LSTM 预测 PM2.5

姚非凡 2019312571 医学院 生物医学工程
yff19@mails.tsinghua.edu.cn

摘要 对北京市多个监测点的空气质量原始数据进行处理，将得到的时间序列数据转换为监督学习数据，借助 Keras 实用 LSTM 对数据进行回归处理，在进行分类。在调整超参数、网络结构、训练样本、数据不平衡处理以及特征组合后，得到最优的模型。最后结果发现当隐藏层神经元数量达到 50，且激活函数为 Sigmoid，训练样本为 2019 年数据，并且只选择 PM2.5 特征，模型的预测能力最优，RMSE 为 10.702，不同类别的分类结果宏平均为 0.991，比最开始的设计的基本模型，有很大提升。

***Keywords*—PM2.5 LSTM RESM 宏平均**

I. 简介

空气质量对环境保护与人类健康有着至关重要的影响。对空气质量的科学预测与准确测控就对人类的生活、工作、娱乐有着指导作用。而空气质量的评估指标也大都为 PM2.5。PM2.5 又称细颗粒物，是指环境空气中空气动力学当量直径小于等于 2.5 微米的颗粒物。其在空气中含量浓度越高，就代表空气污染越严重，是一个反映空气质量的有效指标。对未来时间段空气质量的预测可以向人们的提供出行有效的建议，所以对空气质量的预测具有很强的现实意义。

主流的预测未来时间段的空气质量有数值模式方法和机器学习方法两种。第一种是数值模式方法，即大气预测模型，数值预测方法一般都需要比较详细的时空分布资料和分辨率要求较高的气象模式，而这种数值预测模式目前并不成熟第二种机器学习方法，其应用较为广泛，主要采用决策树、时间序列、神经网络等预测方法，基本都是基于历史气象和空气质量数据进行预测。

使用大气模型预测空气质量其本质上是一种 “生成模型”，即通过模拟大气中的各项活动来预测空气质量状况。国内外的空气质量预报从上世纪 70 年代开始，已经经过了三代的发展，第一代仅考虑十余个化学物种的化学反应以及简单的气液相转化，例如 STEM-I 模式 [1]，第二代比第一代优秀的原因在于它对气液相化学、降雨和云的湿清除过程的处理，并且模式考虑的物种扩展到几十个，例如 STEM-II 模式 [2]，第三代模式基于 “一个大气” 的观念，考虑了气相液相和固相多相化学种类和机制，包含相应的空间位置源排放模式提供化学模式所需的排放清单，代表模型为 Model 3 / CMAQ[3]。

我国最新的空气质量预报系统使用的是集合预测模式，有嵌套网格空气质量预报模式系统（NAQPMS）、Model 3 / CMAQ-4.6、Model 3 / CMAQ-4. 4、CAM，和 WRF Chem 共 5 套模式，其基础数据包括气象资料、下垫面资料、污染物排放源资料 3 个部分。NAQPMS、CMAQ、CAM，3 个模式中气象要素的模拟由第五代中尺度气象模式（MM 5）完成，为空气质量模拟提供气象场，WRF-CHEM 则由气象模式（WRF）和化学模式（Chem）在线完全耦合 [4]。

大气预测模型虽然可以通过模拟大气环境来预测空气质量，但是需要考虑复杂的气象环境因素，需要进行大量的计算，必须运行在高性能机群上，而且因为计算量高所以能提供的预测粒度不高，使用统计学习模型可以避免这些问题。

很多基本的机器学习模型可以应用于空气质量预测，包括回归和分类模型都是基于气象数据与空气质量是统计相关的这一假设来进行预测的，使用的数据基本上是一个区域的历史气象与空气质量数据，数据源和模型结构较为单一。

回归模型通常用来预测具体污染物的数值指标，常用的模型有时间序列分析，支持向量回归，人工神经网络等，例如，Shuojun Wang 等人使用 ARMA 模型来预测加利福利亚一个城市的逐小时空气污染水平 [5]，Prybutok 等人 BP 神经网络方法预测每日臭氧水平 [6]，Soawalak Arampongsanuwat 等人使用支持向量回归预测 PM10 的值 [7]。

