ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 3,May2024,pp. 598－608
DOI:10. 16337/j. 1004-9037. 2024. 03. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

联合张量补全与循环神经网络的时间序列插补法

何　军[1,2]，赖赵远[1]，时　勘[2]

(1. 中国人民大学信息学院，北京 100872；2. 温州大学温州发展模式研究院，温州 325035)

摘　要：现存的插补方法大致分为基于统计的插补法和基于深度学习的插补法。基于统计的插补法只能捕捉线性时间关系，导致无法精准建模时间序列的非线性关系；基于深度学习的插补法往往没有考虑到不同时间序列之间的相关性。针对现有方法的问题，本文提出了联合张量补全与循环神经网络的时间序列插补法。首先，将多元时间序列建模成张量，通过张量的低秩补全捕获不同时间序列之间的关系。其次，提出了一个基于时间的动态权重，将张量插补结果和循环神经网络的预测结果进行融合，避免因为连续缺失导致的预测误差累积。最后，在多个真实的时间序列数据集上对所提方法进行了实验评估，结果显示该模型优于已有相关模型，且基于插补后的时间序列可以提升时间序列预测效果。

关键词：张量补全；时间序列插补；循环神经网络

中图分类号：TP391　　**文献标志码：**A

Time Series Imputation Method Combining Tensor Completion and Recurrent Neural Network

HE Jun[1,2], LAI Zhaoyuan[1], SHI Kan[2]

(1. School of Information, Renmin University of China, Beijing 100872, China; 2. Academy of Wenzhou Model Development, Wenzhou University, Wenzhou 325035, China)

Abstract: The existing imputation methods are roughly divided into statistical methods and deep learning methods. The statistical methods can only capture the linear time relationship, which makes it impossible to accurately capture the relationship of non-linear time series data. The deep learning imputation methods usually donot consider the correlation between different time series. To solve these problems, a new model jointing the tensor completion and the recurrent neural network is proposed. Firstly, the multivariate time series are modeled as a tensor, and the correlation of different time series is captured by low rank tensor completion. Secondly, a time based dynamic weight is proposed to fuse the tensor completion results with the prediction results of the recurrent neural network to avoid the accumulation of prediction error caused by continuous missing. The proposed method is evaluated on several real time series datasets, and the results show that the proposed model outperforms the existing models in term of imputation accuracy, which is helpful for improving classification and regression accuracy.

Key words: tensor completion; time series imputation; recurrent neural network

---

基金项目：浙江省哲学社会科学新兴(交叉)重大项目(21XXJC04ZD)。
收稿日期：2023-09-15；**修订日期：**2023-12-10

引　言

时间序列分析通过对具有时序关系的一组数据进行建模,从而得到这组数据背后的变化规律,具有广泛的应用领域,如气象观测[1]、预测股票价格变动[2]、指示患者的适应度和诊断类别[3]等。多数时间序列分析模型假设时间序列数据是完整的。然而在现实世界中,存在各种情况导致时间序列数据不完整。以 Physoninet Challenge 2012数据集[4]为例,数据的缺失率在80%以上。时间序列数据中的缺失值会破坏时间序列中的时间相关性[5],使得现有的各种时间序列模型难以应用,增加了时间序列预测任务的难度[6]。因此面对有缺失值的时间序列数据,需要研究对时间序列中的缺失值进行插补的方法,从而提高时间序列分析模型的准确性;另一方面对采集成本高的数据可以降低采集成本。虽然已经有对时间序列缺失值插补的研究,但在多数的研究方法中很少同时考虑单个时间序列的时间依赖性和多个时间序列之间的相关性。通过研究单个时间序列的时间依赖性,可以使插补的缺失值更符合整个时间序列的变化规律;通过研究多个时间序列之间的相关性,可以获得时间序列的全局时间模式,例如,空气质量数据集是由多个不同地区的监测站记录数据构成的,其中邻近监测站记录的数据具有较强的相关性。在缺失率较高的时间段,单个时间序列的时间依赖性由于缺失率较高导致其不可靠,此时可以从其他相关的时间序列中获得其变化规律。如何同时考虑这两个问题,对时间序列缺失值的插补十分重要。针对时间序列缺失值插补的任务,本文提出了一种联合张量补全与循环神经网络(Recurrent neural network, RNN)的时间序列缺失值插补模型——张量补全联合双向循环神经网络(Tensor completion bidirectional recurrent neural network, TCBRNN)。在该模型中,将多个时间序列建模成张量,通过张量补全可以捕获时间序列数据的全局模式。补全后的张量与循环神经网络动态结合,建模时间序列的时序信息,获得单个时间序列的时序模式,即局部模式。联合局部和全局的信息从而对多个时间序列的缺失值完成插补。

本文提出了一种联合张量补全与循环神经网络的插补模型。张量补全捕获多个时间序列之间的关系,发现数据的全局模式,循环神经网络建模时间序列的时序关系。为了让张量补全与循环神经网络有机地结合,提出了一个基于连续缺失时间的动态权重因子,避免因连续缺失时间长导致预测误差累积。在多个真实数据集上进行了实验,以评估所提出的缺失值插补模型。实验结果表明,本文模型比现有的模型具有更好的性能。

