计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于残差连接长短期记忆网络的时间序列修复模型
作者：钱斌，郑楷洪，陈子鹏，肖勇，李森，叶纯壮，马千里
收稿日期：2020-05-30
网络首发日期：2020-10-16
引用格式：钱斌，郑楷洪，陈子鹏，肖勇，李森，叶纯壮，马千里．基于残差连接长短期记忆网络的时间序列修复模型[J/OL]．计算机应用.
https://kns.cnki.net/kcms/detail/51.1307.TP.20201016.1117.002.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2020060928

基于残差连接长短期记忆网络的时间序列修复模型

钱斌[1]，郑楷洪[1]，陈子鹏[2]，肖勇[1]，李森[2]，叶纯壮[1]，马千里[2*]

(1.南方电网科学研究院，广州 510663； 2.华南理工大学计算机科学与工程学院，广州 510006)

(*通信作者电子邮箱 qianlima@scut.edu.cn)

摘 要: 传统的时间序列缺失修复方法通常假设数据由线性动态系统产生，然而时间序列更多地表现为非线性。为此，提出了基于残差连接长短期记忆网络的时间序列修复模型，称为 RSI-LSTM，可以有效捕获时间序列的非线性动态特性，挖掘缺失数据和最近的非缺失数据之间的潜在关联。具体地，采用长短期记忆网络对时间序列的非线性动态特性进行建模，同时引入残差连接，挖掘历史值与缺失值的联系，提升模型的修复能力。首先对单变量日供电量数据集的缺失数据进行修复，然后在第九届电工数学建模竞赛 A 题的电力负荷数据集上，引入气象因素作为模型的多变量输入，提升时间序列的缺失值修复效果。此外，还使用了两个通用的多变量时间序列数据集以验证模型的缺失修复能力。实验结果表明， RSI-LSTM 在单变量和多变量数据集上，缺失值的修复效果均优于 LSTM，修复的均方误差总体下降 10%。

关键词: 缺失数据修复；长短期记忆网络；残差连接；时间序列；时序依赖

中图分类号:TP183 **文献标志码**：**A**

Time series imputation model based on long-short term memory network with residual connection

QIAN Bin[1], ZHENG Kaihong[1], CHEN Zipeng[2], XIAO Yong[1], LI Sen[2], YE Chunzhuang[1], MA Qianli[2*]

(1. Electric Power Research Institute, CSG, Guangzhou Guangdong 510663, China;

2. School of Computer Science and Engineering, South China University of Technology, Guangzhou Guangdong 510006, China)

Abstract: Traditional time series imputation methods typically assumed that time series data was derived from a linear dynamic system. However, the real-world time series showed more non-linear characteristics. To this end, in this paper, a time series imputation model based on residual long-short term memory network, called RSI-LSTM, was proposed to capture the non-linear dynamics of the time series and model the potential relation between missing items and recent non-missing items. Specifically, a long-short term memory network was used to model the underlying non-linear dynamic temporal characteristics. Meanwhile, the residual connection was introduced to mine the connection between the historical observations and the missing value to improve imputation capability. Firstly, RSI-LSTM was applied to impute the univariate daily power supply time series, and then the meteorological factors were introduced as the multivariate input of RSI-LSTM to further improve the imputation performance on power load data set of the 9th electrical engineering mathematical modeling contest A. Furthermore, two general multivariate time series data sets were used to verify the model's ability. The experimental results show that compared with LSTM, RSI-LSTM can get better repairing performance on both univariate and multivariate data sets. The mean square error of RSI-LSTM is 10% lower than that of LSTM.

Keywords: missing value imputation; Long Short-Term Memory (LSTM) network; residual connection; time series; temporal dependency

0 引言

时间序列在诸多领域都有着丰富的应用[1-3]。然而由于存在噪声或传感器故障等原因，现实时间序列不可避免地含有

收稿日期: 2020-05-30; 修回日期:2020-07-21; 录用日期: 2020-07-24。

基金项目：国家自然科学基金重点项目（61751205）；国家自然科学基金项目（61872148）

作者简介:钱斌（1989-），男，湖北十堰人，工程师，硕士，主要研究方向：电能计量技术；郑楷洪（1991-），男，广东汕头人，工程师，硕士，主要研究方向：电能计量、电能计量自动化系统、用电技术；陈子鹏（1996-），男，广东揭阳人，硕士研究生，主要研究方向：数据挖掘、神经网络；肖勇（1978-），男，湖南怀化人，高级工程师，博士，主要研究方向：电能计量管理、电能计量自动化系统、用电技术；李森（1994-），男，广东茂名人，硕士研究生，主要研究方向：数据挖掘、神经网络；叶纯壮（1989-），男，海南海口人，工程师，学士，主要研究方向：电力线损管理工作；马千里（1980-），男（通信作者），甘肃宕昌人，教授，博士，主要研究方向：数据挖掘、神经网络。

