(i)}$ represent their corresponding fuzzy sets, also known as premise parameters; $\hat{x}_{kj}^{(i)}$ represents the output of $R^{(i)}$, and $\theta_0^{(i)}, \theta_1^{(i)}, \dots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i)}, \dots, \theta_s^{(i)}$ are consequence parameters.

The global output of TS model is the weighted summation of each rule output, as shown in (2):

$$\hat{x}_{kj} = \sum_{i=1}^c \tilde{\beta}_k^{(i)} \hat{x}_{kj}^{(i)}, \quad \tilde{\beta}_k^{(i)} = \frac{\beta_k^{(i)}}{\sum_{i=1}^c \beta_k^{(i)}}, \tag{2}$$

$$\begin{aligned}
 \beta_k^{(i)} = &\min(u_{A_1^{(i)}}(x_{k1}), \dots, u_{A_{j-1}^{(i)}}(x_{k(j-1)}), \\
 &u_{A_{j+1}^{(i)}}(x_{k(j+1)}), \dots, u_{A_s^{(i)}}(x_{ks})), \tag{3}
 \end{aligned}$$

where $u_{A_s^{(i)}}(x_{ks})$ represents the membership degree of x_{ks} belonging to $A_s^{(i)}$.

III. TS MODELING, MISSING VALUE IMPUTATION, AND ALTERNATE LEARNING

Given a dataset, the attribute relationships among each record are diverse, which can be similar or different. Therefore, it is more appropriate to divide the dataset into several subsets and carry out regression analysis for each subset separately. To make a more precise analysis for incomplete data and thus obtain more reasonable imputations, a dynamic TS modeling-based method on the premise of fuzzy partition is proposed in this section. Meanwhile, focusing on the problem of incomplete model input caused by missing values, we present an

alternate learning strategy that regards those missing values as variables to promote the training of incomplete model. Through this strategy, the model parameters can be adjusted with the feedback of those updated variables, and those variables can also be more adapted to the model with parameter adjustment. The framework of the proposed method is shown in Fig. 1, where $x_k^{(i)}$ represents the k th sample of the i th cluster.

As shown in Fig. 1, given an incomplete dataset, the method first divides it into c subsets by a fuzzy clustering algorithm, and thus realizes premise parameter identification along with the fuzzy partition. After premise parameters of the incomplete data model are determined, consequence parameter identification can be worked out through the training of alternate learning strategy, and missing value imputation can also be achieved accompanying with the training. In this process, consequence parameters are adjusted with the updating of imputations, and imputations are updated in turn with the adjustment of consequence parameters. Through repeated adjustments and updates, consequence parameters and imputations are learned alternately and tend to be practical, which means that the problem of incomplete model input caused by missing values is resolved effectively.

A. TS MODELING OF INCOMPLETE DATA

Similar to complete data-based TS modeling, the realization for incomplete data modeling can also be divided into premise parameter identification and consequence parameter identification. To make a more precise analysis, we add variable selection to regression modeling of each subset after fuzzy partition. Considering the occurrence of missing values, premise parameters are obtained through fuzzy C-means clustering with partial distance strategy (FCM-PDS) and consequence parameters are estimated using the least square method by treating missing values as variables. Besides, a stepwise regression algorithm is utilized for variable selection to describe the regression relationships between attributes in each subset.

1) PREMISE PARAMETER IDENTIFICATION

FCM-PDS [31] algorithm divides the incomplete dataset into c subsets by minimizing the objective function:

$$J = \sum_{k=1}^N \sum_{i=1}^c [u_{A^{(i)}}(\mathbf{x}_k)]^m d_{ik}^2, \tag{4}$$

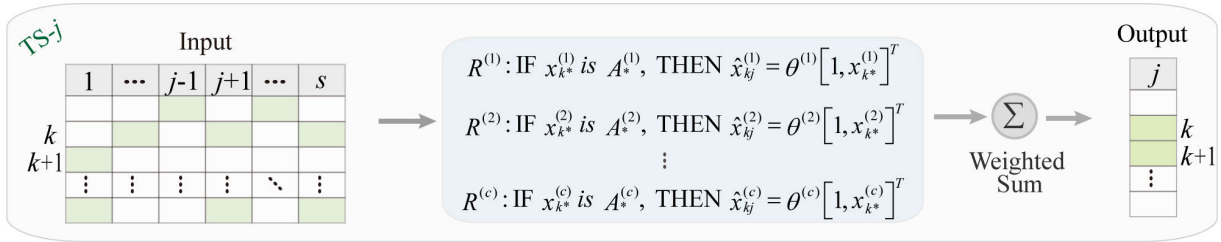


FIGURE 2. Missing value imputation for the j th incomplete attribute in one process.

where $u_{A^{(i)}}(\mathbf{x}_k)$ represents the membership degree of \mathbf{x}_k belonging to fuzzy set $A^{(i)}$, $\sum_{i=1}^c u_{A^{(i)}}(\mathbf{x}_k) = 1$; m represents the fuzzification parameter, $m \in (1, \infty)$; d_{ik} represents partial distance from the k th record to the i th cluster center, which is calculated by

$$d_{ik} = \sqrt{\frac{s}{I_k} \sum_{j=1}^s (x_{kj} - v_{ij})^2 I_{kj}}, \quad (5)$$

where $I_{kj} = \begin{cases} 1, & \text{if } x_{kj} \in X_P \\ 0, & \text{if } x_{kj} \in X_M \end{cases}$, for $j = 1, \dots, s$ and $k = 1, \dots, N$ and $I_k = \sum_{j=1}^s I_{kj}$, in which $X_P = \{x_{kj} | 1 \leq k \leq N, 1 \leq j \leq s, \text{ the value for } x_{kj} \text{ is present}\}$, $X_M = \{x_{kj} | 1 \leq k \leq N, 1 \leq j \leq s, \text{ the value for } x_{kj} \text{ is missing}\}$ represent the set of present values and the set of missing values; v_{ij} represents the j th attribute of the i th cluster center. After the fuzzy partition, premise parameters composed of the membership degree $u_{A_j^{(i)}}(x_{kj})$ can be obtained by projecting $u_{A^{(i)}}(\mathbf{x}_k)$ onto each axis of the input variable [32]. In this paper, we use the Gaussian function, given by

$$u_{A_j^{(i)}}(x_{kj}) = \exp \left\{ -\frac{(x_{kj} - a_{ij})^2}{2\sigma_{ij}^2} \right\}, \quad (6)$$

where a_{ij} represents the center and σ_{ij} represents the standard deviation,

$$a_{ij} = \frac{\sum_{k=1}^N u_{ik} x_{kj}}{\sum_{k=1}^N u_{ik}}, \sigma_{ij} = \frac{2 \sum_{k=1}^N u_{ik} (x_{kj} - a_{ij})^2}{\sum_{k=1}^N u_{ik}}. \quad (7)$$

2) INPUT VARIABLE SELECTION

Stepwise regression algorithm is designed to introduce the variables with significant impact on the output into regression model one by one, which can make the established model contain only all significant variables [33]–[35]. Therefore, we use it for selecting input variables of each regression model in fuzzy rules, so as to improve the model precision while reducing complexity.

