

基于自动编码器的时间序列预测混合模型^①



路强¹, 滕进风¹, 黎杰¹, 凌亮², 丁超², 黄健刚³

¹(合肥工业大学 计算机与信息学院, 合肥 230009)

²(黄山旅游集团水电开发有限公司, 黄山 245899)

³(南京航空航天大学 经济与管理学院, 南京 211106)

通信作者: 路强, E-mail: luqiang@hfut.edu.cn

摘要: 时间序列预测目前在众多领域有着广泛应用. 如果可以准确估计事件或指标的未来发展, 它可以帮助人们做出重要的决定. 然而对不同时间序列建立模型并准确预测已成为最具挑战的应用之一. 因此, 本文提出了一种新颖的混合多步预测模型, 称为 SSA-ConvBiAE. 首先, 通过奇异谱分析 (SSA) 将原始数据分解为不同的趋势分量. 其次, 设计了新的基于卷积长短期记忆 (ConvLSTM) 和双向门控循环单元 (BiGRU) 的自动编码器网络结构. 最后, 将不同的分量分别输入到对应的自动编码器中进行训练和预测并求和预测结果. 为了评价模型的预测性能, 在真实的供水数据集和公开的时间序列数据集上进行了实验, 实验结果表明, 模型的预测结果优于基线方法. 本文已在网站 <https://github.com/VIMLab-hfut/SSA-ConvBiAE> 上发布了源代码.

关键词: 时间序列预测; 混合模型; 奇异谱分析; 数据分解; 自动编码器

引用格式: 路强, 滕进风, 黎杰, 凌亮, 丁超, 黄健刚. 基于自动编码器的时间序列预测混合模型. 计算机系统应用, 2022, 31(7): 55-65. <http://www.c-s-a.org.cn/1003-3254/8581.html>

Hybrid Model for Time Series Prediction Based on Autoencoder

LU Qiang¹, TENG Jin-Feng¹, LI Jie¹, LING Liang², DING Chao², HUANG Jian-Gang³

¹(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China)

²(Huangshan Tourism Group Hydropower Development Co. Ltd., Huangshan 245899, China)

³(School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Time series prediction currently has a wide range of applications in many fields. It can help people make important decisions if they can accurately estimate the future development of events or indicators. However, modeling and accurately predicting time series with different features has become one of the most challenging applications. Therefore, a novel hybrid multi-step prediction model is proposed, called SSA-ConvBiAE. Firstly, the original data is decomposed into different trend components by singular spectrum analysis (SSA). Secondly, we design a new autoencoder network structure based on convolutional long short-term memory (ConvLSTM) and bidirectional gated recurrent unit (BiGRU). Finally, the different components are inputted to the corresponding autoencoders for training and prediction, and the prediction results are fused. To evaluate the predictive performance of our model, we conduct experiments on two real water supply datasets and two publicly available time series datasets. Experimental results show that the proposed model achieves better performance than baseline methods. The source code has been published on <https://github.com/VIMLab-hfut/SSA-ConvBiAE>.

Key words: time series prediction; hybrid model; singular spectrum analysis (SSA); data decomposition; autoencoder

① 基金项目: 安徽省重点研发计划 (201904D07010); 黄山市科技计划 (2019KN-05); 合肥工业大学智能制造研究院科技成果培养 (IMIPY2021022)

收稿时间: 2021-10-11; 修改时间: 2021-11-12; 采用时间: 2021-11-19; csa 在线出版时间: 2022-05-31

时间序列预测在许多工业和商业应用中发挥着重要作用,例如金融市场、网络流量、天气预报和供水行业等。在这些场景中,人们可以利用大量关于过去行为的时间序列数据来预测将来的值^[1-3]。时间序列预测的主要挑战是提高预测能力以满足日益增长的多步预测需求。目前,时间序列预测在实际应用中关系到经济社会生活的各个方面。因此,在这项工作中,本文将研究时间序列的多步预测性能。

传统的线性回归方法主要包括自回归,移动平均,自回归综合移动平均模型 (ARIMA) 等。而非线性机器学习模型主要包括人工神经网络、支持向量回归和梯度提升决策树等^[4-6],它们在时间序列预测方面也取得了良好的效果。目前,深度学习技术可以更加高效的提取数据的特征。经典深度学习框架有深度神经网络,卷积神经网络,循环神经网络 (RNN)^[7]等。RNN 是一种在时间上进行递归的神经网络,但它存在梯度消失和梯度爆炸的问题。长短期记忆网络 (LSTM)^[8]使用 3 个门控机制来解决梯度爆炸和梯度消失问题,它在长序列预测中表现得更好。门控循环单元 (GRU)^[9]作为 LSTM 的变体,GRU 与 LSTM 具有相似的性能,它简化了 LSTM 的门控结构,降低了网络的训练时间。然而这些方法在对序列数据中的长期和复杂关系进行建模时仍然具有局限性。为了提高预测精度很多信号处理方法被用于时间序列预测,在这些模型中,信号处理方法将时间序列进行分解和去噪处理^[10,11]。在时间序列数据中,不同的数据分量可能具有不同的特征,现有的常见混合模型无法充分利用不同分量的特征。

为了解决上述问题,本文提出了一种新型的混合多步预测模型。通过采用奇异谱分析对数据进行预处理,再结合本文设计的自动编码器网络结构对预处理后的数据进行预测,可以实现对复杂数据建模的能力。主要贡献如下:

1) 提出了一种端到端的时间序列预测混合模型,该模型可以用于时间序列数据的多步预测。该混合模型的预测精度优于单一模型。

2) 采用 SSA 将原始时间序列数据分解为不同的趋势分量,SSA 可以从原始数据中提取数据的不同趋势信息。同时设计了新的基于 ConvLSTM 和 BiGRU 自动编码器结构,该模型能够进行短期和长期特征学习。

3) 使用真实世界的 2 个供水数据集和 3 个公开的

时间序列数据集评价所提出的模型。实验结果表明,与所有基线方法相比,该模型在多步预测评价指标下几乎都获得了最好的预测性能,证明了模型在时间序列预测方面的优越性。

本文的其余部分安排如下:第 1 节介绍了相关的工作。第 2 节阐述了所提出模型的设计与实现。第 3 节评价了模型的预测性能,并对实验结果进行了详细讨论。第 4 节对本文进行总结。

