Operations Research and Fuzziology 运筹与模糊学, 2024, 14(2), 1185-1197
Published Online April 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.142216

Hans汉斯

基于改进的Seq2Seq-LSTM模型的空气质量指数预测研究模型

仪　梦[1*]，吴丽丽[2#]

[1]甘肃农业大学理学院，甘肃　兰州
[2]甘肃农业大学信息科学技术学院，甘肃　兰州

收稿日期：2024年2月29日；录用日期：2024年3月19日；发布日期：2024年4月29日

摘　要

针对长短期记忆网络(Long Short-Term Memory, LSTM)模型输入输出时间步长度相等、处理长序列遗忘多、无法按重要程度分配权重等不足，构建了一种基于注意力机制(Attention Mechanism, Attention)改进的Seq2Seq-LSTM组合模型。该模型将序列到序列(Sequence to Sequence, Seq2Seq)模型中编码器、解码器设置为三层LSTM结构，并在解码器输出序列前引入注意力机制对模型进一步优化。为验证改进后的Seq2Seq-LSTM模型的有效性，本研究以山东省青岛市为研究区域，基于历史数据对未来1~7 h的空气质量指数进行模拟预测；预测结果与传统的机器学习模型支持向量机(Support Vector Machines, SVM)以及单一的LSTM模型预测结果进行了对比。结果表明：改进后的Seq2Seq-LSTM模型在中长期空气质量指数预测中的预测效果更突出。说明改进后的Seq2Seq-LSTM模型较单一模型具备更强的预测力，可作为山东省青岛市中长期空气质量指数预测模拟的可靠工具。

关键词

空气质量指数预测，长短期记忆网络，序列到序列模型，注意力机制

The Air Quality Index Prediction Research Model Based on Improved Seq2Seq-LSTM Model

Meng Yi[1*], Lili Wu[2#]

[1]College of Science, Gansu Agricultural University, Lanzhou Gansu
[2]College of Information Science and Technology, Gansu Agricultural University, Lanzhou Gansu

[*]第一作者。
[#]通讯作者。

文章引用: 仪梦, 吴丽丽. 基于改进的 Seq2Seq-LSTM 模型的空气质量指数预测研究模型[J]. 运筹与模糊学, 2024, 14(2): 1185-1197. DOI: 10.12677/orf.2024.142216

Received: Feb. 29th, 2024; accepted: Mar. 19th, 2024; published: Apr. 29th, 2024

Abstract

Since the reform and opening-up, our country is faced with the complex and changeable environmental pollution and governance problems at the same time of economic development, it is of practical value to develop an air quality index (AQI) prediction system to assist the Environmental monitoring work. The AQI prediction model based on the attention mechanism improved Seq2Seq-LSTM model in this paper aims to provide more reliable prediction results for workers, detect pollution ahead of time and reduce the cost of pollution control, then promote our country's ecological civilization construction, promote people's Life Happiness Index. In recent years, data-driven models, represented by deep learning (a classification of machine learning) algorithm, are widely used in AQI prediction due to their advantages of not considering complex parameters, simulating prediction by mining the latent law of data itself, and high simulation accuracy. The commonly used machine learning algorithm has achieved good results in air quality index prediction. Compared with machine learning model, deep learning model can learn the inherent laws and levels of sample data in a faster and more effective way, which greatly improves the prediction accuracy of the model. Long Short-Term Memory (LSTM) and Attention Mechanism (Attention) are commonly used models of deep learning. Among them, LSTM model is widely used in air quality index prediction research because of its simplicity, flexibility, stability and long-term memory Seq2Seq model is widely used in multivariable prediction task because it can handle the non-uniform sequence of input and output step. Attention mechanism shows its excellent performance in the field of natural language processing. Although the LSTM model has been proved to have good performance in air quality index prediction, it still has some defects, it is very important for AQI prediction research to break through the limitation of single model. In order to solve the problems of Long Short-Term Memory (LSTM) model, such as equal length of time step between input and output, Long sequence forgetting and weight distribution in importance, this paper constructs an improved Seq2Seq-LSTM combination model based on Attention Mechanism (Attention). In this model, the encoder and decoder in the Sequence to Sequence (Seq2Seq) model are set as three-layer LSTM structure, and the attention mechanism is introduced before the output Sequence of decoder to further optimize the model. In order to verify the validity of the improved Seq2Seq-LSTM model, this study takes Qingdao City of Shandong province as the research area, and based on the historical data, carries on the simulation forecast to the air quality index of the next 1~7 hours; the predicted results were compared with those of traditional Support Vector Machines (SVM) and single LSTM Support Vector machine. The results show that the improved Seq2Seq-LSTM model is more effective in the medium and long term air quality index prediction. The results show that the improved Seq2Seq-LSTM model is more powerful than the single model, and can be used as a reliable tool for the medium and long-term air quality index prediction simulation in Qingdao City, Shandong province.

Keywords

Air Quality Index Prediction, Long Short-Term Memory, Sequence to Sequence, Attention Mechanism

1. 引言

自改革开放以来，我国现阶段在经济发展的同时也面临着复杂多变的环境污染与治理问题，空气污

染也在慢慢变成我国经济高速健康发展的掣肘[1] [2]。为了处理我国的空气污染问题，习近平[3]总书记在党的十九大上指出，要加快生态文明体制改革，建设美丽中国。所以为了应对未来可能出现的空气污染状况，规划更优的治理方案，研究一套用于辅助环境监测治理相关工作的空气质量指数预测系统是具有一定的实际应用价值的。本文研究的基于注意力机制改进的 Seq2Seq-LSTM 模型的空气质量指数预测模型旨在为工作人员提供更加可靠的预测结果，提前发现污染，降低污染治理的成本，进而推动我国的生态文明建设，提升人民生活的幸福指数。

近年来，以深度学习(机器学习的一种分类)算法为代表的数据驱动模型凭其不需要考虑复杂的空气质量指数参数、通过挖掘数据本身的潜在规律进行模拟预测、模拟精度高等优点在空气质量指数预测中广泛应用[4]。常用的机器学习算法在空气质量指数预测中取得了较好的结果。深度学习模型相较于机器学习模型能够以更快速、更有效的方式学习样本数据的内在规律和层次，大大提高了模型预测精度。长短期记忆网络(Long Short-Term Memory, LSTM)、注意力机制(Attention Mechanism, Attention)等均是较常用的深度学习模型[5]。其中，LSTM 模型以简单、灵活、稳定、具备长时记忆能力等优点被广泛应用于空气质量指数预测研究中[6]；Seq2Seq 模型由于可以处理输入输出步长不统一的序列被广泛用于多变量预测任务中[7]；Attention 机制多在自然语言处理领域展现其优良的模型性能[8]。

