

文章编号:2095-6134(2021)02-0280-08

简报

基于改进神经过程的缺失数据填充算法^{*}

孙晓丽,郭艳[†],李宁,宋晓祥

(中国人民解放军陆军工程大学,南京 210007)

(2019 年 7 月 8 日收稿;2019 年 10 月 8 日收修改稿)

Sun X L, Guo Y, Li N, et al. Missing data imputing algorithm based on modified neural process[J]. Journal of University of Chinese Academy of Sciences, 2021,38(2):280-287.

摘 要 缺失数据填充是数据分析处理领域的一个重要研究课题。特别是在采集数据量较少的情况下,缺失数据填充的难度极大。针对这个问题,提出一种基于改进神经过程模型的缺失数据填充算法,该算法可有效提升小数据集背景下的缺失数据填充性能。首先,将观测到的时间序列进行单一表示,由神经网络得到各自的表征向量;其次,通过神经过程模型获得数据的分布函数,并在训练阶段引入修正系数 α ,从而根据数据缺失率更加精确地确定训练数据的采样率;最后,加入填充过程,通过训练好的模型估计数据缺失值。为检验算法性能,在海洋表面温度数据集以及北京 PM_{2.5} 含量数据集上进行仿真实验,结果表明该算法在小数据集背景下具有良好的填充效果。与其他算法相比,所提算法在高缺失率的情况下具有更低的均方根误差。

关键词 缺失数据填充;时间序列;改进神经过程;修正系数
中图分类号:TN911.7 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2021.02.014

Missing data imputing algorithm based on modified neural process

SUN Xiaoli, GUO Yan, LI Ning, SONG Xiaoxiang

(PLA Army Engineering University, Nanjing 210007, China)

Abstract Missing data imputing is a serious problem in the field of data analysis and process, which is extremely intractable in the case of the small dataset especially. In view of this problem, a missing data imputing algorithm based on modified neural process is proposed, which can improve the imputing performance in the background of the small dataset. Firstly, the observed time series is single-represented and then obtain the symptomatic vector respectively through the neural network. Secondly, it can acquire the distribution function of the data via the neural process and introduce the correction coefficient α to determine the sampling rate more exactly based on missing rate in the training stage. Finally, it imported the imputing process and estimated the missing data via trained model. Experiments are carried out on the sea surface temperature dataset and the Beijing PM_{2.5} dataset to verify the performance of the algorithm. The experiments show that the algorithm has an

^{*} 国家自然科学基金(61871400)和江苏省自然科学基金(BK20171401)资助
[†] 通信作者,E-mail:guoyan_1029@sina.com

excellent performance in the context of small datasets, and it has a lower root mean square error compared with other algorithms.

Keywords missing data imputing; time series; modified neural process; correction coefficient

近年来,数据采集和存储技术飞速发展,为数据分析提供了极大便利,但数据缺失的现象依旧频繁发生,严重影响数据分析的精度。特别是在小数据集背景下,缺失数据的存在将给数据分析工作带来灾难性的影响。因此,如何有效地对缺失数据进行填充成为亟待解决的关键问题。

目前,专家学者们在各个领域进行了大量的研究工作,提出了许多有效的缺失数据填充方法。插值方法^[1-3]将已观测到的值拟合成平滑曲线,而后通过局部插值填充缺失值。该方法会随着时间的推移丢弃变量之间的关系,导致数据填充效果并不理想。另一类则是包括 ARIMA (autoregressive integrated moving model)、SARIMA (seasonal ARIMA)^[4-5]等在内的自回归方法,此类方法消除了时间序列数据中的非平稳部分,拟合出参数化的平稳模型来填充缺失数据。除此之外,文献[6]采用协同过滤的方法对推荐系统中的缺失值进行估计,文献[7]利用基于正则化的矩阵因子分解方法对定时采样的时间序列数据进行缺失值拟合,文献[8]结合迭代模型插值技术提出基于 Aitchison 距离的 k 近邻网络,文献[9]提出一种基于随机森林算法的迭代插值方法,文献[10]提出一种基于傅里叶变换和 KNN (K-nearest neighbor) 算法的时间序列缺失数据补全方法。然而,上述方法只能处理特定的缺失类型和应对较低缺失率的情况。

近年来,基于生成模型的方法在解决缺失数据填充问题上展现了优越的性能。文献[11]提出一种基于卷积自动编码器 (convolutional autoencoder, CAE) 的生理波数据缺失填充算法。该算法能在大量相邻的生理波数据段缺失的情况下进行填充,针对生理波形而言,该模型结构具有较强的一般性与扩展性;文献[12]提出一种去噪自编码器 (denoising autoencoder, DAE) 与生成对抗网络 (generative adversarial network, GAN) 相结合的模型,能够处理含噪声的高缺失率工业物联网数据。但是,该模型无法对高精度的数据进行填充;文献[13]提出 DAE 与堆叠自编码器 (stacked autoencoder, SAE) 结合的模型——堆叠去噪自编码器 (denoising stacked autoencoder,