3) CONSEQUENCE PARAMETER IDENTIFICATION

The least square method has been widely utilized in the parameter calculation of nonlinear regression model, as it obtains the optimal fitting function by minimizing the sum of the squared errors. However, the method fails to estimate when the dataset occurs missing values, thereby we

propose to treat missing values as variables for estimation. Subsequently, those estimated parameters are trained to be more appropriate through an alternate learning strategy. The detailed realization steps are shown in Section III.B.

B. TS MODELING-BASED MISSING VALUE IMPUTATION

1) THE OVERALL STRUCTURE FOR IMPUTATION

In this paper, several incomplete data-based TS models with multiple-input-single-output (MISO) structure are established considering that not only one attribute suffers from missing values in general and a multiple-input-multiple-output (MIMO) problem can be divided into several MISO problems. In each model, one incomplete attribute is taken as output variable and the other attributes are selected as input variables for modeling. For example, taking the j th incomplete attribute as output variable, the process of modeling-based imputation is shown in Fig. 2, where $x_{k*}^{(i)} (i = 1, \dots, c)$ represents the vector of input variables in $R^{(i)}$, $A_*^{(i)}$ represents the set of corresponding premise parameters, and $\theta^{(i)}$ represents the vector of consequence parameters.

As depicted in Fig. 2, establishing an incomplete data model by regarding the j th incomplete attribute as the output variable, where input variables are selected from the other attributes by stepwise regression algorithm. To obtain the imputations represented by corresponding model outputs, the method first assigns a value to each variable in the input vector and then inputs this reconstructed set to the established model for calculation.

2) ALTERNATE LEARNING OF MODEL PARAMETERS AND IMPUTATIONS

After identifying the model parameters and obtaining the input variables for each TS model, we can calculate the output values of these model through these estimated parameters. However, since the data integrity is undermined by missing values, the incomplete data-based TS model usually cannot describe the regression relationships between attributes in each subset well, and then the output values derived from the model are not so appropriate as imputations. To improve the precision of incomplete data model, and thereby enhance the appropriateness of imputations derived from the model, an alternate learning strategy is proposed in this section for training the parameters together with imputations.

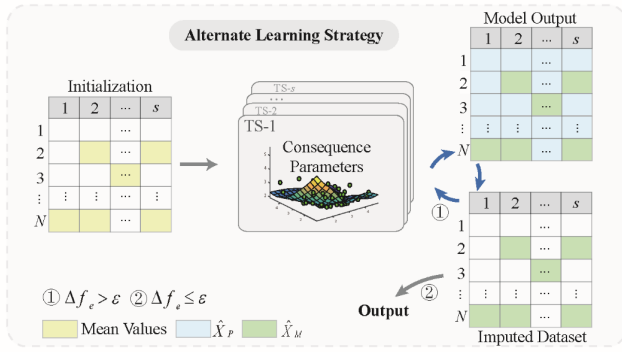


FIGURE 3. The flowchart of the alternate learning strategy.

The strategy is shown in Fig. 3, where \hat{X}_P represents the set of reconstructed values corresponding to present ones in X_P , \hat{X}_M represents that of imputations corresponding to missing values in X_M , ϵ represents the threshold for terminating iterations, and $\Delta f_\epsilon = |f_\epsilon^{(l)} - f_\epsilon^{(l-1)}|$ represents the change of RMSE values between two successive iterations, in which $f_\epsilon^{(l)}$ represents the RMSE value in current iteration and $f_\epsilon^{(l-1)}$ represents that in previous one. The f_ϵ in each iteration is calculated using \hat{X}_P and X_P by (8):

$$f_\epsilon = \sqrt{\frac{1}{|X_P|} \sum_{x_{kj} \in X_P} (x_{kj} - \hat{x}_{kj})^2}, \quad (8)$$

where $|X_P|$ represents the number of values in X_P .

As shown in Fig. 3, by treating missing values as variables, the model parameters and model output values under these parameters can be estimated easily by the method in Section III.A. However, instead of finishing the imputation task after replacing missing values with corresponding estimated values, the strategy is able to decide whether this reconstructed dataset should be output according to the change of reconstruction error f_ϵ . If the error changes within a limited range compared with the previous one, i.e. $\Delta f_\epsilon = |f_\epsilon^{(l)} - f_\epsilon^{(l-1)}| \leq \epsilon$, it means that the fitting ability of this incomplete data model will be no longer changed, and thereby the imputation task can be accomplished with the output of this reconstructed dataset. Otherwise, the dataset should give feedback to the model for the parameter adjustment, so as to update the fitting model and its corresponding output values. In turn, new imputations and reconstruction error can be obtained in response to the model adjustment. Through the alternate learning of model parameters and imputations, the reconstruction error tends to be stable and the imputation task can be finally accomplished with the output of the updated dataset.

In summary, given an incomplete dataset, the alternate learning strategy can be realized through the following steps.

Step1: initialize missing values in X_M ;

Step2: estimate model parameters based on the reconstructed dataset;

Step3: update imputations according to the estimated parameters;

Step4: evaluate the change of reconstruction error. If it is greater than the given threshold, return to Step2, and the current incomplete data model needs to be adjusted according to the updated dataset. Otherwise, end the training and output the updated dataset.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASETS

In this subsection, we select 12 complete benchmark datasets from the UCI Machine Learning Repository [?] for experiments. Their brief descriptions are shown in Table 1. The UCI database is a repository released by the University of California, Irvine. It currently maintains 497 data sets as a service to the machine learning community, which has become a popular database for researchers.

TABLE 1. The brief description of each benchmark dataset.

Dataset	# Record	#Attribute	Dataset	#Record	#Attribute
Iris	150	4	Glass	214	9
Seeds	210	7	Istanbul	536	9
Wireless	2000	7	ILPD	583	9
Ecoli	336	8	Wine	178	13
Abalone	4177	8	Segment	2100	16
Forest fires	503	6	Dow	720	13

To observe the imputation performance of the proposed method under different missing scales, we set 10 missing ratios uniformly in the range of 5% to 50% for each benchmark dataset. Under the constraint of missing ratios, some attribute values are deleted from each benchmark dataset based on Missing Completely At Random (MCAR) and Missing Not At Random (MNAR) mechanisms while keeping the dimension of attributes and the number of records unchanged. In the MCAR mechanism, values are removed uniformly at random, and in the MNAR mechanism, only values higher than the median of the attribute can be removed randomly. These two mechanisms are performed in turn to produce missing data, that is, incomplete datasets under the missing ratios 5%, 15%, 25%, 35% and 45% are generated based on MCAR, and those under missing ratios 10%, 20%, 30%, 40% and 50% are produced based on MNAR. Hence, there are 10 combinations (2 mechanisms of missingness and 5 missing ratios for each mechanism) of missing types. Moreover, 10 incomplete datasets are generated randomly under one combination for each benchmark dataset, which means that a total of 1200 ($12 \times 10 \times 5 \times 2$) incomplete datasets are utilized for experiments.

2) EVALUATION CRITERION

In this paper, we take root mean square error (RMSE) which is calculated by

$$\text{RMSE} = \sqrt{\frac{1}{|\hat{X}_M|} \sum_{\hat{x}_{kj} \in \hat{X}_M} (r_{kj} - \hat{x}_{kj})^2}. \quad (9)$$

and mean absolute percentage error (MAPE) defined by

$$\text{MAPE} = \frac{1}{|\hat{X}_M|} \sum_{\hat{x}_{kj} \in \hat{X}_M} \left| \frac{r_{kj} - \hat{x}_{kj}}{r_{kj}} \right| \quad (10)$$

to evaluate the performance of imputation, where $\hat{x}_{kj} \in \hat{X}_M$ represents the imputation for a missing value x_{kj} , and r_{kj} represents its corresponding actual value.