1　相关工作

缺失值插补在数据挖掘与分析中具有重要的作用,时间序列的缺失值插补有着更为广泛的应用和重要的研究意义。下面对已有的时间序列缺失值的插补方法做一概述分析。

1.1　基于删除的插补法

早期的时间序列缺失值插补通常采用基于删除的插补方法,例如成列删除[7]和成对删除[8]。成列删除是指在进行统计量计算时,把含有缺失值的时间序列记录删除。成对删除适用于两两配对的变量,如果某个时间序列数据含有一个或多个两两配对的变量,其中一个配对变量存在数据缺失,则对这对配对变量的统计量计算时需要将含有缺失值的所有数据删除,在计算其他变量的统计量时不变。

1.2　基于统计学的插补法

基于统计学的插补法是伴随着统计学发展而进步的一种方法。平均值插补法[9]和中位数插补法[10]是两种最简单的统计学插补法。1980年,文献[11]讨论了利用Kalman滤波器计算平稳自回归滑动平均(Auto-regressive moving average model, ARMA)过程的可能性,其中还展示了如何处理平稳序列的缺失观测值。在1982年,Harvey将这些技术推广到非平稳 ARIMA 模型中缺失值的情况[12]。

1.3 基于机器学习和深度学习的插补法

常见的基于机器学习插补法有正则化期望最大化(Expectation-maximum, EM)[13],K近邻(K-nearest neighbor, KNN)[14]和矩阵分解(Matrix factorization, MF)[15]等。其中MF将矩阵分解得到两个因子矩阵,将因子矩阵相乘得到缺失值插补结果。Yu等[16]对矩阵分解的插补法提出了改进,提出了一个支持数据驱动的时间学习和预测的时间正则化矩阵分解(Temporal regularized matrix factorization, TRMF)框架。

随着深度学习的发展,人们开始使用RNN进行时间序列缺失值的插补。Yoon等[17]在2017年提出了M-RNN并在医学问题上取得了突破。Che等[18]在2018年对门循环单元(Gate recurrent unit, GRU)进行了改进,提出了GRU-D。Cao等[19]在2018年提出了BRITS,并且打破了前两个模型在医疗数据集上的限制,取得了最好的插补效果。

1.4 张量补全相关研究

Song等[20]在2019年发表了一篇有关张量补全研究的综述论文,从通用张量补全算法、使用辅助信息的张量补全(多样性)、可扩展张量补全算法(体积)和动态张量补全算法(速度)4个方面介绍有关张量补全近年来的研究进展,还讨论了该研究的主要挑战和可能的一些研究方向。目前有关张量补全理论方面的研究进展不大,大多研究都是针对一些应用领域所做的方法改进。例如,智能交通系统中现有的交通监测方法难以应用于拓扑结构复杂、状态多变的城市交通。Zhao等[21]提出了一种时空约束低秩张量补全(Space time low-rank tensor completion, ST-LRTC)方法,该方法可以增强张量补全模型对复杂城市场景的适应性,从而更好地监测、诊断和优化城市交通状态。Liu等[22]针对物联网流数据的特性,假设数据张量位于时变子空间中,建立了一个基于动态分解的可更新框架,引入了一种称为时态多方面流的算法来解决源自开发模型的优化问题。本文结合张量补全与循环神经网络各自的特点,提出了一种比较通用的联合张量补全与循环神经网络的插补法。

2 问题定义

定义1(多元时间序列) 多元时间序列$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}$是1个有$T$个观测值的序列,第$t$个时间步观测值$\boldsymbol{x}_t\in\mathbf{R}^D$包含$D$个特征$\left\{x_t^1,x_t^2,\cdots,x_t^D\right\}$,其中第$d$个特征对应的时间序列为$\{x_1^d,x_2^d,\cdots,x_T^d\}$。

定义2(掩码向量) 为了表示$\boldsymbol{x}_t$中的缺失值,引入掩码向量$\boldsymbol{m}_t=(m_t^1,m_t^2,\cdots,m_t^D)$,计算公式为

$$m_t^d=\begin{cases}0 & x_t^d\text{未被观测到}\\ 1 & \text{其他}\end{cases}\tag{1}$$

定义3(时间间隔向量) 为了表示时间序列中连续缺失值,引入时间间隔向量$\boldsymbol{\delta}_t=(\delta_t^1,\delta_t^2,\cdots,\delta_t^D)$,其计算公式如下

$$\boldsymbol{\delta}_t^d=\begin{cases}\boldsymbol{s}_t-\boldsymbol{s}_{t-1}+\boldsymbol{\delta}_{t-1}^d & t>1,m_{t-1}^d=0\\ \boldsymbol{s}_t-\boldsymbol{s}_{t-1} & t>1,m_{t-1}^d=1\\ 0 & t=1\end{cases}\tag{2}$$

式中s_t代表第t个时间步的时间戳。

假设$X^1,X^2,\cdots,X^K$为K个多元时间序列,由于现实世界的各种情况,这个多元时间序列的数据包含缺失值。基于该给定的多元时间序列,时间序列插补任务是设计模型,填补所有时间序列中的缺失值,使插补的缺失值与真实值尽量接近。

