SN Computer Science (2021) 2:57
https://doi.org/10.1007/s42979-021-00459-1

ORIGINAL RESEARCH

An Exploration of Online Missing Value Imputation in Non-stationary Data Stream

Wenlu Dong[1] · Shang Gao[1,2] · Xibei Yang[1] · Hualong Yu[1,2]

Received: 31 July 2020 / Accepted: 8 January 2021 / Published online: 29 January 2021

Abstract
Missing value imputation (MVI) is an important data preprocessing technique. In previous decades, MVI technique has been widely studied as well as most MVI approaches have been proposed by means of either statistics or machine learning techniques. However, all previous methods only focus on the static data, but ignore the imputation for the dynamic online data. It is intuitionistic that the imputation errors may be significantly increased when there exists concept drifts in the data stream. In this paper, we investigate the impact of adopting the conventional MVI methods in non-stationary data stream. Meanwhile, two slide time window-based strategies are proposed to alleviate this impact, where one is the plain average strategy, and the other is the logarithmic weighted average strategy that gradually adds the weights of instances along the time axis. Combining with the proposed strategies, three popular MVI techniques, mean imputation (MI), KNN imputation (KNNI) and the Bayesian principal component analysis imputation (BPCAI) are adopted, to indicate the effect of the strategies are irrelevant to the specific MVI technique. The experimental results on three different types' concept drift synthetic data sets and two real-world drifting data sets have presented the effectiveness and feasibility of the proposed strategies. Moreover, the impact of time window size has also been investigated for guiding the parameter settings in future practical applications.

Keywords Missing value imputation · Data stream · Concept drift · Slide time window · Weighting

Introduction

Missing value imputation (MVI) is an important data preprocessing technique, as most machine learning algorithms request the completeness of data [1]. Generally, the missing value is manifested in the loss of some attribute values associating with a specific instance, which can be caused by many reasons, such as the inaccurate measurement, unnecessary measurement, node failure and even the demand of privacy protection. The missing values can be better imputed by taking advantage of their contextual information than simply filling zero which might introduce much wrong information, or directly deleting the instances with missing attributes which might cause much information loss, further making the models are less accurate [2, 3]. The usage of the contextual information can be either statistical [4, 5] or machine learning [6–10]. The existing MVI algorithms have also been widely used in a mass of real-world applications, including Medical Science [11–13], Economics [14], Bioinformatics [15–18], Psychology [19, 20] and even Traffic accident analysis [21].

There have been lots of MVI algorithms, however, the existing ones generally ignore a specific but practical scenario, i.e., the data appears in the form of stream, and meanwhile the data stream has the potential concept drifts. It is intuitionistic that in this scenario, the imputation errors may be significantly increased because of variation of the data distribution, i.e., concept drifts. Considering the requirement of *one pass* in streaming data processing [22, 23], it is difficult to track the variance of the attribute value domain, further to tune the MVI model for providing the accurate imputed values. That is to say, in data stream, the reserved experiences may be incomplete, which is insufficient for changing the MVI model to adapt the concept drift.

✉ Hualong Yu
yuhualong@just.edu.cn

[1] School of Computer, Jiangsu University of Science and Technology, No.2, Mengxi Road, Zhenjiang 212003, Jiangsu, China

[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China

To meet the above practical needs, in this paper, we focus on MVI techniques in context of the streaming data with concept drifts. Firstly, we are curious to two issues as below, (1) whether the imputation quality of the conventional MVI algorithms can be influenced by the non-stationary data stream or not? and (2) if it has an impact, whether the strength of the impact is associated with the types of concept drift or not? Then, based on the investigation about the two issues above, two slide time window-based strategies are proposed to implement online missing value imputation procedure, where one is the plain average strategy, and the other is the logarithmic weighted average strategy. In the time window, some recent received instances are reserved, as well with arrival of a new instance, the oldest instance in the time window will be replaced by it. The strategies can be accepted as at any time, only a few instances are reserved which will not add the burden of storage. In other words, the strategies do not violate the principle of *one pass*.

In our experiments, we adopt three different concept drift types' synthetic data sets and two real-world drifting data sets to observe their impacts to MVI techniques, as well to verify the effectiveness of our proposed strategies. In addition, to present the strategies are irrelevant to the specific MVI techniques, we also randomly choose three popular MVI algorithms, including mean imputation (MI) [4], KNN imputation (KNNI) [6] and Bayesian principal component analysis imputation (BPCAI) [17], in our experiments. The experimental results on five streaming data sets presented the effectiveness and feasibility of the proposed strategies. Moreover, the impact of the time window size has also been investigated for guiding the parameter settings in future practical applications.

To our best knowledge, it is the first work to investigate the MVI procedure in drifting data streams which will become more and more popular in the practical applications. The main contributions of this work are summarized as follows,

1. We investigate the impact of directly adopting the conventional static MVI approaches for imputing non-stationary data stream.
2. We try to propose two slide time window-based strategies which can be combined with the existed MVI algorithms to alleviate the influence of concept drift for imputation quality.

The rest of this paper is organized as follows. "Problem Description" simply describes MVI task and analyzes its difficulties in non-stationary data stream. In "Methods", we introduce the proposed strategies in detail. "Experiments" presents the experimental results and analysis. Finally, "Conclusions" concludes the contributions of this paper and indicates the future work.