缺失值，这使得现有分析算法的推断变得尤为困难[4,5]。因此，如何有效地对缺失数据进行修复具有重要的现实意义。

传统的时间序列缺失修复方法有均值替代、移动平均法、空间自回归、多项式插补、线性插值等等[6-8]，但是，传统的统计方法无法有效提炼缺失数据背后蕴藏的事件信息，这会对缺失修复效果造成一定影响。针对传统缺失修复方法的不足，文献[9]提出了基于 T2 椭圆图的异常数据识别和基于最小二乘支持向量机的缺失修复方法，但该方法以社会经济指标可信为前提条件，无法从时间序列自身挖掘规律。文献[10]提出了一种基于遗传优化算法的时间序列缺失修复方法，该方法考虑时间序列的历史信息，借助遗传算法优化多重插补的参数，寻找最优的修复值。但该方法将时间序列视为线性序列，未考虑时间序列中的非线性，修复效果不佳。

由于神经网络可以较好地建模数据中的非线性特性，可以将其应用于时间序列的缺失修复。文献[11]提出了自适应 BP 神经网络来修复缺失的时间序列，但该方法未对时间序列中的时序依赖关系进行有效的建模。因此，文献[12,13]提出了基于长短期记忆网络(Long Short-Term Memory network, LSTM) [14]的缺失修复方法，可以有效建模时间序列的时序依赖信息，但这些方法需要对数据进行预处理，无法在含缺失值的情况下进行模型训练。而且，不恰当的预处理方法会引入较大偏差，从而误导模型的训练过程，影响后续的缺失修复效果。

针对以上问题，本文提出了基于残差连接 LSTM 网络的时间序列修复模型，使用长短期记忆网络对时间序列中的时序依赖关系以及非线性特性进行建模。并且，在长短期记忆网络的基础上，引入残差连接[15,16]，挖掘缺失数据和它们最近的非缺失数据的潜在关联性，引入残差连接的具体做法是构建一种残差和单元(Residual Sum Unit, RSU)，可以有效聚合历史信息。本文同时考虑了单变量输入和多变量输入的情况，并且，本文的方法无需对缺失数据进行预填补，可以直接在含缺失值的情况下进行模型训练。实验证明，与现有最先进的方法相比，基于残差连接 LSTM 网络的时间序列修复模型可以取得较好的缺失修复效果。本文的主要贡献如下：

1）本文提出了基于残差连接 LSTM 网络的时间序列修复模型，使用长短期记忆网络对时间序列中的时序依赖和非线性特性进行建模，同时构建残差和单元聚合历史信息，进一步提升缺失修复效果。

2）通过在单变量和多变量数据集上定量和定性的实验分析，我们提出的模型均取得比现有最先进方法更好的缺失值修复效果。此外，我们的模型无需进行数据预填补，可以直接在含缺失值的情况下进行训练。

1 长短期记忆网络

长短期记忆网络是循环神经网络的一种变体，能建模时序数据的时间依赖和非线性特性，是当下对时序数据建模的首选模型。长短期记忆网络由记忆单元组成，通过输入、输出和遗忘门来决定流入流出记忆单元的信息多少。长短期记忆网络记忆单元的结构如图 1 所示：

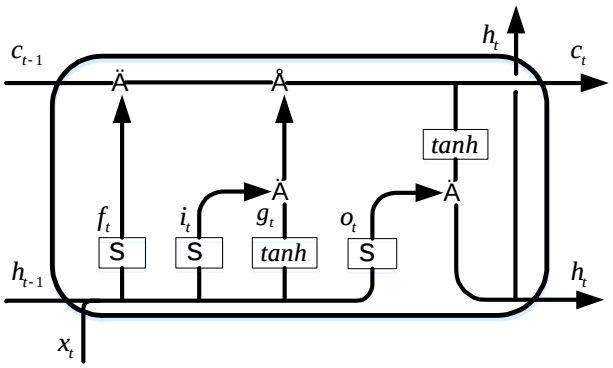

图1 长短期记忆网络记忆单元结构图

Fig. 1 The structure of LSTM unit

在图 1 中，$\boldsymbol{x}_t$ 是时间步 t 的输入数据，$\boldsymbol{h}_t$ 是时间步 t 长短期记忆网络的隐藏状态，$\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$ 分别是长短期记忆网络的输入门，遗忘门和输出门。$\boldsymbol{g}_t$ 是当前加入的信息，$\boldsymbol{c}_t$ 是记忆单元的信息，S表示 sigmoid 激活函数，Ä是逐元素的乘法，Å是逐元素的加法。

给定长度为 T 的输入序列 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T\}$，长短期记忆网络可以将其编码为一个隐藏状态序列 $\boldsymbol{h}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_T\}$，其中，$\boldsymbol{x}_t\in\mathbf{R}^n$，$\boldsymbol{h}_t\in\mathbf{R}^m$,$t$=1,2,…,$T$。在时间步 t,长短期记忆网络的计算公式如下：

$$\boldsymbol{i}_t=\mathrm{S}\left(\boldsymbol{W}_{\mathrm{i}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{i}}\right) \tag{1}$$

$$\boldsymbol{f}_t=\mathrm{S}\left(\boldsymbol{W}_{\mathrm{f}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{f}}\right) \tag{2}$$

$$\boldsymbol{o}_t=\mathrm{S}\left(\boldsymbol{W}_{\mathrm{o}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{o}}\right) \tag{3}$$

$$\boldsymbol{g}_t=\tanh\left(\boldsymbol{W}_{\mathrm{c}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{c}}\right) \tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\ddot{\mathrm{A}}\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\ddot{\mathrm{A}}\boldsymbol{g}_t \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\ddot{\mathrm{A}}\tanh(\boldsymbol{c}_t) \tag{6}$$

