计算机科学
COMPUTER SCIENCE

基于双通道回声状态网络的时间序列补全及单步预测

郑伟楠, 於志勇, 黄昉菀

引用本文

郑伟楠, 於志勇, 黄昉菀. 基于双通道回声状态网络的时间序列补全及单步预测[J]. 计算机科学, 2024, 51(3): 128-134.

ZHENG Weinan, YU Zhiyong, HUANG Fangwan. Time Series Completion and One-step Prediction Based on Two-channel Echo State Network [J]. Computer Science, 2024, 51(3): 128-134.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200055

基于双通道回声状态网络的时间序列补全及单步预测

郑伟楠[1]　於志勇[1,2]　黄昉菀[1,2]
1 福州大学计算机与大数据学院　福州 350108
2 福建省网络计算与智能信息处理重点实验室　福州 350108
(2680684101@qq.com)

摘　要　随着物联网的发展,众多传感器采集到大量具有丰富数据相关性的时间序列,为各种数据挖掘应用提供强大的数据支持。然而,一些客观或主观原因(如设备故障、稀疏感知等)往往会造成采集到的数据出现不同程度的缺失。虽然已有很多方法被提出用于解决这一问题,但这些方法在数据相关性方面或考虑不够全面,或计算成本过高。而且,现有方法仅关注对缺失值的补全,未能兼顾下游应用。针对上述不足,设计了一种兼顾补全与预测任务的双通道回声状态网络。两个通道的网络虽共用输入层,但具有各自的储备池和输出层。两者最大的区别是左/右通道的输出层分别表示输入层前/后一个时刻对应的目标值或预补值。最后将两个通道的估计值进行融合,充分利用来自缺失时刻之前和之后的数据相关性以进一步提升性能。两种缺失现象下(随机缺失和分段缺失)不同缺失率的实验结果表明,所提模型无论是在补全精度还是预测精度上都优于目前流行的各类方法。

关键词:数据相关性;时间序列;外生变量;双通道 ESN;缺失补全;单步预测
中图分类号　TP391

Time Series Completion and One-step Prediction Based on Two-channel Echo State Network

ZHENG Weinan[1], YU Zhiyong[1,2] and HUANG Fangwan[1,2]
1 College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
2 Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350108, China

Abstract　With the development of the Internet of Things, numerous sensors can collect a large number of time series with rich data correlation, providing powerful data support for various data mining applications. However, some objective or subjective reasons (such as equipment failure, sparse sensing) often lead to the loss of collected data to varying degrees. Although many approaches have been proposed to solve this problem, data correlation is either not fully considered or computationally expensive. In addition, existing methods only focus on the completion of missing values, and fail to take into account downstream applications. Aiming at the above shortcomings, this paper designs a two-channel echo state network to achieve both the completion task and the prediction task. Although the two channels share the input layer, they have their own reservoir and output layer. The biggest difference between them is that the output layer of the left/right channels respectively represents the target value or prefilled value corresponding to the moment before/after the input layer. Finally, by fusing the estimates of the two channels, the data correlation from before and after the missing moments is fully utilized to further improve performance. Experimental results of different missing rates with two missing mechanisms (random missing and piecewise missing) show that the proposed model is superior to the current methods in both completion accuracy and prediction accuracy.

Keywords　Data correlation, Time series, Exogenous variables, Two-channel echo state network, Missing value completion, One-step

1　引言

随着物联网的发展,来自各类智能设备上的数十亿个传感器采集到大量的时间序列,为各种数据挖掘应用提供了强大的数据支持。然而,一些不可避免的突发事故往往会造成采集到的数据出现不同程度的缺失[1]。例如,电量不足或元件损坏导致传感器无法采集数据;通信中断或存储故障导致测量数据丢失。除了上述客观原因之外,还有一些主观

到稿日期:2022-12-08　返修日期:2023-04-02
基金项目:国家自然科学基金(61772136);福建省引导性项目(2020H0008);福建省中青年教师教育科研项目(JAT210007)
This work was supported by the National Natural Science Foundation of China(61772136), Fujian Provincial Guiding Project(2020H0008) and Educational Research Project for Young and Middle-aged Teachers in Fujian Province(JAT210007).
通信作者:黄昉菀(hfw@fzu.edu.cn)

原因也会造成数据的缺失。例如，为了降低采集成本，基于稀疏移动群智感知的应用只采集少量的关键数据，然后通过数据相关性估计出未采集的缺失数据[2]。

鉴于数据完整性对下游应用的重要意义，对缺失数据进行估计以提高数据质量往往是众多应用在数据预处理中非常重要的一步[3-4]。虽然已有很多方法被提出用于解决这一问题，但仍然存在以下两点不足：

1）基于物联网采集的时间序列往往存在着丰富多样的数据相关性，如流内相关性、流间相关性、迟滞相关性等[5]。以某个站点监测的细颗粒污染物 PM2.5 时间序列为例，该站点在某个时刻的缺失监测值首先和本站点在其他时刻的监测值有关，称为“流内相关性”；其次又和缺失时刻的相邻站点监测值（跨空间域）和气象条件（跨特征域）等外生变量有关，称为“流间相关性”；最后，考虑到污染物扩散需要一定的时间，它还与过去一段时间的所有站点监测值和气象条件有关，称为“迟滞相关性”[6]。遗憾的是，现有方法或对上述相关性考虑不够全面，或计算成本过高。

2）现有方法仅关注对缺失值的补全，未能兼顾下游应用。当需要同时完成补全和预测两个任务时，常见的做法是为两个不同的任务分别建立不同的模型[7]。这将导致无法充分利用缺失模式中蕴含的有用信息，进而导致预测偏差[8]。