下面结合一个例子说明以上概念。图1(a)表示一个多元时间序列,其中$x_1,x_2,\cdots,x_6$分别为第1天至第6天的观测值,时间戳$\boldsymbol{s}_t=t,t=1,2,3,4,5,6$。图1(a)中的时间序列存在缺失值,根据时间序列的

缺失情况,可以计算出掩码向量如图1(b)所示,观测值存在记为1,不存在记为0。根据掩码向量和时间戳s_t可以计算出时间间隔向量如图1(c)所示,以δ_6^2为例:$\delta_6^2=6-5+\delta_5^2=4$。

序列1			141	6.5	3	9
序列2	211	356				11
序列3		420		2.2	4	13
	x_1	x_2	x_3	x_4	x_5	x_6

(a) Multivariate time series

0	0	1	1	1	1
1	1	0	0	0	1
0	1	0	1	1	1
m_1	m_2	m_3	m_4	m_5	m_6

(b) Mask vector

0	1	2	1	1	1
0	1	1	2	3	4
0	1	1	2	1	1
δ_1	δ_2	δ_3	δ_4	δ_5	δ_6

(c) Time gap vector

图1　多元时间序列、掩码向量和时间间隔向量示意图

Fig.1　Multivariate time series, mask vector and time gap vector

3　联合张量补全的循环神经网络模型

本文所提出的联合张量补全的循环神经网络模型TCBRNN整体结构如图2所示,其中$C_i(i=1,2,\cdots)$为缺失值;$H_i(i=1,2,\cdots)$为隐藏状态变量。首先,将多个多元时间序列建模为一个三维张量,利用张量的低秩补全可以得到一个初步插补结果。然后将多元时间序列和补全后的张量输入到一个双向循环神经网络中。最后,将循环神经网络的预测值和张量补全的预测值通过基于时间的动态权重进行加和,得到最终的插补结果。

下面给出TCBRNN算法的伪代码描述。

输入:n个有缺失值的多元时间序列

$\{X^1,X^2,\cdots,X^n\}$,其中$X^i=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t)$,$x_t\in\mathbf{R}^d$

输出:n个插补后的多元时间序列

$\{X^{1'},X^{2'},\cdots,X^{n'}\}$

(1)初始化张量分解的参数A、B、C,初始化循环神经网络的所有参数W、U、b

(2)将输入的多元时间序列拼接成一个张量

(3)for i in epoch:

(4)　对张量进行分解,更新参数A、B、C

(5)　for t in time:

(6)　　计算时间步$t-1$的估计值$\hat{x}_t$

(7)　　计算时间步$t-1$的插补值c_{t-1}

(8)　通过最小化损失函数l_t来更新循环神经网络的参数W、U、b

(9)返回插补后的多元时间序列$\{X^{1'},X^{2'},\cdots,X^{n'}\}$

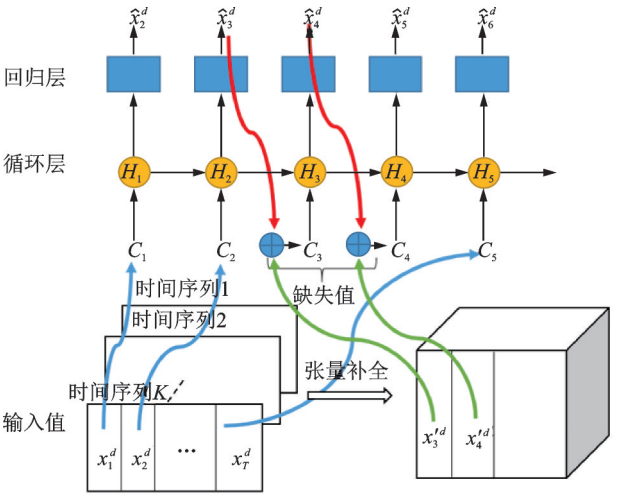

图2　TCBRNN模型结构图

Fig.2　Structure of TCBRNN model

3.1　模型的输入

模型的整体输入是多个多元时间序列,所以需要将多个多元时间序列转化成为一个三维张量。将一个多元时间序列i记为$X^i=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$,其中$\boldsymbol{x}_t\in\mathbf{R}^D$,对于多个多元时间序列将其记为$\{X^1,X^2,\cdots,X^K\}$,采用直接拼接的方式将其转化为一个三维张量$\chi\in\mathbf{R}^{K\times D\times T}$,$\chi$为一个含有缺失值的不完整张量。

3.2　低秩张量补全

多元时间序列的不同特性之间以及不同的时间序列之间通常存在相关性。例如,同一个地点的多

个气象观测指标之间以及相邻地点的气象观测序列时间存在相关性。为了捕获这些关系,本文所提方法的第一步采用张量补全方法,得到一个缺失值的初步插补结果。对有连续缺失值的时间序列数据用时间间隔向量表达,对有相关性的时间序列数据用低秩补全张量补全缺失值。