Problem Description

For a given static data set $\Phi = (x_1, x_{2,\ldots,} x_n)$, where x_i denotes an instance, it can be also seen as a matrix of $n \times m$, where n indicates the number of instances, and m denotes the number of attributes. If Φ is a complete data set without missing value, then it means that for any variable x_{ij}, it has been assigned a specific discrete or continuous value. While if Φ is the data set with missing values, that means there exists some variables that haven't been designated the specific values. Generally, a missing value is represented as ? in the data set or matrix. An example of missing data set can be observed as below.

$$\begin{bmatrix} 1.2 & ? & \cdots & 0.7 \\ ? & 0.9 & \cdots & 1.0 \\ \vdots & \vdots & \ddots & \vdots \\ 1.1 & 0.5 & \cdots & ? \end{bmatrix}_{n\times m}$$

Here, ? indicates the corresponding missing variable. For each ?, the MVI techniques takes advantage of the contextual information reflecting from the same row, the same column or both to find or calculate an appropriate value for replacing it. However, all existing imputation algorithms admit a basic hypothesis, i.e., the distribution of the data is consistent. The hypothesis guarantees the approximately accurate imputation for the missing variables, as the contextual information reflects the data distribution.

Next, let us consider MVI problem in the non-stationary data stream. Non-stationary means the concept drift, as well means the variation of distributions. An example of missing data stream with concept drift can be observed as below.

$$\begin{bmatrix} 1.2 & ? & \cdots & 0.7 \\ ? & 0.9 & \cdots & 1.0 \\ \vdots & \vdots & \ddots & \vdots \\ 10.1 & 4.5 & \cdots & ? \end{bmatrix}_{n\times m} \Bigg\downarrow \text{Time axis}$$

It is clear that when the distribution changes, i.e., the value domains of some attributes change, the old experiences would misguide the imputed results of new received instances, further generate a larger imputation error. In other words, on the drifted data stream, the imputation procedure cannot excessively depend on the old contextual information, otherwise the imputation quality would be not guaranteed.

The other difficulty that applying MVI techniques on non-stationary data stream is the requirement of *one pass*. That is to say, it is impossible to reserve all received instances into the storage, otherwise, the storage would increase sharply until overflow. That means we have to quickly deal with the received new instance, and then delete it immediately.

The variation of distributions caused by concept drift and the requirement of *one pass* both bring challenges for MVI procedure.

Methods

In this section, we first introduce two slide time window-based strategies for transforming the traditional MVI algorithms from the static environment to the dynamic environment. Next, three popular MVI algorithms used in this work are described in brief.

Slide Time Window-Based Strategies

As discussed in Section II, MVI procedure is challenged by the non-stationary streaming data based on two main reasons, one is at any time, the distribution may change and the old experiences will be useless, and the other one is the old instances can't be reserved to prevent an unbearable storage consumption. To address these two problems synchronously, we profit from the idea of slide time window adopted by online learning [24], to design two strategies.

Slide time window, as its name indicates, varies with time, and it only reserves the recently received instances in the window. That means the window should be a queue structure, i.e., when a new instance is received, it will replace the first instance (the oldest instance) emerging in the window. As in the window, all instances are received recently, thereby it can track the variation of the distributions and adapt the concept drift to a maximal extent. Meanwhile, it doesn't violate the rule of *one pass* as the length of the window has been fixed.

In the slide time window, we adopt two different strategies to combine MVI techniques. The first one is a plain average strategy (AS), which treats all instances in the window equally. This strategy equals to that MVI techniques run on the static data. The other one is a logarithmic weighted average strategy (WAS), which considers the newer the instance is, the more important the instance for MVI is. To fairly assign the weights for the instances, we adopt the order number in the window to calculate the weights. Suppose there is an instance x_i in the window, and its order number is l, then its initial weight ω_i can be calculated as,

$$\omega_i = \log_{10}(l + \lambda) \tag{1}$$

where λ is a small constant to avoid the case $\omega_i = 0$ when the order number l equals to be 1. In this paper, we empirically designate $\lambda = 0.5$. In addition, considering the value domain of ω_i is uncontrollable, the initial weight should be regulated by the following normalization function,

$$\varpi_i = \omega_i / \Sigma \tag{2}$$

where ϖ_i denotes the normalized weight, Σ indicates the sum of weights, which is decided by how many instances participating in the procedure of MVI. It is clear that in WAS, if a instance has a larger order number, meaning we received it later, thereby it will be assigned a larger weight. That satisfies the rule of receiving later, contributing more. Of course, the WAS strategy may only associates with some MVI algorithms which can adapt the instance weights, while the AS strategy can be easily combined with any one existing MVI algorithm.

Suppose the baseline strategy is to reserve the first batch of the received instances, then the difference of these three strategies can be observed in Fig. 1.

Three MVI Algorithms Used in this Paper

In this paper, three popular MVI algorithms are used to combine with the proposed strategies and to verify their effectiveness. The algorithms are simply described as follows,

- Mean Imputation (MI) [4]. MI either finds the mode when the attribute is discrete, or calculates the average of the same attribute throughout all instances when the attribute is continuous, to fill the missing attribute value.
- KNN Imputation (KNNI) [6]. KNNI firstly calculates the distances between the missing instance and all other instances which do not miss the same attribute, and then finds its K nearest neighbors, finally only use the average of the same attribute from these K instances to fill the missing attribute.
- Bayesian principal component analysis imputation (BPCAI) [17]. In comparison with MI and KNNI, the BPCAI is a more complex MVI technique. It orderly conducts three elementary processes, (1) principle component regression, (2) Bayesian estimation, and (3) an expectation–maximization (EM)-like repetitive algorithm.

According to the description above, we can observe that both MI and BPCAI take advantage of the global information to impute the missing values, while KNNI only utilizes the local correlation information. In addition, we also find that both MI and KNNI can adapt the instance weights, thus they can be combined with both AS and WAS strategies. However, BPCAI can only be integrated into the AS strategy. In practical applications, any one imputation algorithm can replace these three algorithms into the slide time window.

