基于 LSTM 的 $PM_{2.5}$ 预测模型综述

肖敏志[1]　王淑君[2]　宋巍巍[1]

（1. 生态环境部华南环境科学研究所，广州　510655；
2. 广东省环境技术中心，广州　510308）

摘　要　当下雾霾问题引起了国家领导人的重视和公众广泛关注，预防 $PM_{2.5}$ 污染成为生态环境部主要工作之一。本文介绍了 LSTM 神经网络，以及以此为基础的 $PM_{2.5}$ 预测模型的算法和运算；采用灰色关联分析法计算了气象因子和大气污染物因子与 $PM_{2.5}$ 浓度关联度，并评估了基于 LSTM 的 $PM_{2.5}$ 预测模型的性能。结果显示，基于 LSTM 的 $PM_{2.5}$ 预测模型的准确度相对较高，研究结果为组力打赢蓝天保卫战具有一定的理论价值。

关键词　$PM_{2.5}$　预测　神经网络　LSTM　长短期记忆网络

清洁空气对于保障人类健康至关重要，根据国际能源署的研究报告，每年大气污染导致 650 万人过早死亡。当下，中国尤其是北方地区的主要城市受雾霾影响较大，引起了国家领导人的重视和公众的广泛关注，$PM_{2.5}$ 成为大气环境研究的热点。人工神经网络是一种模仿生物神经元运作的数学模型，是一种强大的非线性建模工具。长短期记忆神经网络（long－shorttermmemory，LSTM）是进行改进的时间循环的人工神经网络，随着深度学习技术的兴起，LSTM 算法也逐渐用于构建 $PM_{2.5}$ 的预测模型。

1　预测模型

1.1　LSTM 神经网络

循环神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络，一般是全连接网络，非相邻的网络之间没有连接，无法有效处理较长时序数据。LSTM 神经网络延续了 RNN 同样的重复连接结构，并增加了 Cell 状态参数以及遗忘门、输入门和输出门。其中，用于代表长期记忆，用于保存序列数据中重要信息并在较长时序中传递，遗忘门和输入门负责对进行更新，从而解决无法有效处理较长时序数据的问题。图 1 为 LSTM 神经网络结构。

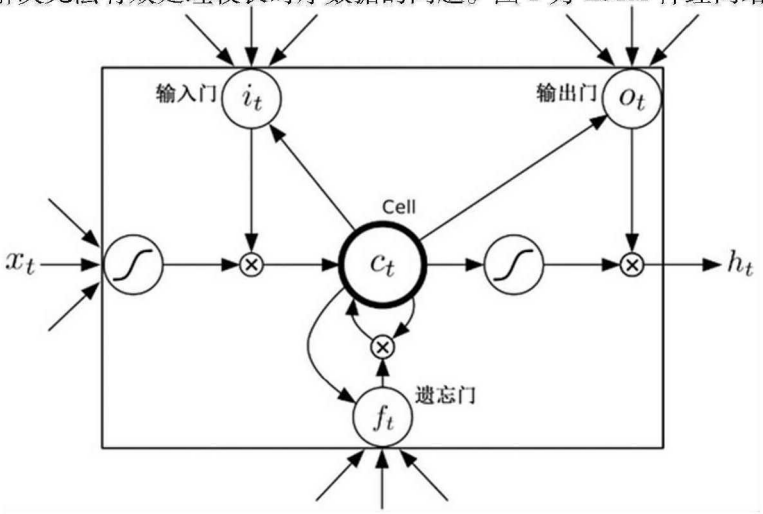

图 1　LSTM 神经网络结构

当接收到前一个时刻隐藏层输出 h_{t-1} 和当前时刻 x_t，使用遗忘门判定 c_t 中信息的重要程度，并决定是否舍弃。遗忘门计算公式如下所示：

$$f_t = \sigma\ (W_x f x_t + W_h f h_{t-1} + W_c f c_{t-1} + b_f) \tag{1}$$

式中：W_x、W_h、W_c、f、b_f 为遗忘门参数；σ 代表 sigmoid 激活函数；f_t 是遗忘门输出向量，其中向量中的每一个元素都在（0，1）范围内。