3) COMPARISON METHODS

In order to verify the effectiveness of clustering-based TS model in missing value imputation and the feasibility of alternate learning strategy in modeling with incomplete data, the following nine comparison methods are designed to carry out experiments.

- (1) The k nearest neighbor imputation (KNNI). Select k nearest neighbors for each incomplete record, then impute missing values with the mean values of corresponding attributes in nearest neighbors [7].
- (2) Exception maximization imputation (EMI). Take the iteration of E-step and M-step to estimate missing values and calculate parameters [11].
- (3) Fuzzy exception maximization imputation (FEMI). Make a fuzzy clustering of the dataset and then perform the EMI algorithm in each cluster [12].
- (4) Regression model-based imputation (REGI). Establish a regression model with variable selection based on complete records, and estimate missing values depending on this complete data-based model.
- (5) TS model-based imputation (TSI). Establish a TS model with variable selection based on complete records, and estimate missing values depending on this complete data-based model.
- (6) Regression model-based imputation with full utilization of present values (REGIf). Establish a regression model with variable selection based on all records instead of complete records, and estimate missing values depending on this incomplete data-based model. In this method, the model parameters are estimated based on a reconstructed dataset where missing values are pre-imputed with mean values of corresponding attributes.
- (7) TS model-based imputation with full utilization of present values (TSIf). Establish a TS model with variable selection based on all records instead of complete records, and estimate missing values depending on this incomplete data-based model. In this method, the premise parameters are obtained by FCM-PDS clustering algorithm together with Gaussian projection, while consequence parameters are estimated based on a reconstructed dataset where missing values in each subset are pre-imputed with mean values of corresponding attributes;
- (8) Regression modeling-based imputation trained by alternate learning strategy (REGIf-AL). On the basis of REGIf, model parameters and imputations are

learned alternately and updated dynamically until the reconstruction error calculated from present values tends to be stable;

- (9) TS modeling-based imputation trained by alternate learning strategy (TSIf-AL). On the basis of TSIf, consequence parameters and imputations are learned alternately and updated dynamically until the reconstruction error calculated from present values tends to be stable.

In TSI, TSIf and TSIf-AL methods, fuzzy rules are generated separately for each dataset by means of TS modeling. Generally, the generation of fuzzy rules is equivalent to the construction of TS model, which contains three steps: premise parameter identification, input variable selection, and consequence parameter identification. For each incomplete dataset, these three steps are conducted to generate fuzzy rules automatically.

B. EXPERIMENTAL RESULTS

In order to make the conclusion more reliable, the average of imputation performance for all comparison methods obtained from the ten incomplete datasets with the same missing ratio, benchmark dataset and missingness mechanism are taken as a set of results. In other words, a benchmark dataset corresponds to only one group of results under the constraint of each combination of missing ratios and mechanisms, as shown in Tables 2 to 13. Moreover, entries in boldface are obviously better than all the other entries in the same column. After obtaining the experimental results for all the methods, we adopt t-test with significance level $p = 0.05$ to determine whether the two results in the same column of the table are significantly different from each other. The minimum result will be underlined only when it is significantly different from all the other results. Based on the distribution of bolded results, we can evaluate the imputation performance from the perspective of statistical significance tests.

C. RESULT ANALYSIS

By observing the experimental results in Tables 2 to 13, it is clear that TSIf-AL significantly has the most optimal results, which indicates that the imputation performance of TSIf-AL is better than the rest methods. Furthermore, we can also find that the RMSE values and MAPE values obtained from REGI and TSI are larger than those obtained from the other regression-based methods. This phenomenon indicates that making full use of present values for incomplete data modeling can enable the relationships between attributes to be described more effectively, and thus the imputation performance can also be enhanced correspondingly. In the following analysis, we discuss the superiority of clustering-based modeling compared to overall modeling, the advantage of utilizing alternate learning strategy by those methods making full use of present values, and the comparison between TSIf-AL and non-regression-based methods.

TABLE 2. The RMSE and MAPE values obtained from each imputation method for Iris.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.354	0.387	0.471	0.558	0.700	0.416	0.563	0.734	1.002	1.169	12.17	15.06	16.27	17.54	38.44	7.42	10.91	12.55	18.74	23.29
EMI	0.383	0.393	0.478	0.532	0.550	0.424	0.596	0.752	0.861	1.144	11.44	16.61	16.04	18.36	23.29	7.70	10.65	12.90	14.78	20.79
FEMI	0.337	0.393	0.471	0.573	0.599	0.432	0.528	0.675	0.740	0.989	11.79	15.77	14.94	17.49	22.26	7.49	10.35	11.74	13.74	18.77
REGI	0.438	0.526	0.617	0.717	0.764	0.456	0.613	0.799	0.969	1.183	13.77	26.40	30.24	39.12	45.23	8.98	11.56	15.09	20.26	25.90
TSI	0.376	0.410	0.485	0.521	0.619	0.433	0.578	0.701	0.761	0.961	12.43	18.29	17.36	19.18	23.93	9.53	10.87	13.14	14.67	17.72
REGIf	0.400	0.492	0.581	0.696	0.776	0.416	0.565	0.788	0.995	1.217	17.38	25.21	28.34	34.23	36.96	8.40	11.13	14.78	18.53	23.18
TSIf	0.330	0.394	0.470	0.525	0.622	0.497	0.590	0.685	0.741	0.946	11.29	13.94	14.72	16.90	23.02	8.28	10.98	12.35	13.71	16.73
REGIf-AL	0.431	0.443	0.516	0.570	0.655	0.485	0.532	0.661	0.841	1.048	16.29	19.00	20.62	23.51	27.78	8.18	10.12	12.01	15.44	19.75
TSIf-AL	0.317	0.368	0.450	0.490	0.598	0.396	0.523	0.625	0.694	0.892	10.66	14.18	14.08	15.66	21.14	7.99	9.79	11.03	12.55	15.82

TABLE 3. The RMSE and MAPE values obtained from each imputation method for Seeds.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.516	0.605	0.716	1.075	11.46	0.697	1.342	1.992	2.179	2.407	6.90	8.11	8.66	11.10	14.82	5.55	9.49	13.87	15.99	19.61
EMI	0.442	0.506	0.762	1.632	2.066	0.573	0.856	2.145	3.909	3.066	5.54	6.43	8.20	16.04	21.35	4.77	6.58	16.08	31.52	24.79
FEMI	0.436	0.518	0.729	0.918	1.387	0.524	0.609	1.500	2.298	2.484	5.55	6.44	7.83	14.51	14.51	3.85	4.40	6.82	10.47	16.60
REGI	0.810	0.816	0.832	0.877	0.928	0.907	1.051	1.540	1.880	2.346	8.85	9.25	9.61	10.05	14.60	6.81	7.62	11.09	14.33	19.06
TSI	0.689	0.729	0.804	0.935	1.455	0.835	1.103	1.218	1.476	2.036	6.85	8.41	9.32	10.19	9.87	5.84	7.89	8.67	11.04	16.29
REGIf	0.655	0.710	0.783	0.856	0.888	0.860	1.021	1.416	1.815	2.319	7.44	8.62	9.16	9.34	10.70	6.49	7.25	10.58	14.28	18.39
TSIf	0.552	0.654	0.748	0.824	0.960	0.771	1.068	1.190	1.440	2.028	6.83	8.41	8.99	9.92	9.67	5.75	7.86	8.62	11.04	16.10
REGIf-AL	0.608	0.707	0.770	0.815	0.879	0.836	0.982	1.180	1.520	2.153	7.05	8.37	8.70	9.01	9.04	6.32	6.92	7.11	11.39	17.29
TSIf-AL	0.499	0.558	0.612	0.719	0.851	0.565	0.738	1.008	1.319	1.915	6.04	7.00	7.35	7.98	9.87	4.17	5.11	8.21	9.75	14.83

