Advances in Applied Mathematics 应用数学进展, 2021, 10(12), 4304-4311
Published Online December 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1012457

Hans汉斯

基于LSTM神经网络的大连市空气质量指数预测

陈虹宇，孙德山

辽宁师范大学数学学院，辽宁 大连

收稿日期：2021年11月16日；录用日期：2021年12月13日；发布日期：2021年12月20日

摘 要

由于大气环境的复杂性和多变性，用传统方法预测空气质量指数精度较低，本文首先对空气质量指数和主要污染物进行相关性分析，再采用基于TensorFlow的长短期记忆神经网络(LSTM)模型，对大连市的空气质量指数进行预测，并进行误差分析。实验结果表明，与支持向量机模型和BP神经网络模型相比，LSTM神经网络模型预测空气质量指数的精度较高。

关键词

空气质量指数，TensorFlow，LSTM神经网络，相关性分析

Prediction of Air Quality Index in Dalian Based on LSTM Neural Network

Hongyu Chen, Deshan Sun

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: Nov. 16th, 2021; accepted: Dec. 13th, 2021; published: Dec. 20th, 2021

Abstract

Due to the complexity and dynamics of atmospheric environment, the accuracy of air quality index prediction by traditional method is low. Firstly, this paper analyzes the correlation between air quality index and main pollutants. Then, the long and short-term memory neural network model based on Tensor Flow is used to predict the air quality index of Dalian and the error analysis id made. The experimental results show that LSTM neural network model is more accurate than support vector machine model and BP neural network model in predicting air quality index.

文章引用: 陈虹宇, 孙德山. 基于 LSTM 神经网络的大连市空气质量指数预测[J]. 应用数学进展, 2021, 10(12): 4304-4311. DOI: 10.12677/aam.2021.1012457

Keywords

Air Quality Index, TensorFlow, LSTM Neural Network, Correlation Analysis

1. 引言

国家实施振兴东北老工业基地政策已久，辽宁沿海经济带蓬勃发展，大连市经济社会持续快速发展，主要污染排放量迅速增加，城市机动车尾气污染问题更加突出，大连市生态环境保护面临巨大挑战。因此，精准预测空气质量，加强空气质量监测，对城市环境建设、引导人们生产生活、改善空气污染有重要的意义。

空气质量指数(AQI)是评估空气质量的重要指数，传统的 AQI 预测模型主要采用计量经济学方法和机器学习。牟敬锋等人[1]使用 ARIMA 模型，对深圳市空气质量指数进行预测，取得了良好的预测效果；常恬君等人[2]构建了 Prophet-随机森林模型，解释性强且精度高，有效预测了上海市空气质量指数；徐乔王等人[3]建立了 MAE-SVM 模型，显著提高了预测速度；尹琪等人[4]将遗传算法与支持向量机结合，构建 GA-SVM 模型，预测效果优于单个模型。随着神经网络的兴起，王云中[5]使用 LSTM 网络预测西安市 $PM_{2.5}$ 浓度；张超利[6]建立了基于改进粒子群优化算法的神经网络，预测河南省 17 市的空气污染物浓度；郑洋洋等人[7]运用基于 Kera 的 LSTM 模型，对太原市的空气质量指数进行预测，提高了预测的精度；陈岑等人[8]提出一种基于信息增益和 LSTM 神经网络的空气质量指数混合预测方法，利用信息增益减少输入变量的数量，该 IG-LSTM 模型具有更低的预测误差和损失值。

与计量经济学、机器学习相比，深度学习采用无监督学习进行逐层特征提取，具有更强大的特征表达能力，也能提高数据预测精度，缓解过拟合问题，具有更强大的泛化能力[9]。作为深度学习中最有效的模型之一，LSTM 神经网络能有效地解决梯度消失和梯度爆炸的问题，在手写体识别等方面发挥了巨大的作用，因此，本文采用基于 TensorFlow 的 LSTM 神经网络模型来预测大连市空气质量指数。

2. 理论介绍

2.1. 空气质量指数

空气污染指数(AQI)是将空气中多种不同污染物的具体浓度通过某种计算方法转化为单一的数值，用来衡量空气质量状况。参与空气质量指数计算的六种污染物分别为 $PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3 和 CO。AQI 的计算公式：

$$\mathrm{IAQI}_P = \frac{\mathrm{IAQI}_H - \mathrm{IAQI}_L}{C_{PH} - C_{PL}}\left(C_P - C_{PL}\right) + \mathrm{IAQI}_L$$

$$\mathrm{AQI} = \max\left\{\mathrm{IAOI}_1, \mathrm{IAQI}_2, \mathrm{IAQI}_3, \cdots, \mathrm{IAQI}_6\right\}$$