其中，$\boldsymbol{W}_{\mathrm{i}}$, $\boldsymbol{W}_{\mathrm{f}}$, $\boldsymbol{W}_{\mathrm{o}}$, $\boldsymbol{W}_{\mathrm{c}}$ 是由可训练参数组成的映射矩阵，$\boldsymbol{b}_{\mathrm{i}}$, $\boldsymbol{b}_{\mathrm{f}}$, $\boldsymbol{b}_{\mathrm{o}}$, $\boldsymbol{b}_{\mathrm{c}}$ 是偏置项。简便起见，可以将一个长短期记忆网络记为函数 $\boldsymbol{F}_{\mathrm{LSTM}}$，在长短期记忆网络的前向传播过程中，隐藏状态的更新公式为：

$$\boldsymbol{h}_t=\boldsymbol{F}_{\mathrm{LSTM}}(\boldsymbol{h}_{t-1},\boldsymbol{x}_t,\boldsymbol{c}_{t-1}) \tag{7}$$

2 残差连接 LSTM 网络

2.1 模型构建

文中使用长短期记忆网络来建模时间序列中的时序依赖和非线性特性，结合残差连接进一步挖掘缺失数据与最近非缺失数据的潜在关联，提高网络修复能力。基于残差连接 LSTM 网络的时间序列修复模型(ReSidual Imputation Long-Short Term Memory network, RSI-LSTM)如图 2 所示：

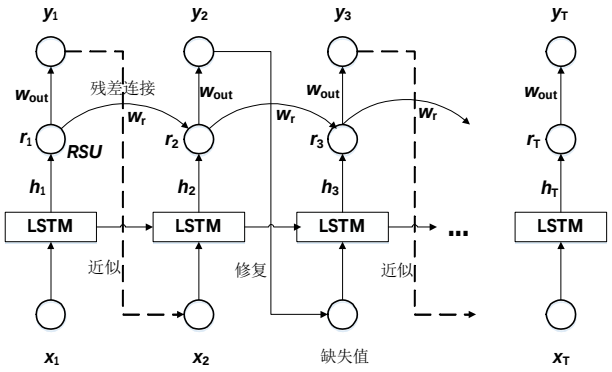

图2 RSI-LSTM 模型结构图

Fig. 2 The structure of RSI-LSTM

首先，模型输入是含缺失值的时间序列 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T\}$，其中，$\boldsymbol{x}_t\in\mathbf{R}^n$，$t$=1,2,…,T。接着，使用长短期记忆网络，将输入序列编码为一个隐藏状态序列 $\boldsymbol{h}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_T\}$，其中，$\boldsymbol{h}_t\in\mathbf{R}^m$,$t$=1,2,…,T 。特别的，在时间步 t，长短期记忆网络可以将输入 $\boldsymbol{x}_t$，编码为一个隐藏状态 $\boldsymbol{h}_t$。

在长短期记忆网络的基础上，引入残差连接的具体做法是构建一种残差和单元，用来聚合长短期记忆网络的隐藏状态和之前时刻的残差信息，有利于挖掘缺失数据和它们最近的非缺失数据的潜在关联性，提高网络修复能力。在时间步 t，残差和 $\boldsymbol{r}_t$ 的计算公式为：

$$\boldsymbol{r}_t=\begin{cases}\boldsymbol{h}_t & \text{if } t=1\\ \boldsymbol{h}_t+\boldsymbol{W}_r\boldsymbol{r}_{t-1} & \text{if } t=2,3,\ldots,T\end{cases} \tag{8}$$

其中，$\boldsymbol{r}_t\in\mathbf{R}^m$，$\boldsymbol{h}_t$ 是时间步 t 长短期记忆网络的隐藏状态，$\boldsymbol{W}_r\in\mathbf{R}^{m'm}$ 是由可训练参数组成的映射矩阵，$\boldsymbol{W}_r\boldsymbol{r}_{t-1}$ 代表了之前时刻的残差信息。