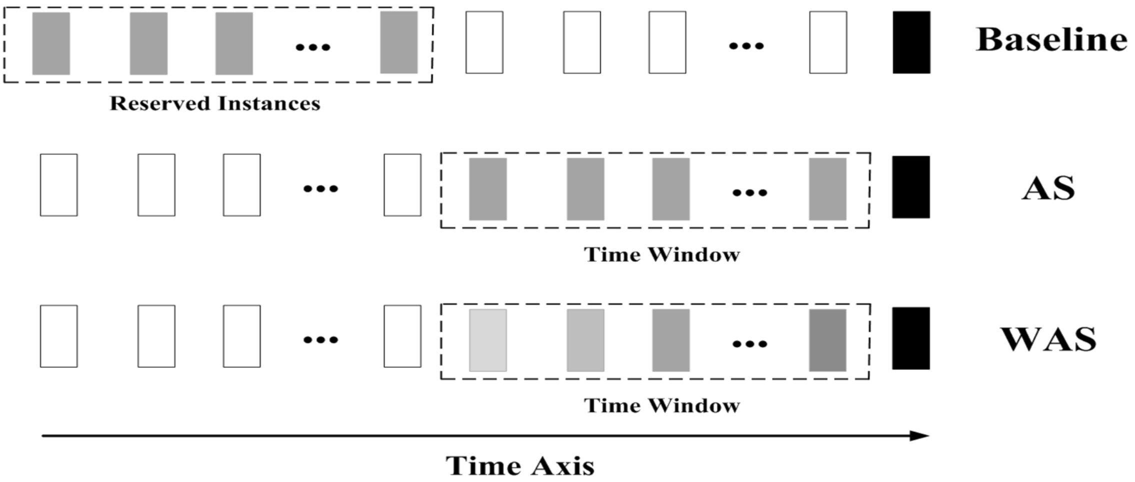


Fig. 1 The schematic diagrams to compare the ideas of baseline, AS and WAS strategies, where the white blocks, black blocks and grey blocks indicate the removed instances, the new received instances and the reserved instances, respectively. Specifically, for WAS, the tonal variation of the grey blocks indicate they have different weights, and the darker the block is, the larger weight it has

Experiments

Data Sets Description

First, to better verify the effectiveness and feasibility of the proposed strategies, three synthetic streaming data sets are generated. Here, each data set covers a concept drift type, namely sudden drift, incremental drift and reoccurring drift, respectively [23, 25–27]. Moreover, to simplify the algorithms comparison and data generation, we suppose all data sets only consist continuous attributes and in each data set, any one data chunk satisfies the Gaussian distribution, as well for each data set, it has 12 chunks, where each chunk has 50 instances. To avoid the instance is completely missing, each instance is assigned ten attributes which satisfy the same distribution. The detailed distribution information for each data set is summarized in Table 1. Furthermore, the distributions' means can be more intuitively observed in Fig. 2.

As for real-world data, we used two data sets, namely electricity and weather, respectively. The electricity dataset was acquired from electricity supplier in New South Wales (NSW), Australia. It consists of 45,312 instances, each with eight attributes. The first attribute is the date, the second attribute is the day of week (1–7), followed by the period of day (1–48). The remaining five attributes represents NSW electricity price, NSW electricity demand, Victoria electricity price, Victoria electricity demand, and the amount of electricity scheduled for transfer between the two states. The electricity data set is available at http://moa.cms.waika to.ac.nz/datasets/. As for the weather data set, it is indeed a subset of noaa data, which contains the real whether records more than 50 years aiming to predict whether rain precipitation was observed on each day. It also contains eight attributes, namely Temperature, Dew Point, Sea level pressure, Visibility, Average wind speed, Maximum sustained wind speed, Maximum temperature and Minimum Temperature, respectively. The weather data set can be downloaded from ftp://ftp.ncdc.noaa.gov/pub/data/gsod. The detailed description about these two data sets is presented in Table 2.

Table 1 Description about the synthetic data sets

Data set	Chunck sequence											
	1	2	3	4	5	6	7	8	9	10	11	12
Sudden	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)
Reoccurring	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)
Incremental	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.5, 0.1)	(0.6, 0.1)	(0.7, 0.1)	(0.8, 0.1)	(0.9, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)	(1.0, 0.1)

The data in the bracket denotes (mean, standard deviation) of a Gaussian distribution

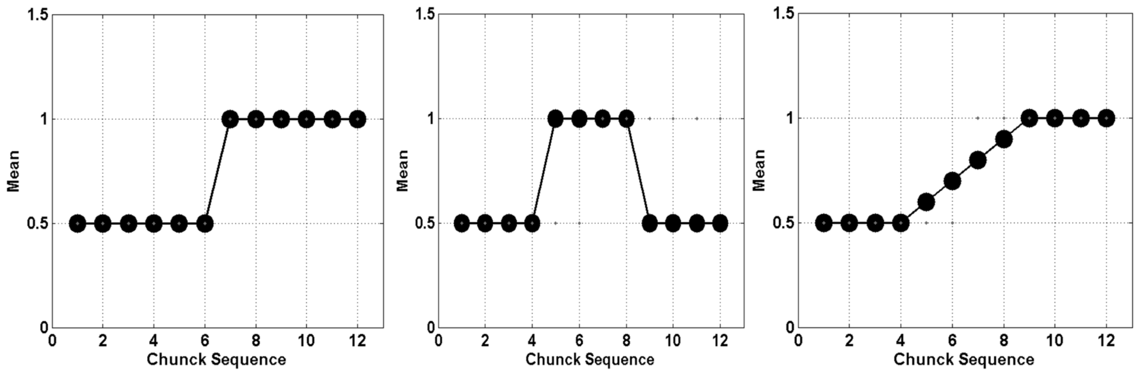


Fig. 2 The variation of the distributions' means on three data sets. From top to down, the synthetic data sets are Sudden, Reoccurring and Incremental, respectively

Table 2 Description about the real-world data sets

Data set	Number of attributes	Number of instances
Electricity	8	45,312
Weather	8	18,159

Experimental Settings

In the experiments, for MI and KNNI, each is combined with three different strategies, i.e., Baseline, AS and WAS, while for BPCAI, it only associates with two strategies, i.e., Baseline and AS, as it is not available for WAS. That means there exist eight different algorithms, namely Baseline-MI, AS-MI, WAS-MI, Baseline-KNNI, AS-KNNI, WAS-KNNI, Baseline-BPCAI and AS-BPCAI respectively. For each MVI algorithm, we independently compare its several combination strategies.

