PM2.5 预测

张斐然

摘要

基于机器学习的空气质量预测近年来显示出很大潜力。本文使用基于长短时记忆网络的模型，实现了基于过去若干小时的空气污染水平，来对未来 6 小时内的 PM2.5 进行预测，达到了不错的准确率。本文进一步分析了不同参数对模型的影响，还尝试了引入注意力机制来进一步提升模型。

关键词

空气质量预测；PM2.5 预测；长短时记忆网络；编码-解码模型；注意力机制

一、简介

天气预报类问题属于时空序列预测问题，它与我们的生活密切相关。其中近些年来空气质量预测，尤其是 PM2.5 预测受到人们很多关注。基于大气动力学、大气环境化学专业知识和数据的数值模型作为传统方法，需要精确的环境数据监测以保证较高的准确率，同时计算量巨大，往往需要超级计算机的参与。而基于对历史数据挖掘的机器学习方法，计算成本相对更低，且能达到很高的准确率，近年来受到人们广泛研究。

在机器学习，包括统计学习方法中，自回归模型（如 ARIMA、VAR）被广泛用于时间序列预测问题中；此外，回归树以及随机森林模型也受到大量应用。而人工神经网络模型，尤其是近年来深度神经网络，如循环神经网络（RNN）、长短时记忆网络（LSTM）等，近年来达到了相对领先的准确率，且不断有新模型、新机制被提出，如编码-解码（encoder-decoder）模型、注意力（attention）机制等，进一步提高了模型的能力。

因此，本文使用了基于 LSTM 的 encoder-decoder 模型，使用历史的空气污染数据进行训练，来基于过去一段时间的空气污染状况进行未来 6 小时内的 PM2.5 预测，达到了不错的准确率。此外，本文还进一步利用了注意力机制，来尝试对模型进行改进。

二、任务定义

作为时空序列预测问题，任务要求使用过去和现在共T小时内，N个地点测得的C类空气污染指数数据，来预测未来 1~6 小时内 N 个地点的 PM2.5 污染指数。因此，模型的输入数据$\boldsymbol{X} \in \mathbb{R}^{T\times N\times C}$，而输出$\boldsymbol{Y} \in \mathbb{R}^{6\times N}$。同时，基于 LSTM 的 encoder-decoder 模型的损失函数是模型输出与真实值的均方误差，即模型目标为：

$$\min\left\|\boldsymbol{Y}-\tilde{\boldsymbol{Y}}\right\|_2^2$$

预测准确率首先通过将 PM2.5 预测值分类到 6 个污染区间，来分别计算各个区间里的准确率，同时将N个地点的准确率取平均，以此来得到 6 个小时、6 个污染区间共 36 个准确率指标。此外，定义 AP（Average Precision）为 36 个准确率的平均值。

三、数据整理

空气污染指数数据共包括 2014 年 1 月~2020 年 5 月以天为单位的北京市内共 35 个检测点的 7 类空气污染数据，包括主要的 3 类（AQI、PM2.5、PM10）和额外的 4 类（CO、NO_2、SO_2、O_3）。爬取的数据中存在不少的异常数据。数据清洗过程首先包括 gzip 文件解压；然后是空文件和 http 错误文件的删除，共 40 个文件。

其次数据中缺失数据较多，需要对缺失值进行处理，包括删除缺失较多的文件，以及对剩余少部分缺失值进行插值处理。删除过程包括：首先删除表格列数没有覆盖全部 24 小时的文件，其次对一天内缺失值超过 20%的文件进行删除；接下来，如果某一天的 all 和 extra 数据只存在其中一个，则将另一个也删除；处理过后从 4491 个文件，变为 3504 个。接下来，插值使用简单的线性插值，即对某一观测点的某一类观测值，使用其之前和之后时间的数据进行线性插值。

最后，考虑到算力的问题，只选取了城区的观测点数据参与后续训练和预测；同时注意到城区内的植物园观测点存在大量缺失（占比近一半），故也将其排除，因此最终观测点数量$N = 11$。

