EURASIP Journal on Advances
in Signal Processing

RESEARCH Open Access

TLGRU: time and location gated recurrent unit for multivariate time series imputation

Ruimin Wang[1], Zhenghui Zhang[2], Qiankun Wang[1] and Jianzhi Sun[1*]

*Correspondence:
sunjz@th.btbu.edu.cn

[1] College of Computer Science, Beijing Technology and Business University, Beijing 100048, China
[2] Beijing Zhongzhao United Bidding and Purchasing Network, Beijing 100048, China

Abstract

Multivariate time series are widely used in industrial equipment monitoring and maintenance, health monitoring, weather forecasting and other fields. Due to abnormal sensors, equipment failures, environmental interference and human errors, the collected multivariate time series usually have certain missing values. Missing values imply the regularity of data, and seriously affect the further analysis and application of multivariate time series. Conventional imputation methods such as statistical imputation and machine learning-based imputation cannot learn the latent relationships of data and are difficult to use for missing values imputation in multivariate time series. This paper proposes a novel Time and Location Gated Recurrent Unit (TLGRU), which takes into account the non-fixed time intervals and location intervals in multivariate time series and effectively deals with missing values. We made necessary modifications to the architecture of the end-to-end imputation model E^2 GAN and replaced Gated Recurrent Unit for Imputation (GRUI) with TLGRU to make the generated fake sample closer to the original sample. Experiments on a public meteorologic dataset show that our method outperforms the baselines on the imputation accuracy and achieves a new state-of-the-art result.

Keywords: Multivariate time series, Missing values imputation, Machine learning, GRUI, GAN

1 Introduction

Time series data refers to data arranged in chronological order, reflecting the state changes of things over time. Typical time series include industrial data, medical data, meteorological data, stock data and traffic data, etc. Due to the complexity of the scene, the collected time series are usually multivariate time series data with diverse features and changing patterns.

Multivariate time series are widely used in industrial equipment monitoring and maintenance, health monitoring, weather forecasting, stock price forecasting and other fields. Due to abnormal sensors, equipment failures, environmental interference and human errors, the collected multivariate time series usually have certain missing values. Missing values imply the regularity of data, and seriously affect the further analysis and application of multivariate time series.

Among the related methods of missing values imputation, the methods, such as Mean, Median, Mode and Last Observed are easy to operate, which makes it difficult to restore the real data attributes and the imputation effect is general [1]. The Regression Imputation [2] is prone to random errors, resulting in large fluctuations in the imputation effect. *k*-Nearest Neighbor (KNN) [3], Clustering [4], Expectation Maximization (EM) [5, 6], and Multiple Imputation (MI) have high computational complexity and low efficiency, which makes it difficult to impute multivariate time series. The imputation methods Recurrent Neural Networks (RNN) [7]-based can learn the latent relationships and regularity of the time series and have been used to impute missing values.

Recently, Generative Adversarial Network (GAN) [8]-based imputation methods have gradually become a research hotspot and have achieved better results in the field of missing values imputation in multivariate time series. In order to learn the latent relationships between observations with non-fixed time intervals, Luo et al. [9] proposed a novel RNN cell called Gated Recurrent Unit for Imputation (GRUI), which can take into account the non-fixed time intervals and fade the influence of the past observations determined by the time intervals. Based on GRUI, Luo et al. [10] further proposed an end-to-end GAN-based imputation model E^2 GAN which consists of a generator and a discriminator. After adversarial training of the generator and the discriminator, the generator can generate complete time series that fits the distribution of the original dataset and is used to impute the missing values. E^2 GAN achieved a better imputation accuracy, however, GRUI only considers the time interval information between two observations of missing time series, ignoring the equally important location interval information.

The contribution of this paper is to make full use of the time interval and location interval information between observations of missing time series based on GRUI, we propose a novel GRU cell called Time and Location Gated Recurrent Unit (TLGRU). The experiments on a real meteorologic dataset show that our method achieves a new state-of-the-art imputation accuracy with similar time efficiency to GRUI.