2.2 模型训练

模型训练的前向传播分为两种情况：近似过程和修复过程。如图 2 所示，虚线表示近似过程，实线表示修复过程。如果下一个时刻输入值 $\boldsymbol{x}_t$ 已知，则使用残差和 $\boldsymbol{r}_{t-1}$ 乘以一个映射矩阵 $\boldsymbol{W}_{\text{out}}$ 得到 $\boldsymbol{y}_{t-1}$，用来近似下一时刻的输入值 $\boldsymbol{x}_t$，目的是利用序列中未缺失的值来指导网络进行有效学习；如果下一个输入值 $\boldsymbol{x}_t$ 是缺失值，则用 $\boldsymbol{y}_{t-1}$ 修复 $\boldsymbol{x}_t$。计算公式为:

$$\boldsymbol{y}_{t-1}=\boldsymbol{W}_{\text{out}}\boldsymbol{r}_{t-1} \tag{9}$$

其中，$\boldsymbol{W}_{\text{out}}\in\mathbf{R}^{n'm}$ 是由可训练参数组成的映射矩阵，可以把残差和 $\boldsymbol{r}_{t-1}$ 映射到维度 n，用于近似或修复当前时刻的输入值 $\boldsymbol{x}_t$。根据当前的输入是否为缺失值，可以用一个统一的形式 $\boldsymbol{u}_t$ 来表示当前的输入值：

$$\boldsymbol{u}_t=(\boldsymbol{x}_t\odot\mathrm{I}\{\boldsymbol{x}_t\text{ 未缺失}\})\oplus(\boldsymbol{y}_{t-1}\odot\mathrm{I}\{\boldsymbol{x}_t\text{ 缺失}\}) \tag{10}$$

其中，$\odot$是逐元素的乘法，$\oplus$是逐元素的加法，$\mathrm{I}\{\boldsymbol{x}_t\}$是逐元素的指示函数，指示向量 $\boldsymbol{x}_t$ 每个维度是否缺失。因此，如果 $\boldsymbol{x}_t$ 已知，则作为网络的输入值；如果 $\boldsymbol{x}_t$ 是缺失值，则使用 $\boldsymbol{y}_{t-1}$ 修复 $\boldsymbol{x}_t$，作为当前时刻网络的输入值。在长短期记忆网络的前向传播过程中，可以将隐藏状态的更新公式重写为：

$$\boldsymbol{h}_t=\boldsymbol{F}_{\text{LSTM}}(\boldsymbol{h}_{t-1},\boldsymbol{u}_t,\boldsymbol{c}_{t-1}) \tag{11}$$

并且，在网络训练的反向传播过程中，如果当前时刻的目标值缺失，则该时刻的损失不可定义。因此，时刻 t 损失函数的计算方式如下：

$$\boldsymbol{L}_t=\|(\boldsymbol{y}_{t-1}-\boldsymbol{x}_t)\odot\mathrm{I}\{\boldsymbol{x}_t\text{未缺失}\}\|_2^2 \tag{12}$$

其中，$\mathrm{I}\{\boldsymbol{x}_t\}$是逐元素的指示函数，指示向量 $\boldsymbol{x}_t$ 每个维度是否缺失。如果定义上标 k 表示样本集合的第 k 个样本 k=1,2,…,N，总的训练损失函数为：

$$\boldsymbol{L}_{total}=\sum_{k=1}^{K}\{\sum_{t=2}^{T}\left\|(\boldsymbol{y}_{t-1}^{(k)}-\boldsymbol{x}_t^{(k)})\odot\mathrm{I}\{\boldsymbol{x}_t^{(k)}\text{未缺失}\}\right\|_2^2+\lambda\boldsymbol{L}_{\text{reg}}(\|\mathrm{W}\|)\} \tag{13}$$

其中，$\boldsymbol{L}_{\text{reg}}(\|\mathrm{W}\|)$是对模型所有参数W的二范数正则项，$\lambda$是常量，用来调节两个损失函数的权重，在实验中设为 1E-4。

2.3 算法流程

1）对原始含缺失值的时间序列进行归一化处理，将数据映射到[0,1]区间之内，得到含缺失值的时间序列 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}$，并按照 7:3 的比例将样本集划分为训练集和测试集。

2）将时间序列 $\boldsymbol{x}$ 逐时刻输入 LSTM，在时间步 t-1，LSTM 将输入 $\boldsymbol{x}_{t-1}$ 编码为隐藏状态 $\boldsymbol{h}_{t-1}$。

3）在时间步 t-1，根据 LSTM 的隐藏状态 $\boldsymbol{h}_{t-1}$ 和前一时刻的残差和 $\boldsymbol{r}_{t-2}$，计算残差和 $\boldsymbol{r}_{t-1}$。

4）如果下一个时刻输入值 $\boldsymbol{x}_t$ 已知，则使用残差和 $\boldsymbol{r}_{t-1}$ 乘以一个映射矩阵 $\boldsymbol{W}_{\text{out}}$ 得到 $\boldsymbol{y}_{t-1}$，用来近似下一时刻的输入值 $\boldsymbol{x}_t$；如果下一个输入值 $\boldsymbol{x}_t$ 是缺失值，则用 $\boldsymbol{y}_{t-1}$ 修复 $\boldsymbol{x}_t$。

5）使用随时间反向传播算法(BackPropagation Through Time, BPTT)[17]更新网络参数，并且，如果当前时刻的目标值缺失，则该时刻的损失不可定义，不计算该时刻的损失。

6）网络迭代直到收敛，最后，如果输入值 $\boldsymbol{x}_t$ 是缺失值，则可以使用上一时刻的预测值 $\boldsymbol{y}_{t-1}$ 作为 $\boldsymbol{x}_t$ 的修复值，将整条时间序列的缺失值修复完毕，即可得到完整的时间序列。