DSAE), 通过将缺失数据和观测数据看作一个整体恢复完整数据,可在不同缺失率情况下保持稳定的误差。但上述方法大都应用于大型数据集,应用到小数据集的效果不理想。

基于以上分析,本文提出一种基于改进神经过程 (modified neural process, MNP) 模型的缺失数据填充算法。该算法利用改进神经过程获得数据的分布函数模型,并通过训练来捕获对未观测点的不确定性,进而对数据缺失值进行估计。在训练阶段,根据缺失率引入采样率的修正系数,以提高高缺失率情况下的填充效果。仿真结果表明,缺失数据的填充值与真实值之间具有较低的平均相对误差;与其他算法相比,所提算法在高缺失率情况下的填充性能更优。

1 神经过程

神经过程^[14] (neural process, NP) 是一种结合了神经网络 (neural network, NN) 与高斯过程 (Gaussian process, GP) 两者优点的模型。基于深度神经网络强大的非线性拟合能力以及学习高斯过程的逼近方法,即学习在函数之上建模分布,能够根据上下文的观测估计其预测的不确定性。因此, NP 又被称为是高斯过程的深度学习版本。

NP 的模型结构如图 1 所示。其中 (x_c, y_c) 为上下文数据集, (x_i^*, y_i^*) 为目标数据集。通过模型结构图,可以将 NP 分解为以下几步:

- 1) 将上下文数据 (x_c, y_c) 通过神经网络映射, 获得其表征向量;
- 2) 将获得的表征向量进行聚合, 得到单个表征值 r ;
- 3) 通过聚合后的表征值 r 对隐向量 z 进行参数化, 使得隐向量 z 满足

$$p(z \mid x_{1:c}, y_{1:c}) = N(\mu_z(r), \sigma_z^2(r)); \quad (1)$$

- 4) 为了估计输入 x_i^* 后的函数值, 对隐向量 z 进行采样, 并与 x_i^* 一同输入到神经网络中, 从而获得估计的 x_i^* 对应的函数值。

NP 近似了两个分布, 隐变量 z 的变分后验分布 $q(z \mid x_{1:n}, y_{1:n})$ 和其条件先验分布 $p(z \mid x_{1:n}, y_{1:n})$ ^[14], 以 KL 散度推导出模型的训练损失函

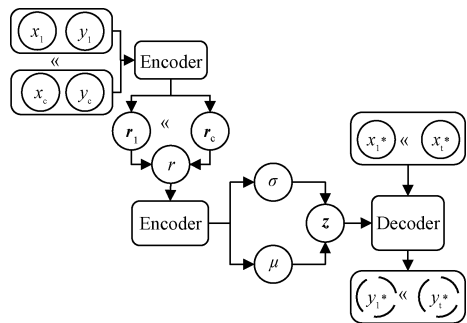


图 1 NP 模型结构图

Fig. 1 NP model structure diagram

数,并称之为证据下界(ELBO),表示为

$$\text{ELBO} = \mathbb{E}_{q(z|x_{1:n}, y_{1:n})} \left[\sum_{j=1}^n \log p(y_j | z, x_j) + \log \frac{p(z)}{q(z|x_{1:n}, y_{1:n})} \right]. \quad (2)$$

设上下文集表示为 $(x_{1:m}, y_{1:m})$, 目标集表示为 $(x_{m+1:n}, y_{m+1:n})$, 又因为条件先验 $p(z|x_{1:m}, y_{1:m})$ 在实验中难以求得, 因此可用变分后验 $q(z|x_{1:m}, y_{1:m})$ 近似代替, 则有

$$\text{ELBO} = \mathbb{E}_{q(z|x_{1:n}, y_{1:n})} \left[\sum_{j=m+1}^n \log p(y_j | z, x_j) + \right.$$

$$\left. \log \frac{q(z|x_{1:m}, y_{1:m})}{q(z|x_{1:n}, y_{1:n})} \right]. \quad (3)$$

2 改进 NP 算法模型

GP 是处理非平稳时间序列的常用方法, 而 NP 是由 GP 和 NN 结合而成的, 是 GP 的深度学习版本。NP 通过 NN 确定核函数, 具有很强的自适应性, 可以很好地满足非平稳时间序列对核函数的要求。数据填充过程和时间序列的预测过程具有很高的相似度, 可以先应用 NP 在已观测数据的基础上对整体数据的分布函数进行拟合, 再利用获得的分布函数对缺失数据进行填充。但为使得在现有数据的基础上更好地反映完整数据的分布, 需要对现有的 NP 进行修正。