TABLE 4. The RMSE and MAPE values obtained from each imputation method for Wireless.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	4.264	4.449	4.865	5.315	5.792	4.236	4.829	6.263	8.671	9.598	5.32	5.50	6.17	6.56	7.22	4.57	5.26	6.80	9.91	11.57
EMI	4.312	4.378	4.851	5.298	5.636	4.241	4.782	5.656	7.407	8.908	5.58	5.68	6.17	6.44	6.85	4.75	5.36	6.46	8.55	9.72
FEMI	4.166	4.367	4.831	5.194	5.956	4.518	5.551	6.408	8.118	9.277	5.40	5.56	6.19	6.41	7.32	4.91	6.04	7.08	8.98	10.92
REGI	4.529	4.757	5.121	5.529	6.151	4.437	5.501	6.874	8.603	9.419	5.75	6.13	6.79	7.41	8.02	4.98	6.33	8.32	10.84	12.19
TSI	5.440	5.211	5.204	5.442	6.044	5.317	5.597	5.991	7.549	8.886	7.06	6.71	6.70	6.89	7.50	6.18	6.44	7.09	9.32	10.54
REGIf	4.429	4.667	5.115	5.524	5.937	4.431	5.474	6.721	8.464	9.448	6.03	6.16	6.60	7.18	8.02	4.96	6.27	8.10	10.67	12.09
TSIf	4.129	4.359	4.896	5.243	5.632	4.143	5.003	5.742	7.716	8.825	5.35	5.65	6.22	6.75	7.19	4.59	5.61	6.56	8.63	10.26
REGIf-AL	4.356	4.656	5.099	5.513	5.753	4.873	5.404	6.171	7.289	8.254	5.72	6.06	6.59	7.10	7.35	5.32	5.98	6.93	8.34	9.52
TSIf-AL	4.089	4.233	4.559	4.928	5.339	4.019	4.472	4.929	6.493	8.148	5.27	5.44	5.81	6.06	6.57	4.27	4.94	5.49	7.15	9.10

TABLE 5. The RMSE and MAPE values obtained from each imputation method for Ecoli.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.125	0.140	0.151	0.158	0.166	0.157	0.185	0.235	0.301	0.319	22.14	25.31	29.41	31.29	32.80	17.18	20.60	27.54	38.32	41.95
EMI	0.123	0.136	0.151	0.175	0.194	0.149	0.182	0.232	0.292	0.313	24.29	27.22	33.29	33.47	35.42	16.65	20.71	27.96	37.54	40.80
FEMI	0.121	0.130	0.141	0.149	0.158	0.140	0.167	0.358	0.312	0.330	23.21	26.01	30.21	30.73	32.46	16.42	19.61	30.69	33.97	41.41
REGI	0.136	0.141	0.151	0.157	0.162	0.154	0.184	0.225	0.280	0.308	25.97	28.14	32.74	32.98	34.43	17.39	21.21	27.74	36.35	40.36
TSI	0.121	0.133	0.143	0.153	0.159	0.138	0.162	0.193	0.230	0.279	23.67	26.47	30.44	31.00	32.54	15.60	18.70	23.27	27.83	36.12
REGIf	0.123	0.134	0.147	0.154	0.156	0.150	0.177	0.214	0.251	0.292	23.87	27.65	32.27	32.25	33.97	17.00	20.51	26.26	32.31	38.84
TSIf	0.117	0.128	0.139	0.149	0.152	0.137	0.161	0.190	0.221	0.276	23.33	25.73	30.25	30.95	32.38	15.82	18.41	22.19	27.97	35.71
REGIf-AL	0.122	0.134	0.143	0.151	0.156	0.156	0.178	0.207	0.242	0.287	23.58	26.96	31.47	31.59	33.60	17.34	20.01	24.78	30.58	37.78
TSIf-AL	0.118	0.122	0.131	0.141	0.143	0.123	0.150	0.175	0.206	0.276	22.72	23.97	28.15	28.55	30.95	14.15	17.25	20.64	25.06	35.07

1) THE EFFECT OF CLUSTERING-BASED MODELING ON IMPUTATION PERFORMANCE
 As illustrated in Tables 2 to 13, there are 12 datasets, 10 combinations per dataset and 2 evaluation criteria, leading to

240 comparisons altogether. The RMSE values and the MAPE values obtained from TSIf are generally smaller than those obtained from REGIf. Among their 240 groups of comparisons, TSIf performs better in 198 out of 240 comparisons

TABLE 6. The RMSE and MAPE values obtained from each imputation method for Abalone.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.106	0.111	0.116	0.121	0.130	0.114	0.129	0.190	0.280	0.312	9.15	9.69	10.80	13.11	14.02	9.00	10.74	17.32	29.33	33.56
EMI	0.094	0.098	0.101	0.107	0.110	0.139	0.142	0.152	0.206	0.217	8.47	9.30	10.16	12.76	12.72	7.43	9.70	13.94	18.24	21.76
FEMI	0.093	0.097	0.101	0.106	0.112	0.122	0.129	0.178	0.295	0.281	7.77	8.32	8.95	10.75	11.12	8.70	9.97	13.35	22.11	25.45
REGI	0.140	0.151	0.159	0.165	0.173	0.174	0.193	0.215	0.276	0.318	25.83	35.17	45.36	56.03	54.87	13.07	21.10	18.63	30.37	36.24
TSI	0.097	0.125	0.146	0.161	0.170	0.147	0.188	0.232	0.199	0.193	27.28	29.76	31.93	35.79	36.96	12.06	17.67	23.61	24.87	24.73
REGIf	0.092	0.103	0.116	0.129	0.138	0.128	0.152	0.186	0.258	0.308	22.94	25.67	27.90	33.92	35.98	9.74	19.22	20.85	27.93	35.41
TSIf	0.090	0.102	0.114	0.128	0.136	0.123	0.156	0.203	0.210	0.216	12.80	19.96	26.81	34.51	35.25	9.83	13.65	19.81	21.58	20.77
REGIf-AL	0.090	0.099	0.105	0.115	0.118	0.123	0.134	0.155	0.170	0.231	22.90	22.95	20.79	23.47	22.16	9.26	10.51	12.49	15.30	22.35
TSIf-AL	0.089	0.092	0.091	0.097	0.099	0.106	0.113	0.135	0.169	0.183	12.06	14.60	16.23	19.07	19.39	8.61	9.12	11.23	14.97	17.35