利用张量的低秩逼近是一种张量补全的方法。一般来说,希望通过最小化张量$\boldsymbol{X}$的秩,通过张量的低秩性质对张量进行补全。最小化张量$\boldsymbol{X}$的秩可以通过凸松弛为最小化张量$\boldsymbol{X}$的核范数,所以这一步需要对最小化张量$\boldsymbol{X}$核范数的目标函数进行求解。给定一个张量$\boldsymbol{X}$,张量补全的优化任务为

$$\begin{cases}\min\limits_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}) \\ \text{s.t.} \quad \boldsymbol{X}_{\Omega}=\boldsymbol{M}_{\Omega}\end{cases} \tag{3}$$

式中$\boldsymbol{X}_{\Omega}=\boldsymbol{M}_{\Omega}$表示补全后的张量与原张量在未缺失值的位置上相等。

张量的秩优化问题是非凸优化不可求解,所以可以进行一个等价代换,通过使张量的迹范数$|\boldsymbol{X}|_{*}$最小化从而近似逼近张量的最小秩。优化问题转化为

$$\begin{cases}\min\limits_{\boldsymbol{X}} |\boldsymbol{X}|_{*} \\ \text{s.t.}\ \boldsymbol{X}_{\Omega}=\boldsymbol{M}_{\Omega}\end{cases} \tag{4}$$

在张量运算中,张量$\chi\in\mathbf{R}^{K\times D\times T}$可以展开为$\chi\in\mathbf{R}^{K\times(DT)}$,$\chi\in\mathbf{R}^{D\times(KT)}$,$\chi\in\mathbf{R}^{T\times(KD)}$三种矩阵形式,分别记为$\chi_{(1)}$,$\chi_{(2)}$,$\chi_{(3)}$,为了获得张量不同特征间的交互信息,1个张量可以写成3个张量展开矩阵的权重和形式$\chi=\sum\limits_{k=1,2,3}\alpha_k\chi_{(k)}$,$\sum\limits_{k=1,2,3}\alpha_k=1$,从而式(4)可以转化为

$$\begin{cases}\min\limits_{\boldsymbol{X}} \sum\limits_{k=1,2,3}\alpha_k\left|\boldsymbol{X}_{(k)}\right|_{*} \\ \text{s.t.}\ \boldsymbol{X}_{\Omega}=\boldsymbol{M}_{\Omega} \\ \quad\ \sum\limits_{k=1,2,3}\alpha_k=1\end{cases} \tag{5}$$

矩阵迹范数的优化问题可以由交替方向乘子法(Alternating direction method of multipliers, ADMM)算法进行求解,并且由Liu等[23]给出了具体的算法过程,通过更新矩阵的元素对矩阵的最小秩进行逼近,在更新矩阵元素的过程中完成张量补全。在张量补全的过程中,通过对张量沿3个方向进行展开,可以学习到张量3个特征之间两两交互的关系。

3.3 基于时间的动态权重

在张量补全的基础上,所提模型进一步利用循环神经网络进行缺失值的预测。循环神经网络的输入是每一个多元时间序列中每个特征的时间序列,假设$\{x_1^d,x_2^d,\cdots,x_T^d\}$代表的是任一个多元时间序列中第$d$个特征的时间序列,其中每个时间步的取值作为循环神经网络的对应时间步的输入。对于时间步t,如果x_t^d缺失,则利用张量补全步骤中得到的值$x_t^{d'}$和循环神经网络上一时间步的预测值$\hat{x}_t^d$,共同计算得到插补值I_t^d。$\hat{x}_t^d$的计算方法如下

$$\hat{x}_t^d=\boldsymbol{W}_x\boldsymbol{h}_{t-1}+\boldsymbol{b}_x \tag{6}$$

式中:$\boldsymbol{h}_{t-1}$为时间步$t-1$的隐层状态向量;$\boldsymbol{W}_x$和$\boldsymbol{b}_x$是要学习的网络参数,分别对应权重向量和偏置变量。

如果存在连续的缺失值,后面的预测值就会累积前面插补产生的误差,随着连续缺失的时间变大会产生累积效应。为了解决此问题,希望连续缺失天数越长,循环神经网络的预测值$\hat{x}_t^d$的权重越小。因此,本模型设计了基于时间的动态权重,公式如下

$$\gamma_t^d=\exp\left\{-\max\left(0,W_\gamma\delta_t^d+b_\gamma\right)\right\} \tag{7}$$

从式(7)可以看出γ_t在缺失天数越大时越接近0,缺失天数越小时越接近1,所以最终插补值的计算

公式为

$$I_t^d = \hat{x}_t^d \times \gamma_t^d + \left(1 - \gamma_t^d\right) \times \hat{x}_t^{d'} \tag{8}$$