针对上述不足，本文设计了一种兼顾补全与预测的回声状态网络集成框架。采用回声状态网络 ESN[9]（Echo State Network）的原因是：1）作为循环神经网络（Recurrent Neural Network，RNN）的一种变体，它可以利用输入连接和循环连接捕捉影响目标时间序列元素值（简称“目标值”）的外生变量带来的流间相关性和迟滞相关性；2）它可以利用其特有的输出层反馈连接捕捉来自未缺失时刻目标值的流内相关性；3）它采用储备池计算的方式可以使用较少的网络参数捕获比传统 RNN 更丰富的数据相关性。为了进行有效的集成，本文设计了双通道的 ESN 架构，用于协同估计缺失值和推断未来值。两个通道的 ESN 网络虽共用输入层但具有各自的储备池和输出层。其中，输入层输入的是每个时刻影响目标值的外生变量。左通道的输出层表示输入层前一个时刻的目标值，可以补全 $t=\{1,2,\cdots,n-1\}$ 时刻中的缺失值（n 为目标时间序列的长度）。而右通道的输出层则表示输入层后一个时刻的目标值，不仅可以补全 $t=\{2,3,\cdots,n\}$ 时刻中的缺失值，还可以预测出 $n+1$ 时刻的目标值。最后通过将两个通道在相同时刻的估计值进行融合，充分利用来自缺失时刻之前和之后的数据相关性以进一步提升补全精度。两种缺失现象下（随机缺失和分段缺失）不同缺失率的实验结果表明，所提出的模型无论是在补全精度还是预测精度上都优于目前流行的各类方法。

2 相关工作

根据所关注的数据相关性的差异，目前常用的时间序列补全方法可大致分为 4 类。

1）插值法。其主要关注时间序列元素值之间的流内相关性，常见的有线性插值、三次样条插值等。作为最原始的方法，线性插值计算简单、稳定性好，但是缺乏光滑性[10]。相比线性插值，三次样条插值拥有较好的光滑性，因此在许多领域得到广泛应用[11]。然而，传统插值技术是利用有限个元素值来估算出缺失值的近似值，对缺失率以及间隔时长非常敏感。针对上述不足，基于压缩感知的时序插值利用时间序列的时域平滑特性，将缺失值补全问题转化成稀疏向量恢复问题，可以较好地恢复高缺失率的时序数据[12]。插值法的缺点在于无法捕捉影响时间序列的外生变量带来的流间相关性和迟滞相关性。

2）归算法。其主要关注同一时刻的时间序列元素值与外生变量的流间相关性。最早提出的线性回归模型虽然思想简单、实现容易，但是对于非线性数据的拟合效果较差[13]。相较而言，回归树模型对于复杂的非线性关系有较好的拟合效果，常被用来补全缺失数据[14]。此外，K 最近邻算法（K Nearest Neighbors，KNN）也被广泛应用于数据补全领域[15]，它具有简单易懂等优点，但性能受 K 值和距离度量函数的影响较大。上述方法还可以通过迭代的方式进一步提升性能，如基于随机森林的迭代方法（MissForest）和使用一系列回归模型的链式方程多元插补（Multiple Imputation by Chained Equations，MICE）等[16-17]。归算法的缺点在于无法捕捉影响时间序列的外生变量带来的迟滞相关性和时间序列元素值之间的流内相关性。

3）矩阵填充法。这类方法是将需补全的时间序列与外生变量拼接成一个矩阵，然后利用矩阵内已有的数据对缺失值进行估计。常见的有低秩矩阵补全（Low-Rank Matrix Completion，LRMC）、矩阵分解（Matrix Factorization，MF）等[18-19]。LRMC 的原理是利用核范数代替矩阵的秩进行正则化约束来找到一个低秩矩阵，若不完整矩阵的未缺失部分与低秩矩阵的相应位置近似，则可以用低秩矩阵的值来填补不完整矩阵的缺失部分。MF 则首先将含缺失值的矩阵分解为两个（或多个）矩阵，然后将这些分解后的矩阵相乘得到原矩阵的近似矩阵，最后用近似矩阵的值来填补原矩阵的缺失部分。矩阵填充法可以在一定程度上兼顾到各种相关性，其缺点是将时序数据视为静态数据且对数据恢复施加严格的假设（如低秩性、时间平衡性、空间稳定性等）[20]。近年来，矩阵填充已拓展到张量领域，如低秩张量补全（Low-Rank Tensor Completion，LRTC）技术[21]。虽然补全精度有所提升，但张量秩的计算复杂度非常高，不适合实时性要求较高的应用。

4）深度学习法。在时间序列缺失补全领域广泛使用的是 RNN 架构，因为其特有的隐藏层自连接结构，其非常擅长捕捉迟滞相关性[22]。但目前常用的 RNN 架构大多基于门控机制，如长短期记忆网络（Long Short-Term Memory，LSTM）和门递归单元（Gated Recurrent Unit，GRU）[23-24]，它们均采用梯度随时间反向传播算法（Back Propagation Through Time，BPTT）更新权重。门的引入虽然可以缓解梯度消失问题，但对训练数据量和计算成本都有很高的要求，这显然不适合具有大量缺失数据或实时性要求较高的应用，而且传统的 RNN 架构并不能捕捉时间序列元素值之间的流内相关性。