1 相关工作

目前,为了更好地提高时间序列预测的准确性,很多信号处理的方法被用于时间序列预测。例如,Zhang 等人^[12]采用小波变换 (WT) 获取交通流的时变和周期性特征,建立了季节性 ARIMA 进行预测。Ahani 等人^[13]提出了一种多步预测系统,集成了经验模式分解 (EEMD),最小二乘支持向量回归 (LSSVR) 和长短期记忆神经网络 (LSTM)。Chang 等人^[14]提出了一种基于小波变换和 Adam 优化的 LSTM 神经网络混合模型,用于时间序列数据预测。此外,很多方法采用了数据分解技术来提取数据不同的趋势信息进行预测。例如,Tang 等人^[15]通过 WT 和 SSA 对金融时间序列进行分解重构去噪,将平滑序列引入 LSTM 得到预测值。Li 等人^[16]利用变分模式分解 (VMD) 将非平稳月降水时间序列分解为几个相对稳定的固有模式函数 (IMFs),然后为每个 IMF 建立一个极限学习机预测模型,最后将预测值累加得到预测结果。Liu 等人^[17]采用 SSA 将原始数据进行分解,同时使用 CNNGRU 和 SVR 分别预测不同的分量数据。Zhu 等人^[18]通过 VMD 来捕捉时间序列的趋势和可变性信息,通过引入双向门控循环单元实现对天然橡胶期货的短期预测。在这些信号处理方法中,WT 可以提取时频信息,EEMD 和 VMD 可以提取趋势信息。然而,SSA 具有严格的数学理论和少量的参数,并且可以有效地提取信号的趋势信息,因此被选为信号处理的方法。

由于深度学习算法在数据特征提取方面的强大功能,目前也有大量的工作在时间序列预测领域取得了令人瞩目的成绩。例如,Sagheer 等人^[19]提出了一种深度长短期记忆 (DLSTM) 模型用来捕捉时间序列数据中的非线性动态和长期依赖关系。Zhang 等人^[20]提出了一种基于双向长短期记忆 (Bi-LSTM) 网络模型用于时间序列数据的预测。Hu 等人^[21]采用了一种将卷积神

经网络与双向长短期记忆网络相结合的模型,用于预测城市的用水需求. Essien 等人^[22]提出了一种用于多步预测的端到端模型,该模型包含一个深度 ConvLSTM 编码器-解码器架构,应用于智能工厂时间序列数据的预测. Salinas 等人^[23]提出了 DeepAR,这是一种产生精确概率预测的方法,基于对大量相关时间序列训练自动回归的递归网络模型.因此,深度学习方法在时间序列预测方面具有广泛的应用.同时,它们使用由多个非线性变换组成的架构来映射数据中的高级抽象.

根据以上文献,信号处理方法和深度学习方法在时间序列预测方面有着广泛的使用.然而,由于不同数据具有不同的特征,单一的深度学习方法在对数据中的长期和复杂关系进行建模时具有一定的局限性.基于此背景,在本研究中,提出了一种新的混合模型.使用奇异谱分析对数据进行预处理,再结合本文设计的自动编码器网络用于对预处理后的数据进行预测,该模型可用于供水数据的预测,同时模型也可以应用于其他时间序列数据的预测任务.

2 方法

2.1 问题描述

在本研究中,多步(即序列到序列)预测是一个重要的时间序列预测问题,本文将时间序列预测制定为有监督的机器学习任务.多步时间序列预测的目标是使用大量关于过去行为的时间序列数据来预测将来的值,给定一个一维的单变量时间序列, $x_t \in \{x_1, x_2, \dots, x_{N-1}, x_N\}$, N 为原始序列的长度.对于多步提前预测,监督机器学习模型的输入 X 是 $x_{t-w+1}, \dots, x_{t-1}, x_t$, 输出 \hat{Y} 是 $y_{t+1}, y_{t+2}, \dots, y_{t+k}$. 本文使用固定长度的滑动时间窗口方法构建 X 、 \hat{Y} , 得:

$$\hat{Y} = (y_{t+1}, y_{t+2}, \dots, y_{t+k}) = f(x_{t-w+1}, \dots, x_{t-1}, x_t) \quad (1)$$

其中, w 是滑动窗口时间序列的长度, k 是需要预测的时间序列的长度.

2.2 总览

在本节中,将描述如何使用 SSA-ConvBiAE 模型实现时间序列数据的预测任务. SSA-ConvBiAE 模型的总体流程如图 1 所示.首先,对原始数据进行处理,主要是对数据的缺失值进行填充和归一化操作,处理后的数据被划分为训练集和测试集;通过数据分解步骤,将训练集和测试集进行分解和重构,来获取不同趋势的分量数据;然后,将获取到的训练集不同的分量数

据分别输入到自动编码器网络用于模型的训练,经过迭代保存训练好的模型.最后,将测试集分量输入到对应训练好的模型中用于预测,将不同分量的预测结果相加得到最终的预测结果.

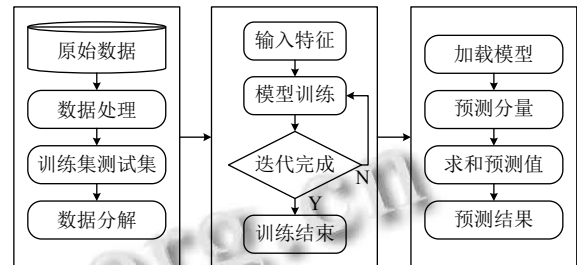


图1 总体预测模型的管线图

2.3 提出的模型

本文提出的 SSA-ConvBiAE 模型的整个过程如图 2 所示,详细说明如下.

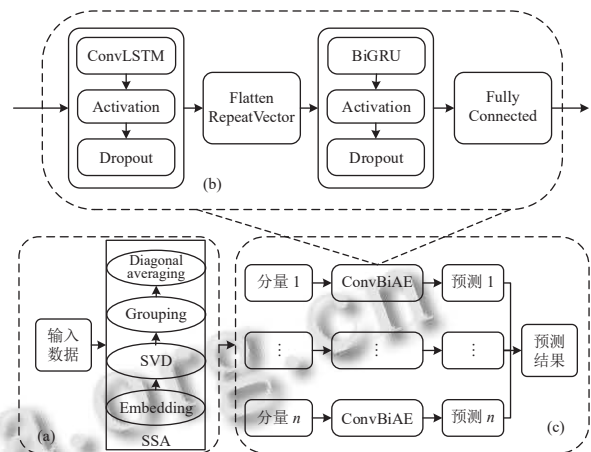


图2 SSA-ConvBiAE 模型结构图

2.3.1 数据分解

为了提高模型预测的准确性,采用 SSA 信号处理方法将原始的时间序列数据进行预处理, SSA 是用于时间序列分析有效的非参数算法, SSA 可以有效地实现信号的分解和重构.此外, SSA 具有严格的数学理论和较少的参数,并且可以很好地识别时间序列的周期、准周期和趋势信息.本文选择这种非参数方法将原始的时间序列分解并从原始数据中提取不同的趋势分量.具体如图 2(a) 所示, SSA 经过 4 个步骤分别是:嵌入 (embedding)、奇异值分解 (SVD)、分组 (grouping) 和对角平均化 (diagonal averaging), 可以将原始的时间序列数据进行分解和重构,并获取数据不

同的趋势信息,用于模型的训练和预测. SSA 的详细步骤描述如下:

1) 嵌入. 首先选择适当的窗口长度 $L (2 \leq L \leq N)$, 将一维的原始时间序列 $X = (x_1, x_2, \dots, x_N)$ 转化成多维的序列 $Z = (Z_1, Z_2, \dots, Z_K), Z_i = (x_i, x_{i+1}, \dots, x_{i+L-1})^T \in R^L$, 其中, $K = N - L + 1$, 可得轨迹矩阵 $Z = (Z_1, Z_2, \dots, Z_K) = (z_{ij})_{i,j=1}^{L,K}$, 表示为:

$$Z = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{pmatrix} \quad (2)$$

可以看出矩阵 Z 为汉克尔矩阵, 即沿对角线元素相等.