本文采用基于改进神经过程模型——MNP 对缺失数据进行填充。模型结构图如图 2 所示。其中, 实线部分为训练过程, 虚线部分为填充过程, 填充过程中的 Encoder_r, Encoder 是经过训练过程学习得到的, 与训练过程中的结构完全相同。图 2 中变量含义如表 1 所示。

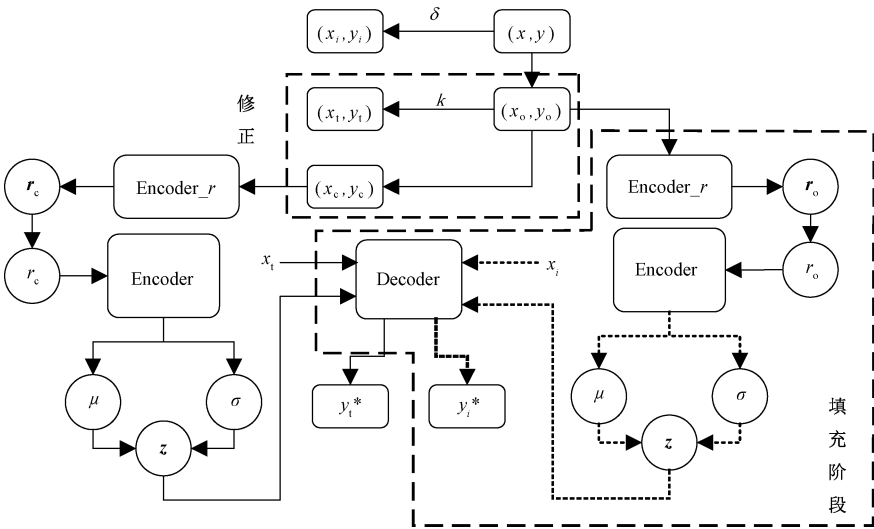


图 2 MNP 结构框图

Fig. 2 MNP model structure diagram

MNP 算法在 NP 的基础上增加填充阶段, 引入修正系数, 使其在高缺失率的情况下能高效地对缺失数据进行填充。训练过程中, 以 k 为采样比对观测数据进行随机采样, 将其分为上下文集和目标集两部分。在计算模型的概率损失函数时, 目标集也参与其中, 这使得损失函数有意义, 也有助于防止模型过拟合。并且在训练的过程中, 反复以 k 为采样比对观测数据进行随机采样,

使得上下文集对观测数据集有更为全面的概括, 从而能更好地实现对观测数据集分布的描述; 对上下文集进行学习, 提取数据的特征, 使得上下文集数据的分布近似观测数据的分布。填充阶段, 将缺失数据近似作目标集, 观测数据近似作上下文集。利用训练好的网络对观测数据进行特征提取, 并根据提取的特征对缺失数据进行填充。具体算法过程如表 2 所示。

表 1 变量说明
Table 1 Variable declaration

变量名	含义
(x,y)	完整数据集
(x_o,y_o)	观测数据
(x_i,y_i)	缺失数据
(x_c,y_c)	上下文 (context) 集
(x_t,y_t)	目标 (target) 集
δ	缺失率
k	修正后采样比
r_c,r_o	(x_c,y_c) 、 (x_o,y_o) 的表征向量
r_c,r_o	r_c,r_o 求均值后的标量值
y_t^*	x_t 经过 Decoder 后的输出
y_i^*	x_i 对应的缺失值

表 2 MNP 算法过程
Table 2 MNP algorithm process

算法 1: MNP
观测数据 (x_o,y_o) , 缺失数据 (x_i,y_i) 迭代次数为 n_step 训练集采样率为 λ 修正系数为 α , 令 $\alpha = f(\delta)$ 输入数据: O: (x_o,y_o)
训练阶段: 1) 设定 $k = \alpha \times \lambda$, 对训练集采样获得 target 和 context, 使其满足 $k = \frac{\text{len}(\text{target})}{\text{len}(\text{target} + \text{context})}$ 2) $(x_c,y_c) \rightarrow \text{Encoder}_r \rightarrow r_c \rightarrow r_c$ $(x_t,y_t) \rightarrow \text{Encoder}_r \rightarrow r_t \rightarrow r_t$ 3) $r_c \rightarrow \text{Encoder} \rightarrow \mu_c, \sigma_c$ $r_t \rightarrow \text{Encoder} \rightarrow \mu_t, \sigma_t$ 4) 参数重构: $\varepsilon \sim N(0,I), z = \mu + \sigma \times \varepsilon$, 使得满足: $p(z \mid x,y) = N(\mu_z(r), \sigma_z^2(r))$ 5) $(z,x_t) \rightarrow \text{Decoder} \rightarrow y_t, \sigma$ 6) 计算 $\text{loss} = \text{ELBO}$ 7) $\text{loss.backward}()$
填充阶段: 1) 经训练过程的步骤 2 后的观测数据 $O \rightarrow$ 训练好的 Encoder $\rightarrow \mu_o, \sigma_o$ 2) μ_o, σ_o 参数重构得到隐变量 z 3) $(z,x_i) \rightarrow$ 训练好的 Decoder $\rightarrow y_i^*$