尽管 LSTM 模型在空气质量指数预测中已被证实具有优良表现，但仍存在一定的缺陷，因此，将模型进行组合突破单一模型的局限性对于空气质量指数预测研究至关重要。

2. 模型原理

2.1. 序列到序列模型(Sequence to Sequence, Seq2Seq)

Seq2Seq 模型是一种将输入序列编码为中间向量再解码为输出序列的模型[9]。其输入输出序列长度自由，基本框架由编码器、解码器和中间向量三部分组成[10]。其中，编码器能够捕捉输入序列 x 的规律，并将 x 压缩成指定长度的中间向量 C，再由中间向量传递最后一个隐藏层的状态或所有隐藏层状态的变换利用解码器进行解码输出，使任意长度的输入序列映射到任意长度的输出序列上。

Seq2Seq 模型的编码器和解码器可根据任务使用不同的神经网络模型。

2.2. 长短期记忆网络(Long Short-Term Memory, LSTM)

长短期记忆网络 LSTM (Long Short-Term Memory)是 RNN 的一种变体，其核心概念在于细胞状态以及“门”结构[11]。理论上讲，细胞状态能够将序列处理过程中的相关信息一直传递下去。因此，即使是较早时间步长的信息也能携带到较后时间步长的细胞中来，这克服了短时记忆的影响。信息的添加和移除通过“门”结构来实现。

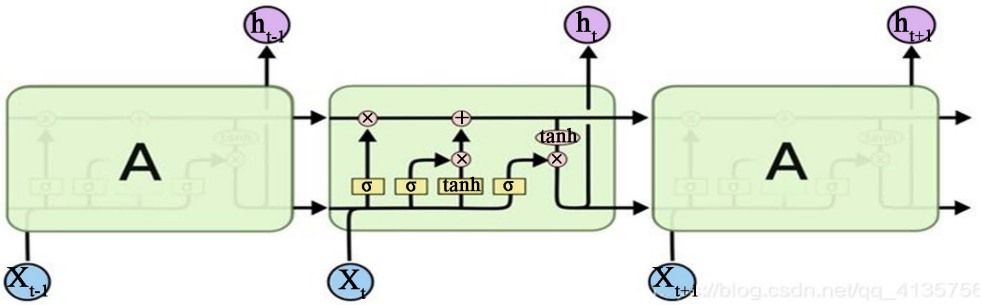

Figure 1. The structure of LSTM model
图 1. LSTM 结构图

如图 1 所示，门(Gate)是一种可选地让信息通过的方式。它由一个 Sigmoid 神经网络层和一个点乘法运算组成。sigmoid 神经网络层输出 0 和 1 之间的数字，这个数字描述每个组件有多少信息可以通过，0 表示不通过任何信息，1 表示全部通过。遗忘门决定应丢弃或保留哪些信息。输入门用于更新细胞状态。输出门用来确定下一个隐藏状态的值，隐藏状态包含了先前输入的信息。

2.3. 注意力机制(Attention Mechanism, Attention)

Attention 机制本质为产生一个输入特征的权重分布，再将此权重分布对映到原特征上，使任务重点放在主要特征上，忽略次要信息，进而提高任务效率。Bahdanau [12]率先将注意力机制应用到机器翻译领域并取得了较好的结果，说明注意力机制能够有效地应用于时间序列任务中。设输入序列向量为 $X=[x_1,x_2,...,x_t]$，则 Attention 机制的计算公式如下：

$$X'=soft\max\left(WX\right)*X \tag{1}$$

其中，W 为权重矩阵，与输入序列 X 做矩阵运算后经过 Softmax 激活函数，最后与输入序列相乘得到新序列 X' [13]。

2.4. 改进后的 Seq2Seq-LSTM 模型

本研究构建的改进后的 Seq2Seq-LSTM 组合模型的原理为通过算法之间的耦合使构建的模型在具备单一模型优势的同时克服单一模型的缺陷，使模型更加完善。LSTM 通过引入门控单元可以从隐藏的长期信息中学习而具备长期记忆功能[14]，并保持了训练过程中梯度下降的稳定性，但每一个输入都产生相应的隐藏状态，输入和输出需要相同的时间步长[15]。在本研究中，需要多变量作为输入进行中长期多步预测，Seq2Seq 模型允许在输入和输出时间步长不同时建立模型，并通过编码器解码器的信息传递减轻了模型的遗忘程度，因此本文将既具有长期记忆功能又允许输入输出步长不同的 Seq2Seq-LSTM 模型为第一步的组合模型[16]。但 Seq2Seq-LSTM 模型在传递过程中隐藏层的权重赋值相同，无法针对性地提取有效信息。针对此问题，将 Attention 机制与 Seq2Seq-LSTM 进行耦合可以在模型传递过程中利用评分函数计算各输入对预测值的影响程度，并为其赋予不同的权重，从而捕捉各隐藏层的有效信息，有利于模型精度的提升[17]。

改进后的 Seq2Seq-LSTM 模型的结构如图 2 所示。编码器经过 m 个时间步更新，最终时间步的隐藏状态为 h_m，解码器初始状态 h_t^* 以 h_m 作为输入值，加强记忆功能、减少遗忘状态；解码器每一时刻隐藏状态 h_t^* 通过当前时刻的输入 h_m 与上一时刻的隐藏状态 h_{t-1}^* 和细胞状态 C_{t-1}^* 更新，其表达式为：

$$h_t^*=f\left(h_n,h_{t-1}^*,C_{t-1}^*\right) \tag{2}$$

最终通过 Attention 机制在解码器结构产生新状态前先读取解码器中所有隐藏层的输出向量 $h=\left[h_1,h_2,...,h_m\right]$并对 h 分配不同比重，使网络能有针对性地捕捉对预测有效的特征信息并通过全连接层输出为最终的预测值序列。具有 m 个时间步的编码器 LSTM 的最终输出可以存储在一个状态向量的单元中，并与之前的隐藏状态一同用作具有 n 个时间步的解码器 LSTM 的输入。