数据可视化（训练集）

考虑到测试集不能参与模型训练，而数据可视化可能能为模型训练时的参数选择提供帮助，故只选取 2015 年后的数据进行可视化。

首先统计各项空气污染的分布如下：

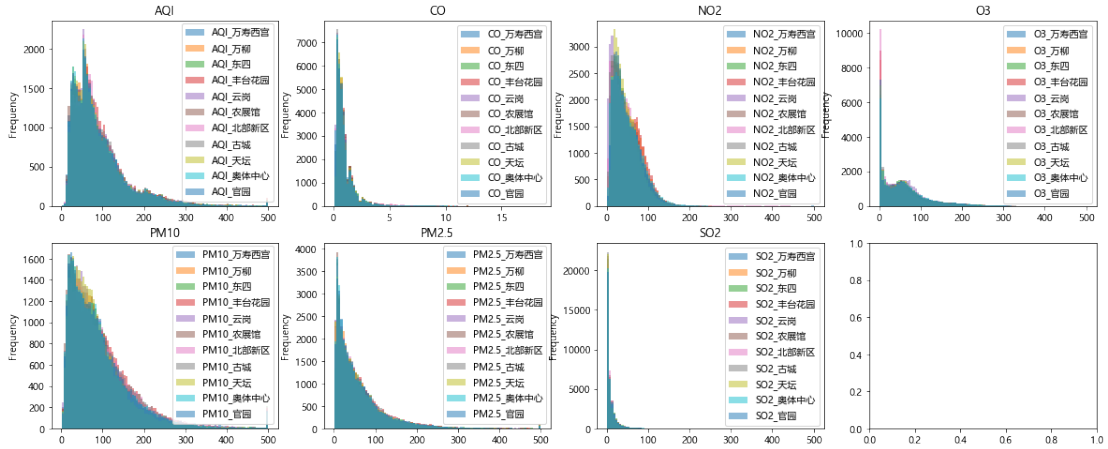

图 1 各项空气污染指标分布

可见污染整体分布类似，均为较低污染最多，随着污染严重而变少。

其次以 2019 年 12 月数据为例，空气污染走势如下：

图 2 2019 年 12 月空气污染走势

可见不同观测点污染变化基本一致；且不同污染之间存在较强相关关系，如果进一步观察其两两之间相关性，如图：

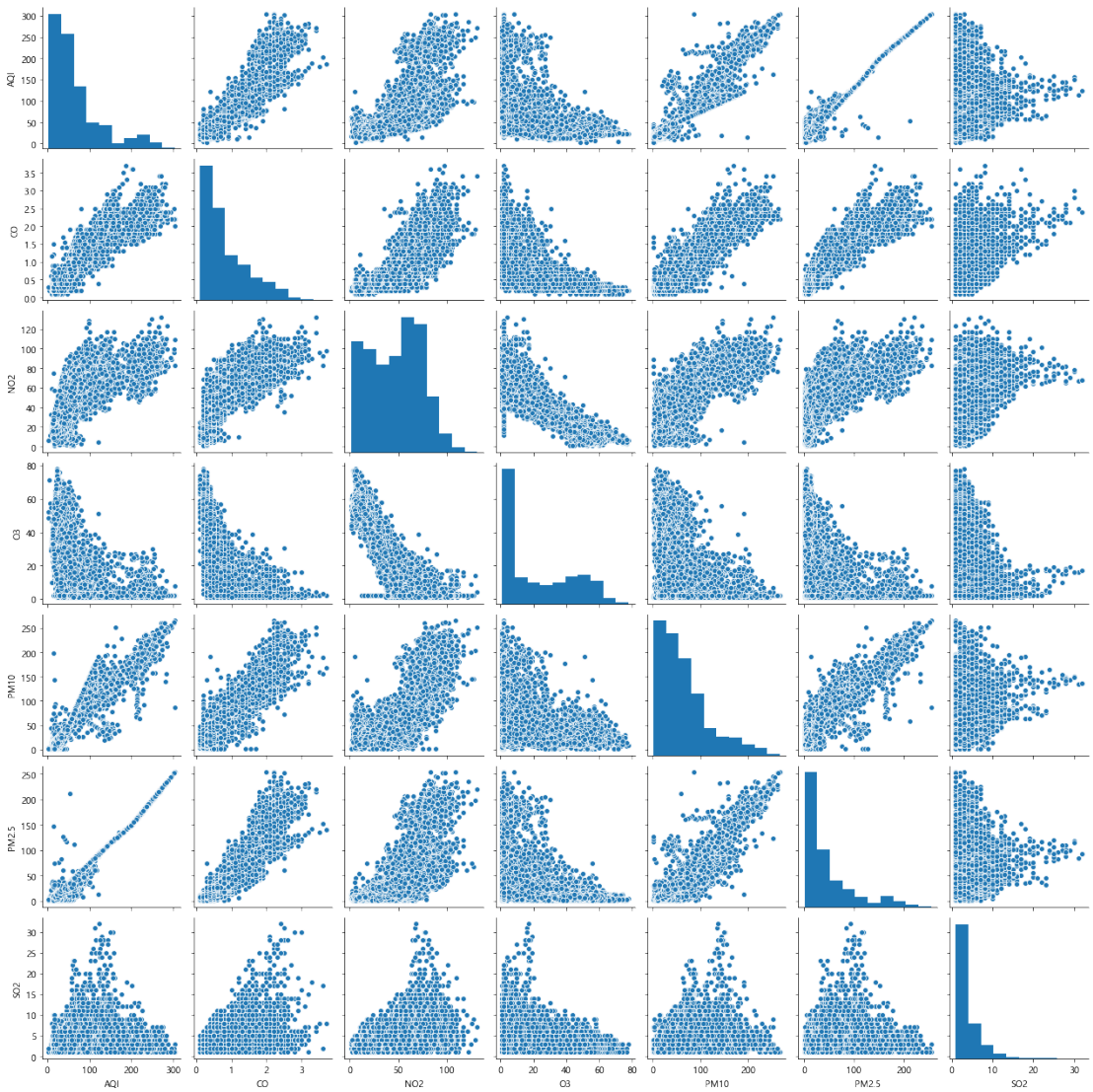

图 3 2019 年 12 月各类污染之间相关性

可见除了 O_3 与其余污染之间负相关、SO_2 与其余污染几乎无关之外，剩余 5 类污染均互相正相关，其中 AQI 基本上由 PM2.5 决定。。

四、特征提取

对于输入模型的数据，首先通过 clip 最大到 500，再除以 500 来进行归一化；其次将不同观测点（11 个）、不同类别（7 个，包含 4 个 extra）污染统一作为特征列输入模型，故特征维度为：使用 extra 数据为 77 列，而不使用则为 33 列。而输入的行则由参数 T（预测基于的历史长度）决定