2 Related works

The research on missing values imputation methods has received extensive attention from researchers, and various theoretical methods have been proposed in the industry. In the statistics-based imputation methods, Kantardzic [11] tried to impute missing values by mean value. Purwar et al. [12] used mode to impute missing values. Amiri et al. [13] used last observation to impute missing values for incomplete data. The imputation methods statistics based does not consider the characteristics of the missing values, the imputation result is affected by the observed values and the imputation accuracy is poor.

In the machine learning-based imputation methods, Hudak et al. [14] used the mean value of *k* nearest neighbors to impute missing values. White et al. [15] proposed the Multiple Imputation by Chained Equations (MICE) to impute the missing values by using an iterative regression model. Hastie et al. [16] proposed an imputation method based on Matrix Factorization (MF), which treats the original dataset as a matrix, and decomposes the original matrix into the product of two matrices using algorithms such as Principal Component Analysis (PAC), and finally imputes the missing values with the product result. Ogbeide [5] proposed a Mode-Related Expectation Adaptive Maximization (MEAM) for obtaining better statistical inferences from multivariate data with missing observations. The method produces initial values closest to the mean of the complete dataset and reduces computation time when solving problems such as missing survey observations, non-response, or missing data. Dzulkalnine et al. [17] proposed a hybrid FPCA-Support Vector Machines-FCM (FPCA-SVM-FCM) imputation method. The feature selection method used in this method is Fuzzy Principal Component Analysis (FPCA), which identifies relevant features in the dataset while considering outliers. The selected features are then classified and irrelevant features are removed using the SVM. After identifying the significant features in the dataset, the missing data is then estimated by Fuzzy c-Means (FCM). Machine learning-based missing values imputation methods are usually computationally complex, inefficient, and unable to learn the latent relationships in time series, making it difficult to impute missing values in multivariate time series.

There are also many RNN-based imputation methods in the field of multivariate time series. Berglund et al. [7] proposed two probabilistic interpretations of bidirectional recurrent neural networks that can be used to reconstruct missing samples efficiently. Che et al. [18] proposed GRUD, which imputes missing values of clinical dataset with a smooth method. GRUD takes the advantage of last observed value and mean value to represent missing patterns of incomplete time series. Cao et al. [19] proposed Bidirectional Recurrent Imputation for Time Series (BRITS), which directly learns the missing values in a bidirectional recurrent dynamical system, without any specific assumption.

The imputation methods GAN-based seek to generate new samples that obey the distribution of the training dataset, have been used to impute missing values, and achieved high imputation accuracy. Yoon et al. [20] proposed Generative Adversarial Imputation Nets (GAIN), which uses a hint vector that is conditioned on what we actually observed to impute missing values. GAIN made tremendous advances in data imputation. Shang et al. [21] proposed a GAN-based missing values imputation algorithm for multimodal data, which can learn the common properties of multimodal data and impute missing values in certain modal data. Luo et al. [10, 22] proposed E^2 GAN, which takes a compressing and reconstructing strategy to automatically learns internal representations of the time series and tries its best to reconstruct this temporal data. E^2 GAN also improves the imputation performance by getting a better feature representation of samples, which contributes to better reconstructed samples and improves the imputation. Optimization of E^2 GAN was achieved by Zhang et al. [23] using real data during the training of the generator to force the imputed values to be close to the real ones.

3 Problem formulation

Given a d-dimensional multivariate time series X, observed at T=$(t_0, t_1, \cdots, t_{n-1})$ and L =$(l_0, l_1, \cdots, l_{n-1})$, is denoted by X=$(x_0, x_1, \cdots, x_{n-1}) \in \mathbb{R}^{d \times n}$, where T is the observing timestamp and L is the observing locationstamp, x_i is the ith observation of X, and x_i^j is the jth dimension of x_i.

Suppose that d-dimensional time series X is incomplete, the $M \in \mathbb{R}^{d \times n}$ is a mask matrix that takes values in $\{0, 1\}$. M means whether the values of X exist or not, if x_i^j exists, m_i^j=1, otherwise, m_i^j=0.