A popular metric called RMSE (Root Mean Square Error) is used for evaluating the performance of MVI and comparing various algorithms. RMSE can be calculated by,

$$\mathrm{RMSE} = \sqrt{\left(\sum (\hat{y} - y)^2\right)/N} \tag{3}$$

where y and $\hat{y}$ indicate the real and estimated values of a specific attribute, respectively. As for N, it is the number of missing attributes. It is clear that, the nearer the real values and the estimated values are, the smaller the RMSE is, and also the better the MVI algorithm is.

Without loss of generality, we designate 5% random missing rate for the data streams. That means on each data set, 5% places are replaced randomly. To simulate the real-world application, the data can be received one by one. The final RMSE results are given in the form of mean value of 50 times' random runnings, i.e., for each running, the missing places in the data set are randomly replaced.

As for several important parameters, the length of the slide time window, it is initially set to be 50, the number of nearest neighbors K in KNNI is empirically set to be 5, and in BPCAI, the parameter K is designated the recommended value, i.e., W-1, where W indicates the number of instances in the slide time window [17].

Results and Discussion

We firstly compared the proposed strategies with Baseline in the context of MI, KNNI and BPCAI algorithms, respectively. The performance comparison of the results can be observed in Fig. 3, Tables 3, 4 and 5.

From the results in Fig. 3, Tables 3, 4 and 5, it is not difficult to draw some conclusions as follows.

- In contrast with the baseline strategy, the proposed AS and WAS strategies acquired significantly better imputation quality, no matter which kind of concept drift was met or which kind of MVI algorithm was adopted. To explore its reasons, it is not difficult to observe that the proposed strategies can track the distribution's variation and abandon the experiences related with the old distributions in time, but the baseline strategy which is adopted in conventional static MVI scenario can't adapt the variation of the distribution. That is to say, adopting the experiences from the old distribution to compensate the missing values in new distribution must decrease imputation quality. This finding also confirms the correctness of our initial hypothesis, i.e., it is harmful for directly using the traditional MVI techniques to impute the missing non-stationary streaming data.
- In the context of synthetic data, compared with the reoccurring drift, two other concept drift types have clearly stronger damages for the imputation quality of the base-

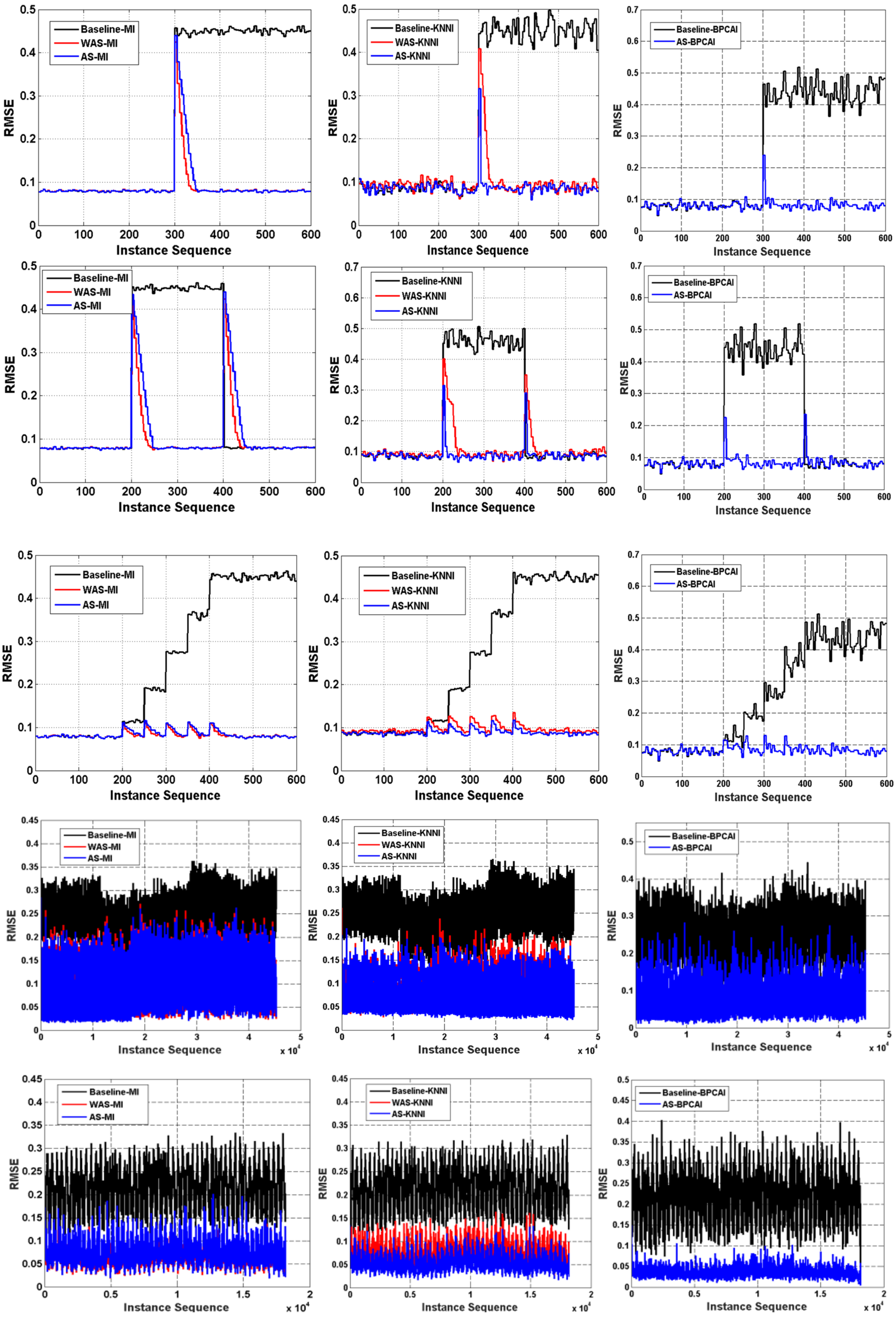


Fig. 3 Performance comparison among Baseline, AS and WAS strategies, where the 1st column associates with MI algorithm, the 2nd column relates with KNNI algorithm and the 3rd column associates with BPCAI algorithm. From top to down, five rows correspond to Sudden, Reoccuring, Incremental, electricity and weather data sets, respectively