TABLE 7. The RMSE and MAPE values obtained from each imputation method for Forest fires.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	69.23	80.84	101.36	83.39	92.24	73.11	106.7	119.3	187.2	245.8	43.24	45.30	39.52	44.35	66.19	15.73	19.62	24.41	38.18	48.63
EMI	80.13	105.91	103.70	89.39	109.81	83.58	115.9	136.9	183.8	214.0	37.50	41.98	43.59	46.37	45.38	14.39	27.14	30.42	33.68	37.05
FEMI	66.87	79.75	96.62	83.45	90.12	69.31	99.82	125.7	165.2	186.1	29.43	33.35	39.82	42.25	46.05	19.87	19.34	22.21	29.16	34.72
REGI	72.50	80.79	99.74	88.41	100.16	77.23	109.9	131.3	149.4	198.8	45.74	51.15	53.49	58.33	62.00	15.55	19.18	23.15	29.11	36.27
TSI	68.44	80.14	96.12	88.49	92.84	74.83	108.9	112.8	114.5	142.9	29.85	36.24	39.63	43.90	42.77	17.55	18.79	21.11	24.90	30.78
REGIf	68.02	80.99	102.30	87.28	101.58	71.30	101.5	105.0	119.0	143.2	51.59	50.94	55.25	61.41	58.13	15.06	16.72	20.49	25.21	30.94
TSIf	66.65	77.22	95.68	83.80	89.74	73.97	103.0	108.4	112.7	141.1	30.14	32.92	38.77	39.04	40.02	16.18	16.63	20.04	24.42	30.67
REGIf-AL	75.14	80.18	95.11	86.33	91.94	75.40	105.4	110.0	121.3	139.7	44.66	48.95	47.66	59.87	48.28	15.09	16.38	18.93	21.61	25.81
TSIf-AL	66.08	74.26	94.99	82.37	83.00	75.36	99.62	103.3	104.1	126.1	23.37	28.09	31.19	35.04	36.33	15.97	15.78	18.05	20.12	24.46

TABLE 8. The RMSE and MAPE values obtained from each imputation method for Glass.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.204	0.352	0.548	0.609	0.992	0.485	0.503	0.619	1.127	—	11.33	11.89	16.96	20.89	34.79	8.26	10.32	15.00	27.85	—
EMI	0.264	0.337	0.523	0.727	0.748	0.204	0.239	0.558	0.888	1.295	10.82	11.78	15.31	21.24	42.78	5.31	11.28	8.53	16.81	19.98
FEMI	0.242	0.412	0.570	1.373	1.605	0.395	0.330	1.189	1.805	2.360	7.98	16.89	20.21	37.45	41.61	4.08	10.33	18.78	18.65	27.52
REGI	0.397	0.363	0.507	0.739	0.681	0.402	0.445	0.575	0.694	1.403	13.23	16.90	17.82	21.47	25.29	12.17	9.20	10.19	14.45	30.01
TSI	0.237	0.407	0.478	0.480	0.561	0.371	0.421	0.591	0.652	0.676	8.89	11.49	11.95	15.44	19.26	7.40	8.95	9.42	14.91	16.46
REGIf	0.390	0.374	0.505	0.537	0.524	0.295	0.372	0.594	0.745	0.745	14.05	15.06	16.76	16.22	25.54	8.00	9.92	11.39	15.38	19.54
TSIf	0.212	0.342	0.490	0.478	0.489	0.315	0.368	0.562	0.668	0.681	9.01	11.08	12.47	14.08	19.19	7.30	9.37	10.64	14.54	17.29
REGIf-AL	0.389	0.354	0.456	0.510	0.503	0.496	0.382	0.562	0.721	0.730	11.88	13.40	16.66	15.54	22.31	8.44	9.52	9.11	14.06	19.28
TSIf-AL	0.200	0.319	0.443	0.450	0.460	0.291	0.316	0.470	0.607	0.658	8.09	9.60	10.93	12.34	17.09	7.14	7.74	7.10	11.12	14.99

TABLE 9. The RMSE and MAPE values obtained from each imputation method for Istanbul.

Imputation methods	RMSE (×10)										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.069	0.071	0.074	0.117	0.125	0.077	0.093	0.111	0.125	0.132	96.65	112.6	111.3	110.5	119.8	40.61	48.72	60.02	70.74	76.95
EMI	0.058	0.061	0.064	0.115	0.129	0.066	0.079	0.092	0.116	0.135	95.08	117.2	108.4	113.4	135.9	31.62	39.07	47.15	58.94	76.26
FEMI	0.064	0.065	0.064	0.111	0.190	0.068	0.083	0.104	0.107	0.183	101.8	117.7	107.9	119.8	164.6	34.19	41.60	54.31	58.49	86.35
REGI	0.059	0.061	0.067	0.109	0.153	0.069	0.079	0.096	0.120	0.130	95.13	111.5	129.4	126.7	195.1	31.59	36.07	47.76	64.81	76.01
TSI	0.066	0.071	0.068	0.101	0.137	0.075	0.084	0.091	0.101	0.119	108.8	122.3	118.9	125.7	139.3	43.00	42.10	50.14	55.51	66.49
REGIf	0.059	0.058	0.069	0.098	0.098	0.067	0.078	0.089	0.105	0.122	117.5	127.2	117.1	114.4	116.0	29.90	36.93	42.69	52.05	67.99
TSIf	0.063	0.060	0.062	0.096	0.098	0.069	0.077	0.086	0.101	0.120	92.39	110.2	106.3	109.5	114.0	33.21	35.04	41.93	52.02	67.16
REGIf-AL	0.069	0.065	0.062	0.095	0.097	0.077	0.080	0.092	0.106	0.122	95.12	110.2	107.0	110.0	112.8	32.27	35.40	42.68	51.64	67.86
TSIf-AL	0.063	0.061	0.061	0.089	0.093	0.067	0.074	0.082	0.096	0.117	90.89	106.3	100.3	103.3	107.6	33.85	34.77	41.28	49.54	64.79

and REGIf performs better in the remaining 42 combinations. For another set of comparisons between TSIf-AL and REGIf-AL, a similar pattern can be drawn that the former has better performance in most cases. According to the above descriptions, analyzing with TS model is more effective in

missing value imputation than that with traditional regression model.

The primary reason why TS modeling-based methods can achieve better performance is that TS model is realized on the premise of fuzzy partition, which considers the differences

TABLE 10. The RMSE and MAPE values obtained from each imputation method for ILPD.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	3.738	2.944	3.687	3.449	4.569	2.716	3.431	4.656	4.550	4.595	54.71	48.93	56.83	61.54	65.41	31.62	31.81	37.50	42.90	46.83
EMI	2.909	2.656	3.207	3.450	4.884	2.409	3.184	5.503	5.619	6.809	40.90	37.61	51.11	63.77	66.44	20.17	64.89	36.44	46.29	54.88
FEMI	2.607	2.633	4.495	4.756	4.275	2.138	2.917	4.197	4.626	4.933	42.56	27.14	52.88	85.85	63.02	18.74	25.68	30.71	40.36	46.87
REGI	2.972	2.756	3.279	3.514	4.700	2.789	3.352	4.127	4.170	4.581	59.86	54.40	62.09	91.71	71.53	36.11	54.60	38.94	47.89	46.68
TSI	2.823	2.681	3.155	3.432	4.456	2.483	2.968	4.058	4.148	4.547	57.70	43.17	51.28	63.26	61.83	20.42	28.28	33.77	37.23	45.55
REGIf	3.100	2.638	3.118	3.404	4.471	2.420	3.299	4.283	4.061	4.525	40.84	44.34	58.21	60.13	63.11	29.53	30.23	30.04	40.76	44.33
TSIf	2.636	2.638	3.003	3.360	4.276	2.118	2.904	3.852	3.803	4.524	48.81	44.06	53.04	60.01	58.16	22.01	24.29	28.64	35.12	44.30
REGIf-AL	3.621	3.191	3.770	3.566	4.278	2.663	2.925	4.163	4.020	4.521	51.72	46.92	57.80	59.58	62.20	25.48	27.71	28.73	40.06	43.79
TSIf-AL	2.656	2.460	2.957	3.004	3.859	2.560	2.778	3.455	3.439	4.427	48.65	40.39	47.55	52.22	56.42	24.74	23.44	27.44	32.30	42.87