式中:$\hat{x}_t^d$为RNN前一个时间步的预测值;$\hat{x}_t^{d'}$为张量补全得到的插补值。

将初始隐藏状态$\boldsymbol{h}_0$初始化为零向量,然后通过以下方式更新模型

$$c_t^d = m_t^d \times x_t^d + \left(1 - m_t^d\right) \times I_t^d \tag{9}$$

$$\boldsymbol{h}_t = \mathrm{RNN}\left(\boldsymbol{h}_{t-1}, c_t^d\right) \tag{10}$$

$$\hat{x}_{t+1}^d = \boldsymbol{W}_x \boldsymbol{h}_t + \boldsymbol{b}_x \tag{11}$$

$$l = \sum_t \sum_d m_t^d \left(x_t^d - I_t^d\right)^2 \tag{12}$$

式(9)计算的是RNN的输入值,未缺失值直接用真实值作为输入,缺失值用张量补全的插补值和循环神经网络的估计值加权和替代。式(10)表示利用循环神经网络RNN计算隐藏状态的计算方式。式(11)计算时间t步的预测值,最后在式(12)中使用平方误差来计算一个时间序列的损失函数。

由于循环神经网络RNN估计缺失值的误差会被延迟到下一次观测的出现。为了减少这种误差延迟带来的影响,模型最终选择双向循环神经网络LSTM作为基础结构,模型在前向和后向的计算方法是相同的,最终输出两个预测值的平均值。损失函数的计算,除了计算两个预测值与真实值的平方误差之外,还计算两个预测值之间的差值的平方,使得两个预测值尽量接近。

4　实　　验

4.1　数据集

4.1.1　空气质量数据集

空气质量数据集包含了2017/11/01～2017/11/15期间北京35个监测站监控的6个空气指标(PM2.5、PM10、NO_2、CO、O_3、SO_2)。整个数据集数据缺失率约为5.2%。在这个数据集上进行插补实验和利用插补后的数据进行回归预测实验。为方便测试,随机消除了10%的数据作为插补实验的测试集,并选取了2017/11/16这一天的数据作为回归实验的测试集。为了训练模型,选取连续的360个时间步作为一个时间序列。

4.1.2　医疗数据集

PhysioNet Challenge 2012[24]医疗数据集包括重症监护室12 000个病人的健康测量数据,每条记录由大约48 h的多元时间序列数据组成,训练集包含4 000条记录,其余记录组成测试集。每条记录包含6个静态特征值(编号、年龄、性别、身高、ICU类型、体重),数据的采集时间(单位秒),37个可以进行多次观测的动态时间序列特征值,这些值记录了患者在住院期间不同时间段的37个指标,分别是:Albumin、ALP、ALT、AST、Bilirubin、BUN、Cholesterol、Creatinine、DiasABP、FiO2、GCS、Glucose、HCO3、HCT、HR、K、Lactate、Mg、MAP、MechVent、Na、NIDiasABP、NIMAP、NISysABP、PaCO2、PaO2、pH、Platelets、RespRate、SaO2、SysABP、Temp、TropI、TropT、Urine、WBC、Weight。这37个指标值观察到值的情况不同,有的只观察到一次,有的多次,也有可能根本观察不到,因而存在很多缺失值。整个数据集数据缺失率约为80.53%。在这个数据集上进行插补和分类预测的实验。数据集处理方式与上个数据集相同。选取连续48个时间步作为一个时间序列。

4.1.3　人类活动定位数据集

UCI-HAR数据集[25]利用智能手机采集,是用于机器学习测试的有关人体活动识别的共计10 299个多变量时间序列传感器样本数据。数据的采集由年龄在19～48岁之间的30名志愿者完成。采集数据时,每名志愿者的左/右脚踝、胸部和腰带上都有传感器,执行了6种基本活动,包括3种静态姿势(站

立、坐着、躺着)和3种动态活动(走路、下楼、上楼),记录的运动数据是来自智能手机的x、y和z三轴的加速度计数据(线性加速度)和陀螺仪数据(角速度),采样频率为50 Hz(每秒50个数据点),共采集19项人体行为活动数据。2015年UCI-HAR数据集进行了更新,增添了在静态姿势之间发生的姿势转换。除此之外,还包括动态活动与静态活动之间发生的姿势转换过程,分别是站到坐、坐到站、坐到躺、躺到坐、站到躺和躺到站。选取连续40个时间步作为一个时间序列。在这个数据集上,消除了不同百分比的数据作为测试集,检验所提方法对不同趋势率的性能变化。

4.2 实验设置

为了检验模型的性能,本文进行了插补准确值的衡量以及后续分析任务分类和回归性能的实验,将本文所提模型与其他基线模型进行了直接和间接的对比,以验证所提模型的优势。对于插补任务,均采用了随机抽取原序列中的数据作为测试集,将插补后的数据与原序列进行比较。为了公平比较,所有抽取的数据都不会在实验训练中出现。模型使用了双向LSTM作为基础的RNN结构,使用Adam优化器训练模型,学习率为0.001,batch size为64。对所有数据集的数据进行了零均值标准化。3个数据集建模的张量大小分别为35×6×360、4000×35×48、4000×3×40。误差的衡量使用均方根误差RMSE和平均相对误差MRE作为插补和回归任务的评价指标,使用ROC曲线下的面积AUC作为分类任务的评价指标。