Table 3 RMSE in the context of MI Algorithm

Data Set	Baseline-MI	AS-MI	WAS-MI
Sudden	0.2648	0.0930	**0.0883**
Reoccurring	0.2025	0.1064	**0.0967**
Incremental	0.2545	0.0844	**0.0827**
Electricity	0.2414	0.0801	**0.0774**
Weather	0.2093	0.0734	**0.0667**

The best results have been highlighted in boldface

Table 4 RMSE in the context of KNNI Algorithm

Data Set	Baseline-KNNI	AS-KNNI	WAS-KNNI
Sudden	0.2674	**0.0871**	0.0991
Reoccurring	0.2103	**0.0898**	0.1071
Incremental	0.2572	**0.0890**	0.0972
Electricity	0.2432	**0.0638**	0.0775
Weather	0.2096	**0.0507**	0.0748

The best results have been highlighted in boldface

Table 5 RMSE in the context of BPCAI Algorithm

Data Set	Baseline-BPCAI	AS-BPCAI
Sudden	0.2588	**0.0808**
Reoccurring	0.1987	**0.0830**
Incremental	0.2506	**0.0829**
Electricity	0.2469	**0.0683**
Weather	0.2129	**0.0360**

The best results have been highlighted in boldface

line strategy. We believe the reason lies in that both sudden drift and incremental drift hold irreversible changes about the distribution, but for the reoccurring drift, the old experiences can be useful intermittently. In addition, the variation of RMSE of the baseline algorithm conforms to the distribution variation of the corresponding concept drift (see Fig. 3).

- On the real-world data sets, although the curves hold the much larger fluctuations which might be caused by both the combinations of various drift types and the frequent data drift, we still observe that the proposed strategies performs significantly better than the baseline strategy. That means the proposed *slide time window*-based strategies are virtually useful for missing data imputation in the drifting data stream.
- For the baseline strategy, the imputation error relies on how much the distribution changes. The larger the difference between the current distribution and the original distribution is, the lower imputation quality the baseline strategy provides.
- When we compare the AS and WAS strategies, we clearly find that they may be suitable for combining different MVI algorithms. For MI algorithm, its combination with WAS generally outperforms the combination between MI and AS. While for KNNI algorithm, its combination with AS performs better than that with WAS. We think the reasons lie in that for MI, all non-missing instances in the window participate in the MVI procedure, hence when we adopt WAS strategy, it would not neglect the old experience, and meanwhile highlight the newest experience. However, KNNI only select a few instances the impute a missing value, thus when the order numbers among the extracted KNN instances have a large difference, the missing value imputation would only rely on the newest neighbor instance, which can bias the imputation results. Therefore, we can say that AS strategy is more appropriate for combining with the global MVI algorithms, while WAS is more suitable for integrating with the local MVI approaches. As for BPCAI, it can't adapt the weighting strategy, hence it can only be integrated with the AS strategy.

Comparison with Time Series Missing Data Imputation Strategies

Next, we consider the proposed scenario is very similar with time series [28], and the difference between these two scenarios lies in that time series generally has a stronger dependence in the context. Therefore, we also compare the proposed strategy with a popular time series missing data imputation algorithm, i.e., local weighted regression (LR) [29]. That means the data stream can be seen as a successive time series. Here, we adopt two different local weighted regression strategies, one constructs weighted regression with all history data, which can be seen as the traditional local weighted regression algorithm, and the other conducts the same weighted rule as the proposed WAS strategy in the slide time window. We call the former as LR-ALL, and the latter as LR-STW. Specifically, for LR-STW, the same length of slide time window as that in the default experiments is used to guarantee the impartiality of the comparison experiments. The comparison results are presented in Fig. 4.

The results in Fig. 4 show that LR strategy is actually effective for dealing with the missing data imputation issue in non-stationary data stream. In fact, LR-STW strategy can be considered to be equivalent to our proposed WAS strategy, which is a combination of least-squares imputation algorithm and the WAS framework. While for LR-ALL strategy, although it has presented the better performance, especially on reoccurring data stream, it violates *one pass* rule, thus it is impractical in real-world applications.

Table 6 Variation of RMSE of AS-MI with the variation of the size of the slide time window

Data Set	The size of slide time window				
	10	20	30	40	50
Sudden	**0.0860**	0.0863	0.0880	0.0903	0.0930
Reoccurring	**0.0891**	0.0921	0.0968	0.1017	0.1064
Incremental	0.0842	**0.0828**	0.0830	0.0836	0.0844
Electricity	**0.0534**	0.0752	0.0855	0.0857	0.0801
Weather	**0.0632**	0.0645	0.0669	0.0698	0.0734

The best results have been highlighted in boldface

Table 7 Variation of RMSE of WAS-MI with the variation of the size of the slide time window

Data Set	The size of slide time window				
	10	20	30	40	50
Sudden	0.0871	**0.0852**	0.0857	0.0869	0.0883
Reoccurring	0.0888	**0.0885**	0.0907	0.0936	0.0967
Incremental	0.0858	0.0826	**0.0821**	**0.0821**	0.0827
Electricity	**0.0387**	0.0574	0.0690	0.0754	0.0774
Weather	**0.0608**	0.0619	0.0633	0.0649	0.0667