TABLE 11. The RMSE and MAPE values obtained from each imputation method for Wine.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.947	1.545	1.653	2.069	3.459	1.666	1.837	2.231	—	—	17.54	22.89	31.84	41.69	64.87	14.82	24.12	28.51	—	—
EMI	1.189	1.831	2.510	2.690	2.186	1.433	2.867	2.002	3.136	4.047	16.96	22.24	40.82	47.57	42.41	13.83	33.90	20.73	34.71	40.04
FEMI	0.996	1.459	1.159	1.794	2.146	1.559	1.774	1.925	2.624	3.432	16.00	25.93	25.28	28.21	31.86	18.30	22.52	20.14	29.78	37.73
REGI	1.131	1.506	1.646	1.829	1.596	1.643	1.915	2.212	2.250	2.533	18.62	24.09	30.94	35.05	30.38	16.92	20.81	27.00	27.74	33.84
TSI	1.132	1.379	1.324	1.457	1.425	1.470	1.411	1.684	1.952	2.450	18.25	22.95	24.78	27.60	26.12	14.50	17.81	19.02	20.87	31.25
REGIf	1.433	1.379	1.316	1.420	1.446	1.555	1.485	1.774	2.109	2.532	18.32	20.87	20.32	22.93	24.09	13.92	15.61	20.37	25.85	34.44
TSIf	0.917	1.152	1.134	1.398	1.374	1.489	1.403	1.600	1.960	2.440	16.13	19.97	19.92	22.64	23.72	13.13	15.51	18.17	22.69	32.11
REGIf-AL	1.130	1.294	1.242	1.409	1.382	1.544	1.429	1.738	2.100	2.505	18.01	20.73	20.29	22.68	22.97	13.44	15.34	19.48	25.34	33.11
TSIf-AL	0.911	1.129	1.118	1.272	1.300	1.394	1.357	1.528	1.801	2.391	15.48	17.72	18.68	21.37	22.13	12.42	14.11	15.99	19.63	30.66

TABLE 12. The RMSE and MAPE values obtained from each imputation method for Segment.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	2.488	4.482	14.16	20.64	20.14	5.252	13.58	29.41	35.42	37.86	12.35	19.43	38.49	87.24	107.85	13.66	25.82	44.52	59.92	69.63
EMI	2.649	3.872	5.871	7.051	23.39	4.789	6.842	11.72	22.35	31.98	15.43	20.80	19.70	23.64	57.99	13.41	17.47	24.29	38.94	51.76
FEMI	1.509	2.871	8.488	10.14	18.67	4.356	6.791	12.91	16.10	34.36	9.11	11.00	12.08	30.49	55.45	8.59	13.18	26.37	31.85	50.15
REGI	5.460	8.675	12.65	13.57	14.65	12.84	13.80	19.22	24.88	31.37	32.68	43.76	50.18	64.74	68.46	16.81	22.53	28.22	36.98	45.73
TSI	5.661	5.936	7.025	7.874	10.13	9.989	10.40	12.92	15.15	25.12	23.85	35.24	32.92	29.96	34.26	29.52	46.63	37.99	40.80	43.43
REGIf	4.860	7.765	10.59	11.31	12.89	10.60	11.55	17.73	23.39	30.49	27.98	41.37	39.10	48.74	53.77	16.93	19.90	25.35	34.28	46.76
TSIf	2.768	4.447	6.836	8.011	10.61	5.863	8.348	12.00	15.34	25.59	15.35	26.12	29.02	34.76	38.32	13.18	17.41	22.40	30.49	43.04
REGIf-AL	4.326	6.145	8.736	9.287	10.78	8.429	11.39	13.67	18.28	26.18	27.87	32.41	30.34	35.15	43.33	14.71	20.96	24.55	29.89	39.69
TSIf-AL	2.354	3.433	5.457	6.893	9.49	4.782	6.437	10.15	13.27	22.21	13.79	18.57	18.70	22.98	29.14	12.94	15.37	18.84	23.82	34.87

TABLE 13. The RMSE and MAPE values obtained from each imputation method for Dow.

Imputation methods	RMSE										MAPE (%)									
	MCAR					MNAR					MCAR					MNAR				
	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%	5%	15%	25%	35%	45%	10%	20%	30%	40%	50%
KNNI	0.076	0.079	0.095	—	—	0.088	0.103	0.181	0.203	—	32.64	37.88	52.88	—	—	17.24	23.98	32.02	43.97	—
EMI	0.068	0.076	0.087	0.088	0.092	0.082	0.092	0.275	0.288	0.164	25.62	30.29	61.55	83.59	136.95	11.39	17.29	50.16	61.66	70.56
FEMI	0.058	0.065	0.083	0.086	0.089	0.096	0.104	0.168	0.166	0.168	19.26	23.75	60.87	79.20	116.70	13.25	27.40	44.39	53.48	62.81
REGI	0.073	0.085	0.102	0.159	0.145	0.105	0.125	0.177	0.193	0.248	41.47	54.04	67.41	117.87	109.44	20.04	22.78	37.29	42.46	66.17
TSI	0.081	0.088	0.097	0.099	0.096	0.099	0.112	0.118	0.144	0.238	40.26	43.17	62.05	68.71	73.63	22.48	26.34	29.45	34.97	59.85
REGIf	0.071	0.079	0.083	0.091	0.095	0.098	0.113	0.156	0.190	0.237	38.56	48.37	59.24	70.09	76.33	15.54	22.19	30.46	41.63	57.51
TSIf	0.070	0.078	0.083	0.087	0.090	0.083	0.095	0.110	0.138	0.233	29.56	54.37	55.44	59.33	53.98	14.92	19.64	24.30	32.57	56.47
REGIf-AL	0.075	0.082	0.076	0.081	0.096	0.090	0.099	0.146	0.184	0.237	46.92	51.13	53.21	61.33	76.68	15.37	19.64	30.26	41.53	56.56
TSIf-AL	0.052	0.062	0.078	0.080	0.084	0.077	0.084	0.100	0.123	0.233	26.83	34.76	40.54	48.06	50.19	13.87	17.08	20.55	26.70	56.21

of attribute relationships between subsets while carrying out regression analysis. Besides, TS model is a nonlinear model in essence, which obtains the model output by weighting

and summing those values derived from each fuzzy rule. In general, these fuzzy rules can reflect the local characteristics of incomplete data, and thus mine the distribution of

association among attributes in different partitions to some extent. Therefore, it is more capable of data estimation than traditional regression model, and thus performing better in missing value imputation.