MNP 模型继续沿用 NP 训练损失函数, 在此处表示为

$$\text{ELBO} = \mathbb{E}_{q(z \mid \text{context}, \text{target})} \left[\sum_{j=m+1}^n \log p(y_j^* \mid z, x_j^*) + \log \frac{q(z \mid \text{context})}{q(z \mid \text{context}, \text{target})} \right]. \quad (4)$$

3 数据集及评价方法

3.1 数据集

海洋表面温度 (sea surface temperature, SST)

是海洋的重要物理参数, 在大气与海洋间的能量交换过程中扮演着重要的角色, 是决定海气相互作用及气候变化的主要因素^[15]。SST 数据集^[16]是由热带大气海洋项目的实测数据组成的, 采样频率为 1 h。从中选取 1 000 个连续采样的数据作为实验数据。SST 数据现已广泛应用于多个领域, 如赤潮研究、气候变化研究、海洋表面特征的解释以及各种勘察结果的解释。因此, 监测海洋表面温度的变化, 提供完整准确的海洋表面温度, 对了解地区气候变化以及各种其他的海洋工作有重要意义。

北京 PM_{2.5} 含量数据集是经由北京大学统计科学中心上传至 UCI^[17] 数据库中, 记录从 2010 年 1 月 1 日至 2014 年 12 月 31 日以 1 h 为采样频率记录的实时北京 PM_{2.5} 含量值。

本文选择使用 SST、北京 PM_{2.5} 含量数据作为实验数据, 以小数据集为背景, 选取其中一个变量属性, 分别从完整数据集中选取 1 000 个连续采样的数据作为时间序列, 以此作为实验数据进行缺失数据填充的实验。

3.2 评价方法

为了更好地反映缺失数据填充的效果, 采用 3 种不同的方法对结果进行评价: 均方误差 (MSE)、平均相对误差 (MRE) 以及平均绝对误差 (MAE), 分别定义为:

$$\text{MSE} = \frac{\sum_N (y - y^*)^2}{N}, \quad (5)$$

$$\text{MRE} = \frac{1}{N} \sum_N \frac{y - y^*}{y}, \quad (6)$$

$$\text{MAE} = \frac{\sum_N y - y^*}{N}, \quad (7)$$

其中: N 表示缺失值的数量, y 为真实值, y^* 为模型估计的填充值。

4 实验结果

本文选取 SST 数据和北京 PM_{2.5} 含量数据作为实验数据, 并对其进行预处理, 然后对数据进行归一化, 并将其按照缺失率为 10%、20%、30%、40%、50%、60%、70%、80%、90% 进行处理。

归一化过程具体表示为

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (8)$$

本文实验分为 3 组, 第 1 组实验为确定修正

系数 α 的大小,进而确定修正后采样比 k 值的大小,即确定训练过程中所用数据集的大小;第 2 组为在确定修正后采样比 k 的情况下,各缺失率条件下的数据填充效果;第 3 组实验为本文模型与现有缺失数据填充模型的对比实验。

4.1 修正后采样比 k 值的确定

定义 k 为

$$k = \alpha \times \lambda, \tag{9}$$

且 k 满足

$$k = \frac{\text{len}(\text{target})}{\text{len}(\text{target} + \text{context})}, \tag{10}$$

其中, λ 为采样率, α 为修正系数, 且 α 满足 $\alpha = f(\delta)$, δ 为缺失率。将训练集按照一定的采样率采样得到目标集 (target), 剩余数据作为上下文集 (context)。

对 α 进行两种情况下的分析:

- 1) 针对不同的缺失率, k 是固定不变的, 即 $\alpha = 1$;
- 2) k 随缺失率的变化而变化, 即 α 是 δ 的分段函数。

在训练过程中, 通过 NP 模型对 context 集进

行学习, 利用 target 集的回归误差更新模型参数, 进而获得整个训练集的数据分布函数。而填充阶段则根据训练阶段所学习的模型参数, 估计需要的缺失值。在这两个过程中, 存在两个比例:

$$1) k = \frac{\text{len}(\text{target})}{\text{len}(\text{target} + \text{context})},$$
$$2) \delta = \frac{\text{len}(\text{缺失值})}{\text{len}(\text{观测值} + \text{缺失值})}.$$