The best results have been highlighted in boldface

Table 8 Variation of RMSE of AS-KNNI with the variation of the size of the slide time window

Data Set	The length of slide time window			
	20	30	40	50
Sudden	0.0879	0.0878	**0.0869**	0.0871
Reoccurring	0.0898	**0.0896**	0.0897	0.0898
Incremental	**0.0871**	0.0877	0.0882	0.0890
Electricity	**0.0510**	0.0574	0.0603	0.0638
Weather	0.0519	0.0505	**0.0503**	0.0507

The best results have been highlighted in boldface

Discussions About the Slide Time Window Size

Furthermore, we also investigate the impact of an important parameter adopted by the proposed strategies, i.e., the size of the slide time window, for the MVI results. For MI and BPCA, we change the size of the slide time window from 10 to 50 with an increment of 10. While for KNNI, to avoid the excessively small window to find the insufficient nearest neighbors with non-missing values, further causing a failed feedback, we promote the floor level of the slide time window and designate the length to vary from 20 to 50 with an increment of 10.

The average results of 50 times' random runnings for AS-MI, WAS-MI, AS-KNNI, WAS-KNNI and AS-BPCAI with different sizes of the slide time window can be observed in Tables 6, 7, 8, 9 and 10. The variation tendencies can be observed in Fig. 4 in more detail.

Table 9 Variation of RMSE of WAS-KNNI with the variation of the size of the slide time window

Data Set	The length of slide time window			
	20	30	40	50
Sudden	**0.0974**	0.0991	0.1001	0.1013
Reoccurring	**0.1031**	0.1054	0.1066	0.1083
Incremental	**0.0937**	0.0947	0.0951	0.0965
Electricity	**0.0688**	0.0739	0.0771	0.0775
Weather	**0.0700**	0.0706	0.0722	0.0748

The best results have been highlighted in boldface

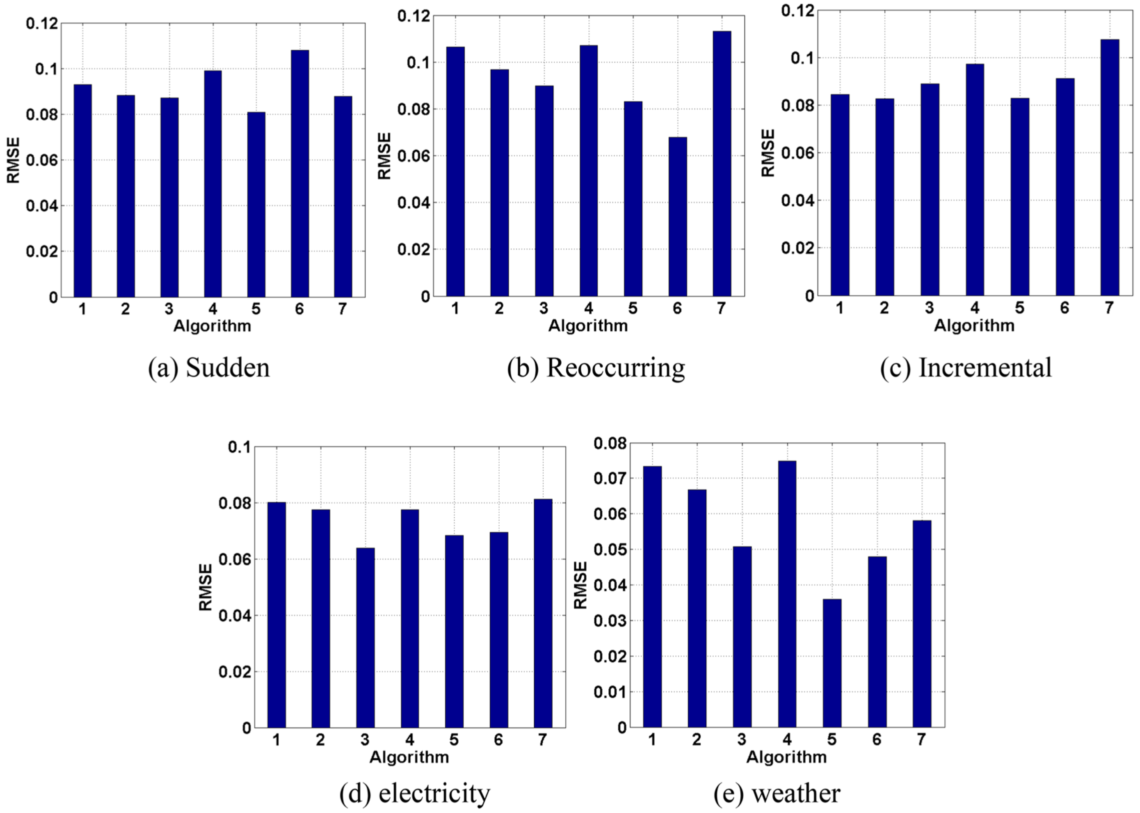
Table 10 Variation of RMSE of AS-BPCAI with the variation of the size of the slide time window

Data set	The size of slide time window				
	10	20	30	40	50
Sudden	0.0841	0.0820	0.0806	**0.0805**	0.0808
Reoccurring	0.0861	0.0836	**0.0828**	0.0829	0.0830
Incremental	0.0848	0.0824	**0.0820**	0.0825	0.0829
electricity	**0.0332**	0.0426	0.0529	0.0634	0.0683
weather	0.0506	0.0417	0.0387	0.0370	**0.0360**

The best results have been highlighted in boldface

The results in Tables 6, 7, 8, 9 and 10 show that the size of the slide time window is not suitable for being assigned an excessively large value. For AS-MI, the best MVI results appear when the length of the slide time window equals to be 10 or 20. For WAS-MI, the best choice of the length is between 10 and 40. For AS-KNNI, the length is suitable for being set to be from 20 to 40. For WAS-KNNI, all best results emerge when the length equals to 20. As for AS-BPCAI, the best results seem to be not constant.