在时间序列预测模型方面，根据研究视角的不同，通常

可以分为3类:传统统计模型、机器学习模型和混合集成模型[25]。但现有的大多数预测模型都需要输入完整的历史时间序列,不能处理含缺失值的不完整时间序列,导致其应用范围有限。常见的做法是将补全和预测分为两个任务,各自建立独立的模型,其弊端是预测模型无法有效地利用缺失模式的隐含信息,从而影响预测精度。因此,更合理的做法是将补全和预测整合到一个综合任务中。例如,将概率张量分解和向量自回归过程集成到一个模型中,利用时间序列数据的全局和局部一致性特征来进行缺失值估算和未来值预测[26]。为了捕捉时间序列的非线性动力学,一些研究人员还试图利用RNN架构同时完成补全和预测任务。例如,Chi等[22]利用基于缺失间隔的衰减,不仅可以优化经验均值和最后目标值之间的估计,而且还可以优化GRU的预测精度。LSTM架构也常被应用于显式结合缺失模式来降低预测残差[27]。上述模型的缺点在于:1)均基于门控机制来缓解RNN在权重训练时易出现的梯度问题;2)将补全作为预测的辅助任务,只采用简单的平滑技术来估计缺失值,因此重点提升的是预测精度而不是补全精度。综上所述,现有方法还需要进一步的改进。

3 ESN简介

为了方便介绍本文所设计的网络结构,有必要先简要介绍传统ESN的结构。传统回声状态网络由输入层、储备池和输出层构成,它们在时刻 t 的值分别用 $\boldsymbol{x}(t)$,$\boldsymbol{u}(t)$ 和 $\boldsymbol{y}(t)$ 表示,其结构如图1所示,其中实线箭头表示随机产生后固定不变的权重,虚线箭头表示需训练的权重。

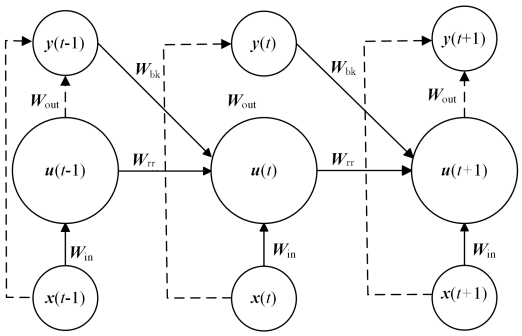

图1 带有前向反馈连接的ESN结构图

Fig. 1 Diagram of ESN with forward feedback connection

图1中 $\boldsymbol{w}_{\text{in}}$ 为输入层到储备池的输入连接权重,$\boldsymbol{w}_{\text{rr}}$ 为前一时刻储备池到当前时刻储备池之间的循环连接权重,$\boldsymbol{w}_{\text{bk}}$ 为前一时刻的输出层到当前时刻的储备池之间的反馈连接权重。这3个权重矩阵均是随机产生后固定不变的,因此在图1中用实线箭头表示。它们的不同点在于,为了得到高维的内部状态,ESN的储备池通常包含上百甚至上千个神经元,为了便于计算,$\boldsymbol{w}_{\text{rr}}$ 要求是稀疏连接的,而 $\boldsymbol{w}_{\text{in}}$ 和 $\boldsymbol{w}_{\text{bk}}$ 是全连接的。此外,前向反馈连接 $\boldsymbol{w}_{\text{bk}}$ 不是必须的,而是由任务所决定的,若不需要,可以删除。为了方便后续表示,本文将具有前向反馈连接的模型简称为ESN-F(ESN-Forward feedback),而不具有反馈连接的模型则称为ESN。无反馈的ESN的内部状态更新公式为:

$$\begin{aligned}&\boldsymbol{u}(t)=\varphi(\boldsymbol{w}_{\text{rr}}\boldsymbol{u}(t-1)+\boldsymbol{w}_{\text{in}}\boldsymbol{x}(t)+\boldsymbol{b})\\&\text{s.t. }\boldsymbol{u}(0)=0\end{aligned}\tag{1}$$

而具有前向反馈的ESN-F的内部状态的更新公式为:

$$\begin{aligned}&\boldsymbol{u}(t)=\varphi(\boldsymbol{w}_{\text{rr}}\boldsymbol{u}(t-1)+\boldsymbol{w}_{\text{in}}\boldsymbol{x}(t)+\boldsymbol{w}_{\text{bk}}\boldsymbol{y}(t-1)+\boldsymbol{b})\\&\text{s.t. }\boldsymbol{u}(0)=\boldsymbol{y}(0)=0\end{aligned}\tag{2}$$

其中,$\boldsymbol{b}$ 是随机产生的偏置向量,φ 为激活函数,常见的有sigmoid或tanh。

图1中虚线箭头表示输出权重 $\boldsymbol{w}_{\text{out}}$(包括输入层到输出层的连接权重和储备池到输出层的连接权重),它是整个网络唯一需要训练的权重矩阵。在训练阶段,首先根据输入和输出,利用式(1)或式(2)计算得到内部状态;然后与输入结合,得到扩展状态 $\boldsymbol{h}(t)=[\boldsymbol{u}(t);\boldsymbol{x}(t)]$,其中$[\cdot;\cdot]$表示两个向量的上下拼接操作。对于训练数据,将所有时刻的 $\boldsymbol{h}(t)$ 和 $\boldsymbol{y}(t)$ 分别按行存储成状态矩阵 $\boldsymbol{H}_1$ 和输出矩阵 $\boldsymbol{Y}_1$,然后利用以下的优化函数求解 $\boldsymbol{w}_{\text{out}}$:

$$\boldsymbol{W}_{\text{out}}=\arg\min\|\boldsymbol{Y}_1-\boldsymbol{H}_1\boldsymbol{W}_{\text{out}}\|_2^2\tag{3}$$