一方面, 要求估计的缺失值与真实值之间的 MRE 最小, 误差越小说明填充效果越好; 另一方面, 从模型本身出发, 通过学习观测值的分布对缺失数据进行填充, 要求训练过程中 target 数据的估计值误差与填充过程中缺失数据的估计值误差差值尽可能小, 差值越小说明填充效果越好。

MSE 与 MAE 描述的是估计值与真实值之间的绝对误差, 而 MRE 描述两者之间的相对误差, 不受数据归一化的影响。因此, 为了更好地体现出所提算法的有效性, 实验以 MRE 为主要评价标准对算法的填充效果进行评价。以 SST 数据和北京 PM_{2.5} 含量数据为实验数据, 对模型的性能进行检验, 实验结果如图 3 所示。

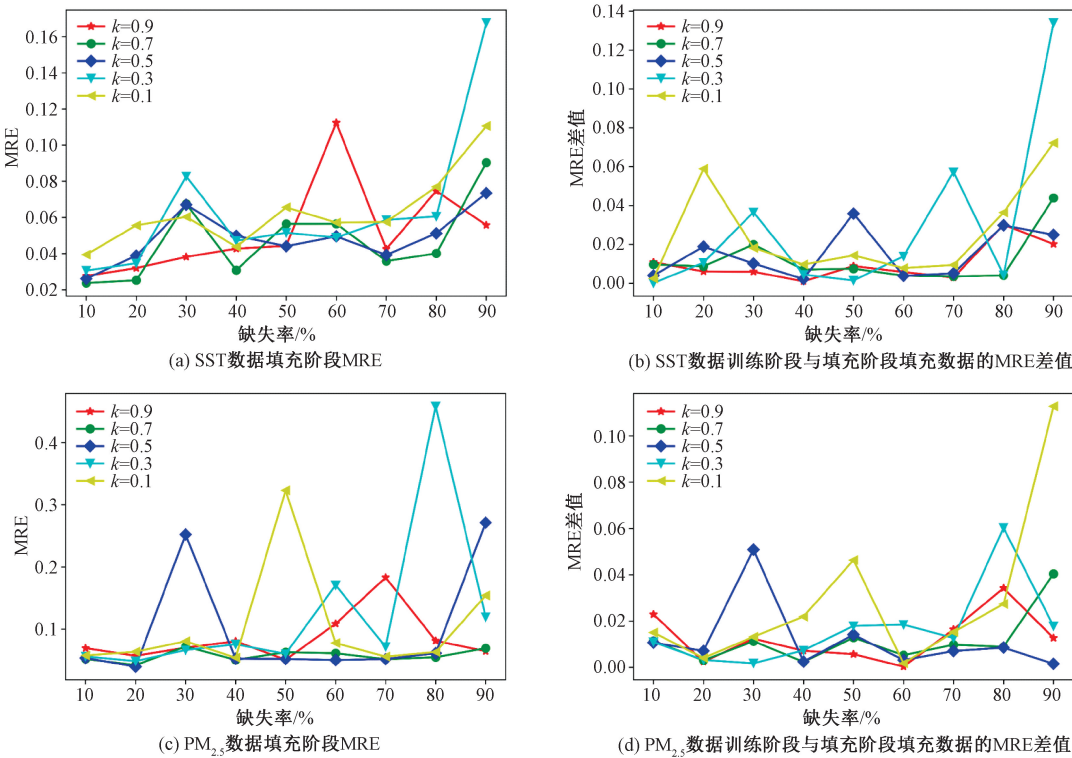


图 3 不同 k 不同缺失率情况下, 训练与填充阶段估计数据的 MRE 误差情况
Fig. 3 MRE errors of data estimated during training and imputing under different k and missing rates