而分类模型通常用来预测污染的等级，常用的模型有分类决策树，支持向量分类等等，例如，Burrow 等人使用 CART 对地表臭氧进行分析并预测臭氧污染的等级 [8]。

可以发现，传统的时间序列模型只考虑到历史空气

质量序列，数据不够充分；其他的单一模型只从时间角度构造模型，使用的是所预测区域的历史数据，很多模型只能以城市为单位进行预测而不能细化到站点，或者只能预测日均的数据，精细粒度不够高；大部分模型预测的是空气质量指数，是回归模型，要预测空气质量等级，除了通过回归预测再转换以外，还可以直接分类。所以这些模型都存在改进空间。[9] 提出了一种 GeoMAN(Multi-level Attention Network) 结构的预测模型，该模型基于 Encoder-Decoder 结构，在时空数据预测问题上首次引入了多层注意力机制，对各传感器之间的动态时空关联性进行建模，并通过在 Decoder 阶段融合传感器对应的兴趣点 (POI) 信息、传感器 ID 和天气预报数据等外部因素显著提升了模型的性能。该模型不仅在 PM2.5 预测上取得了成功，在自来水质预测上也有着同样的出色表现，是一个在地理传感器时间序列预测问题上通用的模型。

本文在以往文献和实验的基础上，利用 LSTM 解决空气质量预测问题。

II. 任务定义

任务为预测未来 1-6 小时逐小时平均观测点的 PM2.5 等级，将 PM2.5 划分为 6 个等级，

$$Grade(P) = \begin{cases} 1, & 0 \le P < 35 \\ 2, & 35 \le P < 75 \\ 3, & 75 \le P < 115 \\ 4, & 115 \le P < 150 \\ 5, & 150 \le P < 250 \\ 6, & 250 \le P \end{cases}$$

$$Grade(P) = [P_{(1)}, P_{(2)}, P_{(3)}, \ldots, P_{(n-2)}, P_{(n-1)}, P_{(n)}]$$

为任务输出，任务输入为

$$X = [x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(t-2)}, x_{(t-1)}, x_{(t)}]$$

其中

$$x_{(t)} = [CO_{(t)}, SO2_{(t)}, O3_{(t)}, NO2_{(t)}]$$

任务的输入可以包括当前和过去一段时间内的观测值。

III. 数据整理

对 2014 年至 2020 年的北京空气质量数据进行清洗与处理，对每小时的特征数据和 PM2.5 数据的各个监测点的数据取中位数，并对各个时间点的数据进行平均，并删去损坏的数据，重新拼接出每日的空气质量数据。对数据重新排列，从而将特征矩阵与标签矩阵做行合并，最后删去缺失值，并对 PM2.5 数据的不同数值设定等级。并以 2015 年空气质量特征数据和标签数据为例，展示各个数据的分布情况，如下：