3 实证分析

3.1 数据来源

本文采用 2016 年 1 月 1 日到 2018 年 6 月 30 日南方某省的区域日供电量数据，供电量数据计量单位为天，因此每个区域有 912 个时间点。随机抽取 10 个区域的序列数据作为样本集，作为模型在单变量情况下的输入。

考虑到气象因素对电力数据的影响，本文还采用了 2012 年 1 月 1 日到 2014 年 12 月 31 日两个地区的电力负荷数据。

数据集来源于第九届电工数学建模竞赛 A 题，除了地区电力负荷数据，该数据集还提供每日的最高温度、最低温度、平均温度、相对湿度和降雨量五个气象因素数据。本文分别抽取每日的 6 点、12 点和 18 点作为原始数据，因此总的样本集有 6 个，分别记为 Electric1~Electric6，每个样本集中包含 1096 个时间点，每个时间点的数据为一个 6 维的向量，包含电力负荷数据以及五个气象因素数据，作为模型在多变量情况下的输入。

同时，本文还使用了两个通用的时间序列数据集 Libras 和 Character Trajectories，数据来源于 UCI 库[18]，以进一步地进行多变量情况下的模型效果验证。

对于每一个数据集的时间序列，我们取前 70%的序列作为训练集，后 30%作为测试集。接着，为了处理输入不同量纲的问题，需要对原始的序列 $\boldsymbol{s}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots,\boldsymbol{s}_T\}$进行归一化：

$$\boldsymbol{x}_t=\frac{\boldsymbol{s}_t-\boldsymbol{s}_{\min}}{\boldsymbol{s}_{\max}-\boldsymbol{s}_{\min}} \tag{14}$$

其中，$\boldsymbol{x}_t\in\mathbf{R}^n$, $\boldsymbol{s}_t\in\mathbf{R}^n$, t=1,2,…,T, $\boldsymbol{s}_{\max}$ 和 $\boldsymbol{s}_{\min}$ 分别表示时间序列的最大值和最小值。并且，对于现实缺失数据，无法获得对应的真实值来进行算法的性能评估。因此，在完整的时间序列的基础上，以一定的缺失率构造含有缺失值的数据。将缺失率设置为 10%，让完整的序列数据按 10%的概率随机缺失，构造出含有缺失值的时间序列，作为模型的输入。同时，缺失值对应的真实值将被用于评估修复算法的性能。

3.2 算例设置

本文同时考虑了单变量输入和多变量输入两种情况。在单变量的情况中，使用南方某省的区域日供电量序列作为模型输入。在多变量的情况中，使用了两个地区的电力负荷数据，结合气象数据作为模型的多变量输入。同时，本文还使用了两个通用的时间序列数据集来辅助进行多变量情况下的模型效果验证。本文采用均方误差(Mean Squared Error, MSE)作为评价指标：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\text{real}}-\boldsymbol{x}_i^{\text{imp}})^2 \tag{15}$$

其中，n 代表序列中缺失值的数量，$\boldsymbol{x}_i^{\text{real}}$ 和 $\boldsymbol{x}_i^{\text{imp}}$ 分别代表第 i 个缺失值对应的真实值和修复值。

对于区域日供电量数据集，模型的输入是单变量的，因而将所提出模型(RSI-LSTM)与常用的单变量缺失修复方法进行对比，对比方法包括卡尔曼滤波(Kalman)[19]、线性插值(Interpolation)[20]、移动平均(Moving Average，MA)[21]和基线模型长短期记忆网络(LSTM)。同时，也将所提出模型与两种先进的缺失修复方法进行对比，分别是生成对抗填补网络(Generative Adversarial Imputation Nets, GAIN)[22]和双向循环填补(Bidirectional Recurrent Imputation for Time Series, BRITS)[23]。表 1 是上述方法的数据缺失修复结果。

从表 1 中可以看出，RSI-LSTM 的修复性能优于 GAIN、BRITS、Kalman、Interpolation 和 MA，表现为均方误差的总体降低，这说明了 RSI-LSTM 可以更好地建模时间序列的信息，修复缺失的时间序列。同时，RSI-LSTM 相比基线模型 LSTM，修复误差有一定的降低，进一步证明了引入残差连接的有效性，因为引入残差连接有利于挖掘缺失数据和它们最近的非缺失数据的潜在关联性，提高网络的缺失修复能力。

对于地区电力负荷数据集以及两个时间序列数据集，输入是多变量的。因此，我们将 RSI-LSTM 与四种近年来最先进的修复方法 GAIN[22]、 BRITS[23]、基于傅里叶的延迟 k 最近邻算法(Fourier-based Lagged k-Nearest Neighbor, FL k-NN)[24]以及动态缺失值的挖掘算法(Dynamics mining with missing values, Dynammo)[25]进行了对比实验，这些方法描述如下：