2) 奇异值分解. 通过 SVD, 轨迹矩阵 Z 可表示为:

$$Z = Z_1 + Z_2 + \cdots + Z_d \quad (3)$$

其中, $Z_i = \sqrt{\lambda_i} U_i V_i^T$, U_i 和 V_i 分别表示轨迹矩阵的左右特征向量, $\sqrt{\lambda_i} (i = 1, 2, \dots, d)$ 为矩阵 Z 的奇异值.

3) 分组. 该过程将区间 $\{1, 2, \dots, d\}$ 划分为多个离散子集 I_1, I_2, \dots, I_m , 其中 m 是组数, 则矩阵 Z 可表示为:

$$Z = Z_{I_1} + Z_{I_2} + \cdots + Z_{I_m} \quad (4)$$

4) 对角平均化. 该过程旨在将分组序列的每个矩阵转换为长度为 N 的新序列, 设 Z 为 $L \times K$ 的矩阵, $L^* = \min(L, K)$, $K^* = \max(L, K)$, $N = L + K - 1$, 且当 $L < K$ 时, $z_{ij}^* = z_{ij}$, 否则, $z_{ij}^* = z_{ji}$, 重构序列 $(rs_1, rs_2, \dots, rs_N)$ 可通过如下计算获得:

$$rs_n = \begin{cases} \frac{1}{n} \sum_{m=1}^n z_{m,n-m+1}^*, & 1 \leq n \leq L^* \\ \frac{1}{L^*} \sum_{m=1}^{L^*} z_{m,n-m+1}^*, & L^* \leq n \leq K^* \\ \frac{1}{N-n+1} \sum_{m=n-K^*+1}^{N-K^*+1} z_{m,n-m+1}^*, & K^* \leq n \leq N \end{cases} \quad (5)$$

2.3.2 自动编码器

自动编码器可以解决序列到序列建模的问题, 其中输入和输出均为序列. 本文设计了基于 ConvLSTM 和 BiGRU 的自动编码器结构, 具体如图 2(b) 所示, 提出的网络模型结构主要分为编码器和解码器.

1) 编码器有一层 ConvLSTM, 一层激活函数 \tanh 和一层 Dropout 函数堆叠在一起作为编码层. ConvLSTM^[24] 作为 LSTM 的一种变体, 虽然 LSTM 可以有利地捕获长输入序列的长期依赖, 但它的各个门

控结构之间采用的是由全连接的方式构建的, 因此模型在训练时需要更多的参数. 而 ConvLSTM 用卷积算子替换了全连接层算子, 使得它在输入到隐藏状态以及与各个门控结构之间的转换都具有卷积操作. ConvLSTM 继承了卷积算子中稀疏连接和参数共享的优点, 可以减少全连接结构的冗余. 此外, ConvLSTM 不仅能够提取时间特征, 通过卷积操作还可以有效对空间信息进行特征提取. 同时, ConvLSTM 通过门控机制克服了梯度消失的问题, 也保留了 LSTM 可以捕获长时间记忆的优点. 因此, 本文选择 ConvLSTM 作为编码器. 它的计算公式如下:

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (6)$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (7)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (8)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ C_t + b_o) \quad (9)$$

$$h_t = o_t \circ \tanh(C_t) \quad (10)$$

其中, σ 、 \tanh 、 $*$ 和 \circ 分别表示 Sigmoid、 \tanh 操作、卷积运算和 Hadamard 积; i_t 、 f_t 、 C_{t-1} 、 C_t 和 o_t 分别代表各种门控机制; x_t 表示输入数据, h_{t-1} 表示前一个 ConvLSTM 单元输出的循环状态, h_t 表示当前时刻的循环状态; W 和 b 分别代表门控机制的权重向量和偏差向量.

2) 解码器由一层 BiGRU、一层激活函数 \tanh 和一层 Dropout 函数堆叠在一起作为解码层. BiGRU 模型一种双向循环神经网络, 它可以利用正向和反向信息. 双向门控循环单元将两个传输方向相反的隐藏层连接到同一个输出层. 因此, 输出层具有输入序列中每个点的过去和未来状态的完整信息, 从而做出更准确的预测. BiGRU 是在 GRU 基础上进行正向和反向训练并将结果进行线性融合, 这样不但可以提取过去的特征, 也可以提取未来的特征, 同时数据复用, 而且每个时刻输出值是由两个方向的 GRU 共同决定. BiGRU 的思想是将常规 GRU 神经元分为前向状态和后向状态. GRU 的计算公式如下所示:

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (11)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (12)$$

$$\tilde{h}_t = \tanh(W_h[r_t * h_{t-1}, x_t] + b_h) \quad (13)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (14)$$

其中, σ 表示激活函数 Sigmoid, \tanh 是候选状态的激活函数, $*$ 是点积操作, x_t 表示输入向量, 更新门 z_t 可以确

定保留多少先前存储的信息,重置门 r_t 可以将多少先前存储的信息与新输入集成在一起。 h_{t-1} 表示上一时刻隐藏层的状态, h_t 表示当前时刻隐藏层状态, \tilde{h}_t 表示候选状态。 W 是与输入相关的权重矩阵, b 表示偏差向量。

为了使编码层的输出可以作为解码层的输入,在编码层之后通过使用一个重复向量层(repeat vector)。该层的主要功能是重复来自编码层的最终输出向量,以便解码层能够重建原始输入序列。最后,在解码层之后通过一个全连接层,用来输出数据的预测值。

2.3.3 预测分量

多个分量的预测过程,具体如图2(c)所示,将SSA获取到的 n 个分量数据分别作为每个自动编码器模型的输入数据进行模型的训练,保存训练好的模型。将不同的测试集分量分别输入到对应的训练好的自动编码器模型中进行预测,最后将不同分量的预测值进行求和,得到最终的预测结果。

SSA-ConvBiAE模型是一个健壮的,可扩展的端到端时间序列多步预测模型。奇异谱分析可以从原始数据中提取数据不同的趋势信息,同时本文设计的自动编码器网络结构能够进行短期和长期特征学习,该混合模型可以用于时间序列的多步预测。