五、模型设计

模型使用了基于 LSTM 的 encoder-decoder 模型。类似于 seq2seq 任务，输入的“seq”为过去若干小时的污染值，而输出的“seq”则为未来 6 小时的 PM2.5 污染值。其中 encoder 将除了 PM2.5 以外的特征通过一个 LSTM 编码为 context 隐向量，而 decoder 利用 PM2.5 的历史值和隐向量来推断得到未来 6 小时的 PM2.5 值。

此外，注意力机制的实现参考了 DA-RNN[1]中的两阶段注意力，其与原先的注意力只用在 decoder 中不同，DA-RNN 还在 encoder 中引入注意力，来自适应的选取特定特征。

六、实验设计及结果

1. 实验设计

数据集划分：通过年份，除了 2014 用于测试集，2020 年用于验证集外，其余年份用于训练。

训练参数：batch 大小为 64；优化器使用 Adam，学习率为 0.002；epoch 轮数为 7，在第 3 和第 6 轮结束后减为原本的 1/10。

PM2.5 值发生突变的标准：如果一段数据的最后 7 小时（即现在和未来 1~6 小时）的最大值与最小值之差超过阈值（设定为 0.2，即归一化之前的 100），则视为突变数据，将所有突变数据集合起来作为突变数据测试集。