常用的求解方法有伪逆法、岭回归法、LASSO算法和弹性网络等[26]。在得到 $\boldsymbol{w}_{\text{out}}$ 之后即可在测试数据上运行ESN。同样先将测试数据的所有 $\boldsymbol{h}(t)$ 按行存储成状态矩阵 $\boldsymbol{H}_2$,然后根据式(4)计算输出:

$$\boldsymbol{Y}_2=\boldsymbol{H}_2\boldsymbol{W}_{\text{out}}\tag{4}$$

4 双通道回声状态网络框架

4.1 集成框架说明

为了兼顾补全和预测任务,本文设计了如图2所示的双通道ESN集成框架(Two-channel ESN,TC-ESN)。该框架的特点是左右通道共用输入层但具有各自的储备池和输出层。其中,t 时刻的输入层为影响时间序列的外生变量 $\boldsymbol{x}(t)$。左通道ESN的输出层对应的是输入层前一个时刻的目标值或预补值。因此,左通道储备池的内部状态更新公式应更改为:

$$\begin{aligned}&\boldsymbol{u}'(t)=\varphi(\boldsymbol{w}_{rr1}\boldsymbol{u}'(t-1)+\boldsymbol{w}_{\text{in1}}\boldsymbol{x}(t)+\boldsymbol{w}_{\text{bk1}}\boldsymbol{y}'(t)+\boldsymbol{b}_1)\\&\text{s.t. }\boldsymbol{u}'(0)=0\end{aligned}\tag{5}$$

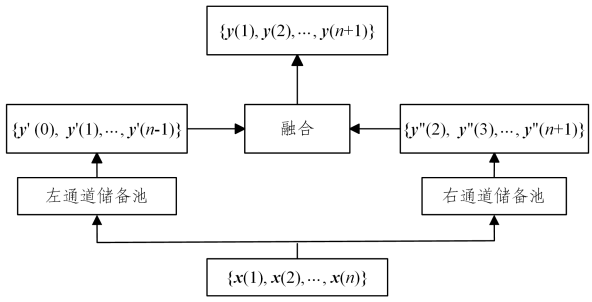

图2 双通道回声状态网络集成框架

Fig. 2 Integrated framework of two-channel echo state network

此时,输入连接权边 $\boldsymbol{w}_{\text{in1}}$ 可以帮助捕捉来自同一时刻外生变量的流间相关性;反馈连接权边 $\boldsymbol{w}_{\text{bk1}}$ 可以帮助捕捉前一时刻目标值的流内相关性;循环连接权边 $\boldsymbol{w}_{\text{rr1}}$ 可以帮助捕捉过去时段外生变量和目标值的迟滞相关性。

从图2中可以看出,不同于左通道,右通道ESN的输出层对应的是输入层后一个时刻的目标值或预补值。因此,右

通道储备池的内部状态更新公式可设置为：

$$\begin{aligned}&\boldsymbol{u}''(t)=\varphi(\boldsymbol{w}_{\mathrm{rr2}}\boldsymbol{u}''(t-1)+\boldsymbol{w}_{\mathrm{in2}}\boldsymbol{x}(t)+\boldsymbol{w}_{\mathrm{bk2}}\boldsymbol{y}''(t)+\boldsymbol{b}_2)\\&\text{s.t. }\boldsymbol{u}''(0)=0\end{aligned}\tag{6}$$

此时，输入连接权边$\boldsymbol{w}_{\mathrm{in2}}$和循环连接权边$\boldsymbol{w}_{\mathrm{rr2}}$的作用和左通道类似，但反馈连接权边$\boldsymbol{w}_{\mathrm{bk2}}$则帮助捕捉后一时刻目标值的流内相关性。这样做的好处是补全任务不会像预测任务那样只能依赖于预测时刻之前的数据，相反它可以通过缺失时刻之前和之后的数据来进行估计。本课题组的前期工作也表明具有前向和后向反馈连接的 ESN 在补全精度上要高于只有前向反馈连接的 ESN。综上所述，双通道 ESN 框架不仅可以共用输入层而且可以共用相同的网络结构，如图 3 所示。

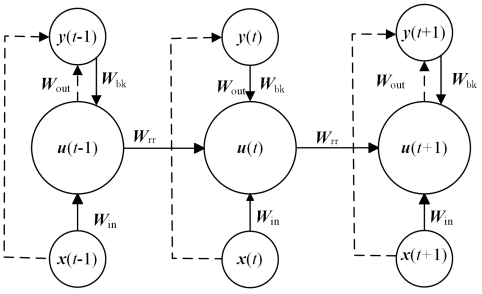

图 3 双通道回声状态网络框架采用的网络结构图

Fig. 3 Network structure diagram of two-channel echo state network

需要强调的是，由于输出层的时间序列存在缺失值，为了保证反馈连接的有效性，在计算储备池状态之前，应该首先对缺失值进行预补。对于左通道储备池而言，为了不中断流内相关性从前往后传播，在计算内部状态之前，应首先对除$\boldsymbol{y}'(0)$之外的缺失值进行前溯最近邻预补，即$\boldsymbol{y}'(t)$用$\boldsymbol{y}'(t-i)$进行预补的前提是从$\boldsymbol{y}'(t-i+1)$至$\boldsymbol{y}'(t)$全都缺失。$\boldsymbol{y}'(0)$由于无法前溯，只能用后续时刻的最近邻真实值预补。而对于右通道储备池而言，为了不中断流内相关性从后往前传播，在计算内部状态之前，应首先对除$\boldsymbol{y}''(n+1)$之外的缺失值进行后续最近邻预补，即$\boldsymbol{y}''(t)$用$\boldsymbol{y}''(t+i)$进行预补的前提是从$\boldsymbol{y}''(t)$至$\boldsymbol{y}''(t+i-1)$全都缺失。$\boldsymbol{y}''(n+1)$由于没有后续值，只能用前溯时刻的最近邻真实值预补。