The findings above seem to be counterintuitive, as in our opinion, more instances can provide more reliable experience for us to provide the accurate imputation. However, when we observe the variation tendency of RMSE in Fig. 5, we can find some cues to explain the phenomenon. Obviously, when concept drift just happens, the old experience will misguide the imputation, then with the increase of the experience from the new distribution, more useful contextual information can be used to accurately impute the missing values. At the boundary of concept drift, the old and new experiences conflict with each other. At this moment, if the size of the time window is small enough, then it can faster adapt the new distribution. Of course, for the stationary data stream, a short time window must take a risk of decrease of the generalization ability as well. In our experiments,



(a) Sudden

(b) Reoccurring

(c) Incremental

(d) electricity

(e) weather

Fig. 4 The comparison of RMSE of seven different algorithms on five different data streams, where 1–7 indicate 1. AS-MI, 2. WAS-MI, 3. AS-KNNI, 4. WAS-KNNI, 5. AS-BPCAI, 6. LR-ALL and 7. LR-STW, respectively

the concept drift generally happens in an interval of 50 instances, and meanwhile, the distribution of the instances is relatively simple, hence a small size of the slide time window can meet the demand of accurate imputation. However, in real-world applications, considering the difference of drift frequence and characteristic of the specific MVI algorithm, this parameter must be set carefully.

Conclusions

This paper explores a novel issue, i.e., online missing value imputation in non-stationary data stream. To our best knowledge, it is the first study which focus on this issue. In this study, we found that the data stream with concept drifts means the variation of data distribution, which will substantially damage the MVI procedure and decrease the MVI quality. To address this problem, we proposed two slide time window-based strategies, and combined them with several existing MVI algorithms. Actually, the strategies are based on a basic hypothesis, i.e., the newest received instances can better reflect the current data distribution. That is to say, we used the proposed strategies to track and adapt the distribution for providing better imputed results.

In our experiments, the strategies were combined with three popular MVI algorithms, mean imputation, KNN imputation and Bayesian principal component analysis imputation. The results show that the combination of the traditional MVI algorithms and our proposed strategies significantly outperforms the static imputation algorithms, further verifying its effectiveness and feasibility.

This paper is only the initial exploration for the issue of online MVI. In future work, we expect to study the issue deeply, and investigate the possibility of directly using the more sophisticated incremental MVI algorithm with the forgetting mechanism, which can adapt the concept drift to impute the missing stream data. Moreover, how to utilize the distribution information to determine the size of window will be investigated as well.