假设x_i为缺失的真实值,$\hat{x}_i$为模型输出的插补值,一共有N个缺失值。RMSE和MRE计算公式分别为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\hat{x}_i)^2}{N}} \tag{13}$$

$$\mathrm{MRE}=\frac{\sum_{i=1}^{N}|x_i-\hat{x}_i|}{\sum_{i=1}^{N}|x_i|}\times 100\% \tag{14}$$

4.3 基线方法

实验中将所提模型与以下基准时间序列插补模型进行比较。Mean:用相应的全局平均值来替换缺失值;MF[15]:将数据矩阵分解为两个低秩矩阵,并通过矩阵补全来填充缺失值;CP[26]:将数据矩阵张量分解为3个低秩矩阵,并通过张量补全来填充缺失值;TRMF[16]:MF与自回归模型的结合,并通过矩阵补全来填充缺失值;LATC[27]:利用张量的低秩逼近与自回归模型的结合,并通过张量补全来填充缺失值;BRITS[19]:利用双向循环神经网络对缺失值进行插补,并结合多元时间序列特征的相关性。

4.4 实验结果分析

4.4.1 多个模型在不同数据集上插补结果的准确性比较

表1展示了所有模型在3个数据集上的插补实验结果。可以看出平均插补法最不准确,在所有的数据集上都取得了最差的效果。一些简单的机器学习插补法,如MF和CP都十分不稳定。MF在人类活动定位数据集上效果很好,但在另外两个数据集上则效果很差;CP则是在前两个数据集上效果不错,在人类活动定位数据集上效果较差。而利用到了自回归的两个模型TRMF、LATC表现虽然稳定且相对于MF和CP有了提升,但是和前面分析的一样,自回归只能建模线性的时间关系,所以它们的表现都不如基于RNN的插补模型。本文所提模型在3个数据集上都优于所有的基线方法,并且TC-BRNN基于BRITS的双向LSTM结构进行改进,这证明通过捕获不同时间序列的相关性可以提高插

补性能。在空气质量数据集和医疗数据集上,TCBRNN取得了较大的提升。这是因为构建这两个数据集的3个特征中有较强的相关性,例如空气质量数据集中不同观测站和不同空气指标,地区邻近观测站的观测指标会有较强的相关性,同一观测站的不同观测指标之间也会有相关性,张量补全的过程中会将观测站和观测指标进行交互,沿时间展开成为一个新的矩阵,更新矩阵的过程就可以学习到这两个特征之间的相关性,从而提升了插补的结果。

表1　不同数据集上的实验结果比较

Table 1　Comparison of experimental results of all algorithms on different training sets

模型	空气质量数据集 RMSE(MRE/%)	医疗数据 RMSE(MRE/%)	人类活动定位数据集 RMSE(MRE/%)
Mean	0.943 9(93.17)	0.750 1(65.48)	2.008 1(99.98)
MF	0.867 3(83.28)	0.922 5(86.73)	0.252 1(9.45)
CP	0.555 5(47.54)	0.780 1(66.80)	0.352 9(15.54)
TRMF	0.440 2(37.15)	0.856 7(81.56)	0.206 7(7.05)
LATC	0.331 7(25.92)	0.634 5(45.55)	0.192 0(6.70)
BRITS	0.259 6(19.33)	0.658 1(39.37)	0.191 0(7.01)
TCBRNN	0.165 0(12.86)	0.186 6(7.29)	0.123 6(5.41)

4.4.2　多个模型的插补结果对提升回归和分类准确性的影响比较

图3展示了回归和分类的实验结果。用一个LSTM模型加全连接层作为回归实验的回归模型,然后用不同方法插补后的空气质量数据对回归模型进行训练,最后发现TCBRNN插补后的数据有助于提升回归模型的性能,且提升效果优于其他的对比模型。用一个简单的随机森林作为分类器,将不同方法插补后的医疗数据集输入随机森林,并把预测病人死亡情况看成一个二分类问题。预测结果同样表明TCBRNN插补后的数据可以提升分类的准确性,且提升效果优于其他模型。

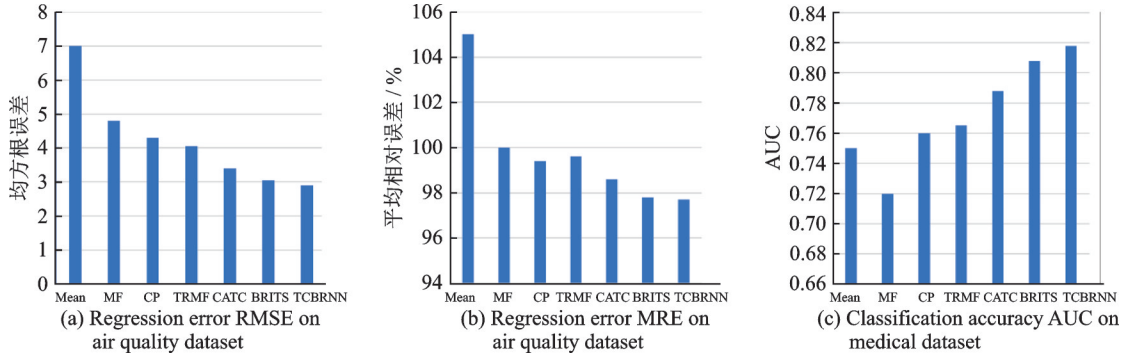

图3　回归和分类实验比较结果

Fig.3　Experimental results of regression and classification

图4展示了所有模型在10%、20%、40%、80%缺失率下人类活动定位数据集的实验结果。可以看出除了MEAN和MF算法外,大部分算法都表现很好,且TCBRNN在所有缺失率下均取得了最好的插补效果。