2.4 损失函数

为了更好地反映出模型的预测值跟真实值的差距,在SSA-ConvBiAE模型中采用的损失函数是均方误差。同时,使用了基于随机梯度的优化算法Adam,每次迭代时更新模型参数。损失函数的计算方法如下:

$$Loss = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (15)$$

其中, \hat{y}_i 表示第 i 时刻的预测值, y_i 表示第 i 时刻的真实值, n 表示数据样本总数。

3 实验分析

3.1 实验环境

本文的实验环境是基于Windows操作系统,使用Python语言对代码进行实现。深度学习框架主要采用了TensorFlow,模型算法的实现是使用Keras,它是基于TensorFlow的深度学习库。硬件环境的配置是Intel(R) Core(TM)i7-7700K CPU @4.20 GHz、32.0 GB内存和NVIDIA GeForce RTX 2080 8 GB GPU。

3.2 数据描述

本文提出的模型在5个数据集进行了实验,其中

包括2个真实的供水数据集和3个公开的时间序列数据集。TH-reservoir和XH-waterworks是由黄山旅游集团水电开发有限公司提供的黄山风景区供水数据集。Milan-air^[25], Delhi-meantemp^[26]和Global-power^[27]是从Kaggle网站获得的公开的时间序列数据集。5个数据集的详细描述如下所示:

1) TH-reservoir: 该数据集是黄山风景区某水库的水位值,数据集包含了黄山风景区某水库2017年1月1日到2019年12月31日每天的水位值。

2) XH-waterworks: 该数据集是黄山风景区某水厂的供水量,数据集包含了黄山风景区某水厂2017年1月1日到2019年12月31日每天的供水量。

3) Milan-air: 该数据集是米兰空气中PM_{2.5}浓度数据,数据集包含2020年7月24日到2020年9月20日每小时PM_{2.5}浓度的平均值。

4) Delhi-meantemp: 该数据集是印度德里市的天气温度数据,数据集包含2013年1月1日至2017年1月1日每天温度的平均值。

5) Global-power: 该数据集是一个家庭在4年内的有功功率电力消耗数据,数据是每分钟收集一次,经过累加计算得出每天的消耗量,数据集包含2006年12月16日至2010年11月26日每天消耗量的值。

上述黄山风景区供水数据集包含少量的缺失数据,因此使用线性插值的方法进行填充。在实验中,将80%的数据用作训练集,其余20%用作测试集。为了评价模型的多步预测性能,在本文中,使用了输入步长为18的滑动窗口机制,进行多步预测。另外,将输入数据进行归一化处理,其公式如下所示:

$$x_{\text{nor}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (16)$$

其中, x_{nor} 表示归一化后的数据, x 表示输入的样本数据中某一个值, x_{max} , x_{min} 分别表示样本数据中的最大值和最小值。

3.3 基线方法

为了验证所提出模型的预测性能,本文进行了5个实验。在每个实验中,将SSA-ConvBiAE模型的预测性能与以下基线方法进行比较:

1) SVR^[5]: 支持向量回归,它使用历史数据来训练模型并对数据进行预测。

2) XGBoost^[6]: 它是一个端到端的梯度提升决策树模型,是一种高效且广泛使用的机器学习方法。

3) GRU^[9]: 门控循环单元, GRU 通过简化门控结构使得网络更易计算, 降低了模型的训练时间。

4) DLSTM^[19]: 通过堆叠多个 LSTM 块, 构建一个深度 LSTM 神经网络。

5) Bi-LSTM^[20]: 主要使用前向和后向 LSTM 来捕捉过去和未来隐含的信息, 然后将两部分结合起来形成最终的预测输出。

6) SSA-ConvLSTM: 消融实验. 奇异谱分析和 ConvLSTM 的混合模型, 详细描述请参见第 2.3 节。

7) SSA-BiGRU: 消融实验. 奇异谱分析和 BiGRU 的混合模型, 详细描述请参见第 2.3 节。

本实验采用 3 种评价指标对模型的预测结果进行评价, 具体如下:

平均绝对误差 (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (17)$$

均方根误差 (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (18)$$

平均绝对百分比误差 (MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (19)$$

其中, \hat{y}_i 表示第 i 时刻的预测值, y_i 表示第 i 时刻的真实值, n 表示数据样本总数. 使用 MAE, RMSE 和 MAPE 度量预测误差: 该值越小, 表明预测效果越好。

3.4 实验结果

在本节中, 为了研究所提出模型的多步预测性能, 对涉及到的所有模型进行 1-step 到 4-step 的预测. 表 1, 表 2 和表 3 给出了 2 个真实的供水数据和 3 个公开时间序列数据在几种不同模型上的预测误差评价结果。

表 1 SSA-ConvBiAE 模型和基线方法对 TH-reservoir 和 XH-waterworks 数据集的预测结果

方法	Metric	TH-reservoir				XH-waterworks			
		1-step	2-step	3-step	4-step	1-step	2-step	3-step	4-step
SVR	MAE	0.2148	0.3014	0.3648	0.4062	15.7708	16.4340	16.6213	16.9048
	RMSE	0.3331	0.4119	0.4897	0.5387	20.1755	20.9013	21.3955	21.7661
	MAPE	4.0307	5.5511	6.7227	7.5221	22.1686	22.8458	23.0520	23.5074
XGBoost	MAE	0.2107	0.2901	0.3492	0.3868	16.5008	16.8854	16.9158	17.2214
	RMSE	0.3310	0.4185	0.4947	0.5421	21.7074	21.8985	21.8063	21.8424
	MAPE	3.8793	5.2964	6.4083	7.1386	22.5624	22.8463	22.7885	23.4139
GRU	MAE	0.2304	0.3053	0.3515	0.3780	14.0561	15.9247	16.5329	17.4905
	RMSE	0.3431	0.4158	0.4688	0.4907	18.9721	20.7614	21.6677	22.4874
	MAPE	4.2354	5.5630	6.3754	6.8397	20.0045	21.8871	22.9554	24.0004
DLSTM	MAE	0.2086	0.2927	0.3545	0.3941	15.6888	16.5521	17.1444	17.9619
	RMSE	0.3324	0.4051	0.4746	0.5076	20.1358	21.2547	22.0329	22.8184
	MAPE	3.8266	5.3095	6.4040	7.0854	20.0847	21.8357	22.7096	23.8186
Bi-LSTM	MAE	0.2193	0.2988	0.3569	0.3806	14.4704	16.1560	16.6158	17.8016
	RMSE	0.3422	0.4139	0.4778	0.4972	19.2699	20.9463	21.6793	22.7279
	MAPE	4.0100	5.4091	6.4612	6.8891	19.8307	21.8053	22.7468	23.8500
SSA-ConvLSTM	MAE	0.0753	0.1223	0.1932	0.2506	4.0829	4.9776	8.4561	12.2600
	RMSE	0.1025	0.1601	0.2529	0.3337	5.0621	6.1267	10.6065	15.7783
	MAPE	1.3817	2.2113	3.4728	4.5687	5.2740	6.1589	11.1632	16.9180
SSA-BiGRU	MAE	0.0883	0.1526	0.1998	0.2533	3.9296	4.8381	7.5997	11.4564
	RMSE	0.1157	0.1940	0.2632	0.3394	4.4895	5.8972	9.6162	14.7939
	MAPE	1.6013	2.7391	3.5876	4.6135	5.4663	6.2221	10.3075	15.4798
SSA-ConvBiAE	MAE	0.0750	0.0953	0.1408	0.2239	4.0902	4.4687	7.2962	11.1657
	RMSE	0.0951	0.1231	0.1943	0.3042	5.0023	5.4573	9.1255	14.2681
	MAPE	1.3721	1.7295	2.5551	4.0622	5.1325	5.8490	9.4062	15.0781