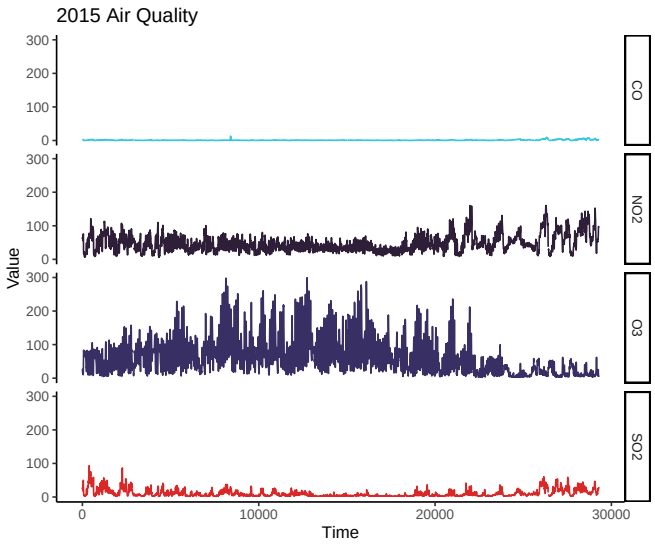

图 1. 2015 年北京空气质量特征分布图

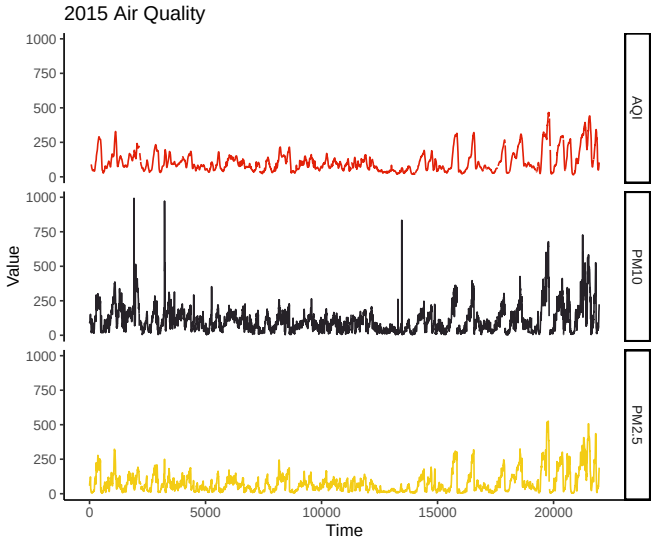

图 2. 2015 年北京空气质量标签分布图

并在 2015 年对这 7 个指标做相关分析，得到互相关结果，发现各个指标间都存在正相关，而 O_3 是负相关，如下：

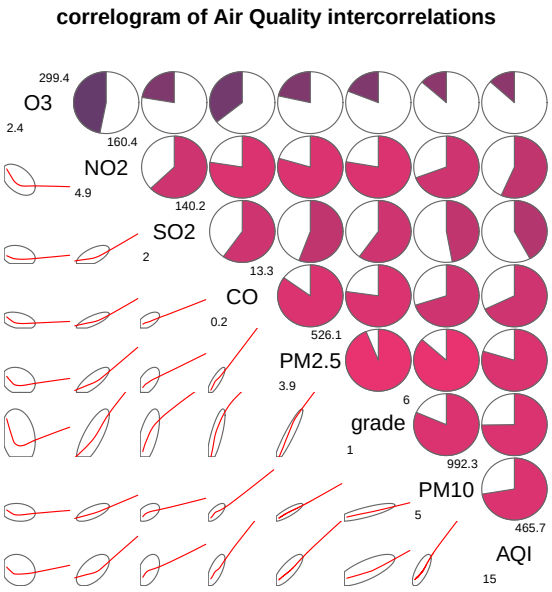

图 3. 2015 年北京空气质量数据相关图（越红代表相关越强）

IV. 特征提取

对特征数据 SO_2 实时浓度、NO_2 实时浓度、O_3 实时浓度、CO 实时浓度和前一小时的 PM2.5 进行提取，并对不同测量点的数据求平均形成每小时的特征。对四个特征进行归一化处理，并根据数据整理的结果选取包含 O_3 与不包含 O_3 的情况。通过将时间序列数据转化为监督学习数据，将后一小时的数据也作为特征，最终确定数据特征维度为 6，分别为 4 种空气化学成分，即 SO_2 实时浓度、NO_2 实时浓度、O_3 实时浓度和 CO 实时浓度与前后 1 小时的 PM2.5 指数，对特征进行归一化处理。

V. 模型设计

基于人工神经网络的模型: 实现一个 Long Short-Term Memory recurrent neural network (LSTM)，按时间顺序处理输入，并按时间顺序输出未来 1-6 小时的预测结果。使用多元 LSTM 预测模型，首先对数据处理，将时间序列数据转化为监督学习数据。选用 Keras 库中的 LSTM 模型，隐藏层有 50 个神经元，输出层 1 个神经元计算出回归值，输入变量是一个时间步（t-1）的特征，损失函数采用 Mean Absolute Error(MAE)，优化算法采用 Adam，模型采用 50 个 epochs 并且每个 batch 的大小为 72。