其中，IAQI_P 为污染物 P 的空气质量分指数，C_P 为测量的污染物浓度，C_{PH} 为表 1 中与 C_P 接近的该污染物浓度高位值，C_{PL} 为表 1 中与 C_P 接近的该污染物浓度低位值，IAQI_H 为 C_{PH} 对应的 IAQI_P，IAQI_L 为 C_{PL} 对应的 IAQI_P，$P = 1$，2，3，4，5，6，对应六种污染物。

Table 1. System concentration limits for air pollutants
表 1. 空气污染物浓度限值

IAQI	$PM_{2.5}$ 浓度 (24 h 平均值) ug/m^3	PM_{10} 浓度 (24 h 平均值) ug/m^3	NO_2 浓度 (24 h 平均值) ug/m^3	SO_2 浓度 (24 h 平均值) ug/m^3	O_3 浓度 (1 h 平均值) ug/m^3	CO 浓度 (1 h 平均值) mg/m^3
0	0	0	0	0	0	0
50	35	50	40	50	160	5
100	75	150	80	150	200	10
150	115	250	180	475	300	35
200	150	350	280	800	400	60
300	250	420	565	1600	800	90
400	350	500	750	2100	1000	120
500	500	600	940	2620	1200	150

2.2. LSTM 神经网络基本原理

长短期记忆神经网络(LSTM)是循环神经网络(RNN)的一种特殊形式，它们都具有递归的特性，也可以说，LSTM 神经网络是 RNN 的改进，它独特的具有记忆和遗忘模式[7]。LSTM 的核心概念在于单元状态(cell state)和“门”结构，有三种类型的“门”结构：遗忘门(forget gate)、输入门(input gate)和输出门(output gate) [10]。LSTM 神经网络的内部展开结构如图 1 所示。

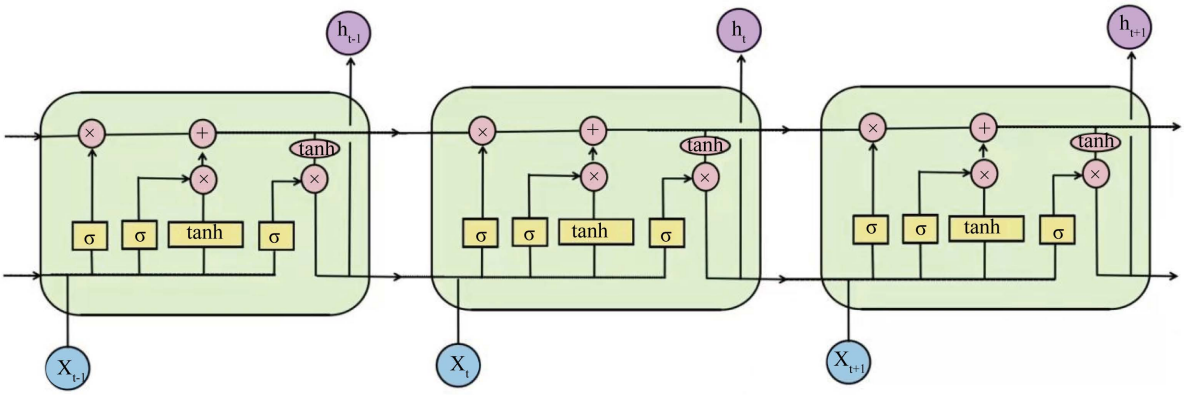

Figure 1. The internal structure of LSTM neural network
图 1. LSTM 神经网络内部展开结构

遗忘门决定应该丢弃或保留哪些信息，来自上一个隐藏状态的信息 h_{t-1} 和当前输入的信息 x_t 同时传递到 sigmoid 函数中，输出值 f_t 介于 0 和 1 之间，越接近 0 代表越该被丢弃，越接近 1 代表越该被保留。f_t 的计算公式：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)$$

其中，W_f、b_f 分别是遗忘门的权重矩阵、偏置向量。

输入门用于更新细胞状态，将上一个隐藏状态的信息 h_{t-1} 和当前输入的信息 x_t 传递到 sigmoid 函数中来判断要更新哪些信息，输出值 i_t 介于 0 和 1 之间，越接近 0 代表越不重要，越接近 1 代表越重要。还

要将 h_{t-1} 和 x_t 传递到 tanh 函数中，创造一个新的向量 $\widetilde{C}_t$。i_t 和 $\widetilde{C}_t$ 的计算公式：