3. 实验

3.1. 研究区概况

青岛为海滨丘陵城市，地势东高西低，南北两侧隆起，中间低凹，地处沿海。受海路风的影响，青岛的空气质量物可以输送的方式被清除，海风也会大大降低城市内的污染物浓度，所以相对于山东其他城市而言，空气质量还算是不错的。但是由于经济的发展，排放增多(新建工厂、居民生活、汽车尾气、建筑施工等等)，使得局地空气质量有所下降[18]。

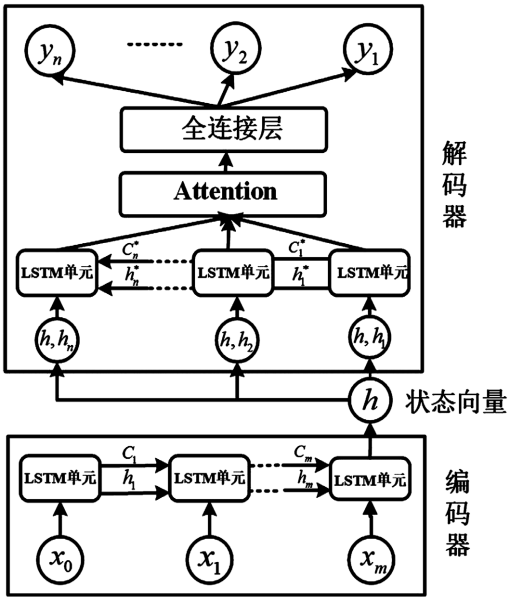

Figure 2. The structure of Improved Seq2Seq-LSTM model
图 2. 改进后的 Seq2Seq-LSTM 模型结构

以山东省为例，淄博、东营、聊城等等都属于重污染城市，必然在特定气象条件下对青岛的空气质量产生影响。此外，由于机动车的增加，氮氧化物排放增多，一种新型的污染类型逐渐开始出现——光化学污染。这种污染的直接效果是大气浑浊，臭氧超标，对人体危害大，青岛也会有类似情况。

最后，由于青岛湿度较大，高湿度更加利于颗粒物的凝结增长，产生的直接后果是——能见度恶化。另外，臭氧作为一种特殊的污染物，其在城市中的特点是白天生成，夜晚同氮氧化物反应消除。但是由于海陆风的存在，傍晚的陆风将城市的臭氧吹到海面上面“保存”起来，而第二天的海风又将污染物从海面吹回来，继续存留在空气中，导致污染的不断累积[19]。

本文从山东省青岛市下设的每个县市选取 1~2 个空气站，共计 11 个空气站。统计了每个站点 2021~2022 年每小时空气质量指数数据，共计 1 万 7 千多条数据作为本文的研究数据。各站点分布如图 3 所示。

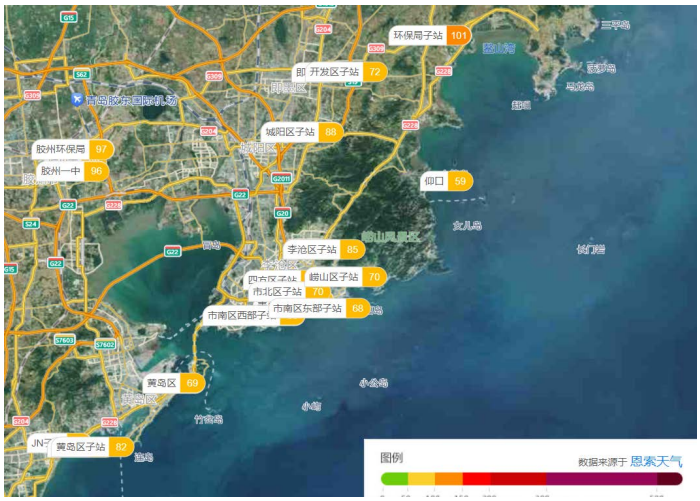

Figure 3. Distribution of air stations
图 3. 空气站点分布

由于空气质量指数变动幅度较大，为加快模型收敛速度，使预测结果具有一定的可信度，需要对各站点数据采取归一化处理。其公式为：

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3}$$

其中，X为原始数据；$X_{\min}$和$X_{\max}$分别是原始数据的最小值和最大值。

3.2. 模型构建

空气质量指数在时间序列数据预测中，一个关键问题是确定输入变量的最佳时滞数，但是目前还没有确定的方法或者确定的标准来确定时滞数[15]。本文将最佳时滞设为 6 [16]，模型输入具有当前时刻和前 5 个时间步数据序列，即以(t − 6)~(t − 1) h 的 11 个站点数据作为模型输入，输出为(t + 1) h、(t + 3) h、(t + 5) h、(t + 7) h 的空气质量指数预测值。

通过穷举法[17]以及控制变量法多次确定模型最优参数，此时模型的编码器和解码器均为三层 LSTM 结构，编码层隐藏层单元数量为 256、256、128，解码层隐藏层单元数量为 256、256、128。训练过程以均方误差为损失函数，学习率设为 0.00001，batch-size 为 64，最大训练轮数为 20，在此基础上使用 Adam 算法作为降低损失函数值的优化器，最后对模型输出进行反归一化处理，输出最终预测结果。

此外，为验证模型的有效性，采用 SVM、LSTM 模型输入与该模型相同特征序列与时滞，得出预测结果并进行对比研究。SVM 核函数参数 γ 为 0.1、LSTM 网络为三层，神经元个数分别为 256、256、128。具体参数如图 4 所示。