2. 模型最好结果（测试集）

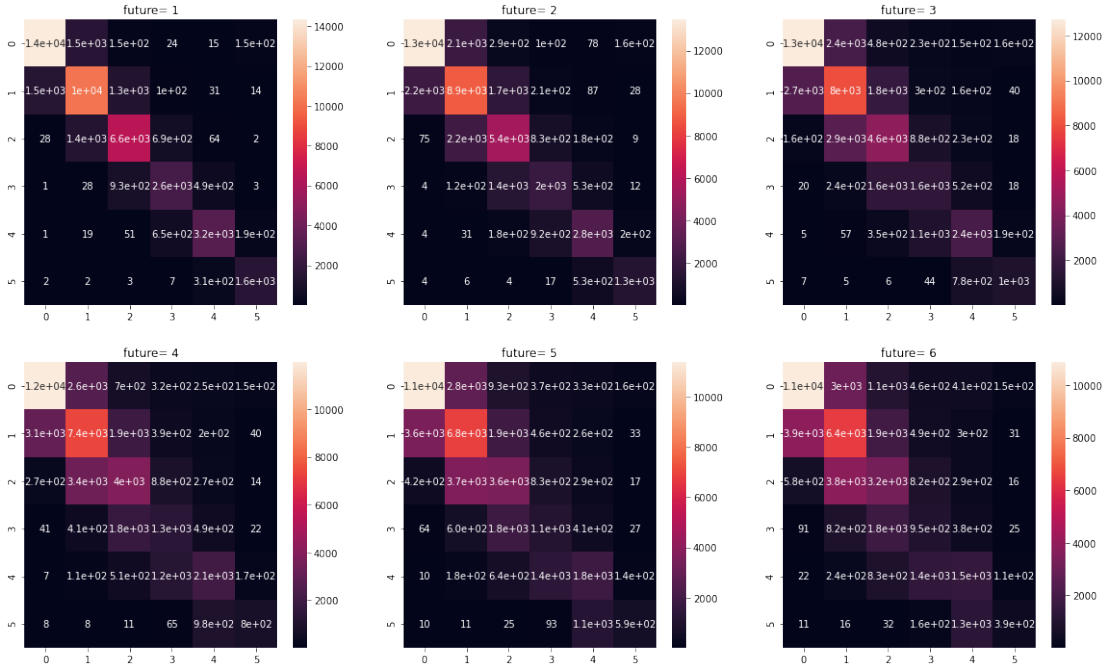

图 4 混淆矩阵结果

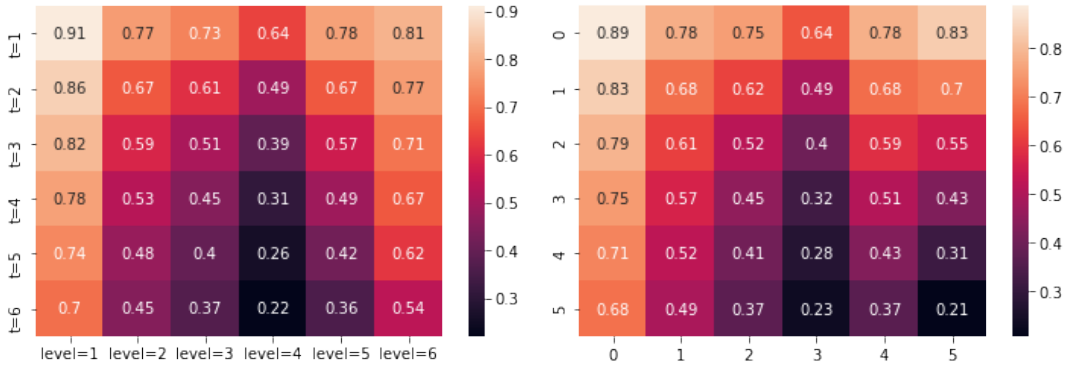

图 5 准确率（左）与召回率（右）结果

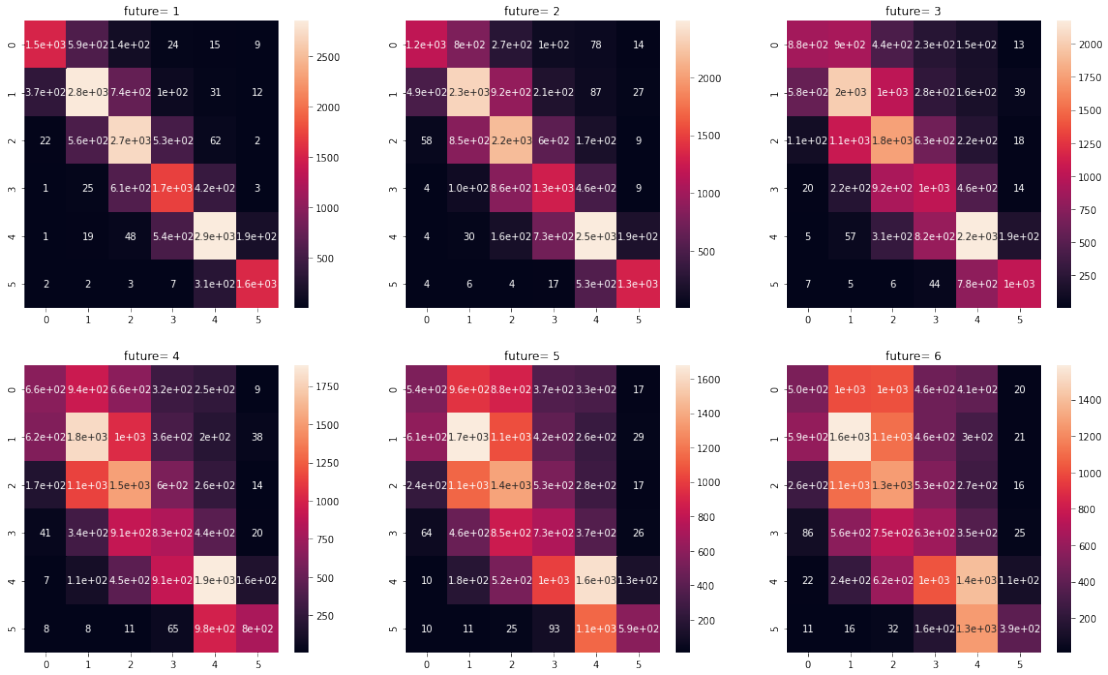

图 6 混淆矩阵结果（PM2.5 突变数据下）

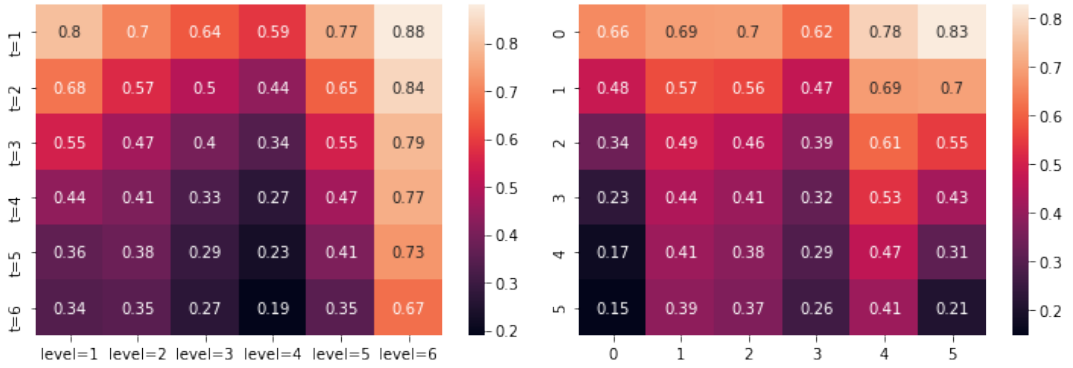

图 7 准确率（左）与召回率（右）结果（PM2.5 突变数据下）

3. 其他实验结果（验证集）

1) 是否使用 extra 数据

	使用	不使用
AP	0.601	0.602

可见 extra 数据并没有够提供额外的信息来供 LSTM 预测使用。

2) 是否使用 teacher force

Teacher force 是指在 decoder 中预测未来 PM2.5 时，LSTM 的每一时刻的输入是使用真实值，还是模型自身在上一时刻的预测值。如果 teacher force 不为 0 或 1，则以一定概率决定是否使用真实值。