输入门主要是确定当前时刻哪些位置的信息需要更新以及更新后的数值，计算公式如公式（2）（3）所示：

$$i_t = \sigma\ (W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \tag{2}$$

式中：W_{xi}、W_{hi}、W_{ci}为输入门参数。

$$c'_t = \tanh\ (W_{xc} + x_t + W_{hc} h_{t-1} + b_x) \tag{3}$$

式中：x_t 为输入数据，c'_t 为当前时刻候选的 Cell 状态。

当遗忘门和输入门运算完成，LSTM 会更新当前时刻 Cell 状态，计算公式如下所示：

$$c_t = f_t c_{t-1} + i_t c'_t \tag{4}$$

式中：f_t 为遗忘门输出，决定哪些信息需要舍弃，哪些信息需要保留。

最后 LSTM 会计算出最终隐藏层的输出。公式（5）计算需要输出的 Cell 状态数值，公式（6）将 Cell 状态值通过 tanh 函数处理，得到一个（-1，1）范围的值，并与 O_t 相乘得到当前时刻隐藏层输出。

$$O_t = \sigma\ (W_{xo} x_t + W_{ho} h_{t-1} + b_o) \tag{5}$$

$$h_t = O_t \tanh\ (c_t) \tag{6}$$

式中：W_{xo}、W_{ho}为输入门参数，h_t 为当前时刻隐藏层输出。

1.2 $PM_{2.5}$预测算法

输入历史时刻 $PM_{2.5}$的测量数据 x_1、x_2、……、x_t，依次根据公式（1）至公式（6）计算隐藏层向量 H_1、H_2、……、h_t，经过合并得到编码向量，合并计算如公式（7）所示。公式（8）对编码向量作线性变化得到模型最终的特征向量。输出门的结果用向量形式表示，符号为 $PM2.5_O$，$PM2.5_O$ 向量的维数跟预测时刻的数量一致。

$$h = Concat\ (h_1,\ h_2,\ \cdots\cdots,\ h_t) \tag{7}$$

$$m = \sigma\ (W_{hm} h + b_{hm}) \tag{8}$$

$$PM2.5_O = W_{mo} m + b_{mo} \tag{9}$$

式中：W_{hm}、b_{hm}分别代表权重和偏置。

1.3 模型运算步骤

模型运算步骤包括：①初始化 LSTM 神经网络参数，即公式（1）至公式（6）所有参数；②初始化公式（8）参数，即 W_{hm}、b_{hm}；③初始化（9）参数，即 W_{mo}、b_{mo}；④根据 $PM_{2.5}$历史测量数据定义 batch 数，并采样训练样本序列 x，y；⑤根据公式（1）至公式（6）计算出隐藏层向量 h_1、h_2、……、h_t；⑥根据公式（7）至公式（9）计算出预测结果的向量值，即 $PM2.5_O$，神经网络参数值更新。

2 关联度分析

$PM_{2.5}$浓度会受到多种因素的影响，主要包括气象因子和大气污染物因子，且这些影响具有一定的不确定性。灰色关联分析方法能够根据因素之间发展趋势的相似或相异程度来衡量因素间的关联程度，可用于研究分析对 $PM_{2.5}$浓度影响较强的影响因子。

灰色关联分析计算公式如下：

$$\xi_{oi}(k) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_{oi}(K) + \rho\Delta_{\max}} \tag{10}$$

$$r_i = \frac{1}{N}\sum_{k=1}^{N}\xi_{ok}(k) \tag{11}$$

式中：$\Delta_{\min}$、$\Delta_{\max}$分别表示各时刻两序列绝对差的最小值和最大值；ρ 为分辨系数，取值范围为（0，1）；Δ_{oi}（k）为 k 时刻两序列的绝对差；ξ_{oi}（k）为子序列 i 和母序列 o 在 k 时刻的关联度；N 为数据序列长度；r_i 为平均关联度，各个时刻关联度 ξ_{oi}（k）的平均值。