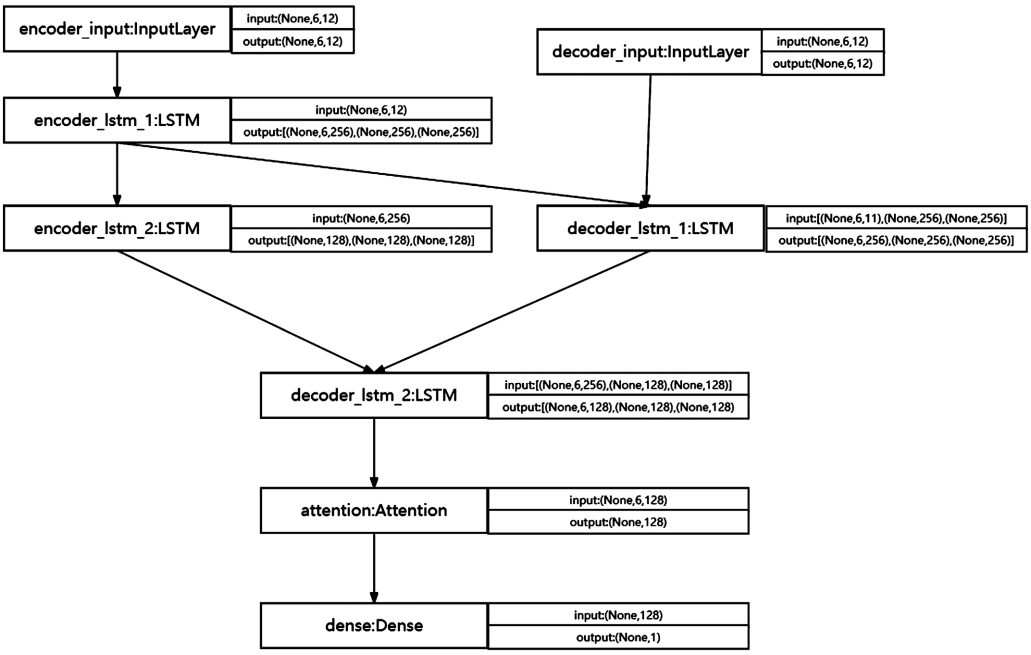

Figure 4. The parameter diagram of the improved Seq2Seq-LSTM model
图 4. 改进后的 Seq2Seq-LSTM 模型参数图

4. 结果与讨论

4.1. 改进后的 Seq2Seq-LSTM 模型结果分析

本文采用拟合优度 R 方、纳什效率系数(Nash-Sutcliffe Efficient, NSE)、均方根误差(Root Mean Square

Error, RMSE)定量评价各模型的模拟效果。R 方越接近 1，拟合效果越好[18]。NS 的取值范围为−∞~1，NS 值越接近 1，模型可信度越高。RMSE 可反映空气质量指数预测值和实测值之间的偏差，RMSE 越接近 0，实测值和预测值的误差越小[19]。改进后的 Seq2Seq-LSTM 模型在(t + 1)~(t + 7) h 的预测结果如表 1 所示。

Table 1. The air quality index prediction results of the improved Seq2Seq-LSTM model at (t + 1)~(t + 7) h
表 1. 改进后的 Seq2Seq-LSTM 模型在(t + 1)~(t + 7) h 的空气质量指数预测结果

时间	训练期			测试期		
	R 方	NS	RMSE	R 方	NS	RMSE
t + 1	0.884	0.884	6.643	0.908	0.996	8.52
t + 3	0.653	0.985	11.494	0.717	0.988	14.895
t + 5	0.52	0.979	13.519	0.684	0.986	15.754
t + 7	0.416	0.975	14.913	0.536	0.98	19.085

由表 1 可知，改进后的 Seq2Seq-LSTM 模型在训练期的模拟精度要优于测试期，但无论是训练期还是测试期模型的模拟精度都是可接受的。具体来看，改进后的 Seq2Seq-LSTM 模型在测试期(t + 1)~(t + 7) h 的 R 方值均大于 0.5，说明拟合效果较好。RMSE 值在(t + 1) h 为 8.52，在(t + 7) h 为 19.085，说明模型预测值和实测值之间误差不大。预测时间的间隔越大，改进后的 Seq2Seq-LSTM 模型的预测精度相应降低，但仍有较高的预测精度。如测试期该模型的 NS 值在(t + 1) h 为 0.996，在(t + 7) h 为 0.978，减少了 0.018。尽管改进后的 Seq2Seq-LSTM 模型随着预测时间变长精度有所降低，但是由于总体精度较高，在长时间空气质量指数预测有很大的发展空间。

由图 5 可知，改进后的 Seq2Seq-LSTM 模型在(t + 1)~(t + 7) h 的预测值与实测值变化趋势基本一致，在预测时间增加时预测精度随之降低，但模型在(t + 7) h 的模拟效果仍较好，表明该模型在中长期空气质量指数预报中展现良好的预测能力。

4.2. 模型预测精度对比分析

为验证改进后的 Seq2Seq-LSTM 模型的有效性，与相同输入的 SVM、LSTM 模型的预测结果进行对比研究。表 2、表 3 为 SVM、LSTM 模型在(t + 1)~(t + 7) h 的空气质量指数预测结果。对 SVM、LSTM、改进后的 Seq2Seq-LSTM 在更为重要的测试期的表现进行对比分析。由表 1~3 可知，测试期 SVM、LSTM 和改进后的 Seq2Seq-LSTM 在(t + 1) h 的 NS 值均在 0.99 以上，表明各模型在(t + 1) h 时的预测具有非常高的预测精度；在(t + 7) h 时 NS 值均达到了 0.97 以上，说明各模型在(t + 7) h 时预测精度依旧很高，可用于长期空气质量指数预报。

总体来看，改进后的 Seq2Seq-LSTM 模型的预测效果优于 SVM 和 LSTM。具体以三个模型在(t + 1) h、(t + 7) h 上的预测结果为例进行说明。

在(t + 1) h 时，改进后的 Seq2Seq-LSTM 的 R 方值为 0.908，比 SVM 的 R 方值高了 1%，RMSE 值为 8.52，低于 SVM 的 RMSE 值。但是改进后的 Seq2Seq-LSTM 的 R 方值低于 LSTM，RMSE 值高于 LSTM。通过分析可知，LSTM-Seq2seq-Attention 在(t + 7) h 上的各评价指标均优于 SVM 但不如 LSTM，说明该模型在短期空气质量指数预测效果比 SVM 模型好，但却不如 LSTM 模型。

在(t + 7) h 时，改进后的 Seq2Seq-LSTM 模型的 R 方值为 0.536，高于 LSTM 的 0.506 与 SVM 的 0.483；改进后的 Seq2Seq-LSTM 模型的 NS 值为 0.98，高于 LSTM 的 0.978 与 SVM 的 0.977；改进后的