由实验结果可以看出, SST 数据集与北京 PM_{2.5} 含量数据集在不同缺失率情况下的实验效

果较为一致, 填充效果均随 k 值的变化而改变。随着缺失率的升高, 整体的 MRE 呈现上升趋势,

在个别 k 值处出现效果不好的点。

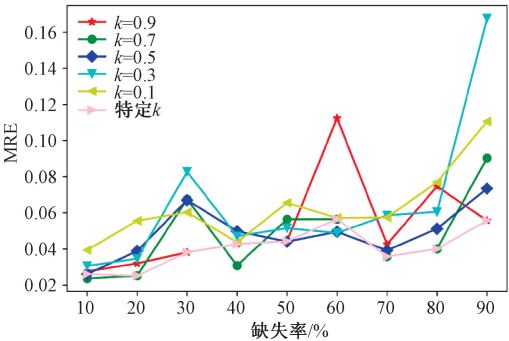
在图 3(a)、3(b) 中,当 $k = 0.1$,在缺失率为 20%、90% 时训练阶段与填充阶段估计数据的误差过大; $k = 0.3$,缺失率为 30%、70%、90% 时的差值同样过大,且在 30%、90% 时缺失数据的估计值 MRE 误差过大; $k = 0.9$ 在缺失率为 60% 时缺失数据估计值的 MRE 过大。图 3(c)、3(d) 中,当 $k = 0.1$ 时,算法在缺失率为 50% 处的填充误差较大,且在 50%、80%、90% 处的训练阶段与填充阶段估计数据的误差过大;当 $k = 0.3$ 时,算法在缺失率为 60%、80% 处的填充误差较大,且在 80% 处的差值同样存在较大的情况;当 $k = 0.5$ 时,在

缺失率为 30%、90% 处的填充误差较大,且在 30% 处的训练阶段与填充阶段估计数据的误差过大。由此可以看出,在固定 k 不变时,算法不能保证多个缺失率情况下缺失数据的填充效果。当 k 随缺失率变化,即 α 是 δ 的分段函数时,算法的适应性更强。结合实验结果,可在每个缺失率处找到一个合适的 k 值。结果如表 3 所示,选定 k 值后结果如图 4 所示。

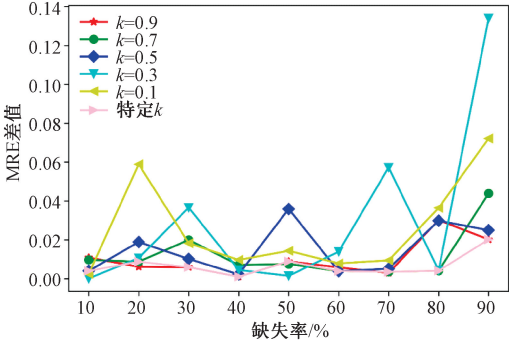
表 3 不同缺失率情况下 k 的取值情况

Table 3 k values at different miss rates

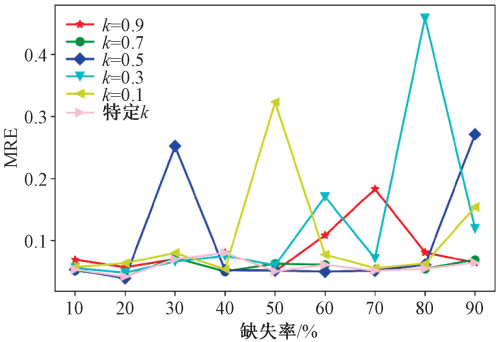
缺失率/%	10	20	30	40	50	60	70	80	90
k	0.5	0.7	0.9	0.9	0.9	0.7	0.7	0.7	0.9



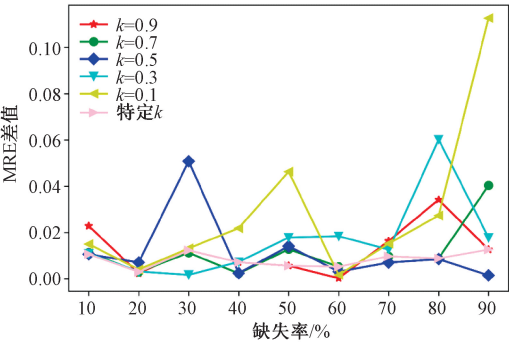
(a) 改进后 SST 数据填充阶段 MRE



(b) 改进后 SST 数据训练阶段与填充阶段填充数据的 MRE 差值



(c) 改进后 PM_{2.5} 数据填充阶段 MRE



(d) 改进后 PM_{2.5} 数据训练阶段与填充阶段填充数据的 MRE 差值

图 4 不同缺失率情况下,固定 k 值与特定 k 值,训练与填充阶段估计数据的 MRE 误差情况

Fig. 4 MRE error of data estimated during training and imputing phase in the case of fixed k value and specific k value at different missing rates

4.2 确定修正后采样比 k 情况下的填充效果

通过实验 1 可确定在不同缺失率的情况下 k 的取值。由图 4 可以看出,在两种不同数据集上得到的实验数据是一致的,因此在本实验中,仅在 SST 数据的基础上针对不同缺失率对缺失数据进行填充,实验效果如表 4 所示。

由实验结果可以看出,所提算法能够在不同缺失率情况下实现缺失数据的填充。伴随着缺

失率的升高,MAE 与 MSE 的变化相对较为稳定,MRE 随着缺失率的升高有所增大,但涨幅并不大,在可接受范围内。因此,该实验证明所提算法在不同缺失率情况下均能取得很好的效果。

4.3 对比实验

为验证所用模型的有效性,在两种不同的数据集上做了以下对比试验。所对比的方法有: 1) 稀疏贝叶斯算法^[18] (sparse bayesian learning,