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 90)                34560
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 91
=================================================================
Total params: 34,651
Trainable params: 34,651
Non-trainable params: 0
```

图 4. 基本的神经网络结构

输入数据的 80% 为训练集，输入数据的 20% 为验证集。记录训练集和验证集的损失，并在完成训练和验证后绘制损失图：

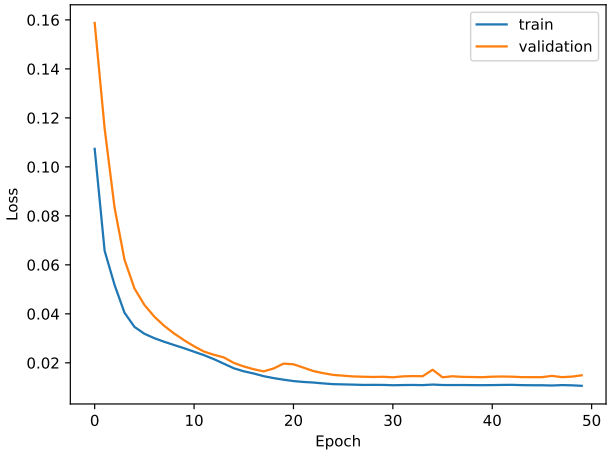

图 5. 2015 年数据为训练样本（80%）和验证样本的损失曲线（20%）

VI. 实验设计及结果

由于算力不足的原因，mini 数据集，即对各个气象站的数据进行平均，保留气象站全年的数据。将 2014 年 4/29 号至 2014 年 12/30 的数据作为测试数据。训练和验证数据分别选用 2015、2016、2017、2018、2019 年数据，其中输入数据的 80% 为训练集，输入数据的 20% 为验证集。

采用均方根误差对模型输出的回归值与真实 PM2.5 值进行评估，值为 11.031。

对预测的回归值逆归一化以同样的任务规则划分等级，形成 1-6 的标签数据，与真实值做混淆矩阵，并计算 6 类的分类正确率，计算平均值为宏平均为 0.931。

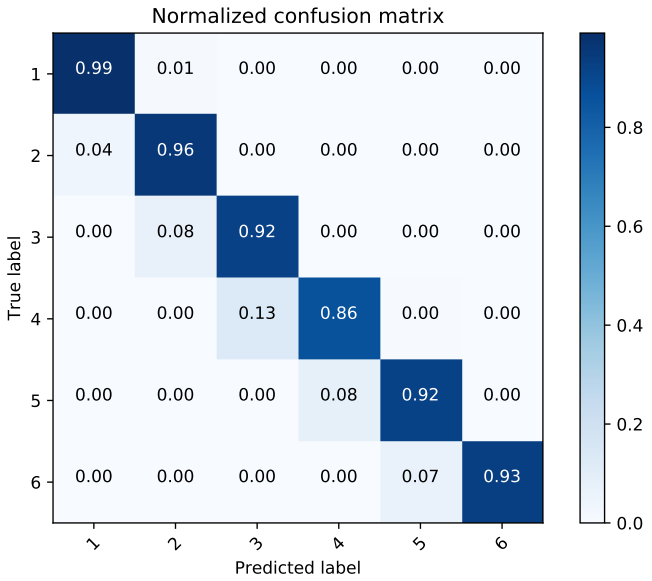

图 6. 2015 年数据作为训练样本的混淆矩阵

对模型设计加入 Dropout 和 Attention，发现正确率没有上升，所以在设计网络结构时没有加入。选择模型的不同超参数，得到结果如下：

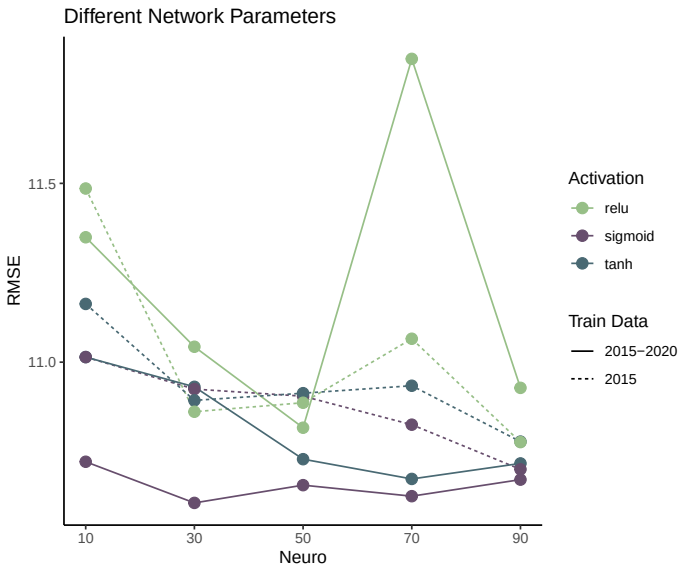

图 7. 不同隐藏层神经元数与激活函数对 RMSE 的影响

可以发现，模型的最优超参数设计在全部数据作为训练数据时为 10.606，参数为隐藏层有 30 个神经元，且激活函数为‘sigmoid’，在 2015 年数据作为训练数据时为 10.700，参数为隐藏层有 90 个神经元，且激活函数为‘sigmoid’。

VII. 实验结果分析

改变不同训练样本量，求得在最优超参数下的正确率，发现样本量与结果没有关系，全部数据量并不能提高正确率，发现以 RMSE 和宏平均为指标，则 2019 年为训练数据时，两者均最小，进一步发现混淆矩阵中分类出错主要出现在‘4’，推测可能是由于样本不均衡导致，故将 2015、2016、2017、2018、2020 的‘4’类数据与 2019 年数据进行拼接，发现并没有提高，发现‘4’、‘5’、‘6’三类样本数均过少，将 2015、2016、2017、2018、2020 的‘4’、‘5’、‘6’三类数据与 2019 年数据进行拼接，发现准确率也没有比单独使用 2019 年数据高，所以在训练样本选择上选择 2019 年数据：