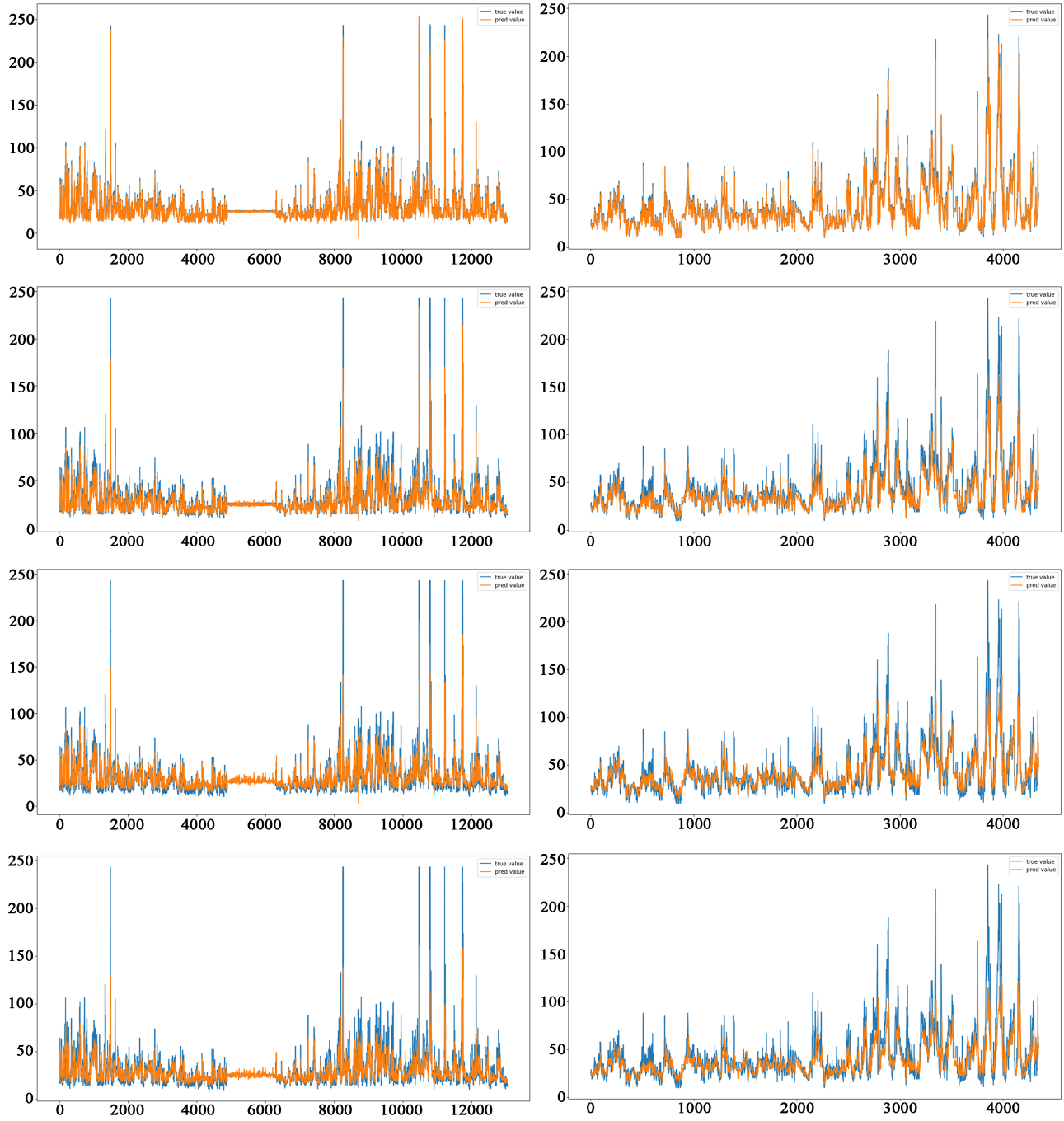

Figure 5. The predicted values of the improved Seq2Seq-LSTM at (t + 1)~(t + 7) h are compared with the measured values
图 5. 改进后的 Seq2Seq-LSTM 在(t + 1)~(t + 7) h 的预测值与实测值对比

Table 2. The air quality index prediction results of SVM in (t + 1)~(t + 7) h
表 2. SVM 在(t + 1)~(t + 7) h 的空气质量指数预测结果

时间	训练期			测试期		
	R 方	NS	RMSE	R 方	NS	RMSE
t + 1	0.917	0.996	5.623	0.898	0.996	8.962

续表

t + 3	0.757	0.989	9.61	0.714	0.987	14.986
t + 5	0.905	0.983	12.261	0.579	0.982	18.192
t + 7	0.5	0.978	13.801	0.483	0.977	20.15