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right)$$

其中，W_i、W_c 分别是输入门、单元状态的权重矩阵，b_i、b_c 分别是输入门、单元状态的偏置向量。

将遗忘门的输出和输入门的输出进行线性运算，更新 t 时刻的细胞状态 C_t：

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t$$

输出门用来确定下个隐藏状态的值，将上一个隐藏状态的信息 h_{t-1} 和当前输入的信息 x_t 同时传递到 sigmoid 函数中，得到输出值 o_t，再将细胞状态 C_t 传递给 tanh 函数，使其与相乘 o_t，最终确定隐藏状态应该携带的信息 h_t。o_t 和 h_t 的计算公式：

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

其中，W_o、b_o 是输出门的权重矩阵、偏置向量。

3. 模型建立

3.1. 数据来源

本文使用的数据来自真气网(https://www.aqistudy.cn/historydata/)发布的大连市空气质量历史数据，采样时间是 2018 年 1 月 1 日至 2021 年 9 月 30 日，采样频率为一天一次，由于仪器故障等原因使得 2 天的数据缺失，缺失值在整个数据集中占的比例非常小，可以直接删除，删除缺失值后，共计 1367 组数据。数据信息包含大连市每日的 AQI、$PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3 和 CO 的浓度，这里列举前 6 天数据，如表 2 所示。

Table 2. Air quality index and concentration of main pollutants in Dalian (part)
表 2. 大连市空气质量指数和主要污染物浓度(部分)

日期	AQI	$PM_{2.5}$	PM_{10}	NO_2	SO_2	O_3	CO
2018 年 1 月 1 日	62	40	74	43	26	37	1.1
2018 年 1 月 2 日	41	22	41	24	21	50	0.7
2018 年 1 月 3 日	38	16	38	19	16	62	0.6
2018 年 1 月 4 日	39	18	39	19	18	54	0.6
2018 年 1 月 5 日	40	22	40	26	20	57	0.7
2018 年 1 月 6 日	41	23	41	28	22	57	0.8

3.2. AQI 和污染物间的相关性分析

本文使用 python 中的相关性分析函数对空气质量的相关数据进行相关性分析，并画出相关性分析热力图，如图 2 所示。可以直观的看出，大连市的 AQI 与 $PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3、CO 均呈正相关，AQI 与 $PM_{2.5}$ 的相关性高达 80%，AQI 与 PM_{10} 的相关性高达 77%。因此，为了有效改善大连市的空气质量，相关部门可以从治理 $PM_{2.5}$ 和 PM_{10} 入手，加大管控和治理的力度，降低空气中 $PM_{2.5}$ 和 PM_{10} 的浓度，从而改善空气质量。

从相关性分析热力图也可以得出，AQI 与 6 个污染物浓度之间有一定的关联，因此，本文将前面时刻的 6 个污染物浓度和 AQI 作为模型的输入特征，下一时刻的 AQI 作为模型的输出。

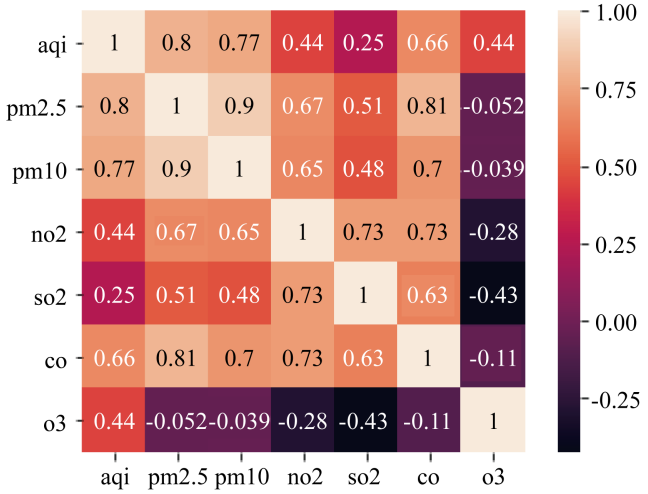

Figure 2. Correlation analysis thermal diagram
图 2. 相关性分析热力图

3.3. 实验设计及结果分析

3.3.1. 数据处理

本文选取 2018 年 1 月 1 日至 2021 年 6 月 30 日的数据为训练集，共 1276 组数据，选取 2021 年 7 月 1 日至 2021 年 9 月 30 日的数据作为测试集，共 91 组数据。由于数据的量纲不同，避免因输入数据的数量级差别较大，而造成预测的误差较大，因此，本文采用线性函数归一化(Min-Max scaling)的方法，将原始数据转化到[0, 1]的范围，归一化公式：