表1 单变量数据集修复误差(MSE)对比

Tab. 1 Comparison of imputation errors (MSE) on univariate time series datasets

Data set	RSI-LSTM	LSTM	GAIN [22]	BRITS [23]	Kalman [19]	Interpolation[20]	MA [21]
1	**8.90E-04**	9.30E-04	3.06E-03	3.84E-02	9.40E-03	1.25E-02	1.49E-02
2	**4.06E-03**	5.18E-03	7.30E-03	1.88E-02	7.82E-02	7.28E-02	7.64E-02
3	**8.37E-03**	8.38E-03	2.56E-02	2.28E-02	5.10E-02	6.75E-02	6.46E-02
4	**2.66E-03**	3.73E-03	6.42E-03	1.42E-02	4.24E-02	1.73E-02	2.04E-02
5	**7.00E-04**	7.50E-04	5.73E-03	3.24E-02	2.26E-02	7.40E-03	7.40E-03
6	**3.78E-03**	4.09E-03	7.12E-03	2.40E-02	1.37E-02	2.09E-02	1.62E-02
7	**8.57E-03**	9.80E-03	1.20E-02	1.53E-02	3.17E-02	4.67E-02	3.69E-02
8	**7.51E-03**	7.81E-03	7.73E-03	2.32E-02	1.75E-02	2.62E-02	2.54E-02
9	**5.05E-03**	6.25E-03	1.65E-02	3.16E-02	4.63E-02	4.36E-02	4.81E-02
10	**9.03E-03**	9.28E-03	9.15E-03	2.07E-02	3.69E-02	6.68E-02	4.51E-02

表2 多变量数据集修复误差(MSE)对比

Tab. 2 Comparison of imputation errors (MSE) on multivariate time series datasets

Data set	RSI-LSTM	LSTM	GAIN [22]	BRITS [23]	FL k-NN [24]	Dynammo [25]
Electric1	**4.23E-03**	6.44E-03	4.41E-03	4.98E-03	6.79E-03	5.82E-03
Electric2	**1.20E-02**	1.43E-02	1.79E-02	1.51E-02	1.50E-02	1.45E-02

Electric3	**1.43E-02**	1.73E-02	3.71E-02	1.96E-02	1.76E-02	1.77E-02
Electric4	**1.99E-03**	2.48E-03	6.64E-03	3.91E-03	4.6E-03	2.37E-03
Electric5	**4.45E-03**	6.78E-03	5.85E-03	9.45E-03	8.42E-03	7.39E-03
Electric6	**7.08E-03**	8.52E-03	7.21E-03	8.84E-03	8.45E-03	8.97E-03
Libras	**9.22E-03**	1.18E-02	1.14E-02	2.32E-02	2.97E-02	1.42E-02
Character Trajectories	**7.10E-03**	7.68E-03	8.20E-03	7.23E-03	8.12E-03	9.52E-03

1) GAIN：使用生成对抗网络来进行缺失修复，并提出一种提示向量来辅助模型训练，但训练数据较少时训练困难。

2) BRITS：该方法使用双向循环神经网络来进行时间序列的缺失值修复，但在连续缺失的情况下效果较差。

3) FL k-NN：结合滞后的 k 最近邻方法和傅立叶变换的集成方法，该方法较为复杂，需要大量的人工选择的超参数。

4) Dynammo：该方法基于期望最大化方法和卡尔曼滤波。它会在存在缺失值的情况下学习线性动力学系统，并对缺失值进行估计。该方法假设时间序列具有潜在的线性动力学特性，然而时间序列更多地表现为非线性。

表 2 是上述模型的修复结果。可以看出，RSI-LSTM 的修复性能明显优于 GAIN、BRITS、FL k-NN 以及 Dynammo，修复效果相比这几种方法有较大的提升。通过与这四种近年来最先进的方法作对比，RSI-LSTM 可以取得当前最好的结果。同时，相比基线模型 LSTM，RSI-LSTM 的修复效果也有一定的提升，验证了所提出模型的有效性。