通过表 1 的实验结果, 可以发现: 在供水数据案例研究中, SSA-ConvBiAE 模型可以实现准确的多步预测结果, 混合模型的预测精度优于单一模型. 这表明

SSA 对数据的分解, 可以获取数据不同趋势的分量信息, 从而降低了时间序列的复杂度, 提高了模型的预测精度. 与其他几种基线方法相比, SSA-ConvBiAE 模型

的多步预测几乎具有最好的预测精度,这表明设计的基于 ConvLSTM 和 BiGRU 结构的自动编码器模型可以有效地提取数据特征,适用于时间序列预测。同时,为了验证模型不仅可以用于供水数据预测,本文也在 3 个公开的时间序列数据进行了实验,通过表 2 和表 3 的实验结果可以发现,所提出的模型在公开的数据集

上都有优异的表现,表明模型具有一定的泛化能力。

表 1, 表 2 和表 3 显示了 SSA-ConvBiAE 模型和其他基线方法在 5 个时间序列数据集上的多步预测任务的性能。可以看出, SSA-ConvBiAE 模型在多步预测范围的评价指标下几乎都获得了最好的预测性能。模型与其他基线方法预测结果相比,详细描述如下。

表 2 SSA-ConvBiAE 模型和基线方法对 Milan-air 和 Delhi-meantemp 数据集的预测结果

方法	Metric	Milan-air				Delhi-meantemp			
		1-step	2-step	3-step	4-step	1-step	2-step	3-step	4-step
SVR	MAE	4.0273	4.7397	5.2387	5.7201	1.3028	1.5254	1.6435	1.7267
	RMSE	5.3833	6.2017	6.8107	7.3278	1.6849	1.9693	2.1015	2.1720
	MAPE	16.4039	19.9924	22.6056	24.9536	4.6044	5.3939	5.8726	6.3452
XGBoost	MAE	4.5597	5.0486	5.4746	5.8706	1.2982	1.5577	1.6780	1.7881
	RMSE	5.8903	6.4470	6.9882	7.5210	1.6700	1.9873	2.1517	2.2806
	MAPE	18.7956	21.3365	23.4955	25.3287	4.5608	5.4778	5.9581	6.4796
GRU	MAE	3.8591	4.8680	5.5715	6.0236	1.3588	1.5569	1.6935	1.7761
	RMSE	5.0071	6.2108	7.1241	7.5341	1.7536	2.0335	2.1844	2.2659
	MAPE	17.0842	21.5169	24.6730	26.9503	4.8136	5.4863	6.0537	6.4880
DLSTM	MAE	3.8704	5.2239	5.9987	6.2409	1.2841	1.5249	1.6850	1.7747
	RMSE	5.0131	6.6093	7.5250	7.7459	1.6816	1.9898	2.1502	2.2444
	MAPE	17.2026	22.9217	26.3463	27.5334	4.6342	5.4411	6.0572	6.5354
Bi-LSTM	MAE	3.4784	4.7412	5.3466	5.6334	1.3685	1.5829	1.7375	1.8239
	RMSE	4.6494	6.0811	6.8411	7.1539	1.7587	2.0597	2.2269	2.3236
	MAPE	15.2156	20.5404	23.1114	24.5163	4.8734	5.6258	6.2571	6.7123
SSA-ConvLSTM	MAE	0.6895	1.1380	1.7412	2.8244	0.2428	0.4356	0.7269	1.0090
	RMSE	0.8586	1.3728	2.0823	3.4567	0.2981	0.5344	0.8902	1.2743
	MAPE	3.1635	5.1916	7.6204	12.0154	0.8754	1.5822	2.6334	3.7716
SSA-BiGRU	MAE	0.4754	0.9366	1.4520	2.3340	0.3304	0.4639	0.6983	0.9730
	RMSE	0.6056	1.1637	1.8236	2.9402	0.3925	0.5725	0.8657	1.2366
	MAPE	2.2221	4.3080	6.7008	10.5619	1.1430	1.6146	2.4534	3.5537
SSA-ConvBiAE	MAE	0.6981	0.8552	1.0886	1.8585	0.3620	0.3615	0.5522	0.8219
	RMSE	0.8773	1.0752	1.3759	2.3944	0.4319	0.4483	0.6889	1.0505
	MAPE	3.0741	3.8722	4.9224	8.2525	1.2692	1.3006	1.9731	3.0350

1) 预测精度. 以 TH-reservoir 数据集为例, 从表 1 中可以看出: 所提出的模型 SSA-ConvBiAE 极大地提高了多步预测的性能, 这表明该模型在长期预测能力方面取得了优异的效果。例如, 对于 1-step 的供水数据预测任务, SSA-ConvBiAE 模型和 SVR 和 XGBoost 模型相比, RMSE 误差降低了约 71.45% 和 71.27%。这主要是由于 SVR 和 XGBoost 等方法难以处理复杂的非平稳时间序列数据。SSA-ConvBiAE 模型和 GRU 模型、DLSTM 模型和 BiLSTM 模型相比, RMSE 误差降低了约 72.28%、71.39% 和 72.21%, 这主要是由于针对具有不同特征的时间序列数据, 这几种模型的预测结果会有所降低。从表 1 可以看出, 经过 SSA 进行数据分解, 预测精度得到了明显提高, 这表明提出的混合

模型是有效的。

2) 消融分析. 为了验证 SSA-ConvBiAE 模型采用自动编码器网络结构的有效性, 将 SSA-ConvBiAE 与 SSA-ConvLSTM 和 SSA-BiGRU 模型进行了比较。以 TH-reservoir 数据集为例, 从图 3 中可以看出, 在 RMSE 和 MAPE 两种不同的实验评价指标下, 可以清楚地看到提出的 SSA-ConvBiAE 比 SSA-ConvLSTM 和 SSA-BiGRU 具有更好的预测结果。SSA-ConvBiAE 与 SSA-ConvLSTM 和 SSA-BiGRU 模型相比, 从 1-step 预测可以看出, RMSE 分别降低了约 7.22% 和 17.80%。实验表明, ConvLSTM 充分捕捉了序列数据的时间和空间分布, 同时模型中结合 BiGRU, 使输出层具有输入序列中每个点的过去和未来状态的完整信息, 可以融

合更多的序列特征,提高了模型的预测性能.因此,表明了本文提出的自动编码器网络结构可以有效地提取数据特征.