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

其中，X_{norm} 是归一化后的数据，X 为原始数据，$X_{\min}$、$X_{\max}$ 分别为原始数据集的最小值、最大值。

3.3.2. LSTM 神经网络的构建

本文实验采用 python3.7 编程实现。构建的 LSTM 网络结构如图 3 所示，由两个 LSTM 层，Dropout 层和 Dense 层构成，第一个 LSTM 层包括个 40 节点，第二个 LSTM 层包括个 50 节点，在每个 LSTM 层后加入 Dropout 层。批处理参数 batch_size 设置为 32，时间步长 time_step 设置为 10，迭代次数 epoch 为 100，采用 Adam 算法进行优化，学习率设为 0.01。

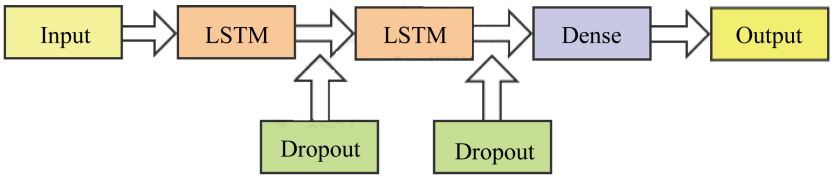

Figure 3. Structure of LSTM neural network
图 3. LSTM 神经网络结构

增加 LSTM 层可以降低误差，提高精度，但是过多的 LSTM 层也会使网络更加复杂，模型难以收敛，甚至出现过拟合的现象。Dropout 是指在深度神经网络的训练过程中，按照一定的概率将神经网络单元暂时丢弃，可以有效的防止过拟合的问题。Dense 层能够对高维信息进行降维处理，同时保留有用的信息。

4. AQI 预测及结果分析

4.1. AQI 预测

根据搭建的 LSTM 模型对测试集进行预测，再将得到的预测值进行反归一化处理，从而得到真实值和预测值的拟合曲线，其中，红色曲线代表真实值，蓝色曲线代表预测值，如图 4 所示。

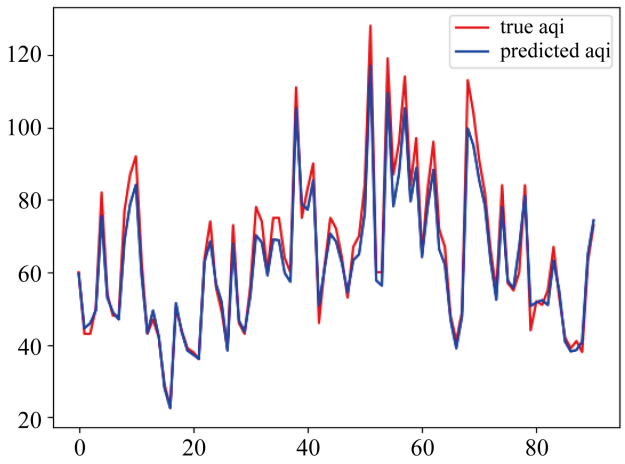

Figure 4. The fitting curve of real value and predicted value based on LSTM neural network model
图 4. 基于 LSTM 神经网络模型的真实值和预测值拟合曲线

为了进行比较，本文还采用 SVM 模型、BP 神经网络模型进行实验，得到的拟合结果如图 5、图 6 所示。

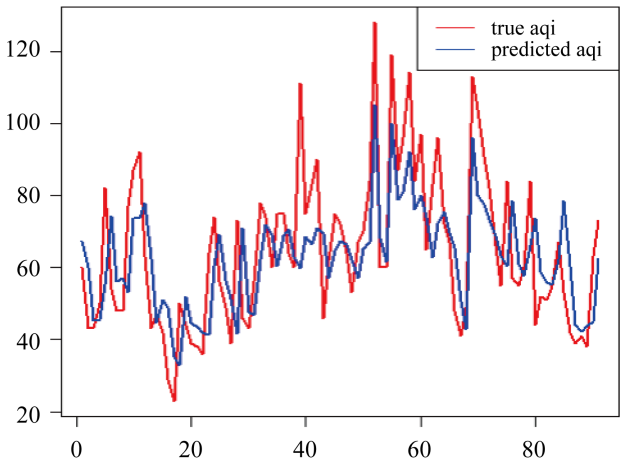

Figure 5. The fitting curve of real value and predicted value based on SVM model
图 5. 基于 SVM 模型的真实值和预测值拟合曲线