In the following is an example of a 3-dimensional multivariate time series X and its corresponding$M$, T and$L$. "/" is missing value.

$$X = \begin{bmatrix} 1 & 4 & / & 5 & \\ / & / & 6 & / & \cdots \\ 2 & 3 & 2 & 10 & \end{bmatrix}$$

$$M = \begin{bmatrix} 1 & 1 & 0 & 1 & \\ 0 & 0 & 1 & 0 & \cdots \\ 1 & 1 & 1 & 1 & \end{bmatrix}$$

$$T = (0, 1, 6, 11, \cdots)$$

$$L = (0, 1, 2, 3, \cdots)$$

We define a matrix $\delta_t \in \mathbb{R}^{d \times n}$ that records the time interval between current value and last observed value. The following part shows the calculation and a calculated example of δ_t.

$$\delta^j_{t_i} = \begin{cases} 0, & i == 0 \\ t_i - t_{i-1}, & m^j_{i-1} == 1 \text{ and } i > 0 \\ \delta^j_{t_{i-1}} + t_i - t_{i-1}, & m^j_{i-1} == 0 \text{ and } i > 0 \end{cases} \tag{1}$$

$$\delta_t = \begin{bmatrix} 0 & 1 & 5 & 10 & \\ 0 & 1 & 6 & 5 & \cdots \\ 0 & 1 & 5 & 5 & \end{bmatrix}$$

We define a matrix $\delta_l \in \mathbb{R}^{d \times n}$ that records the location interval between current value and last observed value. The following part shows the calculation and a calculated example of δ_l.

$$\delta^j_{l_i} = \begin{cases} 0, & i == 0 \\ l_i - l_{i-1}, & m^j_{i-1} == 1 \text{ and } i > 0 \\ \delta^j_{l_{i-1}} + l_i - l_{i-1}, & m^j_{i-1} == 0 \text{ and } i > 0 \end{cases} \tag{2}$$

$$\delta_l = \begin{bmatrix} 0 & 1 & 1 & 2 & \\ 0 & 1 & 2 & 1 & \cdots \\ 0 & 1 & 1 & 1 & \end{bmatrix}$$

4 Approach

In this part, we show the details of the TLGRU and the method E^2 GAN-based for multivariate time series missing values imputation. The overall architecture of the proposed method is shown in Fig. 1. We replaced GRUI with TLGRU in the architecture of E^2 GAN and achieved a new state-of-the-art imputation accuracy.

The imputation method consists of a generator (G) and a discriminator (D). The generator is composed of an auto-encoder and recurrent neural networks. We take a compressing and reconstructing strategy to compress the input incomplete time series X into a low-dimensional vector z by the encoder. Then we use vector z to reconstruct a complete time series $X^{'}$ by the decoder. The discriminator tries to distinguish the original

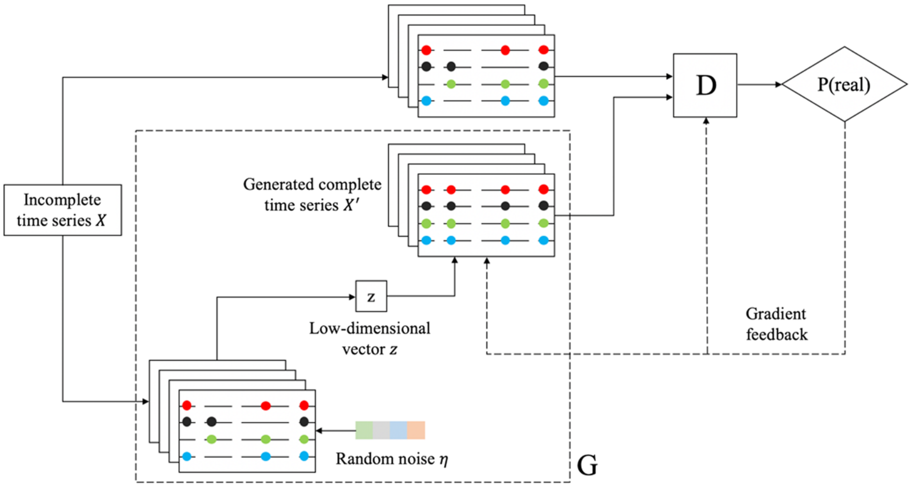
D
P(real)
Incomplete time series X
Generated complete time series X'
z
Low-dimensional vector z
Gradient feedback
Random noise η
G

Fig. 1 The architecture of the proposed method

incomplete time series X and the fake but complete sample $X^{'}$. After the adversarial training, the generator generates complete time series $X^{'}$ that can fool the discriminator, and the discriminator can best determine the authenticity of $X^{'}$.