表 I
不同训练样本下的最优结果（超参数已经调整）

Data	*Sample*	*RMSE*	*Mean*
2015	8455	10.7003	0.9366
2016	8413	10.6415	0.9747
2017	7139	10.6346	0.9693
2018	8348	10.8474	0.9784
2019	8543	10.6017	0.9850
2015~2020	43946	10.6064	0.9280
2019+‘4’	10735	10.6553	0.9738
2019+‘4’+‘5’+‘6’	14256	10.7479	0.9688

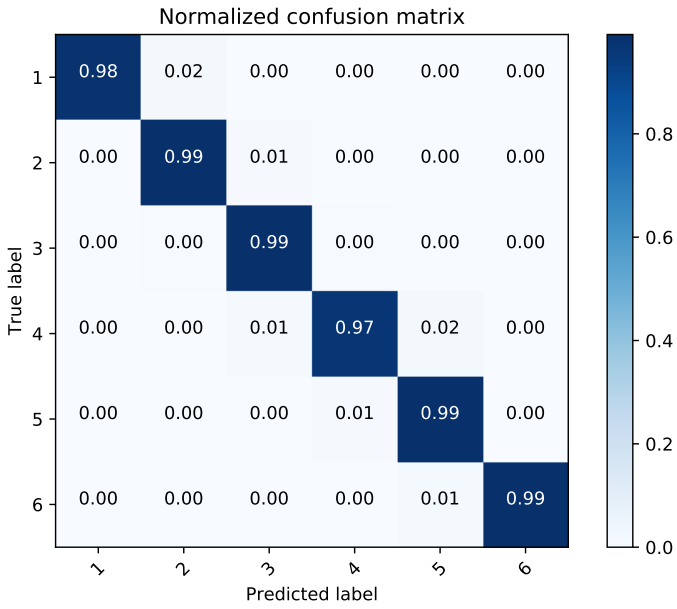

图 8. 最优训练样本在最优超参数下的混淆矩阵（2019 年数据-隐藏层 70 神经元-Sigmoid）

通过数据整理发现 O_3 数据相关性不足，考虑将其移出特征维度，比较结果移出 O_3 特征对回归结果不好。

将所有化学成分特征数据移出，发现移出的特征越多，回归数据越不准确，但 6 分类宏平均数据越高：

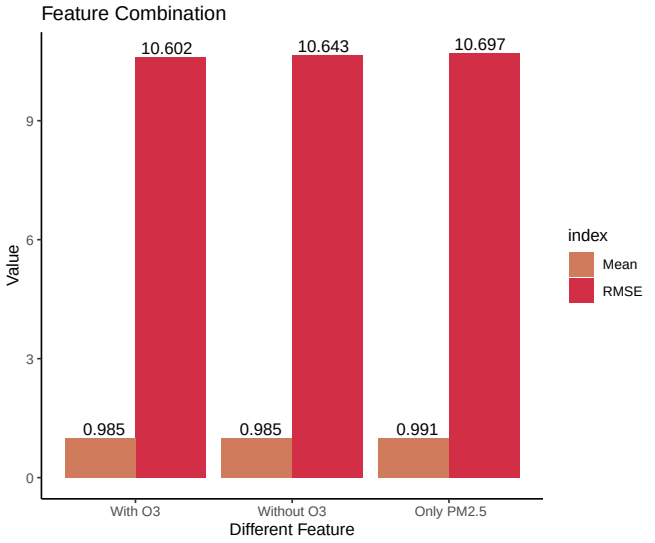

图 9. 不同特征组合对空气预测的影响

VIII. 代码接口

见附件中。

从https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/中借鉴了时间序列数据的转化和基本 LSTM 模型的搭建。

IX. 小组成员贡献

姚非凡独立完成。

X. 结论

通过 LSTM 对空气质量进行预测，最优的模型结果为 RMSE：10.701，宏平均：0.991。主要的分类错误集中在第 4 类数据错误分类为第 5 类数据，这可能由于对 PM2.5 数据中 115 至 150，区间的数据过少有关，之后的改进可以通过不将多个监测点数据进行平均，尝试更多复杂的 LSTM 设计，在时间序列上，不单单取间隔一个小时的数据，也将不同间隔数据放进 LSTM 进行建模，从而增加模型的复杂性，从而提高预测准确性。

参考文献

[1] Carmichael, G. R. . (1986). A second generation model for regional-scale transport/chemistry/deposition. Atmospheric Environment, 20(1), 173-188.

[2] Carmichael, G. R. , Peters, L. K. , & Saylor, R. D. . (1991). The stem-ii regional scale acid deposition and photochemical oxidant model—i. an overview of model development and applications. Atmospheric Environment Part A General Topics, 25(10), 2077-2090.