表3 SSA-ConvBiAE模型和基线方法对Global-power数据集的预测结果

方法	Metric	Global-power			
		1-step	2-step	3-step	4-step
SVR	MAE	249.8204	259.8261	270.3639	277.3588
	RMSE	328.6983	338.9298	351.6436	354.8314
	MAPE	20.3915	21.3555	22.3821	22.7836
XGBoost	MAE	248.7647	267.3514	273.2386	273.9314
	RMSE	331.7338	348.4573	350.9789	349.3888
	MAPE	20.0822	21.5613	22.1403	22.0714
GRU	MAE	249.0471	258.0371	259.4382	264.2936
	RMSE	332.5221	335.9406	338.1379	338.2258
	MAPE	19.2328	20.1282	20.7071	21.7013
DLSTM	MAE	240.6523	262.3337	276.2036	265.8219
	RMSE	326.6338	343.8437	357.6111	344.9340
	MAPE	18.6140	21.3528	22.9073	21.5067
Bi-LSTM	MAE	239.1285	255.5482	265.8352	264.0310
	RMSE	329.1918	343.8356	348.6273	341.0071
	MAPE	18.2622	19.8988	21.6603	21.3600
SSA-ConvLSTM	MAE	97.8648	96.2532	172.9475	253.5218
	RMSE	126.7232	129.2711	231.1604	331.6199
	MAPE	7.5148	7.2990	13.1184	19.3042
SSA-BiGRU	MAE	85.7732	96.0599	161.7639	251.3836
	RMSE	108.5563	127.1634	219.038	328.3125
	MAPE	6.6407	7.0613	12.2081	20.0782
SSA-ConvBiAE	MAE	85.0859	83.8347	146.8099	244.5662
	RMSE	112.4816	112.0768	203.0153	333.1809
	MAPE	6.4662	6.2925	10.8272	18.0097

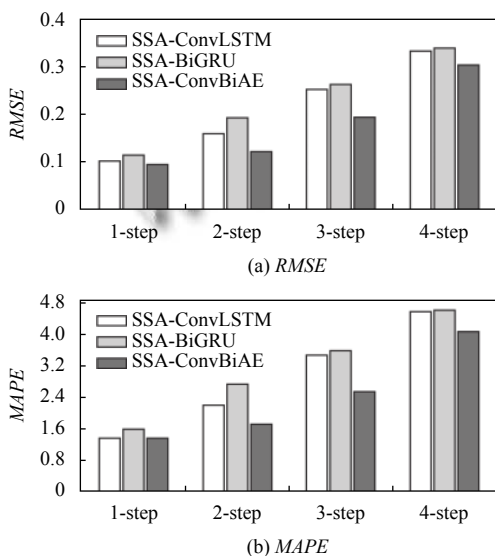


图3 消融实验预测结果对比

3) 多步预测.从以上的实验结果可以看出,在5个不同的数据集上SSA-ConvBiAE模型预测精度在多步预测范围内明显要高于其他基线方法.以TH-reservoir和XH-waterworks数据集为例,从表1中可以看出,本文提出的SSA-ConvBiAE模型在1-step到4-step的预测结果几乎都获得了最好的预测性能,这表明提出的混合模型是可行的.因此,该模型可以用于供水数据和其他时间序列任务的预测,同时本文提出的SSA-ConvBiAE模型不仅可以用于短期预测,还可以用于中长期预测.

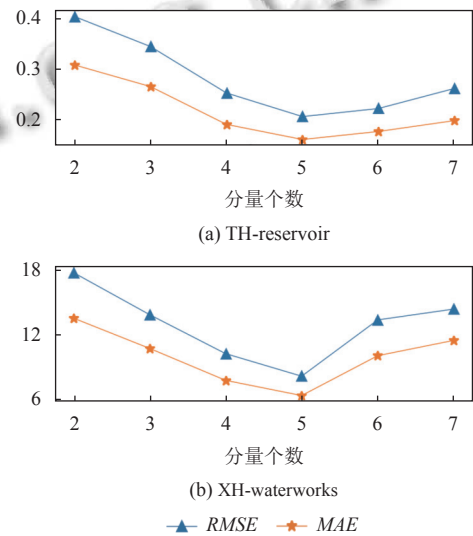


图4 不同的SSA分量在TH-reservoir和XH-waterworks数据集上的预测性能比较

3.5 模型参数选择

在本节中,将选择SSA-ConvBiAE模型的相关参数.SSA-ConvBiAE模型的超参数主要包括:批次大小、训练次数、隐藏层单元数、卷积核的尺寸、Dropout和数据分量个数.在实验中,将批次大小设置为32,训练次数设置为100,隐藏层单元设置为128,卷积核的尺寸设置为1×3,Dropout参数设置为0.1.

数据分量的个数是SSA-ConvBiAE模型的一个非常重要的参数,本文通过尝试不同的数据进行对比预测实验来选择最佳值.在实验中,从[2, 3, 4, 5, 6, 7]中选择不同的分量个数并分析预测精度的变化.如图4和图5所示,横轴代表不同的分量个数,纵轴代表RMSE和MAE两种不同的评价指标的变化.图4(a)和图4(b)分别显示了数据集TH-reservoir和XH-waterworks不同分量个数的RMSE和MAE的预测结果.可以看出,当分量个数为5时,误差最小.同样图5(a)和图5(b)分

别显示了数据集 Milan-air 和 Delhi-meantemp 的预测结果, 当个数为 6 时, 结果达到最小值, 图 5(c) 显示了数据集 Global-power 的预测结果, 当个数为 4 时, 结果达到最小值. 这表明不同的分量个数会极大地影响预测精度, 因此需要通过不同参数的对比实验来选择最佳值.

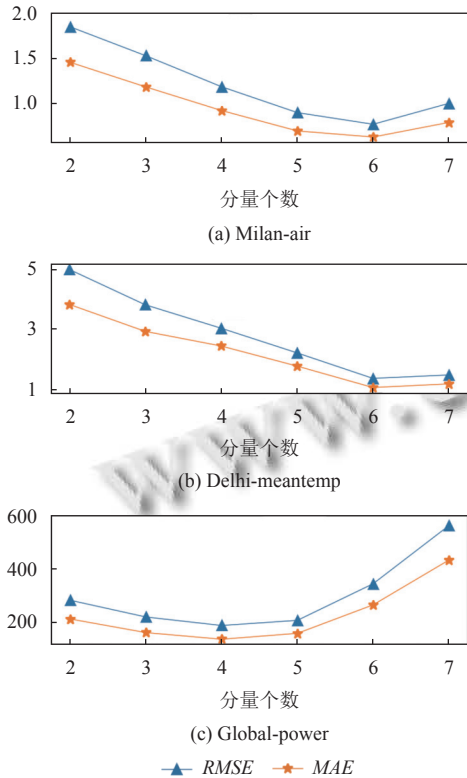


图5 不同的SSA分量在Milan-air, Delhi-meantemp和Global-power数据集上的预测性能比较

3.6 模型解释

为了更好的理解 SSA-ConvBiAE 模型的预测结果, 选择对 TH-reservoir 的部分测试集样本数据进行预测结果可视化, 图 6(a)–图 6(d) 分别显示了模型 1-step 到 4-step 的预测范围的可视化结果.