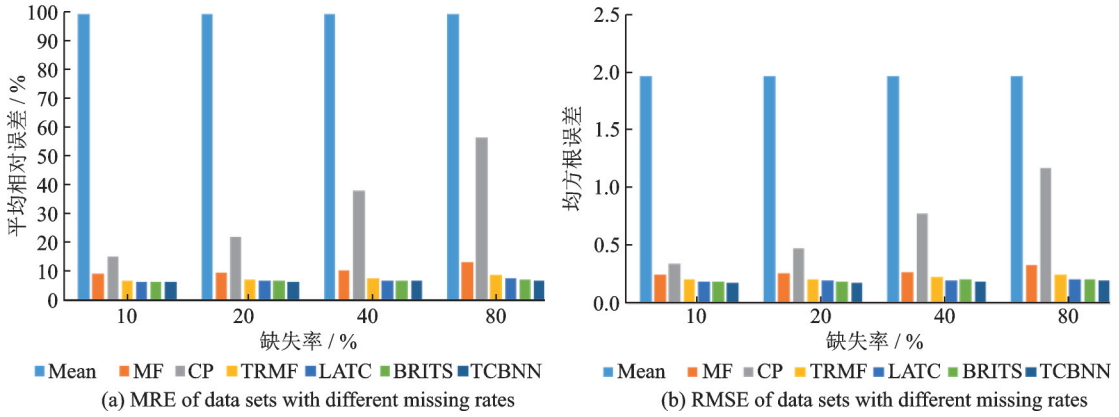

图 4 人类活动定位数据集下不同缺失率的数据集不同模型实验结果

Fig.4 Experimental results of different models with different missing rates in human activity location dataset

4.4.3 节点数及序列长度对插补结果准确性的影响

循环神经网络模型隐藏层的节点个数是可变的，图 5 展示了在空气质量数据集上选择不同节点个数的实验结果，选择了 24、48、64、128 四个不同的隐藏层节点数对 TCBRNN 进行了训练。图 5 结果表明，随着节点数的增大，误差在降低，当隐藏层节点数为 64 时，插补效果最好。隐藏层节点数大于 64 时没有明显提升。

对于空气质量数据集，本文变换时间序列的长度，分析了长度对插补效果的影响，结果如图 6 所示。图 6 结果表明序列长度在 120 时效果比较差，随着长度的增大误差在降低，在 360 时插补效果最好，在 360 附近变化不显著。

从对实验结果的分析中可以看出，对时间序列的时序性建模越精确，模型的插补性能越好；模型中张量补全模块可以捕获不同时间序列的相关性，利用相关性可以提升模型的插补效果。

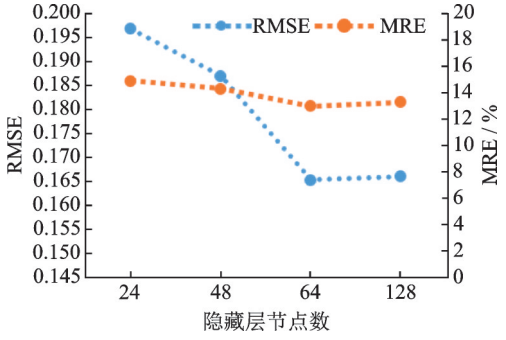

图 5 空气质量数据集不同隐藏层节点数的实验结果

Fig.5 Experimental results of different numbers of hidden layer nodes in air quality dataset

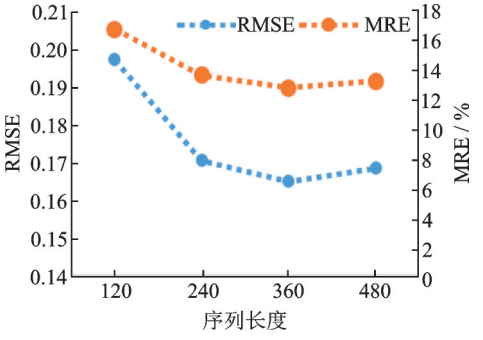

图 6 空气质量数据集下不同序列长度的实验结果

Fig.6 Experimental results of different series length in air quality data set

5 结束语

本文提出了一种联合张量补全与循环神经网络的时间序列缺失值插补模型 TCBRNN，可以有效地对多个多元时间序列中的缺失值进行插补。该模型同时考虑不同时间序列之间的相关性和复杂时序关系，直接学习出缺失值。在多个数据集上对所提出的模型进行了性能评估，包括对缺失值的插补值准确性进行了评估，并对插补后的数据进行了回归与分类模型的构建，测试其对后续时间序列分析

任务的效果。实验结果表明,和其他基线模型相比,本文所提模型在插补上有更准确的效果,并且插补后的数据可以提升回归、分类模型的性能。未来工作可以对模型的插补结果进行可视化处理,这样可以直观地展示模型插补结果与真实数据的差异,还可以尝试用其他方法捕获不同时间序列的相关性来获得更好的效果,例如利用深度学习的方法来代替张量补全。

参 考 文 献:

[1] CHATFIELD C. The analysis of time series: An introduction[M]. New York, USA: CRC Press, 2016.

