doi: 10.3969/j.issn.1674-8425(z).2018.08.021

本文引用格式: 张春露,白艳萍. 基于 TensorFlow 的 LSTM 模型在太原空气质量 AQI 指数预测中的应用[J]. 重庆理工大学学报(自然科学),2018(8):137-141.

Citation format: ZHANG Chunlu, BAI Yanping. Application of LSTM Prediction Model Based on TensorFlow in Taiyuan Air Quality AQI Index[J]. Journal of Chongqing University of Technology(Natural Science),2018(8):137-141.

基于 TensorFlow 的 LSTM 模型在太原空气质量 AQI 指数预测中的应用

张春露,白艳萍

(中北大学,太原 030051)

摘 要: 由于空气质量 AQI 指数受多个难以确定的和非线性的因子的影响,经常用到的回归预测方法效率和精度都比较低,基于长短期记忆单元(long short-term memory,LSTM)的递归神经网络模型却能有效利用时序数据中长距离依赖信息的能力,精准地预测空气质量 AQI 指数。首先,利用 Ri386 3.3.3 分析出空气中各种污染物质与 AQI 指数的相关性;然后基于 Python3.5.2 和 TensorFlow,结合近几年空气质量的各种影响因素的走势,对太原空气质量的 AQI 指数进行预测;最后使用均方误差(MSE)对预测的数据和原始数据进行误差分析。最终得出结论:基于 TensorFlow 的 LSTM 神经网络能较精准地预测空气质量 AQI 指数。

关 键 词: 空气质量;相关性因素分析;TensorFlow;LSTM 神经网络

中图分类号: X823 **文献标识码:** A **文章编号:** 1674-8425(2018)08-0137-05

Application of LSTM Prediction Model Based on Tensor Flow in Taiyuan Air Quality AQI Index

ZHANG Chunlu, BAI Yanping

(North Central University, Taiyuan 030051, China)

Abstract: We often use regression methods to predict the air quality AQI index. While the AQI index is affected by a number of factors, which are non-linear and are difficult to determine. According to previous studies, the recursive neural network model based on Long-Short-Term-Memory(LSTM) can predict the AQI index accurately. We use Ri386 3.3.3 to analyze the correlation between various air pollutants and the AQI index. Then we estimate the AQI index of Taiyuan air quality based on Python 3.5.2 and TensorFlow. In the end, the mean square error (MSE) is used to analyze the predicted data and the original data. We conclude that the LSTM neural network, based on TensorFlow, can accurately perdict the air quality AQI index.

Key words: air quality; analysis of correlation factors; TensorFlow; LSTM neural network

收稿日期: 2018-03-10

基金项目: 山西省自然科学基金资助项目(201701D22111439)

作者简介: 张春露,女,硕士,主要从事数据挖掘与分析研究,E-mail: 1060185296@qq.com。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

当今社会，随着人们生产、生活的发展，环境问题逐渐成为人们关注的焦点。山西省太原市是一座拥有 2 500 年历史的文化古城。长久以来，山西省作为全国的煤炭大省在为国家发展做贡献的同时，也不可避免地造成了一定的环境影响。如今环境问题日益严重，特别是在冬季供暖时期，山西主要依靠的是烧煤供暖，环境问题十分严峻。山西省想要恢复碧水蓝天，尤其是作为省会城市的太原，更加需要以新的面貌去迎接未来的挑战和发展。太原市经过多年的整治，但环境改善并没有达到预期的效果，为了进一步揭示和治理太原空气质量的污染情况，必须了解空气变化趋势，掌握及时、准确、全面的空气质量信息，对空气质量 AQI 指数进行精准地预测。基于此，本文提出了一种基于 TensorFlow 的 LSTM(递归神经网络)时间序列模型来预测太原空气质量的 AQI 指数。

1 理论介绍

1.1 LSTM 简介[1]

长短期记忆人工神经网络(long-short term memory，LSTM)是一种改进的时间循环神经网络(recurrent neural network，RNN)。

LSTM 的关键是单元状态[2](cell state)，它像是传送带一样，将信息从上一个单元传递到下一个单元，与其他部分只有很少的线性相互作用。LSTM 通过“门”(gate)来控制丢弃或增加信息，从而实现遗忘或记忆的功能。“门”是一种使信息选择性通过的结构，由一个 sigmoid 函数和一个点乘操作组成。sigmoid 函数的输出值在 [0,1] 区间，“0”代表完全丢弃，“1”代表完全通过。1 个 LSTM 单元有 3 个这样的门，分别是遗忘门(forget gate)、输入门(input gate)、输出门(output gate)。