2) THE EFFECT OF ALTERNATE LEARNING ON IMPUTATION PERFORMANCE

According to the RMSE values in Tables 2 to 13, TSif-AL performs better than TSif, and the performance of REGif-AL is also superior to the REGif, which indicates that the estimated values derived from the incomplete data model are more approximate to their actual values after using the alternate learning strategy. Taking the *Abalone* datasets with different missing ratios as examples, when applying the proposed strategy, the accuracies of regression modeling-based imputation are increased by more than 15% in most cases, and the performance of TS modeling-based imputation also has a further improvement. Therefore, the feasibility and effectiveness of alternate learning strategy can be verified based on those descriptions.

The feasibility and effectiveness lie in the following two aspects. On the one hand, model output values are able to approximate to their actual values with the adjustment of model parameters, which means that imputations can be more reasonable with the optimization of incomplete data modeling. On the other hand, model parameters can reflect the real attribute relationships with the development of data quality, and thus further enhancing the reliability of those imputations. In summary, the accuracy of incomplete data modeling and the effectiveness of missing value imputation can be enhanced in a collaborative way.

3) THE COMPARISON BETWEEN TSIF-AL AND NON-REGRESSION-BASED METHODS

Comparing the imputation performance of TSif-AL, KNNI, EMI and FEMI in Table 2 to 13, it can be seen that TSif-AL has a higher imputation accuracy in most cases. In a total of 240 set of results, TSif-AL outperforms the other methods in 197 sets. Moreover, TSif-AL has an even better imputation results when the missing ratio of the incomplete data is large. Specifically, the results of TSif-AL are totally better than those of KNNI, EMI and FEMI except for the *Iris* dataset, the *Abalone* dataset and the *Dow* dataset when the missing ratio is not less than 35%, and TSif-AL has a better imputation performance in 92 out of 96 sets. The above analysis shows that TSif-AL can impute missing values in an effective way. Furthermore, taking into account the relationship among attributes may contribute to the imputation accuracy, especially in the case of high missing ratios.

D. FURTHER EVALUATION

1) CONVERGENCE OF ALTERNATE LEARNING

Convergence is one of the key concerns for iterative algorithms. In this subsection, the convergence of alternate learning strategy is verified using present values due to the

consideration that only those values are available in real-world datasets. Taking the *Segment* datasets with different missing ratios as an example, the RMSE values obtained in each iteration are shown in Fig. 4.

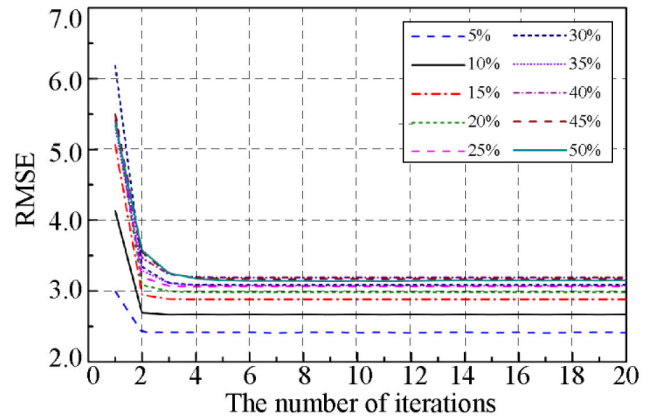


FIGURE 4. The variation of RMSE values in alternate learning process.

As shown in Fig. 4, all the curves present the same trend in general, which drops rapidly at the beginning and then tends to be stable. Specifically, the RMSE sharply goes to a small value within the first 3 iterations for each missing ratio, then the convergence rate decreases gradually and remains unchanged. Therefore, it can be easily concluded that the alternate learning strategy has a fast convergence speed and good stability.

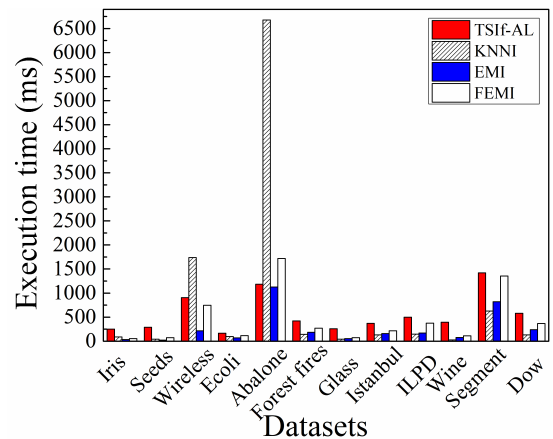


FIGURE 5. Average execution time (ms) for methods on 12 datasets.

2) TIME EFFICIENCY

The average execution time for 10 datasets (2 mechanisms and 5 different incomplete datasets per mechanism) for each benchmark dataset is shown in Fig. 5. In order to make the results more reliable, we use the same machine to carry out the experiments. As shown in Fig. 5, TSif-AL generally takes more time than KNNI, EMI and FEMI in order to pay the cost of apparently better imputation accuracy. We can

also find that in respect of *Wireless* and *Abalone* datasets, the time consumption of KNN is obviously larger than the other methods due to the increase in data size, and for *Segment* dataset, the gap of execution time between TSif-AL and FEMI is not obvious. These results indicate that the problem of time consumption for TSif-AL can be neglected to some extent when the size of data gets larger, and the reason lies in that TSif-AL can obtain the ideal performance of imputation compared with the other methods.

Next, we analyze the time complexity from the perspective of theoretical analysis. TSif-AL is realized by premise parameters identification, input variables selection and the alternate learning of model parameters with imputations. Let N , c , s and l represent the numbers of records, clusters, attributes and iterations for alternate learning respectively. Since we take FCM-PDS algorithm with the complexity of $O(Nc^2s)$ to identify the premise parameters for all the TS models and utilize stepwise regression algorithm which has the complexity of $O(Ns^2)$ to select the input variables for each TS model, the complexity of the above step can be described as $O(Nc^2s + Ns^3)$. In each iteration of alternate learning, the consequence parameters of each TS model are obtained through the least square method with the complexity of $O(c^3s^3)$. Therefore, the complexity of TSif-AL is $O(Nc^2s + Ns^3 + lc^3s^4)$. Generally, l , c are chosen to be numbers significantly smaller than N [12], and thus the complexity of TSif-AL can be simplified to $O(Ns^3 + s^4)$. Additionally, the complexity of KNNI, EMI and FEMI are $O(Ns)$, $O(Ns^2 + s^3)$ and $O(Ns^2 + s^3)$, respectively. Although TSif-AL needs higher computation time compared with the other methods, imputation accuracy generally has a higher priority in the imputation of missing values especially when the difference of time computation is not obvious.

V. CONCLUSION

Taking the differences in regression relationships among subsets into consideration, we propose a method of incomplete data modeling based on TS model for imputing missing values. The method performs regression analysis on each subset obtained by fuzzy clustering algorithm and takes the weighted sum of the regression models to build the global model, which has higher precision and better imputation performance than traditional regression model. Meanwhile, concentrating on the problem of incomplete model input caused by data corruption, this paper carries out an alternate learning strategy for training model parameters together with imputations, in which missing values are treated as variables to promote training. Through this strategy, a collaborative improvement of model accuracy and imputation accuracy can be realized additionally as the problem of incomplete model input is resolved. Experiments on 12 UCI datasets with different missing ratios and mechanisms demonstrate that precise TS modeling with the consideration of differences among subsets is capable of missing value imputation, which derives more appropriate estimation values than the traditional regression model. Furthermore, the effectiveness

of incomplete data modeling is enhanced by engaging all the present values in modeling, and the performance of missing value imputation is further improved when training with alternate learning strategy.