每个通道在完成所有时刻的内部状态计算后，都需要利用未缺失的真实目标值先训练得到输出权重$\boldsymbol{W}_{\mathrm{out1}}$和$\boldsymbol{W}_{\mathrm{out2}}$，然后再利用它们与缺失时刻的内部状态相乘得到缺失时刻的估计值用于替换预补值。此时，左通道可以补全$\boldsymbol{y}(1)$至$\boldsymbol{y}(n-1)$中的缺失值；而右通道不仅可以补全$\boldsymbol{y}(2)$至$\boldsymbol{y}(n)$中的缺失值，还可以得到预测值$\boldsymbol{y}(n+1)$。最后，为了进一步提升补全精度，需要将两个通道在相同时刻的估计值进行融合，充分利用来自缺失时刻之前或之后的数据相关性。输出层的融合公式如下：

$$\boldsymbol{y}(t)=\begin{cases}\boldsymbol{y}'(t), & t=1\\ \alpha\boldsymbol{y}'(t)+(1-\alpha)\boldsymbol{y}''(t), & 1<t<n\\ \boldsymbol{y}''(t), & t=n\text{ or }t=n+1\end{cases}\tag{7}$$

其中，$\alpha\in(0,1)$为融合因子，可以通过验证集得到或直接设置为 0.5。

4.2 具体操作流程

为了更好地解释双通道 ESN 实现补全和预测的过程，下面说明其具体的操作流程。假设有v条时间序列，每条序列的长度为n，缺失值个数为s（缺失位置可不同）。假设这些序列在每个时刻具有共同的外生变量$\boldsymbol{x}(t)\in R^{m\times 1}$，可采用以下步骤对这$v$条时间序列进行缺失补全以及单步预测。

步骤 1 构建一个如图 2 所示的双通道 ESN，其中输入层的神经元个数为m（即输入为外生变量），左/右储备池的神经元个数均为p，左/右输出层的神经元个数均为v（即输出为v条时间序列的元素值）。

步骤 2 分别对左/右通道输出层中每条时间序列的缺失值进行前溯/后续最近邻预补。

步骤 3 使用式(5)和式(6)求出左/右储备池在所有时刻的内部状态$\boldsymbol{u}'(t)\in R^{p\times 1}$和$\boldsymbol{u}''(t)\in R^{p\times 1}$，然后根据各自的输入和输出得到各自的扩展状态$\boldsymbol{h}'(t)=[\boldsymbol{y}'(t);\boldsymbol{u}'(t);\boldsymbol{x}(t)]\in R^{(v+p+m)\times 1}$和$\boldsymbol{h}''(t)=[\boldsymbol{y}''(t);\boldsymbol{u}''(t);\boldsymbol{x}(t)]\in R^{(v+p+m)\times 1}$。

步骤 4 对左/右通道 ESN 分别执行以下循环($z=1;z\leqslant v;z^{++}$)

1)删除$\boldsymbol{h}'(t)/\boldsymbol{h}''(t)$中的第$z$个元素，得到$\boldsymbol{h}_z'(t)/\boldsymbol{h}_z''(t)\in R^{(v-1+p+m)\times 1}$；

2)将未缺失时刻的$\boldsymbol{h}_z'(t)/\boldsymbol{h}_z''(t)$按行存储成矩阵$\boldsymbol{H}_{z1}'/\boldsymbol{H}_{z1}''\in R^{(n-s)\times(v-1+p+m)}$；

3)取出未缺失时刻的$\boldsymbol{y}'(t)/\boldsymbol{y}''(t)$中的第$z$个元素，并按行存储成向量$\boldsymbol{Y}_{z1}'/\boldsymbol{Y}_{z1}''\in R^{(n-s)\times 1}$；

4)利用式(8)和式(9)分别计算出左/右通道对应于第z条时间序列的输出权重$\boldsymbol{w}_{\mathrm{out1}}^z/\boldsymbol{w}_{\mathrm{out2}}^z\in R^{(v-1+p+m)\times 1}$：

$$\boldsymbol{w}_{\mathrm{out1}}^z=\arg\min\|\boldsymbol{Y}_{z1}'-\boldsymbol{H}_{z1}'\boldsymbol{w}_{\mathrm{out1}}^z\|_2^2\tag{8}$$

$$\boldsymbol{w}_{\mathrm{out2}}^z=\arg\min\|\boldsymbol{Y}_{z1}''-\boldsymbol{H}_{z1}''\boldsymbol{w}_{\mathrm{out2}}^z\|_2^2\tag{9}$$

5)将缺失时刻的$\boldsymbol{h}_z'(t)/\boldsymbol{h}_z''(t)$按行存储成矩阵$\boldsymbol{H}_{z2}'/\boldsymbol{H}_{z2}''\in R^{s\times(v-1+p+m)}$；

6)利用式(10)和式(11)分别计算左/右通道对应于第z条时间序列缺失时刻的估计值：

$$\boldsymbol{Y}_{z2}'=\boldsymbol{H}_{z2}'\boldsymbol{w}_{\mathrm{out1}}^z\tag{10}$$

$$\boldsymbol{Y}_{z2}''=\boldsymbol{H}_{z2}''\boldsymbol{w}_{\mathrm{out2}}^z\tag{11}$$

7)将$\boldsymbol{Y}'_{z2}$和$\boldsymbol{Y}''_{z2}$中的各个元素根据式(7)进行融合，即可得到缺失时刻的最终估计值和未来时刻的预测值。

5 实验

5.1 实验设置说明

为了验证该集成框架的有效性，本文选取 UCI 机器学习库提供的中国城市 PM2.5 数据集中沈阳市 2014 年 11 月的数据进行了一系列实验。选取该数据子集的原因是在原始数据集中，该月份仅包含 0.46%的缺失观测，可近似认为原始数据是完整的，便于进行性能评估。该数据子集中包含了太原街、小河沿、美国领事馆 3 个站点的 PM2.5 时间序列，存在一定的空间相关性。数据采集的时间间隔为 1 h，故每条时间序列的长度$n=720$。考虑到气象条件对 PM2.5 的影响很大，本文将对应时刻的气象数据（包括露点(℃)、温度(℃)、湿度(%rh)、累计风速(km/h)、压力(Pa)）作为外生变量。表 1 列出了沈阳数据集中某一天的数据。