SBL); 2) 递归神经网络^[19] (recurrent neural network, RNN); 3) 神经过程 (NP)。4 种算法在不同缺失率情况下的 RMSE 对比如表 5 所示, 高缺失率情况下的 RMES 比较如图 5 所示。

表 4 不同缺失率情况下缺失值填充误差
Table 4 Imputing error of missing value with different missing rates

缺失率/%	填充误差		
	MRE	MAE	MSE
10	0.026 0	0.007 5	0.000 1
20	0.025 1	0.006 9	0.000 08
30	0.038 1	0.010 0	0.000 2
40	0.042 6	0.012 0	0.000 3
50	0.044 2	0.007 3	0.000 09
60	0.056 4	0.011 2	0.000 2
70	0.035 8	0.007 8	0.000 1
80	0.040 0	0.010 3	0.000 2
90	0.055 6	0.011 9	0.000 3

表 5 不同缺失率下 4 种填充算法的 RMSE 对比
Table 5 RMSE comparison of four imputing algorithms under different missing rates

缺失率/%	SST				PM _{2.5}			
	SBL	RNN	NP	MNP	SBL	RNN	NP	MNP
10	0.082 5	0.010 0	0.020 0	0.010 0	0.069 3	0.013 5	0.012 3	0.012 2
20	0.087 7	0.013 8	0.020 0	0.008 9	0.080 0	0.015 0	0.013 1	0.012 6
30	0.095 4	0.015 5	0.018 2	0.014 1	0.094 9	0.015 8	0.014 3	0.013 7
40	0.094 9	0.016 7	0.023 7	0.017 3	0.082 5	0.015 2	0.014 9	0.014 1
50	0.102 5	0.016 4	0.023 2	0.009 5	0.091 1	0.021 6	0.015 3	0.014 5
60	0.131 5	0.012 2	0.044 7	0.014 1	0.102 5	0.019 1	0.016 1	0.012 3
70	0.169 4	0.013 0	0.019 2	0.010 0	0.158 7	0.022 0	0.016 7	0.015 7
80	0.172 6	0.019 0	0.018 7	0.014 1	0.209 0	0.021 4	0.018 1	0.013 1
90	0.283 4	0.023 7	0.027 9	0.017 3	0.331 1	0.021 2	0.019 4	0.015 9

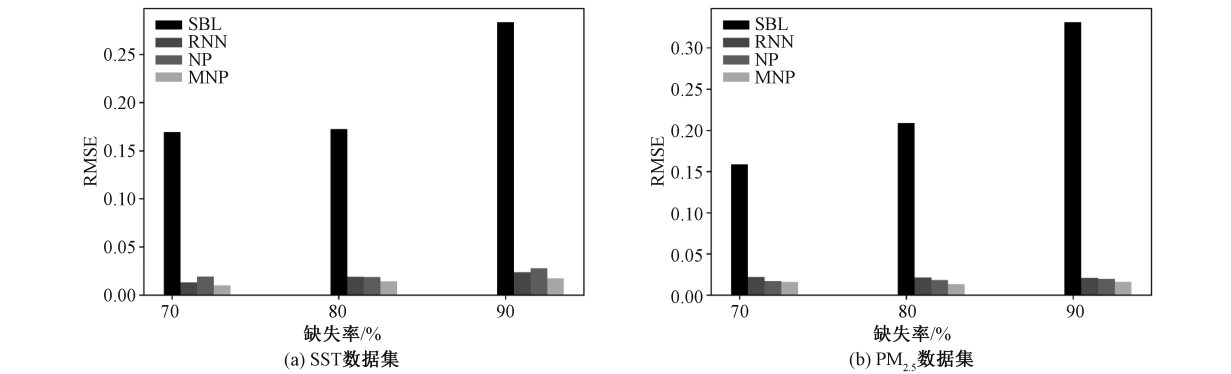


图 5 高缺失率条件下 4 种算法的 RMSE 比较
Fig. 5 RMSE comparison of four algorithms with high missing rates

果相近, 而 SBL 在高缺失率情况下的填充效果不佳; 并且在高缺失率情况下, MNP 在不同的实验数据集上具有相同的优异效果。

总体而言, 基于 MNP 的填充算法能够在不同缺失率情况下有相对稳定的误差, 且在高缺失率

由实验结果发现, 在两种不同的小数据集上, 4 种算法均可实现缺失数据的填充, 但是填充效果存在较大的差异。SBL 算法可以处理缺失率较低的数据集, 且精度要求不能过高; RNN 算法在不同缺失率下的填充效果相较于 SBL 而言更优; 基于 NP 模型的填充算法, 其效果时而介于两者之间, 时而优于 RNN。不难发现, 虽然 SBL、RNN、NP 3 种算法填充的缺失数据误差都随着缺失率的升高而增加, 但 MNP 算法的填充误差能在不同缺失率的情况下保持相对稳定, 且在每个缺失率下的误差均为最小; 特别是在高缺失率的情况下, MNP 的优势更加显著。同时, 可以看出 MNP 的性能较 NP 算法有所提升, 提高了填充精度, 体现了修正系数的作用。并且, 将 4 种算法应用于不同的数据集进行实验, 由图 5 可以看出, 所提算法在高缺失率情况下, 填充效果最优, NP 与 RNN 效