In addition, the variation of RMSE values in alternate learning process indicates the ideal convergence of TSif-AL, and the comparison among TSif-AL with KNNI, EMI and FEMI on time complexity shows that TSif-AL requires higher computation time, but obtains the obviously better imputation performance. From the perspective of execution time, the gaps in time consumption between the proposed method and comparison methods are not obvious when the size of dataset gets large, and can be neglected to some extent since imputation accuracy generally has a higher priority in the imputation of missing values.

REFERENCES

- [1] R. J. A. Little and D. B. Rubin, "Statistical analysis with missing data," *Technometrics*, vol. 45, no. 4, pp. 364–365, 2002.
- [2] F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein, "Missing data in medical databases: Impute, delete or classify?" *Artif. Intell. Med.*, vol. 58, no. 1, pp. 63–72, May 2013.
- [3] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, 2006.
- [4] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.
- [5] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–115, 2010.
- [6] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression," *Neurocomputing*, vol. 251, pp. 26–34, Aug. 2017.
- [7] C.-F. Tsai and F.-Y. Chang, "Combining instance selection for better missing value imputation," *J. Syst. Softw.*, vol. 122, pp. 63–71, Dec. 2016.
- [8] S. Song, Y. Sun, A. Zhang, L. Chen, and J. Wang, "Enriching data imputation under similarity rule constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 275–287, Feb. 2020.
- [9] C.-F. Tsai, M.-L. Li, and W.-C. Lin, "A class center based approach for missing value imputation," *Knowl.-Based Syst.*, vol. 151, pp. 124–135, Jul. 2018.
- [10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.
- [11] M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 389–422, Feb. 2016.
- [12] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang, "Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation," *Knowl.-Based Syst.*, vol. 132, pp. 249–262, Sep. 2017.
- [13] Z. Qi, H. Wang, J. Li, and H. Gao, "FROG: Inference from knowledge base for missing value imputation," *Knowl.-Based Syst.*, vol. 145, pp. 77–90, Apr. 2018.
- [14] S. Zhang, "Shell-neighbor method and its application in missing data imputation," *Int. J. Speech Technol.*, vol. 35, no. 1, pp. 123–133, Aug. 2011.
- [15] M. D. Samad and L. Yin, "Non-linear regression models for imputing longitudinal missing data," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Xi'an, China, Jun. 2019, pp. 1–3.
- [16] K. Lavanya, L. S. S. Reddy, and B. E. Reddy, "A study of high-dimensional data imputation using additive LASSO regression model," in *Computational Intelligence in Data Mining*. Singapore: Springer, 2019, pp. 19–30.
- [17] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: Local least squares imputation," *Bioinformatics*, vol. 22, no. 11, pp. 1410–1411, Jun. 2006.

- [18] K. O. Cheng, N. F. Law, and W. C. Siu, "Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data," *Pattern Recognit.*, vol. 45, no. 4, pp. 1281–1289, Apr. 2012.
- [19] J. Shah, G. N. Brock, and G. Gaskins, "BayesMetab: Treatment of missing values in metabolomic studies using a Bayesian modeling approach," in *Proc. Int. Conf. Intell. Biol. Med.*, 2019, vol. 20, no. 24, pp. 1–13.
- [20] B. van Stein and W. Kowalczyk, "An incremental algorithm for repairing training sets with missing values," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Cham, Switzerland: Springer, 2016, pp. 175–186.
- [21] W. Zhang, Y. Yang, and Q. Wang, "Using Bayesian regression and EM algorithm with missing handling for software effort prediction," *Inf. Softw. Technol.*, vol. 58, pp. 58–70, Feb. 2015.
- [22] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci.*, vol. 233, pp. 25–35, Jun. 2013.
- [23] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," *Expert Syst. Appl.*, vol. 115, pp. 68–94, Jan. 2019.
- [24] S. Rastegar, R. Araújo, and J. Mendes, "Online identification of Takagi–Sugeno fuzzy models based on self-adaptive hierarchical particle swarm optimization algorithm," *Appl. Math. Model.*, vol. 45, pp. 606–620, May 2017.
- [25] X. Xie, L. Lin, and S. Zhong, "Process Takagi–Sugeno model: A novel approach for handling continuous input and output functions and its application to time series prediction," *Knowl.-Based Syst.*, vol. 63, pp. 46–58, Jun. 2014.
- [26] B. M. Al-Hadithi, A. Jiménez, and F. Matía, "A new approach to fuzzy estimation of Takagi–Sugeno model and its applications to optimal control for nonlinear systems," *Appl. Soft Comput. J.*, vol. 12, no. 1, pp. 280–290, 2012.
- [27] C. Fantuzzi and R. Rovatti, "On the approximation capabilities of the homogeneous Takagi–Sugeno model," in *Proc. IEEE 5th Int. Fuzzy Syst.*, vol. 2, Sep. 1996, pp. 1067–1072.
- [28] R. J. Almeida, U. Kaymak, and J. M. C. Sousa, "A new approach to dealing with missing values in data-driven fuzzy modeling," in *Proc. Int. Conf. Fuzzy Syst.*, Jul. 2010, pp. 1–7.
- [29] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 116–132, Jan. 1985.
- [30] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 31, no. 5, pp. 735–744, Oct. 2001.
- [31] L. D. Chambers, *The Practical Handbook of Genetic Algorithms: Applications*, 2nd ed. London, U.K.: Chapman & Hall, 2000.
- [32] S. Abraham, M. Raisee, G. Ghorbaniasl, F. Contino, and C. Lacor, "A robust and efficient stepwise regression method for building sparse polynomial chaos expansions," *J. Comput. Phys.*, vol. 332, pp. 461–474, Mar. 2017.
- [33] K. Y. Chan, H. K. Lam, T. S. Dillon, and S. H. Ling, "A stepwise-based fuzzy regression procedure for developing customer preference models in new product development," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 5, pp. 1728–1745, Oct. 2015.
- [34] M. Antonelli, P. Ducange, F. Marcelloni, and A. Segatori, "On the influence of feature selection in fuzzy rule-based regression model generation," *Inf. Sci.*, vol. 329, pp. 649–669, Feb. 2016.
- [35] D. Dua and C. Graff. (2019). UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine, Irvine, CA, USA. [Online]. Available: <http://archive.ics.uci.edu/ml>



XIAOCHEN LAI received the B.S., M.S., and Ph.D. degrees from the Dalian University of Technology, Dalian, China, in 1999, 2003, and 2016, respectively. He is currently working as an Associate Professor with the School of Technology, Dalian. His major research directions involve deep learning, data modeling, control theory and engineering, body sensor networks, and wireless sensor networks.



LIYONG ZHANG received the B.S. degree in automation, the M.S. degree in control theory and control engineering, and the Ph.D. degree in control theory and control engineering from the Dalian University of Technology, Dalian, China, in 1999, 2002, and 2018, respectively. He is currently working as a Lecturer with the School of Control Science and Engineering, Dalian University of Technology. He has published more than 50 research articles, and has five Chinese invention patents issued. He is a coauthor of three books. His current research interests include data modeling, fuzzy clustering, feature selection, and granular computing.



XIN LIU received the M.S. degree from the Dalian University of Technology, in 2019. She is currently pursuing the Ph.D. degree with Central South University. Her main research topic is devoted to data mining.

• • •