1) 遗忘门：遗忘门是以上一单元的输出 h_{t-1} 和本单元的输入 x_t 为输入的 sigmoid 函数，输出一个 0 到 1 之间的数值 f_t，将其赋值给当前的细胞的状态 C_{T-1}，其中 f_t 计算公式如下：

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

2) 输入门：输入门和一个 tanh 函数配合控制有哪些新信息被加入。函数产生一个新的候选向 $\tilde{C}_t$，输入门为 $\tilde{C}_t$ 中的每一项产生一个在 [0,1] 内的值，控制新信息被加入的多少。至此，已经有了遗忘门的输出 f_t 用来控制上一单元被遗忘的程度，也有了输入门的输出 i_t，用来控制新信息被加入的多少，就可以更新本记忆单元的单元状态了，此过程计算公式如下：

$$C_t = f_t \cdot C_t + i_t \cdot \tilde{C}_t \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

3) 输出门：该层输出基于细胞的状态，但也是一个过滤后的版本。首先，运行一个 sigmoid 层来确定细胞状态的输出部分；接着，将细胞状态通过 tanh 进行处理(得到介于 -1 到 1 之间的值)，并将其与 sigmoid 门的输出相乘，最终确定输出部分。该层计算公式为

$$h_t = O_t \cdot \tanh(C_t) \tag{5}$$

如图 1 中所示，输出时如果已达到阈值，就将该阀门的输出与当前层的计算结果相乘，并把得到的结果作为下一层的输入(此处相乘是在指矩阵中的逐元素相乘)；如果未达到阈值，则遗忘输出结果。每一层及阀门节点的权重都会在每一次的模型反向传播训练过程中得到更新。

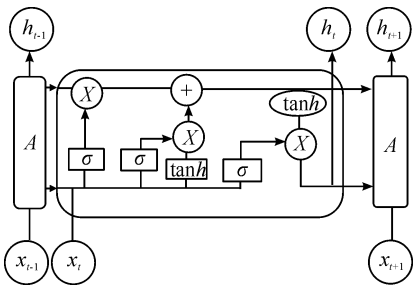

图 1 LSTM 流程

1.2 TensorFlow 简介

在 2015 年的 9 月，谷歌发布了其第 2 代人工智能系统：TensorFlow，一个开源的机器学习软件资源库[3]。TensorFlow 既支持深度学习算法[4]，也实现了很多其他算法，例如回归预测、聚类分析、决策树、关联分析等多种算法。TensorFlow 的发布也让人工智能又一次成为大众关注的焦点。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在 TensorFlow 中提供了 LSTMCell 操作来支持 LSTM 模型的搭建，在 TensorFlow 的内部封装了 LSTM 的隐藏层，其中包含遗忘门、输入门和输出门等结构，但是一般隐含层数目需要用户根据自己的实际情况来设定，本文设置的隐含层数目为 10。

在使用 TensorFlow 搭 LSTM 建神经网络的过程中[5]，不再以神经网络中的节点为单位进行布局，而是以层为基础来进行考虑和搭建。在 TensorFlow中，LSTMCell 就好比 LSTM 模型里面的隐藏层，因此包含多个节点的输入层和输出层也都用向量的形式来表示，向量的长度即为该层节点的个数。

在搭建和训练模型的时候，模型里面的参数初始化是一个十分重要的过程，模型中的训练参数的初始标准化会对训练效果产生很大的影响。本文对训练集标准化的理方式为：normalized_train_data = (data_train - np.mean(data_train,axis = 0)) /np.std(data_train,axis = 0)；对测试集标准化处理为：normalized_test_data = (data_test - mean) / std，并用 orthogonal_initializer 方法对 LSTMCell 中的遗忘门、输入门和输出门的参数进行初始化。本研究使用批量随机梯度下降法进行训练。

2 模型建立

2.1 数据的采集和预处理

数据下载于太原空气质量网，从 2013 年 11 月 1 日至 2017 年 6 月 30 日总计下载 1 433 条数据（某些数据存在缺失）。这里只列出前 10 天的数据，主要格式及污染物形式如表 1 所示。

2.2 太原空气 AQI 指数分析

通过 R 语言编程对太原 2013 年 11 月 1 日至 2017 年 6 月 30 日的空气质量 AQI 指数进行聚合，得到其时间序列分布折线图如图 2 所示。

通过图 2 可以看出：太原空气质量 AQI 指数的分布有一定的季节性波动，大概在每年的冬季 AQI 指数均较高，空气质量较差。主要原因是冬季山西主要依靠烧煤取暖，供暖系统会增加颗粒物污染，并且冬季天气干燥，不利于形成降水，降雨量少且持续时间较短，风速和风力较小，对空气中污染物的冲刷效果不明显，因此冬季的空气质量相对较差。