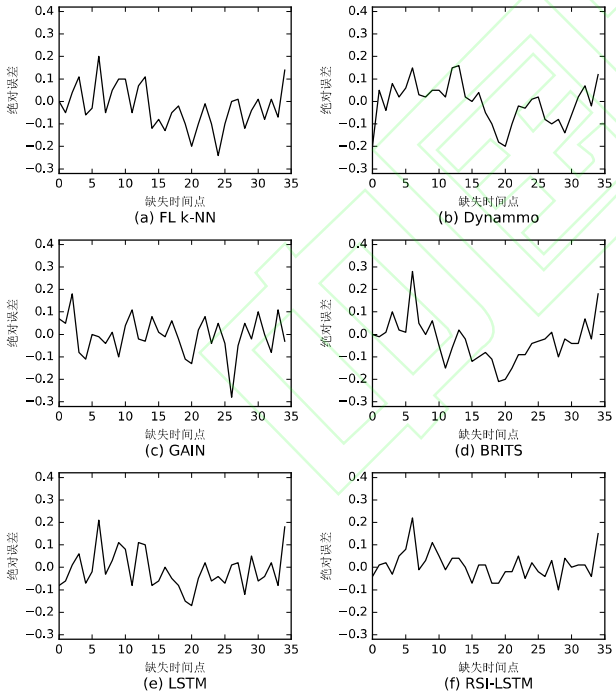

图3 绝对误差对比

Fig. 3 Comparison of absolute errors

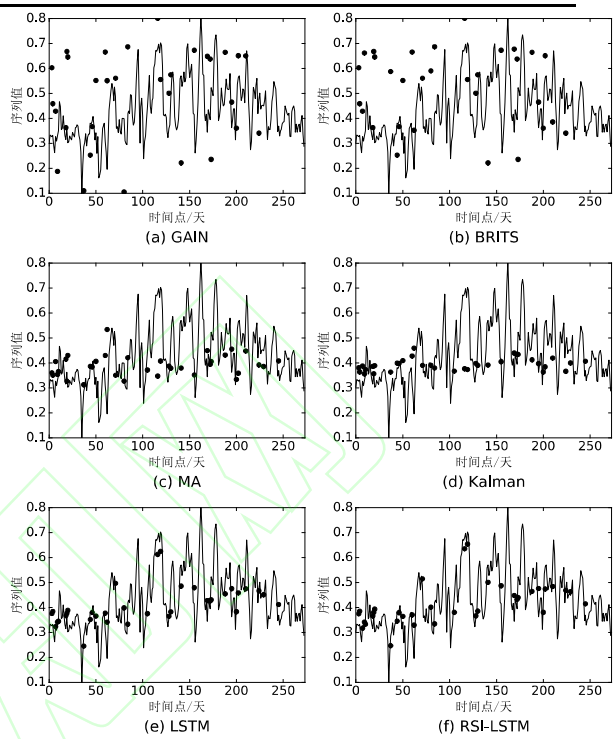

图4 修复结果对比

Fig. 4 Comparison of imputation result

更进一步的，我们随机抽取其中 1 个区域的电力负荷数据序列，使用上述各种多变量缺失修复方法对其进行修复。我们计算数据序列中缺失时间点的修复值和真实值的绝对误差，绝对误差在 0 附近波动，偏离 0 越远，说明误差越大。结果如图 3 所示。由图 3 可以看出，对于电力负荷数据序列绝大多数的缺失时间点，RSI-LSTM 相比其它方法，可以取得更好的修复效果。具体地，我们的模型的误差曲线相较其他对比方法更为平滑，在 0 附近波动较小。

为了能更直观地展示缺失值修复效果，我们随机抽取 1 个区域的日供电量序列，画出上述各种单变量缺失修复方法的修复结果（由于 Interpolation 效果较差，在此不做可视化）。如图 4 所示，对于绝大多数的缺失时间点，RSI-LSTM 相比其它方法，可以取得较好的修复效果，修复值大多和原始的时间序列重合。而相比基线模型 LSTM，RSI-LSTM 在峰值处可以取得较好的效果。因为残差和单元可以更好地挖掘缺失数据和它们最近的非缺失数据的潜在关联性，提高网络的修复能力。

4 结语

1) RSI-LSTM 使用长短期记忆网络对时间序列中的时间依赖和非线性特性进行建模，并且引入残差连接，挖掘缺失数据和它们最近的非缺失数据的潜在关联性。同时，该模型可以直接在含缺失值的情况下进行模型训练；

2) 本文同时考虑了单变量输入和多变量输入两种情况，实验结果证明了该模型对时间序列缺失修复的有效性；

3）目前 RSI-LSTM 只是应用于时间序列的缺失值修复上，以后的研究工作中，将考虑将该模型扩展到含缺失数据的时间序列预测或分类等问题。

参考文献

[1] KAMPOURAKI A, MANIS G, NIKOU C. Heartbeat time series classification with support vector machines[J]. IEEE Transactions on Information Technology in Biomedicine, 2009, 13(4):512-518.

[2] ELMOAQET H, TILBURY D M, RAMACHANDRAN S K. Multi-step ahead predictions for critical levels in physiological time series[J]. IEEE Transactions on Cybernetics, 2017, 46(7):1704-1714.

[3] ANWAR T, LIU C, VU H L, et al. Capturing the spatiotemporal evolution in road traffic networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8):1426-1439.

[4] RUBIN D B. Inference and Missing Data[J]. Biometrika, 1976, 63(3): 581-592.

[5] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553):436–444.

[6] GRACIA-LAENCINA P J, SANCHO-GOMEZ J L, FIGUEIRAS-VIDAL A R. Pattern classification with missing data: a review[J]. Neural Computing & Applications, 2010, 19(2):263-282.

[7] KREINDLER D M, LUMSDEN C J. The Effects of the Irregular Sample and Missing Data in Time Series Analysis[J]. Nonlinear Dynamics Psychology & Life Sciences, 2006, 10(2):187-214.