表 1 沈阳数据集中 2014 年 11 月 5 日的数据展示

Table 1 Data in Shenyang dataset on November 5, 2014

时刻/h	露点/℃	湿度/%rh	压力/Pa	温度/℃	累积风速/(km/h)	领事馆 $PM_{2.5}$/(mg/m^3)	小河沿 $PM_{2.5}$/(mg/m^3)	太原街 $PM_{2.5}$/(mg/m^3)
0	0	46.58	1015	11	107	80	79	104
1	−2	43.01	1015	10	111	74	69	83
2	−3	37.36	1015	11	115	64	69	83
3	−2	40.23	1014	11	120	69	59	69
4	−1	46.29	1014	10	124	69	61	71
5	0	46.58	1014	11	129	67	65	77
6	1	50.08	1014	11	133	74	70	84
7	1	50.08	1014	11	137	79	75	96
8	1	46.88	1014	12	141	87	81	101
9	1	41.14	1015	14	146	87	89	112
10	2	41.43	1014	15	154	99	96	115
11	3	41.73	1014	16	164	102	93	124
12	4	42.02	1014	17	172	84	78	88
13	4	42.02	1014	17	179	79	72	76
14	4	44.78	1014	16	187	69	69	77
15	2	38.86	1014	16	193	70	69	80
16	−2	30.98	1015	15	196	80	57	62
17	−5	30.09	1016	12	2	58	32	65
18	−8	23.89	1017	12	5	51	41	64
19	−7	25.82	1018	12	10	44	32	65
20	−6	31.84	1019	10	12	37	31	32
21	−7	29.48	1020	10	19	38	27	28
22	−8	31.23	1022	8	26	32	20	25

考虑到偶发性故障和持久性故障会分别造成随机缺失和分段缺失，为了验证模型的泛化能力，本文进行了不同缺失率(20%，40%，60%，80%)下的两组实验：第一组是将 3 个站点的监测值分别按照相同的缺失率进行随机缺失；第二组是在相同缺失率下，将某个站点的数据设置成分段缺失，其他两个站点仍然设置为随机缺失。为了避免缺失位置不同带来的偏差，本文为每一种缺失机制下的每一个缺失率都设置了 5 个不同的缺失位置索引，最后取 5 次实验结果的平均值来评估模型的性能。

5.2 实验参数说明

本文涉及的所有参数和调参范围区间及分辨率如表 2 所列，利用文献[27]提供的遗传算法搜索最优参数。隐藏层的激活函数为 tanh，输出权重的求解方法为岭回归法。

表 2 TC-ESN 模型的调参范围区间及分辨率

Table 2 Range and resolution of parameters in TC-ESN model

模型参数	最小值	最大值	分辨率
储备池大小	100	500	5
储备池谱半径	0.5	1.4	0.09
储备池稀疏度	0.25	0.5	0.025
输入层到储备池的输入连接权重缩放因子	0.1	0.9	0.08
输出层到储备池的反馈连接的缩放因子	0.1	0.9	0.08
岭回归的正则化因子	0.001	1	0.1

5.3 实验结果分析

本小节主要展示在不同缺失机制以及不同缺失率下，基准模型与 TC-ESN 在缺失补全与单步预测的性能差异。对于缺失补全任务，本文对比的基准模型包括：第一类插值法中的线性插值(Linear)、三次样条插值(Spline)和压缩感知恢复(CS)；第二类归算法中的 KNN，MICE 和 MissForest；第三类矩阵填充法中的矩阵补全(softImpute)、矩阵分解(MF)和低秩张量补全(LRTC)；第四类深度学习方法中的 Elman RNN(ERNN)，LSTM，GRU。为了公平起见，这 3 种 RNN 模型均采用图 2 的左右通道集成框架。

为了能从不同角度比较本文模型与基准模型的优劣，本文采用了两种评价指标，分别为：均方根误差 RMSE(Root Mean Square Error)和确定系数 R^2(R-Square)。它们的计算公式为：

$$RMSE=\sqrt{\left(\frac{1}{S}\sum_{i=1}^{S}\tilde{y}_i-y_i\right)^2} \tag{12}$$

$$R^2=1-\frac{\sum_{i=1}^{s}(\tilde{y}_i-y_i)^2}{\sum_{i=1}^{s}(\bar{y}-y_i)^2} \tag{13}$$

其中，s 为每个站点的缺失值个数，$\tilde{y}_i$ 为估计值，y_i 为真实值，$\bar{y}$ 为 s 个 y_i 的均值。$RMSE$ 越小代表估计值越贴近真实值。而 R^2 越大代表准确度越高，当 $R^2=0$ 时表示达到了均值补全的效果；当 $R^2<0$ 时，则表示不如均值补全。由于是对 3 个站点的缺失值进行补全，所以对 3 个站点的指标进行平均后，得到平均 $RMSE$ 和平均 R^2。实验结果如表 3−表 6 所列。根据实验结果，可得出以下结论：

1)在随机缺失机制下，当缺失率低于 50%时，插值法的表现明显优于归算法，说明对于该数据集而言，流内相关性的重要性高于流间相关性。当缺失率高于 50%时，插值法的误差上升幅度要大于归算法，说明当缺失率较高时，即使是随机缺失，也有很大的概率出现连续缺失的情况，这对流内相关性的破坏是非常大的。传统矩阵补全的效果在 4 类方法中表现最差，说明此类方法将时间序列视为静态数据后，根本无法捕捉与时间因素有关的相关性(如流内或迟滞相关性)。张量领域的补全方法 LRTC 虽然在低缺失率下补全精度能够接近于深度学习方法，但是随着缺失率的提升，其补全效果依然和深度学习方法有很大的差距。深度学习的方法由于均采用左右通道集成架构，充分考虑了缺失时刻之前和之后的数据相关性，所以表现要明显优于其他 3 类方法。最后，本文提出的 TC-ESN 由于具有反馈连接，可以捕捉比其他 RNN 架构更丰富