表 1 太原空气污染物总表

质量等级	AQI 排名	PM2.5	PM10	So2	No2	Co	O3	AQI
4	93	132	185	113	58	2.56	5	175
3	77	109	160	157	50	2.64	13	144
2	51	71	130	104	41	1.96	14	99
2	65	70	131	158	52	2.05	10	98
3	73	81	128	153	47	1.97	19	110
3	46	58	118	107	36	1.88	24	107
2	24	67	147	153	52	1.66	14	78
3	71	100	195	192	63	2.35	15	124

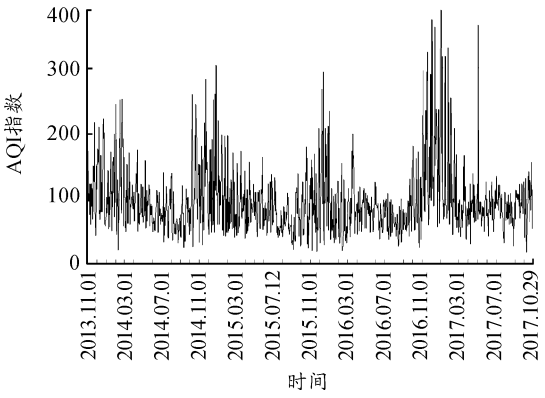

图 2 AQI 指数时间序列分布折线

2.3 AQI 指数污染物相关性分析

利用 R 语言中的相关性分析包 PerformanceAnalytics，并运行下列代码：

```
Library(PerformanceAnalytics)
chart.Correlation(newdata,histogram = TRUE,pch = 5)
```

得到空气质量 AQI 指数和空气污染物的相关性分析结果，如图 3、4 所示。

图 3、4 对角线中的是变量自身分布的曲线图；在下三角形（对角线的左下方），给出了两个属性相关性的散点图，上三角形（对角线的右上方），数字表示两个属性的相关性值（数字越大两个属

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性的相关性越高)，型号表示显著程度(星星颜色越深表示越显著)。分析图3和图4各种污染物与AQI指数的相关性如下:

PM10和AQI指数呈正相关，且相关性较大，为0.95;

So2与AQI指数呈正相关，且相关性相对居中，为0.68;

No2与AQI指数呈正相关，且相关性较小，为0.36;

Co与AQI指数呈正相关，且相关性相对较小，为0.45;

O3与AQI指数呈负相关，其负相关程度相对较小，为-0.25。

虽然每种污染物与空气质量AQI指数相关性存在差异性，但是都与其有着一定的关联，因此将AQI指数排名、PM2.5、PM10、So2、No2、Co2、O3均作为模型的输入特征，AQI指数作为测试标签(lable)。

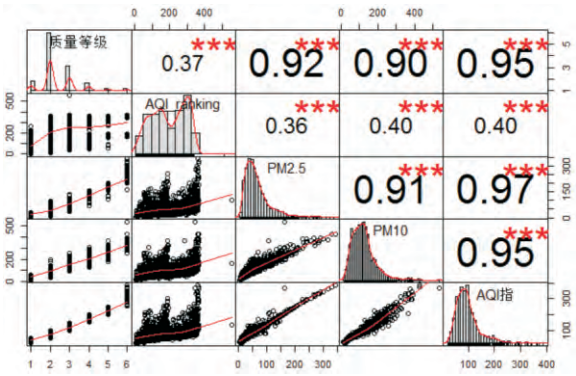

图3 质量等级、AQI排名、PM2.5、PM10、AQI相关性分析

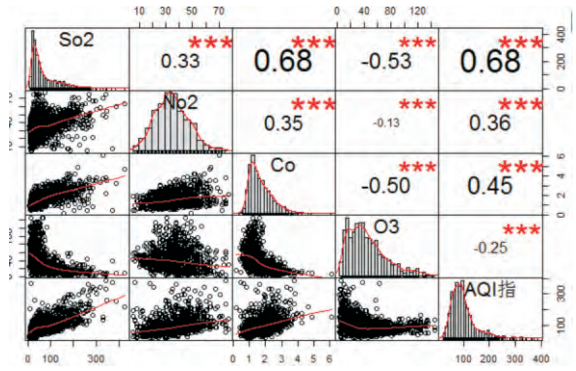

图4 So2、No2、Co、O3、AQI相关性分析

2.4 实验设计及结果分析

根据太原空气质量AQI指数预测过程，反复迭代实验，由于本文是基于TensorFolw的多特征时间序列模型[5]，将通过多变量输入特征拟合LSTM。

首先设置模型的隐含层数为10，输入层和输出层分别为8和1;然后取前1 400条数据为训练集，后36条数据为测试集，并分别对数据进行标准化处理[6]。将每批次训练样本数(batch_size)设置为40，时间步长(time_step)设置为10，学习率(lr)设置为0.000 6，然后定义神经网络变量，在输入特征时需要将tensor转成二维进行计算，计算后的结果作为隐藏层的输入，最后再将tensor转成三维作为lstm cell的输入[7]。