从图 6 中 1-step 和 2-step 预测的可视化结果来看, SSA-ConvBiAE 模型在数据的局部最小值和局部最大值处的预测结果都非常的接近真实值, 这表明 SSA 将数据分解成不同的趋势分量后, 自动编码器结构中 ConvLSTM 充分捕捉了序列数据的空间和时间分布, 具有自动特征提取方面的优势, 而 BiGRU 可以融合更多的序列特征, 提高了模型的预测性能. 从图 6 中 3-step 和 4-step 的预测可视化结果来看, 模型预测的总体趋势接近真实值, 但在局部最小值和局部最大值处的预测结果要低于前两步预测. 尽管 ConvLSTM, BiGRU

可以捕捉数据集中的长期依赖性, 但针对复杂数据的预测, 长期预测的结果也会有所降低. 从图 6 中的 1-step 到 4-step 的预测结果, 可以看出奇异谱分析与自动编码器网络的结合是有效的, 该模型可以提取序列数据的时间特征, 可以很好地拟合数据, 以及准确预测数据的变化趋势. 同时, 在 5 个数据集上的预测结果可以看出, 在针对具有不同特征的时间序列数据, 本文提出的模型在预测结果方面都具有优异的表现.

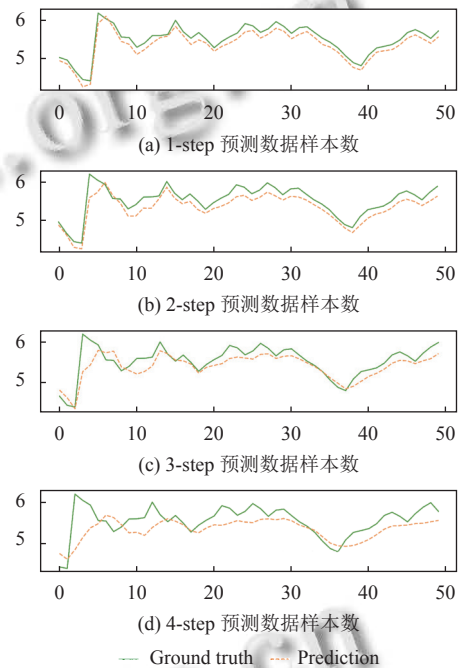


图6 SSA-ConvBiAE模型的可视化预测结果(其中, 纵坐标表示水位(m))

为进一步表明模型的预测结果, 选择 SSA-ConvBiAE 模型与其他基线方法的预测结果进行可视化对比, 如图 7(a)–图 7(d) 所示, 不同的模型对 TH-reservoir 数据集进行 1-step 到 4-step 预测结果可视化, 可以看出, 采用 SSA 对数据进行处理后, 混合模型的预测结果明显优于单一模型, 预测结果更接近真实值, 与 SSA-ConvLSTM 和 SSA-BiGRU 模型相比, SSA-ConvBiAE 模型在多步预测范围内获得了更好的预测性能. 同时本文也选择了对 Milan-air 数据集进行预测结果可视化, 如图 8(a)–图 8(d) 所示, 展示了不同模型的预测结果.

4 结论

本文提出的一种端到端的时间序列预测混合模型, SSA-ConvBiAE 模型是一个健壮的, 可扩展的端到端

模型可以用于时间序列数据的多步预测. 在提出的 SSA-ConvBiAE 模型中, 采用 SSA 将原始时间序列数据分解为不同的趋势分量, SSA 可以从原始数据中提取数据不同的趋势信息, 同时本文设计了新的基于 ConvLSTM 和 BiGRU 自动编码器结构, 模型能够进行短期和长期特征学习, 该混合模型的预测精度优于单一模型. 在真实世界的 2 个供水数据集和 3 个公开的时间序列数据集上进行实验来评价所提出的模型. 实

验结果表明, 与所有基线方法相比, 该模型在多步预测评价指标下几乎都获得了最好的预测性能, 证明了 SSA-ConvBiAE 模型在时间序列预测方面的优越性. 这表明提出的模型不仅可以应用于供水预测领域, 同时对于其他时间序列数据预测也具有一定的适用性. 在未来的研究计划中将继续深化研究 SSA-ConvBiAE 模型在时间序列数据方面的预测性能, 特别是针对多变量的时间序列预测问题.

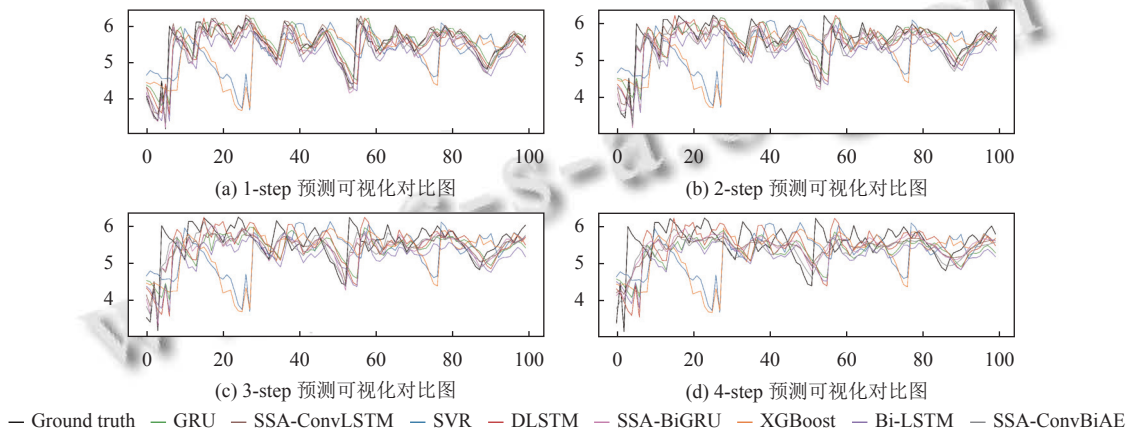


图7 SSA-ConvBiAE 模型与基线方法对 TH-reservoir 数据集的预测结果可视化对比图 (其中, 纵坐标表示水位 (m), 横坐标表示样本数目)