Traditional GAN is difficult to maintain long-term stable training and is prone to mode collapse. Arjovsky et al. [24] proposed the Wasserstein GAN (WGAN), which can improve learning stability and get away from the problem of mode collapse. In our method, we use WGAN instead of GAN. The following are the loss functions of WGAN.

$$L_G = \mathbb{E}_{z\sim P_g}[-D(G(z))] \tag{3}$$

$$L_D = \mathbb{E}_{z\sim P_g}[D(G(z))] - \mathbb{E}_{x\sim P_r}[D(x)] \tag{4}$$

4.1 Time and location gated recurrent unit

Multivariate time series have certain latent relationships and regularity between adjacent observations in the same dimension and observations in different dimensions. When imputing multivariate time series missing values, not only the relationships between missing values and observations of the same dimension, but also the relationships between missing values and observations of different dimensions should be considered. Most of the current missing values imputation methods lack consideration of the relationships between observations and are difficult to be used for imputing missing values.

In multivariate time series, due to the existence of missing values, two adjacent observations have non-fixed time intervals and location intervals. If the data in one dimension is missing continuously for a long time, the time interval and location interval between two valid observations in that dimension will be larger than in other dimensions. The GRUI decreases the memory of the Gated Recurrent Unit (GRU) by introducing the time interval matrix. The TLGRU is improved based on the GRUI. We consider the time interval and location interval information between observations, and we introduce a decay vector β to decrease the memory of GRU. The following is the update function of β.

$$\beta_i = \frac{1}{e^{\max(0, w_\beta(\alpha_t\delta_{t_i} + \alpha_l\delta_{l_i}) + b_\beta)}} \tag{5}$$

where δ_t, δ_l are the time interval matrix and location interval matrix, and the hyper-parameters α_t, α_l are the time weight and location weight. The values of α_t, α_l are determined by the principle of random initialization and by combining a large number of experiments. w_β, b_β are training parameters. The formulation of β guarantees that with the increase in time interval matrix δ_t and location interval matrix δ_l, the value of β decreases. The smaller δ_t and δ_l, the bigger β. This formulation also make sure that $\beta \in (0, 1]$.

The TLGRU of proposed method is shown at Fig. 2. The decay vector β is a core part of the TLGRU. Before each TLGRU iteration, we update the hidden state h_{i-1} by decay vector β. The following are the update functions of the TLGRU.

$$h'_{i-1} = \beta_i \odot h_{i-1} \tag{6}$$

$$z_i = \sigma(W_z[h'_{i-1}, x_i] + b_z) \tag{7}$$

$$r_i = \sigma(W_r[h'_{i-1}, x_i] + b_r) \tag{8}$$

$$\widetilde{h}_i = tanh(W_{\widetilde{h}}[r_i \odot h'_{i-1}, x_i] + b_{\widetilde{h}}) \tag{9}$$

$$h_i = (1 - z_i) \odot h'_{i-1} + z_i \odot \widetilde{h}_i \tag{10}$$

where z is update gate, r is reset gate, $\widetilde{h}$ is candidate hidden state, h is current hidden state, σ is the sigmoid activation function, $\odot$ is an element-wise multiplication, W_z, b_z, W_r, b_r, $W_{\widetilde{h}}$, $b_{\widetilde{h}}$ are training parameters.