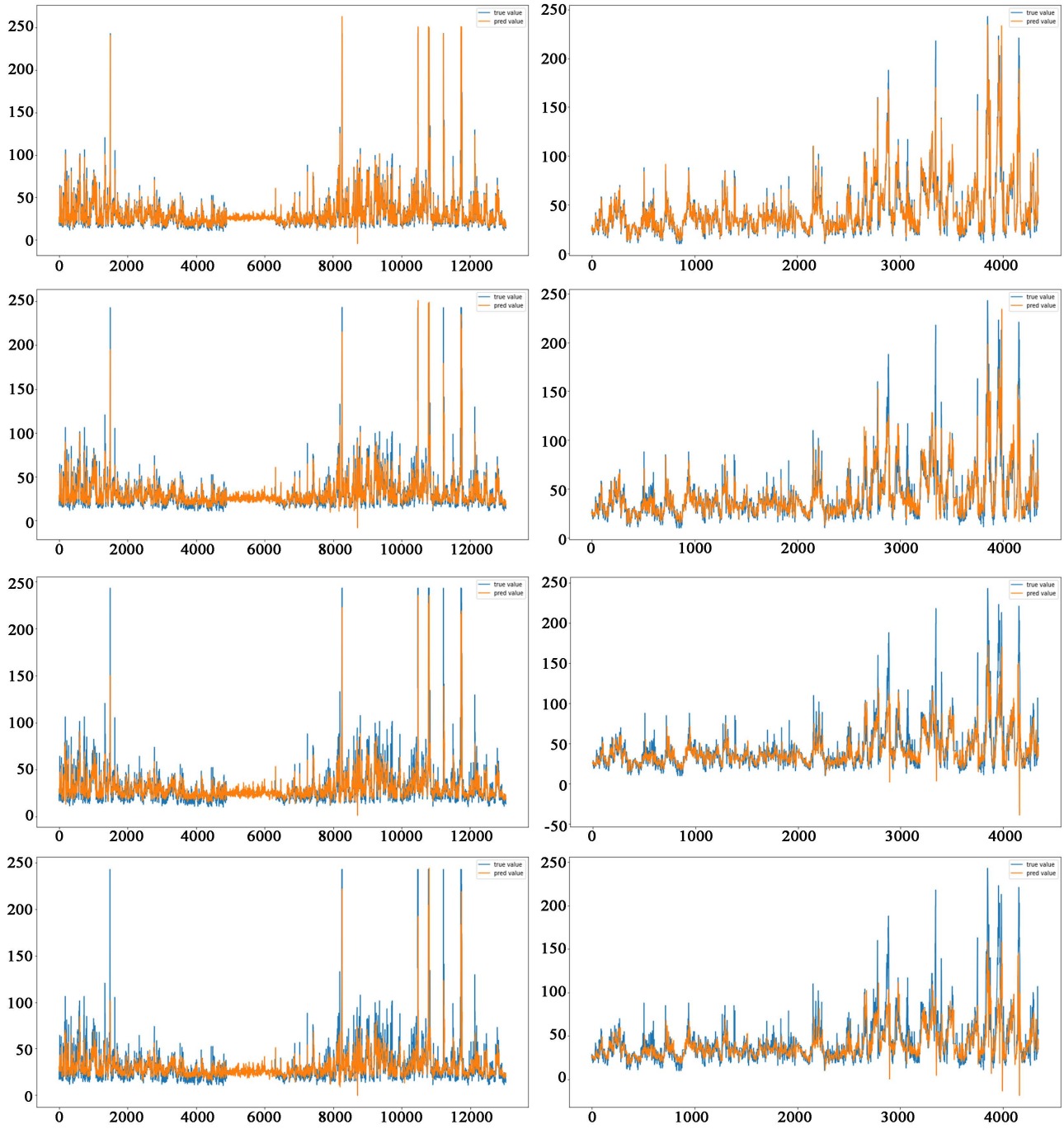

Figure 6. The comparison of predicted and measured values of SVM in (t + 1)~(t + 7) h
图 6. SVM 在(t + 1)~(t + 7) h 的预测值与实测值对比

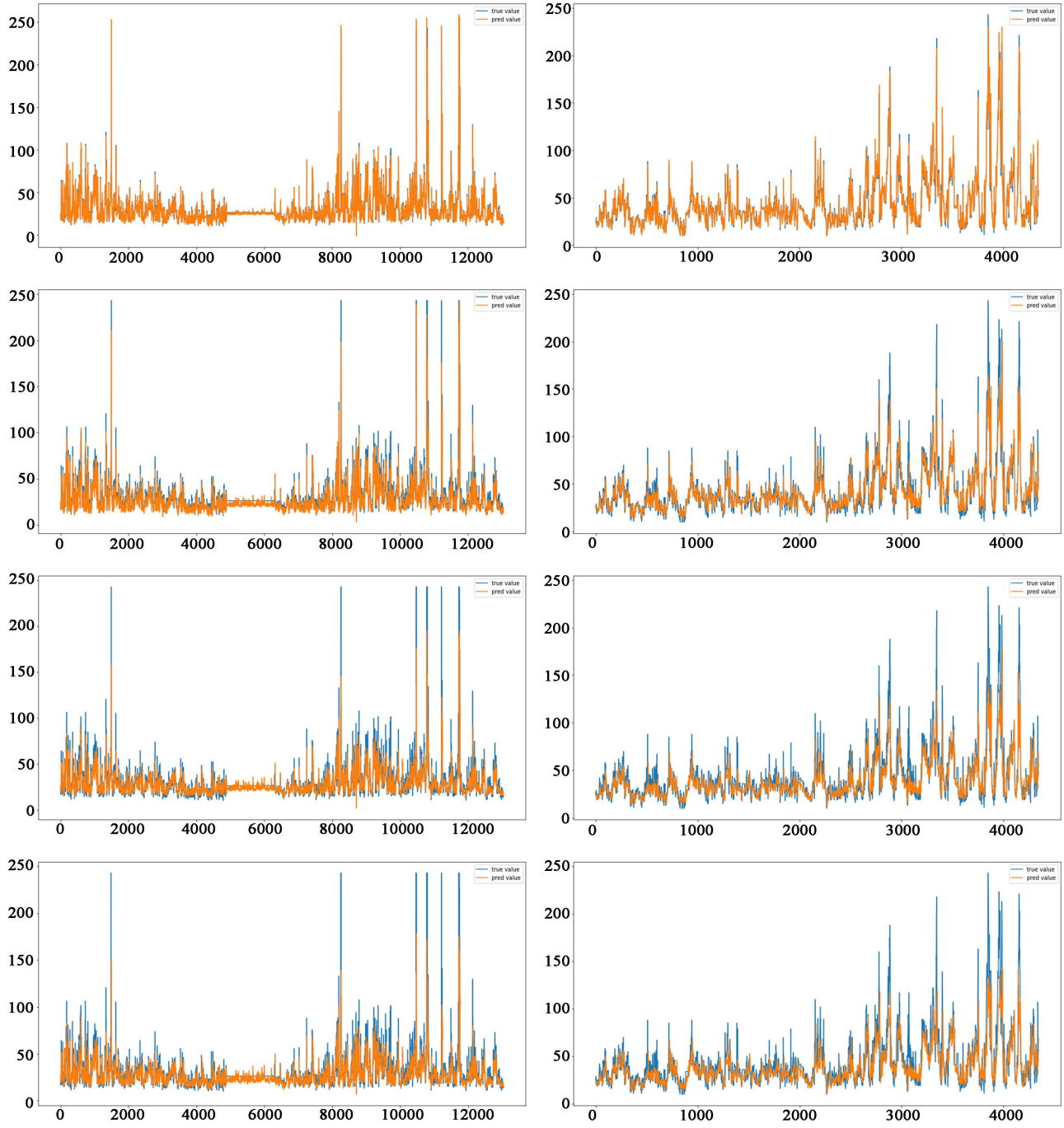

Figure 7. The comparison of predicted and measured values of LSTM at (t + 1)~(t + 7) h
图 7. LSTM 在(t + 1)~(t + 7) h 的预测值与实测值对比

Table 3. The air quality index prediction results of LSTM at (t + 1)~(t + 7) h
表 3. LSTM 在(t + 1)~(t + 7) h 的空气质量指数预测结果

时间	训练期			测试期		
	R 方	NS	RMSE	R 方	NS	RMSE
t + 1	0.895	0.995	6.316	0.916	0.996	8.155