[8] 王立斌,李振东,姚杨,等. 一种用电信息采集系统异常电量数据的识别与修复方法[J]. 电力大数据, 2018, 21(03)：74-78. (WANG L B, LI Z D, YAO Y, et al．A method of identifying and repairing abnormal electric energy data in electric information acquisition system[J]．Power Systems and Big Data, 2018, 21(03)：74-78)

[9] 毛李帆,姚建刚,金永顺,等. 中长期负荷预测的异常数据辨识与缺失数据处理[J]. 电网技术, 2010, 34(07)：148-153. (MAO L F, YAO J G, JIN Y S, et al．Abnormal data identification and missing data filling in medium-and long-term load forecasting[J]．Power System Technology, 2010, 34 (7)：148-153)

[10] 王一蓉,王瑞杰,陈文刚,等. 基于遗传优化的调控系统缺失数据填补算法[J]. 电力系统保护与控制, 2016, 44(21)：182-186. (WANG Y R, WANG R J, Chen W G, et al．A missing data padding algorithm for control system based on genetic optimization[J]．Power System Protection and Control, 2016, 44(21)：182-186)

[11] 许子龙. 基于自适应 BP 神经网络的风电数据缺失数据处理[C]. 第十五届沈阳科学学术年会论文集, 沈阳：沈阳市科学技术协会, 2018：382-387. (XU Z L．A missing data processing method for wind power data based on adaptive BP neural network[C]// Proceedings of the 15th Shenyang Science Academic Conference. Shenyang: Shenyang Association for Science and Technology, 2018: 382-387)

[12] 尹玲,尹京苑,孙宪坤,等. 缺失GPS时间序列的神经网络补全[J]. 测绘科学技术学报, 2018, 35(04): 331-336. (YIN L, YIN J Y, SUN X K, et al. Reconstruction of Gappy GPS Coordinate Time Series Based on Long Short-Term Memory Networks[J]. Journal of Geomatics Science and Technology, 2018, 35(04): 331-336)

[13] 辜超,白德盟,王晶,等. LSTM 在输变电设备缺失值填补中的应用[J]. 电测与仪表, 2019, 056(005):63-69,142. (GU C, BAI D M, WANG J, et al．Application of LSTM in filling missing value of power transmission and transformation equipment[J]．Electrical Measurement & Instrumentation：2019, 056(005):63-69,142.)

[14] HOCHREITER S, SCHMIDHUBER J．Long short-term memory[J]．Neural computation, 1997, 9(8)：1735-1780

[15] HE K, ZHANG X, REN S, et al．Deep residual learning for image recognition[C]// Proceedings of the 2016 IEEE conference on computer vision and pattern recognition．Piscataway: IEEE, 2016：770-778.

[16] HE K, ZHANG X, REN S, et al．Identity mappings in deep residual networks[C]// Proceedings of the 14th European Conference on Computer Vision．Berlin: Springer, 2016：630-645.

[17] WITASZEK J. Backpropagation: Theory, architectures, applications: By Yves Chauvin and David E. Rumelhart (eds). Lawrence Erlbaum, Hillsdale, NJ, Hove, UK, 1995. ISBN 0-8058-1258-X, pp. 561[J]. Neurocomputing, 1995, 9(3):358–359.

[18] DHEERU D and TANISKIDOU E K. UCI machine learning repository[EB/OL].[2019-11-10]. http://archive.ics.uci.edu/ml

[19] GREWAL M S. Kalman filtering [M]. Berlin: Springer, 2011: 705-708.

[20] GREENSTEIN D S. Interpolation and Approximation[J]. Siam Review, 1965, 7(1): 151-152.

[21] NELSON C R. The first-order moving average process: Identification, estimation and prediction [J]. Journal of Econometrics, 1974, 2(2): 121-141.

[22] JINSUNG Y. GAIN: Missing data imputation using generative adversarial nets[C]// Proceedings of the 2018 International Conference of Machine Learning. New York: ACM, 2018: 5675-5684.

[23] WEI C, DONG W, JIAN L, et al. BRITS: Bidirectional Recurrent Imputation for Time Series[C]// Proceedings of the 32nd Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2018: 6775-6785.

[24] RAHMAN S A, HUANG Y, CLAASSEN J, et al. Imputation of Missing Values in Time Series with Lagged Correlations[C]// Proceedings of the 2014 International Conference on Data Mining Workshop (ICDMW). Piscataway: IEEE, 2014: 753-762.

[25] LI L, MCCANN J, POLLARD N S, et al. DynaMMo: mining and summarization of coevolving sequences with missing values[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 507-516.

This work is partially supported by the Key Projects of National Natural Science Foundation of China (61751205), the National Natural Science Foundation of China (61872148).

QIAN Bin, born in 1989. Master, Engineer. His main research interest is electric energy metering technology.
ZHENG Kaihong, born in 1991. Master, Engineer. His main research interests include Electric energy metering, electric energy metering automation system and electricity technology.
CHEN Zipeng, born in 1996. postgraduate. His main research interests include data mining, machine learning, and neural networks.

XIAO Yong, born in 1978. PhD, Senior engineer. His main research interests include Electric energy metering, electric energy metering automation system and electricity technology.
LI Sen, born in 1994. postgraduate. His main research interests include data mining, machine learning, and neural networks.
YE Chunzhuang, born in 1989. Bachelor, Engineer. His main research interest is power line loss management.
MA Qianli, born in 1980. PhD, Professor. His main research interests include data mining, machine learning, and neural networks.