4.2 Generator architecture

The generator of the proposed method is shown in Fig. 3. The generator is an auto-encoder based on the TLGRU cell, including an encoder and a decoder. The generator can not only compress the incomplete time series X into a low-dimensional vector z by the encoder, but also reconstruct the complete time series X' from z by the decoder. Different from traditional auto-encoder, we just add some noise to destroy original samples rather than drop out some values. The random noise η is sampled from a standard distribution $\mathcal{N}(0, 0.01)$, and can avoid the loss of data information in traditional auto-encoder and reduce over-fitting to a certain extent. The following are the update functions of denoising auto-encoder.

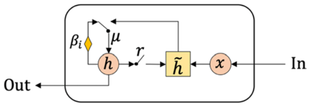

Fig. 2 TLGRU cell

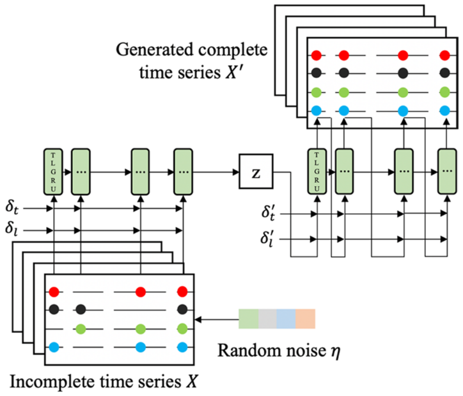

Fig. 3 The architecture of the generator. The generator is a denoising auto-encoder which is mainly composed by the TLGRU cell

$$z = Encoder(X + \eta) \tag{11}$$

$$X^{'} = Deconder(z) \tag{12}$$

Since both the encoder and the decoder use the TLGRU cell to process multivariate time series, we need input corresponding time interval matrix δ_t, $\delta_t^{'}$, and location interval matrix δ_l, $\delta_l^{'}$ in the process of multivariate time series compression and reconstruction. The δ_t and δ_l represent the time interval and location interval of the original incomplete time series. The $\delta_t^{'}$ and $\delta_l^{'}$ represent the time interval and location interval of the reconstructed complete time series.

The generator tries to produce a new sample $X^{'}$ that is most similar to X, we add a squared error loss to the loss function of the generator. The following is the loss function of the generator, where λ is a hyper-parameter that controls the weight of the discriminative loss and the squared error loss.

$$G_L = -D(X^{'}) + \lambda \| X \odot M - X^{'} \odot M \|_2 \tag{13}$$

First, we use zero value to replace the missing values of X at the input stage of TLGRU. Then we feed the TLGRU cell with the incomplete time series X and its interval matrix δ_t and δ_l. After recurrent processing of the input time series, the last hidden state of the recurrent neural network will flow to a fully connected layer. The output of this fully connected layer is the compressed low-dimensional vector z.

Next, we take z as the initial input of another fully connected layer. Then we use this output as the first input of another TLGRU cell. The current output of this TLGRU cell will be fed into the next iteration of the same TLGRU cell. At the final stage, we combine all the outputs of this TLGRU cell as the generated complete time series $X^{'}$.

4.3 Discriminator architecture

The discriminator is composed of TLGRU cells and a fully connected layer. The task of the discriminator is to distinguish between fake complete time series $X^{'}$ and true incomplete time series X. The output of the discriminator is a probability that indicates the degree of authenticity. We try to find a set of parameters that can produce a high probability when we feed true incomplete time series X, and produce a low probability when we feed fake complete time series $X^{'}$. The following is the loss function of the discriminator.

$$D_L = -D(X) + D(X^{'}) \tag{14}$$

With the help of the TLGRU cell, the multivariate time series can be successfully handled. The last hidden state of the TLGRU cell is fed into one fully connected layer that outputs the p of being true. We also use the sigmoid function to make sure that $p \in (0, 1)$.

4.4 Imputation

For each true incomplete time series X, we try to map it into a low-dimensional vector z and reconstruct a fake complete time series $X^{'}$ from z, so that the fake time series $X^{'}$ is most close to the X. We use the corresponding values of $X^{'}$ to impute in the missing values of X. The imputation formula can be summarized as follows.

$$X_{imputed} = M \odot X + (1 - M) \odot X^{'} \tag{15}$$

5 Experiments

In this part, we will present the dataset and experiment results. In order to facilitate comparison with E^2 GAN, we also selected a meteorologic dataset as the experimental dataset. The experiments on the dataset show that our method achieves a new state-of-the-art imputation accuracy.