	0	0.5	1
	0.600	0.603	0.601

可见与利用 seq2seq 模型进行翻译任务不同，引入 teacher forcing 不能帮助模型。

3) 历史长度 T 和是否使用 attention 机制：

T	6	12	18	24	48
AP (with attention)	0.610	0.588	0.611	0.603	0.605
AP (no attention	0.592	0.609	0.619	0.595	0.591

可见 attention 机制在增多计算量的同时并没有提高模型准确率。而历史长度对模型的影响同样不大。

七、实验结果分析

1. 训练参数分析：

实验发现 batch 过大会使得模型不收敛或收敛缓慢，而过小则训练缓慢，故设定为 64。

实验发现 Adam 优化器效果优于 SGD，故使用 Adam。

Epoch 轮数对结果的影响不大，往往前一两个 epoch 就能达到最低 loss，同时验证集上准确率也能达到最高，后续 epoch 结果一般都在上下浮动。

2. 最好结果分析：

从混淆矩阵结果可看出，模型整体表现不错，即使出错也往往是在邻近的区间。而从准确率结果可以明显看出，准确率随未来时间严格递减，在第 6 小时最差已经和随机差不多了，说明模型比较依赖短期时间内的数据。此外，不同 PM2.5 区间的准确率也有区别，处于中间的第 3、4 个区间最难预测，而无污染和严重污染都相对容易预测。其中前者应该因为无污染数据相对最多，模型对其的学习最充分；而后者则由于 PM2.5 值对应的区间大（150~250、250~500），使得准确率的计算相对更宽松，而不像其余区间长度只有 35、40 那样的严格。

而如果考虑突变数据，从混淆矩阵可看到 5、6 小时的预测结果已经变得较差了，而前几个小时的准确率也基本下降了约 0.1，除了严重污染的准确率反而上升了。整体上来说，突变数据的预测也是基于历史数据的机器学习方法的弊端，当然如果能够整合更多环境数据，如周围的风向、风速等，应该可以提高 PM2.5 突变时的预测准确率。

3. 其他实验结果分析：

考虑到多次运行本身就存在较大的随机性，三组不同参数下的实验都与随机情况没有显著差异。当然如果时间允许，利用多次运行后取平均的方法应该可以能够得出不同参数间的差异所在。其中预期的差异，首先是 extra 数据的引入能够提高模型预测能力，因为从可视化中可以看出其余污染指标与 PM2.5 均存在相关性，其中很可能也有因果性，如果模型能够学习到的话则能用于帮助预测。其次是历史事件 T 对模型的影响。从 GeoMAN[2]的论文

结果来看，模型在$T = 12$时结果最优，其理由为空气质量不存在长时间的依赖性。最后注意力的引入也没有为模型带来提升，说明在历史长度较短时，注意力机制没有明显的优势。

当然，实验结果不具有显著性也与所用的评测指标有关，在将 PM2.5 划分到 6 个区间后，模型预测时的精确程度的较小差异就被掩盖掉了，而不同参数对模型的影响也并没有非常大。如果时间允许，使用 RMSE 或 MAE 作为指标，应该可以得到差异性。

八、参考代码

model.py：学习、参考了：

https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

https://github.com/Seanny123/da-rnn

utils.py：参考了 https://github.com/pytorch/vision

九、参考文献

[1] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dualstage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971, 2017.

[2] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, Yu Zheng, GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. In International Joint Conference on Artificial Intelligence (IJCAI), 2018.