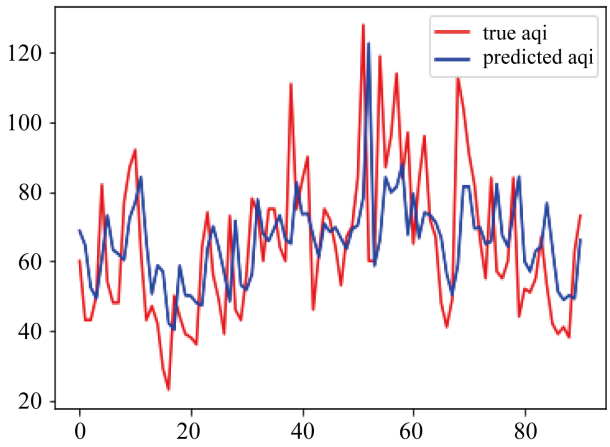

Figure 6. The fitting curve of real value and predicted value based on BP neural network model
图 6. 基于 BP 神经网络模型的真实值和预测值拟合曲线

4.2. 评价指标

建立模型拟合效果的评价指标来对比所建立的模型的优劣，包括均方根误差(RMSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)。

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\widehat{y_i}\right)^2}$$

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|\widehat{y_i}-y_i\right|$$

$$\mathrm{MAPE}=\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\widehat{y_i}-y_i}{y_i}\right|\times 100\%$$

其中，m 表示预测值的个数，y_i 表示真实值，$\widehat{y_i}$ 表示真实值。

LSTM 神经网络模型、SVM 模型和 BP 神经网络模型的评价指标如表 3 所示。可以看出，SVM 模型和 BP 神经网络模型对 AQI 的预测效果相似，SVM 模型稍优于 BP 神经网络模型，但 LSTM 模型的各项评价指标都明显的优于 SVM 模型和 BP 神经网络模型。

Table 3. Evaluation indexes of each model
表 3. 各模型的评价指标

	LSTM 神经网络模型	SVM 模型	BP 神经网络模型
RMSE	5.62	15.43	16.29
MAE	4.50	12.52	15.01
MAPE	8.76%	20.19%	23.91%

5. 结束语

空气质量是人们普遍关心的问题，为了提高空气质量指数的预测精度，本文采用基于 TensorFlow 的 LSTM 神经网络对大连市的空气质量指数进行预测，LSTM 神经网络可以有效解决梯度消失或梯度爆炸

问题，也可以解决传统循环神经网络长时依赖问题。同时，采用 SVM 模型、BP 神经网络模型与之作对比，结果表明，SVM 模型和 BP 神经网络对空气质量指数的预测效果都不理想，而 LSTM 模型的均方根误差、平均绝对误差和平均绝对百分比误差都优于 SVM 模型和 BP 神经网络，拟合效果更好，因此，用 LSTM 神经网络预测大连市空气质量指数是可行的。

参考文献

[1] 牟敬锋, 赵星, 樊静洁, 严宙宁, 严燕, 曾丹, 罗文亮, 范志伟. 基于 ARIMA 模型的深圳市空气质量指数时间序列预测研究[J]. 环境卫生学杂志, 2017(2): 102-107+117.

[2] 常恬君, 过仲阳, 徐丽丽. 基于 Prophet-随机森林优化模型的空气质量指数规模预测[J]. 环境污染与防治, 2019, 41(7): 758-761+766.

[3] 徐乔王, 胡红萍, 白艳萍, 王建中. 基于 MEA_SVM 空气质量指数预测[J]. 重庆理工大学学报(自然科学), 2019, 33(12): 150-155.

[4] 尹琪, 胡红萍, 白艳萍, 王建中. 基于 GA-SVM 的太原市空气质量指数预测[J]. 数学的实践与认识, 2017(12): 113-120.

[5] 王云中. 基于神经网络的 PM_(2.5)浓度预测研究与实现[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2018.

[6] 张超利. 基于神经网络的河南省空气污染预测研究[D]: [硕士学位论文]. 郑州: 华北水利水电大学, 2019.

[7] 郑洋洋, 白艳萍, 侯宇超. 基于 Keras 的 LSTM 模型在空气质量指数预测的应用[J]. 数学的实践与认识, 2019(7): 138-143.

[8] 陈岑, 田晓丹, 武文星. IG-LSTM 模型在空气质量指数预测中的应用[J]. 华北科技学院学报, 2020, 17(4): 85-91.

[9] 闫洪举. 基于深度学习的金融时间序列数据集成预测[J]. 统计与信息论坛, 2020, 35(4): 33-41.

[10] 刘恒. 基于神经网络模型的股票时间序列预测研究[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2019.