5.1 Dataset

The KDD dataset is a public meteorologic dataset that comes from the KDD CUP Challenge 2018. The KDD dataset contains air quality and weather data which is hourly collected between 2017/1/30 to 2018/1/30 in Beijing. The records have a total of 12 variables which include PM2.5(ug/m^3), PM10(ug/m^3), CO(mg/m^3), weather, temperature, and so on. One task of the KDD dataset is the imputation accuracy task. We selected 11 common air quality and weather data observatories for our experiments. We first performed the operation of randomly dropping out 30% of records for all observatories to obtain an experimental dataset with non-fixed collection timestamps. We then randomly dropped out p percent on the variables of the experimental dataset, where $p \in \{10, 20, \cdots, 80\}$. Finally, we imputed these time series and calculated the imputation accuracy by the mean squared error (MSE) between original values and imputed values.

5.2 Training settings

5.2.1 Network details

We performed one task on a real public meteorologic dataset. For the KDD dataset, the input dimension is 132, the batch size is 16, the hidden unit number of all TLGRU cells is 64, and the dimension of the low-dimensional vector z is 128.

5.2.2 Baseline methods

We adopted different imputation methods to carry out experiments on the KDD dataset, the following is an introduction of the methods.

- Median: We use median value to impute missing values simply.
- Mean: We use mean value to impute missing values simply.
- MF: MF imputation is used to factorize the incomplete matrix into low-rank matrices and impute the missing values.
- KNN: The missing values are imputed by using k nearest neighbor samples.
- ISVD: IterativeSVD can impute the missing values by iterative low-rank SVD decomposition.
- GAIN: GAIN is a GAN-based imputation method that uses a hint vector to impute the missing values.
- E^2 GAN [10]: E^2 GAN is an end-to-end GAN-based imputation model.

5.3 Results and discussions

5.3.1 Experimental results on KDD dataset

Table 1 is the imputation results on the KDD dataset by using the proposed method and other baseline methods such as Median imputation, Mean imputation, KNN imputation, MF imputation, ISVD imputation, GAIN imputation, and E^2 GAN imputation. The first column of Table 1 is the missing rate which indicates how many percent values are dropped, and the other columns are MSE. The parameters of our method on the KDD dataset are: pretrain epoch is 10, epoch is 15, learning rate is 0.005, λ takes values in $\{0.5, 1, 5, 10, 20, 40\}$, the values of α_t and α_l are visible in Table 1. We can see that in all cases, our method is one of the best methods and wins others methods in most cases.

To further compare the performance of the proposed method with E^2 GAN model, we evaluated the effects of the discriminator and random noise η on both models. Table 2

Table 1 The MSE results of the proposed method and other imputation methods on the KDD dataset. In most cases, our method owns the best imputation accuracy

Missing rate	Median	Mean	KNN	MF	ISVD	GAIN	E^2 GAN	Ours
10%	0.519	0.386	0.344	0.361	0.356	0.359	**0.334**	α_t=1,α_l=0.5 0.336
20%	0.627	0.522	0.538	0.526	0.544	0.523	0.517	α_t=0.8,α_l=0.3 **0.510**
30%	0.683	0.608	0.628	0.621	0.620	0.616	0.584	α_t=0.3,α_l=0.3 **0.578**
40%	0.755	0.659	0.682	0.681	0.708	0.663	0.641	α_t=1,α_l=0.2 **0.637**
50%	0.816	0.735	0.733	0.760	0.740	0.714	**0.694**	α_t=0.5,α_l=0.3 0.697
60%	0.821	0.736	0.734	0.727	0.714	0.724	0.705	α_t=0.4,α_l=0.2 **0.702**
70%	0.855	0.776	0.807	0.766	0.759	0.768	0.745	α_t=0.6,α_l=0.2 **0.741**
80%	0.883	0.816	0.831	0.792	0.824	0.787	0.756	α_t=1,α_l=0.8 **0.747**

Bold values indicates the minimum MSE at different missing rates shows the results of the MSE on the KDD dataset. The first row is the missing rate. The second and third rows show the results for the proposed method and E^2 GAN. The fourth and fifth rows show the results for both models without the discriminator. The sixth and seventh rows show the results for both models without the random noise η.