续表

t + 3	0.684	0.986	10.96	0.755	0.989	13.865
t + 5	0.514	0.979	13.598	0.623	0.983	17.202
t + 7	0.409	0.974	15.003	0.506	0.978	19.69

Seq2Seq-LSTM 模型的 RMSE 值为 19.085，低于 LSTM 的 19.69 与 SVM 的 20.15。因此，经过进一步对比分析可知，LSTM-Seq2seqAttention 在(t + 7) h 上的各评价指标均优于 SVM 与 LSTM，说明该模型在中长期空气质量指数预测的效果优于 SVM、LSTM 模型。

图 6、图 7 为 SVM、LSTM 模型在(t + 1)~(t + 7) h 上的预测精度拟合曲线。由图 5~7 可知，SVM、LSTM、改进后的 Seq2Seq-LSTM 模型在测试期(t + 1)~(t + 7) h 上拟合效果较好，但预测时间的间隔越大，预测值与实测值的误差逐渐变大，且在空气质量指数高值处的预测存在一定误差，预测时间天数越长，峰值处的拟合效果越差。

同时，为更直观反映改进后的 Seq2Seq-LSTM 模型与单一模型 SVM、LSTM 结果的预测效果，按(t + 1)~(t + 7) h 的时间顺序观察各模型的预测情况。如表 4 所示。

Table 4. The air quality index prediction results of SVM, LSTM and modified Seq2Seq-LSTM at (t + 1)~(t + 7) h during test period
表 4. 测试期 SVM、LSTM 和改进后的 Seq2Seq-LSTM 在(t + 1)~(t + 7) h 的空气质量指数预测

时间	模型	R 方	RMSE	NS
(t + 1) h	SVM	0.898	8.962	0.996
	LSTM	0.916	8.155	0.996
	改进后的 Seq2Seq-LSTM	0.908	8.52	0.996
(t + 3) h	SVM	0.714	14.986	0.987
	LSTM	0.755	13.865	0.989
	改进后的 Seq2Seq-LSTM	0.717	14.895	0.988
(t + 5) h	SVM	0.579	18.192	0.982
	LSTM	0.623	17.202	0.983
	改进后的 Seq2Seq-LSTM	0.684	15.574	0.986
(t + 7) h	SVM	0.483	20.15	0.977
	LSTM	0.506	19.69	0.978
	改进后的 Seq2Seq-LSTM	0.536	19.085	0.98

可以看出，改进后的 Seq2Seq-LSTM 模型的虽然在(t + 1) h 和(t + 3) h 的优势不明显，但在(t + 5) h 和(t + 7) h 的各项指标均优于 SVM、LSTM 模型，进一步说明了改进后的 Seq2Seq-LSTM 模型在长期空气质量指数预测时，不仅预测精度最高且产生的误差最小。

进一步对比分析发现，虽然 SVM、LSTM、改进后的 Seq2Seq-LSTM 模型的预测值与实测值的拟合程度随着模拟时间的增加而变差，但改进后的 Seq2Seq-LSTM 模型在(t + 7) h 时仍达到了可接受的预测精度。说明改进后的 Seq2Seq-LSTM 模型在(t + 1)~(t + 7) h 的空气质量指数预测均具有较高的预测精度。

综上所述，各方面综合对比分析了测试期 SVM、LSTM 和改进后的 Seq2Seq-LSTM 模型的预测结果，

得出结论为：改进后的 Seq2Seq-LSTM 模型中长期空气质量指数预测的各方面预测效果均优于 SVM 与 LSTM，而 SVM 在各评价指标上预测效果最差。说明深度学习组合模型相对单一模型以及机器学习模型具有更高的预测性能。

5. 讨论

本研究将单一深度学习模型LSTM、Seq2seq、Attention机制进行耦合，构建了改进后的Seq2Seq-LSTM组合模型对青岛市空气质量指数进行模拟预测，并与机器学习模型 SVM、单一模型 LSTM 在多方面结果进行对比。得出结果为：本研究构建的改进后的 Seq2Seq-LSTM 模型在对中长期空气质量指数预测时精度最高误差最小。这是因为：其一，LSTM 在对长时序列进行预测时，其细胞状态与隐藏状态能够保存并传递空气质量变化过程中的重要信息，并通过门结构选择性遗忘次要信息，能够利用有限资料对变化过程进行预测模拟；其二，Seq2seq 模型输入输出序列长度自由，同时将编码器解码器为三层 LSTM 结构，使编码器中每一层 LSTM 的信息既可以传递到下一层又作为解码器的输入，增加了模型参数，减缓了网络遗忘速度；其三，解码过程中引入 Attention 机制能够在信息解码时按隐藏状态的重要程度对权重进行分配。因此将三个算法进行组合能够打破 LSTM 输入输出等长的限制并解决 Seq2Seq 模型信息传递时权重相等的问题，使组合模型取得更精准的预测结果。

尽管改进后的 Seq2Seq-LSTM 模型在中长期空气质量指数预测的效果较单一的 SVM、LSTM 模型更好，但改进后的 Seq2Seq-LSTM 模型仍存在一些不足。首先，模型的预测精度随预测时间增加而降低且在峰值处的预测值低于实测值。这可能是由于增加了滞后输入，使得输入数据变量之间的相关性减弱，导致提取变量之间的非线性关系变得更为困难。也可能是输入变量不足，本文在 24 个站点中仅选取了 11 个站点进行输入，虽能捕捉到空气质量指数变化的一般趋势，但难以准确地反映空气质量指数变化的具体特征[20]。同时，本研究输入数据量不足，深度学习模型的构建往往需要大量数据作为输入，而本文只选取 2021~2022 年共一万七千多个数据作为输入，数据量过少可能导致深度学习算法难以发挥其优势，从而使得预测精度受到影响[21]。最后，本文基于三种单一深度学习进行组合建模，由于模型复杂度提高参数也随之增加，由于组合模型为自主构建无法凭经验在短时间内择出最优参数。针对上述问题，在日后的研究工作中可以尝试从增加数据量、优化模型参数、提高模型计算力以及设置自主调参方法等方面提高模型预测精度，从而更准确地反映空气质量指数变化规律。

本研究针对机器学习组合模型不具备记忆功能、无法自动筛选重要信息等问题，提出了一种基于 Attention 机制改进后的 Seq2Seq-LSTM 深度学习组合模型，对青岛市(t + 1)~(t + 7) h 的空气质量指数进行了预测，并与单一的 SVM 与 LSTM 模型进行了对比分析，得到以下结论：