的情况下, MNP 的性能优势更加显著。

5 总结

本文提出一种基于改进神经过程的缺失数据填充算法, 该算法将神经网络和高斯过程推理结

合起来,弥补了两者的缺点,使其在小数据集背景下能够逼近多种不同的随机分布;该算法在 NP 的基础上增加填充过程,通过对观测数据的学习得到合适的分布,以此分布对缺失数据进行填充;算法通过引入修正系数 α ,提高了高缺失率情况下缺失数据的填充精度。该算法能在不同缺失率情况下有相对稳定的误差,且在高缺失率的情况下, MNP 的性能优势更加显著。

由于修正系数为 MNP 模型的超参数,无法在训练阶段灵活地修改,使其缺少自适应性。并且,算法的应用背景是在单变量时间序列,缺乏多变量间的信息互反馈能力。因此,下一步的研究方向为使修正系数更具有自适应能力,增加模型的信息互反馈过程,从而解决多变量时间序列的缺失数据填充问题。

参考文献

- [1] 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.
- [2] Balouji E, Salor Ö, Ermis M. Exponential smoothing of multiple reference frame components with GPUs for real-time detection of time-varying harmonics and interharmonics of EAF currents [J]. *IEEE Transactions on Industry Applications*, 2018, 54(6): 6566-6575.
- [3] Kozera R, Wilkołazka M. Natural spline interpolation and exponential parameterization for length estimation of curves [C] // AIP Conference Proceedings. Rhodes: AIP Publishing, 2017, 1863(1): 400010.
- [4] Newsham G R, Birt B J. Building-level occupancy data to improve ARIMA-based electricity use forecasts [C] // Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building. Zurich: ACM, 2010: 13-18.
- [5] Lippi M, Bertini M, Frasconi P. Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2013, 14(2): 871-882.
- [6] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C] // Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston: ACM, 2006: 501-508.
- [7] Yu H F, Rao N, Dhillon I. Temporal regularized matrix factorization for high-dimensional time series prediction [C] // Advances in Neural Information Processing Systems. Barcelona: NIPS, 2016: 847-855.
- [8] Hron K, Templ M, Filzmoser P. Imputation of missing values for compositional data using classical and robust methods [J]. *Computational Statistics & Data Analysis*, 2010, 54 (12): 3095-3107.
- [9] Stekhoven D J, Bühlmann P. MissForest: non-parametric missing value imputation for mixed-type data [J]. *Bioinformatics*, 2012, 28(1): 112-118.
- [10] Jia Z J, Song T W, Wang J X, et al. A time-series missing data completion method based on Fourier transform and kNNI algorithm [J]. *Software Engineering*, 2017, 20(3): 9-13.
- [11] Miller D, Ward A, Bambos N, et al. Physiological waveform imputation of missing data using convolutional autoencoders [C] // 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom). Ostrawa: IEEE, 2018: 1-6.
- [12] Wang H, Yuan Z L, Chen Y B, et al. An industrial missing values processing method based on generating model [J]. *Computer Networks*, 2019, 158: 61-68.
- [13] Duan Y J, Lü Y S, Kang W W, et al. A deep learning based approach for traffic data imputation [C] // 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). Qingdao: IEEE, 2014: 912-917.
- [14] Garnelo M, Schwarz J, Rosenbaum D, et al. Neural processes [J]. arXiv preprint arXiv:1807.01622, 2018.
- [15] Wentz F J, Gentemann C, Smith D, et al. Satellite measurements of sea surface temperature through clouds [J]. *Science*, 2000, 288(5467): 847-850.
- [16] NOAA/Pacific Marine Environmental Laboratory. Tropical atmosphere ocean [DB/OL]. [2019-06-06]. http://www.pmel.noaa.gov/tao/proj_over/proj_over.html.
- [17] Dua D, Graff C. UCI Machine learning repository [DB/OL]. [2019-07-30]. <https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>.
- [18] Zhang Z L, Rao B D. Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2011, 5(5): 912-926.
- [19] Strauman A S, Bianchi F M, Mikalsen K, et al. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks [C] // 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). Las Vegas: IEEE, 2018: 307-310.