5.3.2 Discussions

The experimental results on the KDD dataset with different percentage missing rates are shown in Table 1. We can see that in the vast majority of cases, the proposed method achieved the smallest MSE compared to other baseline methods. In the baseline methods for imputing multivariate time series, the methods GAN based, such as GAIN and E^2 GAN have a better performance than Median, Mean, KNN, MF, and ISVD. And further, the methods GRUI based, such as E^2 GAN can take into account the non-fixed time interval and fade the influence of the past observations determined by the time interval matrix, improving the imputation accuracy effectively. In the proposed method, we optimized GRUI by introducing the location interval matrix. The weights of the time interval matrix and location interval matrix are controlled by introducing the hyper-parameters α_t and α_l. Our proposed method wins the new state-of-the-art imputation accuracy for all percentages except 10% and 50%. We can also see that the choice of values for the hyper-parameters α_t and α_l will affect the imputation accuracy. We selected different hyper-parameters and conducted a large number of experiments, and finally obtained the experimental results in Table 1. However, the values of the hyper-parameters in Table 1 may not be the optimal results, and we only use these values to improve the imputation accuracy. The specific numerical selection of hyper-parameter needs further research.

The effects of the discriminator and random noise η on the proposed method and E^2 GAN are shown in Table 2. As we can see, the discriminator and random noise η have an impact on the imputation accuracy of both models. In particular, the proposed method outperforms E^2 GAN in the vast majority of cases in all three experiments.

6 Conclusion

In order to learn the latent relationships between observations with non-fixed time intervals and location intervals in multivariate time series, we propose a novel TLGRU cell for dealing with missing values. We made necessary modifications to the architecture of the end-to-end missing values imputation model E^2 GAN by replacing GRUI with TLGRU to make the generated fake sample closer to the original one, and the

Table 2 The effects of the discriminator and random noise η on the proposed method and E^2 GAN

Method/Missing	10%	20%	30%	40%	50%	60%	70%	80%
Ours	0.336	0.510	0.578	0.637	0.697	0.702	0.741	0.747
E^2 GAN	0.334	0.517	0.584	0.641	0.694	0.705	0.745	0.756
Ours-no-D	0.354	0.514	0.606	0.665	0.727	0.729	0.783	0.789
E^2 GAN-no-D	0.357	0.519	0.608	0.671	0.735	0.742	0.784	0.788
Ours-no-noise	0.347	0.517	0.611	0.655	0.723	0.730	0.768	0.782
E^2 GAN-no-noise	0.353	0.518	0.614	0.667	0.736	0.729	0.779	0.791

generated fake but complete sample can be used to impute missing values. Experiments on a public meteorologic dataset show that our method outperforms the baselines on the imputation accuracy and achieves a new state-of-the-art result.

Abbreviations

GRUI	Gated recurrent unit for imputation
E^2 GAN	An end-to-end GAN-based imputation model
TLGRU	Time and location gated recurrent unit
KNN	K-nearest neighbor
EM	Expectation maximization
MI	Multiple imputation
RNN	Recurrent neural networks
GAN	Generative adversarial networks
MICE	Multiple imputation by chained equations
MF	Matrix factorization
PAC	Principal component analysis
MEAM	Mode-related expectation adaptive maximization
FPCA-SVM-FCM	FPCA-support vector machines-FCM
FPCA	Fuzzy principal component analysis
SVM	Support vector machines
FCM	Fuzzy c-means
BRITS	Bidirectional recurrent imputation for time series
GAIN	Generative adversarial imputation nets
WGAN	Wasserstein GAN
GRU	Gated recurrent unit

Acknowledgements

Not applicable.

Author contributions

All authors have contributed toward this work as well as in compilation of this manuscript. All authors read and approved the final manuscript.

Funding

This research is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61972010.