的数据相关性（即所有站点的流内相关性），因此表现最优。

2）在分段缺失机制下，插值法的表现普遍差于归算法，说明连续缺失的确会严重影响流内相关性，这与随机缺失时插值法在高缺失率下表现不佳的结论一致。此外，由于分段缺失对时间因素的破坏要大于随机缺失，所以矩阵补全与其他3类方法的性能差距要小于随机缺失下的性能差距。最后，不管是随机缺失还是分段缺失，深度学习的方法在绝大多数情况下都是表现最好的。其中，LRTC，LSTM 和 GRU 在缺失率为 20%时的表现虽然略优于 TC-ESN，但在更高的缺失率下，TC-ESN 的表现仍然是最好的。这充分说明 TC-ESN 在缺失补全任务中的优势。

表 3 随机缺失下不同方法对缺失补全任务的平均 RMSE

Table 3 Average RMSE of completion tasks of different methods in the case of random missing

方法	20%	40%	60%	80%
Linear	37.192	36.512	50.846	63.049
Spline	38.220	36.848	51.524	91.993
CS	33.551	26.326	47.597	51.216
KNN	46.124	51.955	67.064	56.299
MICE	36.035	39.894	56.420	48.434
MissForest	33.139	35.629	48.604	43.797
softImpute	41.532	45.838	64.771	55.549
MF	58.850	59.509	70.795	66.954
LRTC	26.286	36.530	58.834	63.874
ERNN	31.233	29.805	38.815	44.094
LSTM	29.967	30.190	37.113	42.746
GRU	30.037	29.959	37.264	42.708
TC-ESN	**25.512**	**25.613**	**31.133**	**41.253**

表 4 随机缺失下不同方法对缺失补全任务的平均 R^2

Table 4 Average R^2 of completion tasks of different methods in the case of random missing

方法	20%	40%	60%	80%
Linear	0.829	0.817	0.639	0.389
Spline	0.818	0.813	0.630	−0.564
CS	0.853	0.902	0.681	0.594
KNN	0.756	0.621	0.380	0.237
MICE	0.853	0.774	0.561	0.446
MissForest	0.876	0.819	0.674	0.535
softImpute	0.799	0.704	0.422	0.265
MF	0.659	0.495	0.297	−0.134
LRTC	0.913	0.809	0.520	0.378
ERNN	0.872	0.874	0.793	0.703
LSTM	0.881	0.871	0.811	0.721
GRU	0.881	0.873	0.810	0.722
TC-ESN	**0.914**	**0.907**	**0.867**	**0.741**

表 5 分段缺失下不同方法对缺失补全任务的平均 RMSE

Table 5 Average RMSE of completion tasks of different methods in the case of piecewise missing

方法	20%	40%	60%	80%
Linear	47.058	47.514	69.456	73.999
Spline	47.053	48.548	72.090	77.061
CS	43.415	46.477	62.198	63.915
KNN	40.179	46.710	61.402	68.464
MICE	33.576	35.686	50.174	57.797
MissForest	32.032	34.548	48.646	58.324
softImpute	36.677	41.305	56.716	65.432
MF	50.782	54.219	70.244	85.970
LRTC	**27.640**	32.241	49.587	60.794
ERNN	30.746	31.488	41.408	49.978
LSTM	29.126	31.638	39.907	49.445
GRU	29.347	31.593	40.347	49.589
TC-ESN	30.184	**29.792**	**39.362**	**48.234**

表 6 分段缺失下不同方法对缺失补全任务的平均 R^2

Table 6 Average R^2 of completion tasks of different methods in the case of piecewise missing

方法	20%	40%	60%	80%
Linear	0.732	0.705	0.356	0.145
Spline	0.724	0.693	0.309	0.064
CS	0.765	0.731	0.465	0.358
KNN	0.732	0.621	0.426	0.239
MICE	0.815	0.779	0.616	0.455
MissForest	0.826	0.787	0.637	0.440
softImpute	0.784	0.709	0.513	0.305
MF	0.601	0.503	0.229	−0.235
LRTC	**0.887**	0.830	0.627	0.400
ERNN	0.848	0.824	0.737	0.593
LSTM	0.861	0.823	0.753	0.602
GRU	0.859	0.823	0.748	0.600
TC-ESN	0.839	**0.832**	**0.754**	**0.614**

鉴于在补全任务中，深度学习的方法明显优于其他 3 种方法。在单步预测任务中，本文重点比较了基于双通道集成框架的 RNN，LSTM，GRU，TC-ESN 的优劣。由于完成的是单步预测，无法计算R^2，所以评价指标仅有平均 RMSE。实验结果如表 7－表 8 所列。

表 7 随机缺失下不同方法对单步预测任务的平均 RMSE

Table 7 Average RMSE of one-step prediction tasks by different methods in the case of random missing

方法	20%	40%	60%	80%
ERNN	18.983	21.550	29.852	38.448
LSTM	17.301	21.794	24.857	37.909
GRU	18.138	22.256	26.370	40.361
TC-ESN	**14.942**	**18.385**	**19.276**	**13.944**

表 8 分段缺失下不同方法对单步预测任务的平均 RMSE

Table 8 Average RMSE of one-step prediction tasks of different methods in the case of piecewise missing

方法	20%	40%	60%	80%
ERNN	19.921	23.668	33.064	43.892
LSTM	18.192	21.517	25.616	41.898
GRU	19.098	23.722	28.832	43.564
TC-ESN	**17.233**	**20.113**	**19.921**	**20.152**