在训练模型时迭代次数可以改变，次数越大效果越精确，但需要的时间也越长[8]。本文的迭代次数设置为900，最终得到的平均偏差为2.031，均方误差为5.625，得到的预测值(红色线)与真实值(蓝色线)的对比如图5所示。

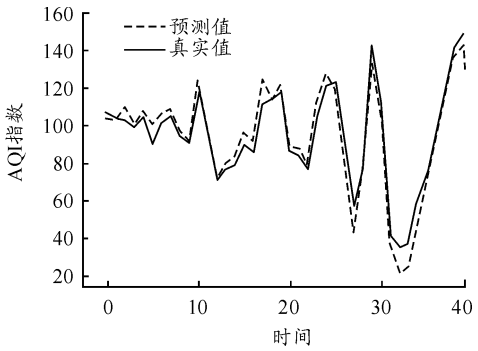

图5 太原空气质量AQI指数预测值与真实值对比

得到如表2的太原空气质量AQI指数在LSTM模型[9]下迭代900次的误差分析。

表2 太原空气质量AQI指数误差分析结果

检验模型	平均偏差	均方误差	迭代次数	耗费时间/s
LSTM	2.013	5.625	900	38.63

从图5(太原空气质量AQI指数预测值与真实值对比)和表2(太原空气质量AQI指数误差分析结果)可以看出:利用LSTM模型预测太原的空

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

气质量 AQI 指数可以得到较精确的结果。

3 结束语

本文通过 R 语言中的相关分析包 PerformanceAnalytics 得出与太原空气质量 AQI 指数相关联的影响因素,并将其作为模型的输入特征,然后利用 LSTM 循环神经网络对 AQI 指数进行预测。由于以往适合多输入变量的神经网络模型[10]一直存在着缺陷和不足,大多数古典的线性方法难以适应多变量或多输入的预测问题,而基于 LSTM 的循环神经网络却几乎能完美解决这个困扰已久的多输入时间变量问题。

通过本文实证表明: 基于 TensorFlow 的 LSTM 时间序列模型预测太原的空气质量 AQI 指数具有精度高、预测时间范围长、自适应高等优点,并且能够充分逼近非线性映射。该方法具有通用性,能够适用于其他多变量输入的时间序列预测问题[11],在生产和生活中有着广泛的应用。

不足之处: 由于本研究数据量有限,且对空气质量 AQI 指数的影响除了大气污染物还与气候条件、工作日与休息日车流量、工厂聚集地等多种因素有关,因此要得到更精确的预测效果还需要加入这些影响因子,这也是未来本研究需要努力的方向。

参考文献:

[1] 陆泽楠,商玉林. 基于 LSTM 神经网络模型的钢铁价格预测[J]. 科技视界,2017(13):116-117.

[2] 陈亮,王震,王刚. 深度学习框架下 LSTM 网络在短期电力负荷预测中的应用[J]. 电力信息与通信技术,2017,15(5):8-11.

[3] 李剑风. 融合外部知识的中文命名实体识别研究及其医疗领域应用[D]. 哈尔滨: 哈尔滨工业大学,2016.

[4] 杨晓峰. 多层随机森林算法在电信离网预测中的应用[D]. 苏州: 苏州大学,2016.

[5] 杨煜,张炜. TensorFlow 平台上基于 LSTM 神经网络的人体动作分类[J]. 智能计算机与应用,2017,7(5):41-45.

[6] 王鑫,吴际,刘超,等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报,2018,44(4):772-784.

[7] WU Y,YUAN M,DONG S,et al. Remaining useful life estimation of engineered systems using vanilla LSTM neural networks[J]. Neurocomputing,2018,275:167-179.

[8] ZHOU F,JIAO J R,LEI B. A linear threshold-hurdle model for product adoption prediction incorporating social network effects[J]. Information Sciences,2015,307:95-109.

[9] XIONG F,LIU Y,ZHANG Z,et al. An information diffusion model based on retweeting mechanism for online social media[J]. Physics Letters A,2012,376(30/31):2103-2108.

[10] SOUSA J C,JORGE H M,NEVES L P. Short-term load forecasting based on support vector regression and load profiling[J]. International Journal of Energy Research,2014,38(3):350-362.

[11] YU B,LAM W H K,TAM M L. Bus arrival time prediction at bus stop with multiple routes[J]. Transportation Research Part C: Emerging Technologies,2011,19(6):1157-1170.

(责任编辑 陈 艳)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net