Availability of data and materials

The original dataset used for this study is available in: KDD CUP 2018 (http://www.kdd.org/kdd2018).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

We agree to the publication of the paper.

Competing interests

The authors declare that they have no competing interests.

Received: 18 July 2022 Accepted: 24 August 2022
Published online: 06 September 2022

References

1. T. Schneider, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. **14**(5), 853–871 (2001)
2. A.N. Baraldi, C.K. Enders, An introduction to modern missing data analyses. J. Sch. Psychol. **48**(1), 5–37 (2010)
3. H. De Silva, A. S. Perera. Missing data imputation using evolutionary k-Nearest neighbor algorithm for gene expression data, in *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2016: 141–146
4. B.M. Patil, R.C. Joshi, D. Toshniwal, Missing value imputation based on k-mean clustering with weighted distance. Commun. Comput. Inf. Sci. **94**, 600–609 (2010)
5. E.M. Ogbeide, A new iterative imputation method based on adaptive expectation maximization. SAU Sci.Tech. J. **3**(1), 133–142 (2018)
6. R. Razavi-Far, B. Cheng, M. Saif et al., Similarity-learning information-fusion schemes for missing data imputation. Knowl. Based Syst. **187**, 104805 (2020)

7. M. Berglund, T. Raiko, M. Honkala et al., Bidirectional recurrent neural networks as generative models. Adv. Neural Inf. Process. Syst. **28**, 856–864 (2015)
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets. Adv. Neural Inf. Process. Syst. 2672–2680 (2014)
9. Y. Luo, X. Cai, Y. Zhang et al., Multivariate time series imputation with generative adversarial networks. Adv. Neural Inf. Process. Syst. **31**, 1601–1612 (2018)
10. Y. Luo, Y. Zhang, X. Cai, et al., E2gan: end-to-end generative adversarial network for multivariate time series imputation, in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019: 3094–3100
11. M. Kantardzic, Data mining: concepts, models, methods, and algorithms. (John Wiley & Sons, 2011)
12. A. Purwar, S.K. Singh, Hybrid prediction model with missing value imputation for medical data. Expert Syst. Appl. **42**(13), 5621–5631 (2015)
13. M. Amiri, R. Jensen, Missing data imputation using fuzzy-rough methods. Neurocomputing **205**, 152–164 (2016)
14. A.T. Hudak, N.L. Crookston, J.S. Evans et al., Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens. Environ. **112**(5), 2232–2245 (2008)
15. I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. **30**(4), 377–399 (2011)
16. T. Hastie, R. Mazumder, J.D. Lee et al., Matrix completion and low-rank SVD via fast alternating least squares. J. Mach. Learn. Res. **16**, 3367–3402 (2015)
17. M.F. Dzulkalnine, R. Sallehuddin, Missing data imputation with fuzzy feature selection for diabetes dataset. SN Appl. Sci. **1**(4), 1–12 (2019)
18. Z. Che, S. Purushotham, K. Cho et al., Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018)
19. W. Cao, D. Wang, J. Li, et al., Brits: bidirectional recurrent imputation for time series. Adv. Neural Inf. Process. Syst. 6776–6786 (2018)
20. J. Yoon, J. Jordon, M. Schaar, Gain: missing data imputation using generative adversarial nets, in *International Conference on Machine Learning*. PMLR, 2018: 5675–5684
21. C. Shang, A. Palmer, J. Sun, et al., VIGAN: missing view imputation with generative adversarial networks, in *2017 IEEE International Conference on Big Data*. IEEE, 2017: 766–775
22. W. Zhang, Y. Luo, Y. Zhang et al., SolarGAN: multivariate solar data imputation using generative adversarial network. IEEE Trans. Sustain. Energy **12**(1), 743–746 (2020)
23. Y. Zhang, B. Zhou, X. Cai et al., Missing value imputation in multivariate time series with end-to-end generative adversarial networks. Inf. Sci. **551**, 67–82 (2021)
24. M. Arjovsky, S. Chintala, et al., Wasserstein generative adversarial networks, in *International Conference on Machine Learning*. PMLR, 2017: 214–223

Publisher's Note

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