与补全任务结论一致，无论是随机缺失还是分段缺失，TC-ESN 的预测误差都是最小的。

结束语 随着智慧城市建设进程的推进，大量基于环境监测的传感器收集到海量的数据，这些数据间存在着大量的流间相关性、流内相关性和迟滞相关性，如何充分利用这些相关性提升各类数据挖掘任务的准确性已成为研究热点。本文针对具有外生变量的时序数据，开展了兼顾补全及预测问题的研究，设计了一种具有双通道的回声状态神经网络 TC-ESN。它巧妙地利用左右通道，不仅可以融合缺失时刻之前和之后的数据相关性从而提升补全精度，而且可以完成单步预测任务。在未来的工作中，可以通过修改右通道的结构，将单步预测推广到多步预测。此外，还可以借助 ESN 的在线训练，进行在线缺失补全与预测的研究。

参 考 文 献

[1] LIN W C，TSAI C F. Missing value imputation：a review and analysis of the literature（2006－2017）[J]. Artificial Intelligence

Review, 2020, 53(2): 1487-1509.
[2] WANG L, ZHANG D, WANG Y, et al. Sparse mobile crowdsensing: challenges and opportunities[J]. IEEE Communications Magazine, 2016, 54(7): 161-167.
[3] WANG S, ZHENG F, ZHAO D. Research on causal network of high-dimensional time series with insufficient information[J]. Journal of Chinese Computer Systems, 2023, 44(5): 981-990.
[4] HUANG F, ZHENG W, GUO W, et al. Estimating missing data for sparsely sensed time series with exogenous variables using bidirectional-feedback echo state networks[J]. CCF Transactions on Pervasive Computing and Interaction, 2023, 5(1): 45-63.
[5] YOON J, ZAME W R, VAN DER SCHAAR M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks[J]. IEEE Transactions on Biomedical Engineering, 2018, 66(5): 1477-1490.
[6] LIU Y, ZHAO N, VANOS J K, et al. Effects of synoptic weather on ground-level PM2.5 concentrations in the United States [J]. Atmospheric Environment, 2017, 148: 297-305.
[7] BOQUET G, MORELL A, SERRANO J, et al. A variational utoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection[J]. Transportation Research Part C: Emerging Technologies, 2020, 115: 102622.
[8] ZHANG Z, LIN X, LI M, et al. A customized deep learning approach to integrate network-scale online traffic data imputation and prediction[J]. Transportation Research Part C: Emerging Technologies, 2021, 132: 103372.
[9] GRIGORYEVA L, ORTEGA J P. Echo state networks are universal[J]. Neural Networks, 2018, 108: 495-508.
[10] NOORN M, AL BAKRI ABDULLAHM M, YAHAYAA S, et al. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set [C]// Materials Science Forum. Trans. Tech. Publications Ltd., 2015, 803: 278-281.
[11] KARIM S A A, ISMAIL M T, OTHMAN M, et al. Rational cubic spline interpolation for missing solar data imputation[J]. Journal of Engineering and Applied Sciences, 2018, 13(9): 2587-2592.
[12] SONG X, GUO Y, LI N, et al. Missing data prediction based on compressive sensing in time series[J]. Computer Science, 2019, 46(6): 35-40.
[13] JADHAV A, PRAMOD D, RAMANATHAN K. Comparison of performance of data imputation methods for numeric dataset [J]. Applied Artificial Intelligence, 2019, 33(10): 913-933.
[14] LOH W Y, ZHANG Q, ZHANG W, et al. Missing data, imputation and regression trees[J]. Statistica Sinica, 2020, 30(4): 1697-1722.
[15] SANTOS M S, ABREU P H, WILK S, et al. How distance metrics influence missing data imputation with k-nearest neighbours [J]. Pattern Recognition Letters, 2020, 136: 111-119.
[16] ZHANG S, GONG L, ZENG Q, et al. Imputation of GPS coordinate time series using MissForest[J]. Remote Sensing, 2021, 13(12): 2312.
[17] LAQUEUR H S, SHEV A B, KAGAWA R M C. SuperMICE: an ensemble machine learning approach to multiple imputation by chained equations[J]. American Journal of Epidemiology, 2022, 191(3): 516-525.
[18] XIONG Z, WEI Y, XU R, et al. Low-rank traffic matrix completion with marginal information[J]. Journal of Computational and Applied Mathematics, 2022, 410: 114219.
[19] YU H F, RAO N, DHILLON I S. Temporal regularized matrix factorization for high-dimensional time series prediction[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 847-855.
[20] KONG L, XIA M, LIU X Y, et al. Data loss and reconstruction in wireless sensor networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 25(11): 2818-2828.
[21] CHEN X, YANG J, SUN L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation[J]. Transportation Research Part C: Emerging Technologies, 2020, 117(8): 102673. 1-102673-12.
[22] KIM Y J, CHI M. Temporal belief memory: imputing missing data during RNN training[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 2326-2332.
[23] YUAN H, XU G, YAO Z, et al. Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks[C]// Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018: 1293-1300.
[24] CHE Z, PURUSHOTHAM S, CHO K, et al. Recurrent neural networks for multivariate time series with missing values[J]. Scientific Reports, 2018, 8(1): 1-12.
[25] YU Z, ZHENG X, HUANG F, et al. A framework based on sparse representation model for time series prediction in smart city[J]. Frontiers of Computer Science, 2021, 15(1): 1-13.
[26] CHEN X, SUN L. Bayesian temporal factorization for multidimensional time series prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 14(9): 4659-4673.
[27] TIAN Y, ZHANG K, LI J, et al. LSTM-based traffic flow prediction with missing data[J]. Neurocomputing, 2018, 318: 297-305.

ZHENG Weinan, born in 1997, postgraduate. His main research interests include data completion and so on.

HUANG Fangwan, born in 1980, Ph.D, senior lecturer, is a member of CCF (No. D3015M). Her main research interests include computational intelligence, machine learning and big data analysis.

(责任编辑:何杨)