[2] HSIEH T J, HSIAO H F, YEH W C. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm[J]. Applied Soft Computing, 2011, 11(2): 2510-2525.

[3] 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.

[4] SILVA I, MOODY G, SCOTT D J, et al. Predicting in-hospital mortality of ICU patients: The physionet/computing in cardiology challenge 2012[C]//Proceedings of 2012 Computing in Cardiology. [S.l.]: IEEE, 2012: 245-248.

[5] CAO W, WANG D, LI J, et al. BRITS: Bidirectional recurrent imputation for time series[J]. Advances in Neural Information Processing Systems, 2018, 31: 6775-6785.

[6] BERGLUND M, RAIKO T, HONKALA M, et al. Bidirectional recurrent neural networks as generative models[J]. Advances in Neural Information Processing Systems, 2015, 28: 856-864.

[7] MCKNIGHT P E, MCKNIGHT K M, SIDANI S, et al. Missing data: A gentle introduction[M]. New York, USA: Guilford Press, 2007.

[8] WOTHKE W. Longitudinal and multigroup modeling with missing data[M]. England, UK: Psychology Press, 2000: 205-224.

[9] KANTARDZIC M. Data mining: Concepts, models, methods, and algorithms[M]. New Jersey, USA: John Wiley &. Sons, 2011.

[10] ACURNA E, RODRIGUEZ C. The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications[C]//Proceedings of the Meeting of the International Federation of Classification Societies (IFCS). Chicago: Springer Berlin Heidelberg, 2004: 639-647.

[11] FUNG D S. Methods for the estimation of missing values in time series[D]. Western Australia: Edith Cowan University, 2006.

[12] HARVEY A C. Forecasting, structural time series models and the Kalman filter[D]. Cambridge, UK: Cambridge University Press, 1990.

[13] NELWAMONDO F V, MOHAMED S, MARWALA T. Missing data: A comparison of neural network and expectation maximization techniques[J]. Current Science, 2007, 93(11): 1514-1521.

[14] NATICK M. The mathworks[M]. United States: Meuwissen, 2010.

[15] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[16] YU H F, RAO N, DHILLON I S. Temporal regularized matrix factorization for high-dimensional time series prediction[C]//Proceedings of NIPS. Barcelona, Spain:Urran Associates Inc., 2016: 847-855.

[17] YOON J, ZAME W R, VAN DER SCHAAR M. Multi-directional recurrent neural networks: A novel method for estimating missing data[C]//Proceedings of Time Series Workshop in International Conference on Machine Learning. Sydney: [s.n.], 2017.

[18] 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.

[19] CAO W, WANG D, LI J, et al. BRITS: Bidirectional recurrent imputation for time series[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada:Curran Associates Inc., 2018: 6776-6786.

[20] SONG Q, GE H, CAVERLEE J, et al. Tensor completion algorithms in big data analytics[J]. ACM Transactions on Knowledge Discovery from Data, 2019, 13(1): 1-48.

[21] ZHAO Z, TANG L, FANG M, et al. Toward urban traffic scenarios and more: A spatio-temporal analysis empowered low-

rank tensor completion method for data imputation[J]. International Journal of Geographical Information Science, 2023, 37(9): 1936-1969.

[22] LIU C, WU T, LI Z, et al. Robust online tensor completion for IoT streaming data recovery[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 10178-10192.

[23] LIU J, MUSIALSKI P, WONKA P, et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 208-220.

[24] SILVA I, MOODY G, MARK R, et al. Predicting mortality of ICU patients: The physioNet/computing in cardiology challenge 2012[EB/OL]. (2012-01-20)[2022-05-23]. https://www.physionet.org/content/challenge-2012/1.0.0/.

[25] FRANK A, ASUNCION A. UCI machine learning repository[EB/OL]. (2010-01-28)[2022-05-23]. https://www.archive.ics.uci.edu/.

[26] KARATZOGLOU A, AMATRIAIN X, BALTRUNAS L, et al. Multiverse recommendation: *n*-dimensional tensor factorization for context-aware collaborative filtering[C]//Proceedings of the fourth ACM Conference on Recommender Systems. Barcelona, Spain: Association for Computing Machinery, 2010: 79-86.

[27] CHEN X, LEI M, SAUNIER N. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 12301-12310.

作者简介:

何军(1962-),**通信作者**,男,博士,教授,博士生导师,研究方向:数据挖掘、社交网络分析、推荐系统等,E-mail:hejun@ruc.edu.cn。

赖赵远(1996-),男,硕士研究生,研究方向:时间序列分析。

时勘(1949-),男,博士,教授,博士生导师,研究方向:生态环境监测、风险认知等。

(编辑:刘彦东)