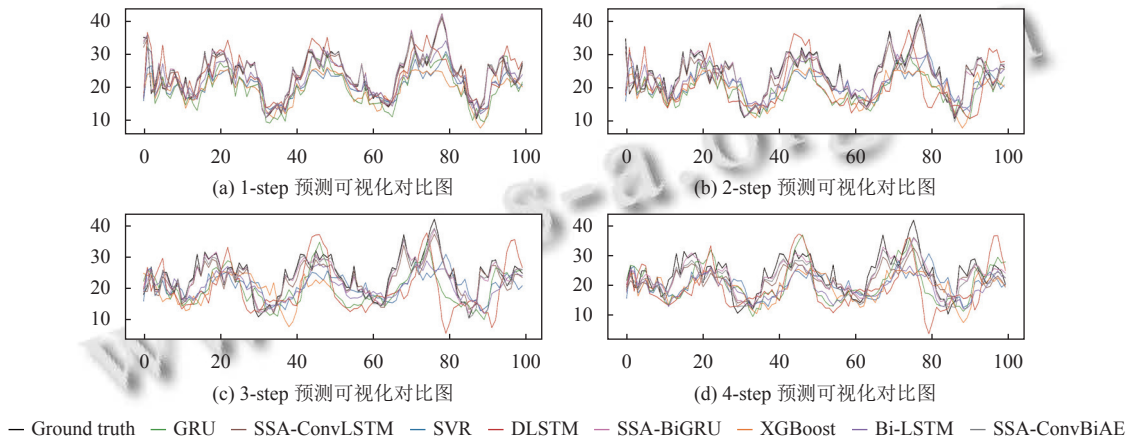


图8 SSA-ConvBiAE 模型与基线方法对 Milan-air 数据集的预测结果可视化对比图 (纵坐标表示: $PM_{2.5}$ 浓度 ($\mu g/m^3$), 横坐标表示样本数目)

参考文献

1 Tealab A. Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing and Informatics Journal, 2018, 3(2): 334–340. [doi: 10.1016/j.fcij.2018.10.003]

2 Lim B, Zohren S. Time-series forecasting with deep learning: A survey. arXiv: 2004.13408, 2020.
3 Han Z, Zhao J, Leung H, et al. A review of deep learning models for time series prediction. IEEE Sensors Journal, 2019, 21(6): 7833–7848. [doi: 10.1109/JSEN.2019.2923982]

- 4 Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. *International journal of forecasting*, 1998, 14(1): 35–62. [doi: [10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)]
- 5 Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and computing*, 2004, 14(3): 199–222. [doi: [10.1023/B:STCO.0000035301.49549.88](https://doi.org/10.1023/B:STCO.0000035301.49549.88)]
- 6 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM. 2016. 785–794.
- 7 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*, 1986, 323(6088): 533–536. [doi: [10.1038/323533a0](https://doi.org/10.1038/323533a0)]
- 8 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*, 1997, 9(8): 1735–1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)]
- 9 Cho K, Van Merriënboer B, Bahdanau D, *et al.* On the properties of neural machine translation: Encoder-decoder approaches. arXiv: 1409.1259, 2014.
- 10 Xu M, Han M, Lin H. Wavelet-denoising multiple echo state networks for multivariate time series prediction. *Information Sciences*, 2018, 465: 439–458. [doi: [10.1016/j.ins.2018.07.015](https://doi.org/10.1016/j.ins.2018.07.015)]
- 11 Zhang Y, Chen B, Pan G, *et al.* A novel hybrid model based on VMD-WT and PCA-BP-RBF neural network for short-term wind speed forecasting. *Energy Conversion and Management*, 2019, 195: 180–197. [doi: [10.1016/j.enconman.2019.05.005](https://doi.org/10.1016/j.enconman.2019.05.005)]
- 12 Zhang H, Wang X, Cao J, *et al.* A multivariate short-term traffic flow forecasting method based on wavelet analysis and seasonal time series. *Applied Intelligence*, 2018, 48(10): 3827–3838. [doi: [10.1007/s10489-018-1181-7](https://doi.org/10.1007/s10489-018-1181-7)]
- 13 Ahani I K, Salari M, Shadman A. An ensemble multi-step-ahead forecasting system for fine particulate matter in urban areas. *Journal of Cleaner Production*, 2020, 263: 120983. [doi: [10.1016/j.jclepro.2020.120983](https://doi.org/10.1016/j.jclepro.2020.120983)]
- 14 Chang Z, Zhang Y, Chen W. Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. *Energy*, 2019, 187: 115804. [doi: [10.1016/j.energy.2019.07.134](https://doi.org/10.1016/j.energy.2019.07.134)]
- 15 Tang Q, Shi R, Fan T, *et al.* Prediction of financial time series based on LSTM using wavelet transform and singular spectrum analysis. *Mathematical Problems in Engineering*, 2021, 2021: 1–13. [doi: [10.1155/2021/9942410](https://doi.org/10.1155/2021/9942410)]
- 16 Li G, Ma X, Yang H. A hybrid model for monthly precipitation time series forecasting based on variational mode decomposition with extreme learning machine. *Information*, 2018, 9(7): 177. [doi: [10.3390/info9070177](https://doi.org/10.3390/info9070177)]
- 17 Liu H, Mi X, Li Y, *et al.* Smart wind speed deep learning based multi-step forecasting model using singular spectrum analysis, convolutional gated recurrent unit network and support vector regression. *Renewable Energy*, 2019, 143: 842–854. [doi: [10.1016/j.renene.2019.05.039](https://doi.org/10.1016/j.renene.2019.05.039)]
- 18 Zhu Q, Zhang F, Liu S, *et al.* A hybrid VMD–BiGRU model for rubber futures time series forecasting. *Applied Soft Computing*, 2019, 84: 105739. [doi: [10.1016/j.asoc.2019.105739](https://doi.org/10.1016/j.asoc.2019.105739)]
- 19 Sagheer A, Kotb M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing*, 2019, 323: 203–213. [doi: [10.1016/j.neucom.2018.09.082](https://doi.org/10.1016/j.neucom.2018.09.082)]
- 20 Zhang M, Wu D, Xue R. Hourly prediction of PM 2.5 concentration in Beijing based on Bi-LSTM neural network. *Multimedia Tools and Applications*, 2021. [doi: [10.1007/s11042-021-10852-w](https://doi.org/10.1007/s11042-021-10852-w)]
- 21 Hu P, Tong J, Wang J, *et al.* A hybrid model based on CNN and Bi-LSTM for urban water demand prediction. *IEEE Congress on Evolutionary Computation (CEC)*. Wellington: IEEE. 2019. 1088–1094.
- 22 Essien A, Giannetti C. A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders. *IEEE Transactions on Industrial Informatics*, 2020, 16(9): 6069–6078. [doi: [10.1109/TII.2020.2967556](https://doi.org/10.1109/TII.2020.2967556)]
- 23 Salinas D, Flunkert V, Gasthaus J. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. arXiv: 1704.04110, 2019.
- 24 Shi X, Chen Z, Wang H, *et al.* Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press. 2015. 802–810.
- 25 Milan-air 数据集下载官方网址. <https://www.kaggle.com/wiseair/air-quality-in-milan-summer-2020>.
- 26 Delhi-meantemp 数据集下载官方网址. <https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data>.
- 27 Global-power 数据集下载官方网址. <https://www.kaggle.com/uciml/electric-power-consumption-data-set>.

(校对责编:牛欣悦)