SVM、LSTM 和改进后的 Seq2Seq-LSTM 均可作为短期空气质量指数预测的有效工具，但改进后的 Seq2Seq-LSTM 模型由于综合考虑了 LSTM、Seq2seq、Attention 三个模型的优势，在(t + 5) h、(t + 7) h 的空气质量指数的预测精度较 SVM 和 LSTM 更高，因此改进后的 Seq2Seq-LSTM 在中长期空气质量指数预测中取得更好的预测效果。由于改进后的 Seq2Seq-LSTM 模型在具备长期记忆功能的基础上允许输入输出序列长度不受约束，并能在信息传递过程中按重要程度提取信息，从而使空气质量指数预测值与实测值的拟合效果更好，能更加真实准确地反映流域空气质量指数的变化规律，可在资料缺乏时为青岛市空气质量指数提供综合可靠预报结果，是中长期空气质量指数预测的有效工具。

参考文献

[1] 李勇, 白云, 李川. 大气污染物 SO2 预测模型研究综述[J]. 四川环境, 2016, 35(1): 144-148.

[2] 张炳彩. 基于残差修正 GM(1,1)模型的银川市空气质量指数预测分析[J]. 绿色科技, 2019(12): 118-122.

[3] 张玉丽, 何玉, 朱家明. 基于多元线性回归模型 PM2.5 预测问题的研究[J]. 安徽科技学院学报, 2016, 30(3): 92-97.

[4] 张勤, 郭进利. 基于机器学习的上海市空气质量指数预测方法研究[J]. 软件导刊, 2022, 21(8): 33-38.

[5] 冀东, 刘祖涵, 王莉莉, 等. 基于粒子群算法和注意力机制的 LSTM 的 PM2.5 预测研究[J/OL]. 西华师范大学学报(自然科学版): 1-10. https://kns.cnki.net/kcms2/article/abstract?v=5UWSsHjGZiGRmk4uVBMEvFKSqQZO3pRnYTqL7z8oNiLmzthVzegTEXVAHpHqNAmSpu-VeW68G22W9odFaMENI-cF1W4UmnpXWbPWYzUf480FWHVHSfQIzDmnDqBUb92J0DfveYYEqcr0=&uniplatform=NZKPT&language=CHS, 2023-12-26.

[6] 赵煜, 韩旭昊. 基于 CEEMDAN-LSTM 组合的兰州空气质量指数预测[J]. 安徽师范大学学报(自然科学版), 2023, 46(5): 433-439, 447.

[7] 黄光球, 王紫薇, 陆秋琴. 融合频域信息和 Seq2Seq 模型的 VOCs 浓度预测研究[J/OL]. 经营与管理: 1-16. https://kns.cnki.net/kcms2/article/abstract?v=5UWSsHjGZiE_ouIjc8LpIFQlcolUPIWjFxDlIyrE1m76RL-TIgaeGLetYe7sLYeuXj1D4SeQ5fMrlZPVnZ79GttHaxPYbFPCuO51uw6uvIp6yP6w_yii5xn3BYuCO5sR9Z-BkTCzIl4=&uniplatform=NZKPT&language=CHS, 2023-12-21.

[8] 黄强. 自然语言描述下的目标检测算法研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2023.

[9] 史超. 基于机器学习的空气质量预测方法研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2021.

[10] 刘艳, 张婷, 康爱卿, 等. Seq2Seq 模型的短期水位预测[J]. 水利水电科技进展, 2022, 42(3): 57-63.

[11] 朱菊香, 任明煜, 谷卫, 等. 基于 CEEMDAN-IGWO-CNN-LSTM 空气质量指数预测建模[J/OL]. 计算机仿真: 1-11. https://kns.cnki.net/kcms2/article/abstract?v=5UWSsHjGZiHMbwdNAivrckgA-xn3MBmGTBIps8GA7g4Z-JQAee7z2Ez4ki-eLmr-lfWKQt2-qG4mMgt-PRYB1p1ZngUIdVfd-NHgBETSavr-G7niqip_n9J-UaSMafbZJAKnvrZ1hc4=&uniplatform=NZKPT&language=CHS, 2023-12-25.

[12] Bahdanau, D., Cho, K. and Bengio, Y. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: 1409.0473.

[13] 马思远, 焦佳辉, 任晟岐, 等. 基于注意力机制的城市多元空气质量数据缺失值填充[J]. 计算机工程与科学, 2023, 45(8): 1354-1364.

[14] Shoaib, M., Shamseldin, A.Y., Melville, B.W., *et al*. (2016) A Comparison between Wavelet Based Static and Dynamic Neural Network Approaches for Runoff Prediction. *Journal of Hydrology*, **535**, 211-225. https://doi.org/10.1016/j.jhydrol.2016.01.076

[15] Xiang, Z., Yan, J. and Demir, I. (2020) A Rainfall-Runoff Model with LSTM-Based Sequence-to-Sequence Learning. *Water Resources Research*, **56**, e2019WR025326. https://doi.org/10.1029/2019WR025326

[16] 周恩. 基于 Seq2seq 模型的空气质量指数预测系统设计与实现[D]: [硕士学位论文]. 郑州: 郑州大学, 2022.

[17] 王克玲. 基于人工神经网络的城市空气质量指数评价与预测[D]: [硕士学位论文]. 乌鲁木齐: 新疆大学, 2022.

[18] 郑秀苹, 姜颂月, 蒋楠, 等. 青岛空气质量变化特征及影响因素分析[J]. 河北环境工程学院学报, 2023, 33(6): 69-76.

[19] 贾智海, 黄燕, 李新, 等. 2022 年青岛市沿海区域臭氧污染特征及传输影响分析[J]. 环境监控与预警, 2023, 15(5): 121-127, 133.

[20] 周婷, 温小虎, 冯起, 尹振良, 杨林山. 基于 BMA 多模型组合的疏勒河径流预测研究[J]. 冰川冻土, 2022, 44(5): 1606-1619.

[21] Wen, X.H., Feng, Q., Ravinesh, C.D., *et al*. (2019) Two-Phase Extreme Learning Machines Integrated with the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise Algorithm for Multi-Scale Runoff Prediction Problems. *Journal of Hydrology*, **570**, 167-184. https://doi.org/10.1016/j.jhydrol.2018.12.060