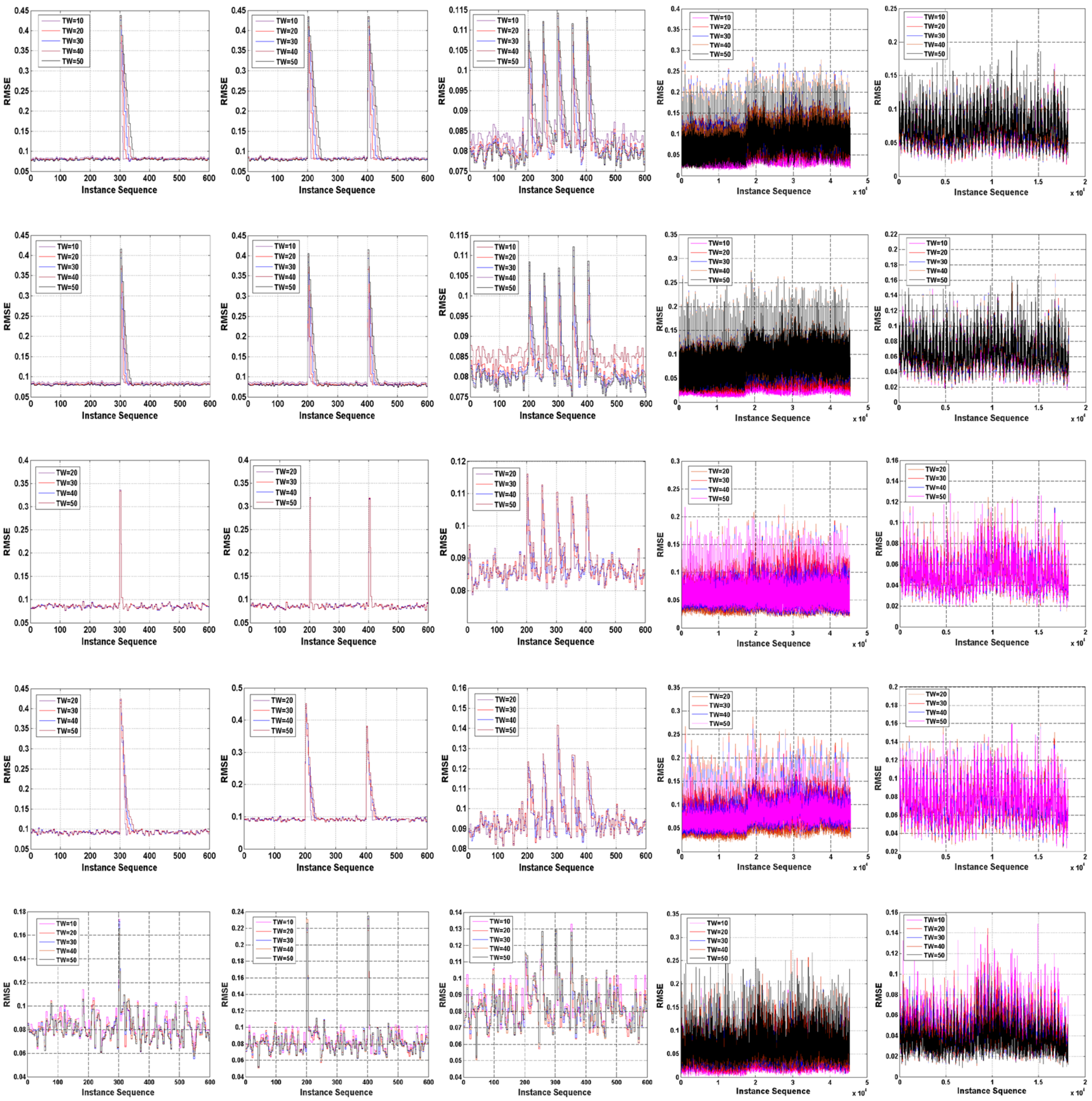


Fig. 5 The variation tendency of various MVI algorithms based on different sizes of the slide time window. From top to down, five rows correspond to AS-MI, WAS-MI, AS-KNNI, WAS-KNNI and AS-BPCAI, respectively. From left to right, five columns correspond to Sudden, Reoccurring, Incremental, electricity and weather data sets, respectively

Acknowledgements This work was supported by Natural Science Foundation of Jiangsu Province of China under Grant No. BK20191457, and Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grants No. KYCX20_3147, Open Project of Artificial Intelligence Key Laboratory of Sichuan Province under Grant No. 2019RYJ02, Nature Science Foundation of the Jiangsu Higher Education Institute of China under grant No.18KJB520050, National Natural Science Foundation of China under Grants No. 61305058 and No. 61572242, China Postdoctoral Science Foundation under Grants No. 2013M540404 and No. 2015T80481, and Jiangsu Province 333 Project.

Compliance with Ethical Standards

Conflict of Interest The authors have declared that no conflict of interest exists.

References

1. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recogn. 2008;41(12):3692–705.
2. Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2019. https://doi.org/10.1007/s10462-019-09709-4.
3. Brown ML, Kros JF. Data mining and the impact of missing data. Industr Manag Data Syst. 2003;103(8):611–21.
4. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087–91.
5. Little RJ, Rubin DB. Statistical analysis with missing data. 3rd ed. Wiley John & Sons; 2019.

6. Dixon JK. Pattern recognition with partly missing data. IEEE Trans Syst Man Cybern. 1979;9(10):617–21.
7. Tsai CF, Chang FY. Combining instance selection for better missing value imputation. J Syst Softw. 2016;122:63–71.
8. Rahman MG, Islam MZ. Missing value imputation using decision trees and decision forests by splitting and merging records: two novel techniques. Knowl-Based Syst. 2013;53:51–65.
9. Sefidian AM, Daneshpour N. Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. Expert Syst Appl. 2019;115:68–94.
10. Zhu X, Zhang S, Jin Z, Zhang Z, Xu Z. Missing value estimation for mixed-attribute data sets. IEEE Trans Knowl Data Eng. 2010;23(1):110–21.
11. García-Laencina PJ, Abreu PH, Abreu MH, Afonoso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput Biol Med. 2015;59:125–33.
12. Purwar A, Singh SK. Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl. 2015;42(13):5621–31.
13. Abawajy J, Kelarev A, Chowdhury M, Stranieri A, Jelinek HF. Predicting cardiac autonomic neuropathy category for diabetic data with missing values. Comput Biol Med. 2013;43(10):1328–33.
14. Grittner U, Gmel G, Ripatti S, Bloomfield K, Wicki M. Missing value imputation in longitudinal measures of alcohol consumption. Int J Methods Psychiatr Res. 2011;20(1):50–61.
15. Wang A, Chen Y, An N, Yang J, Li L, Jiang L. Microarray missing value imputation: a regularized local learning method. IEEE/ACM Trans Comput Biol Bioinf. 2018;16(3):980–93.
16. Hossain A, Chattopadhyay M, Chattopadhyay S, Bose S, Das C. A bicluster-based sequential interpolation imputation method for estimation of missing values in microarray gene expression data. Curr Bioinform. 2017;12(2):118–30.
17. Oba S, Sato MA, Takemasa I, Monden M, Matsubara KI, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003;19(16):2088–96.
18. Farswan A, Gupta A, Gupta R, Kaur G. Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. Front Oncol. 2020;9:1442.
19. Roth PL. Missing data: a conceptual review for applied psychologists. Pers Psychol. 1994;47(3):537–60.
20. Di Nuovo AG. Missing data analysis with fuzzy c-means: a study of its application in a psychological scenario. Expert Syst Appl. 2011;38:6793–7.
21. Deb R, Liew AWC. Missing value imputation for the analysis of incomplete traffic accident data. Inf Sci. 2016;339:274–89.
22. Sun Y, Tang K, Minku LL, Wang S, Yao X. Online ensemble learning of data streams with gradually evolved classes. IEEE Trans Knowl Data Eng. 2016;28(6):1532–45.
23. Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M. Ensemble learning for data stream analysis: a survey. Inf Fus. 2017;37:132–56.
24. Kim HG, Park YH, Cho YH, Kim MH. Time-slide window join over data streams. J Intell Inf Syst. 2014;43(2):323–47.
25. Brzezinski D, Stefanowski J. Reacting to different types of concept drift: the accuracy updated ensemble algorithm. IEEE Trans Neural Netw Learn Syst. 2013;25(1):81–94.
26. Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing concept drift. Data Min Knowl Disc. 2016;30(4):964–94.
27. Yu H, Webb GI. Adaptive online extreme learning machine by regulating forgetting factor by concept drift map. Neurocomputing. 2019;343:141–53.
28. Andiojaya A, Demirhan H. A bagging algorithm for the imputation of missing values in time series. Expert Syst Appl. 2019;129:10–26.
29. Conti PL, Marella D, Scanu M. Evaluation of matching noise for imputation techniques based on nonparemetric local linear regression estimators. Comput Stat Data Anal. 2008;53(2):354–65.

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