[3] K. L. Schere, & R. A. Wayland. (1989). Epa (environmental protection agency) regional oxidant model (rom2. 0): evaluation on 1980 neros data bases. report for january 1987april 1989.

[4] 王茜, 伏晴艳, 王自发, 王体健, 刘萍, & 陆涛等. (2010). 集合数值预报系统在上海市空气质量预测预报中的应用研究. 环境监控与预警, 002(004), 1-6,11.

[5] Shuojun Wang. (2007). Time series analysis of air pollution in the city of bakersfield, california.

[6] Prybutok, V. R. , Yi, J. , & Mitchell, D. . (2000). Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentration. European Journal of Operational Research, 122(1), 31-40.

[7] Arampongsanuwat, S. , & Meesad, P. . (0). Prediction of PM_(10) using Support Vector Regression. International Conference on Information & Electronics Engineering.

[8] William R. Burrows, Mario Benjamin, Stephen Beauchamp, Edward R. Lord, Douglas McCollor, & Bruce Thomson. (2010). Cart decision-tree statistical analysis and prediction of summer season maximum surface ozone for the vancouver, montreal, and atlantic regions of canada. Journal of Applied Meteorology, 34(8), 1848-1862.

[9] Liang, Y. , Ke, S. , Zhang, J. , Yi, X. , & Zheng, Y. . (2018). GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18.