以 $PM_{2.5}$浓度作为参考序列，影响因素作为比较数列，计算 r_i。r_i 越接近 1，说明关联度越大，其对 $PM_{2.5}$浓度的影响也就较强。

选取日照时数、最大风速、极大风速、最高气压、最高气温、平均风速、平均气压、平均气温、最低气压、最低气温、蒸发量、降水量、相对湿度、O_3、NO_2、SO_2、PM_{10}、CO 等 18 个影响因素进行关联度分析，分析结果见表 1。

表 1　$PM_{2.5}$影响因素关联度

序号	影响因素	关联度	序号	影响因素	关联度
1	日照时数	0.510	10	最低气温	0.629
2	最大风速	0.585	11	蒸发量	0.711
3	极大风速	0.612	12	降水量	0.789
4	最高气压	0.596	13	相对湿度	0.765
5	最高气温	0.597	14	O_3	0.653
6	平均风速	0.702	15	NO_2	0.801
7	平均气压	0.605	16	SO_2	0.832
8	平均气温	0.602	17	PM_{10}	0.859
9	最低气压	0.617	18	CO	0.884

3　模型性能评估

采用平均绝对误差（MAE）和均方根误差（RMSE）分别对基于 RNN 和基于 LSTM 的 $PM_{2.5}$预测模型的性能进行测试评估。平均绝对误差和均方根误差的计算公式如下：

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|\mathrm{PM2.5} - \mathrm{PM2.4}_O\right| \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(\mathrm{PM2.5} - \mathrm{PM2.5}_O)^2}{N}} \tag{13}$$

式中：PM2.5 为 k 时刻 $PM_{2.5}$的测量值，$\mathrm{PM2.5}_O$ 为 k 时刻 $PM_{2.5}$的预测值。

表 2 为各个预测模型性能的测试结果。测试结果显示，基于 LSTM 的 $PM_{2.5}$预测模型准确度相对较高。

表 2　模型性能测试结果

误差		测试值			均值
平均绝对误差	LSTM	18.522	13.179	18.142	16.614
	RNN	18.978	15.321	20.787	18.362

续表

误差		测试值			均值
均方根误差	LSTM	25.268	29.633	17.220	24.040
	RNN	26.389	31.396	16.6367	24.807

4 结论

跟 RNN 神经网络相比，LSTM 神经网络增加了 Cell 状态参数以及遗忘门、输入门和输出门，解决了无法有效处理较长时序数据的问题；从模型性能的角度进行分析，基于 LSTM 的 $PM_{2.5}$ 预测模型的平均绝对误差和均方根误差均比较小，预测准确度较高。影响因子方面，CO、PM_{10}、SO_2、NO_2、降水量、相对湿度、蒸发量和平均风速等 8 个因子跟 $PM_{2.5}$ 浓度关联度比较大。

参 考 文 献

[1] Liu X, Liu Q, Zou Y, et al. A Self - organizing LSTM - Based Approach to PM2.5 Forecast [J]. 2018: 683 -693.

[2] Fan, Junxiang, et al. A spatiotemporal prediction framework for air pollution based on deep rnn [J]. Remote Sensing and Spatial Information Sciences 4 (2017): 15.

[3] Guo, Tian, Tao Lin, and Yao Lu. An interpretable LSTM neural network for autoregressive exogenous model [C] // ICLR 2018, 2018: 1 -6.

[4] 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（32），No.386（08）：143 -147.

[5] 杨国田，张涛，王英男，et al. 基于长短期记忆神经网络的火电厂 NO_x 排放预测模型 [J]. 热力发电，2018，47（10）：16 -21.

[6] 杨训政，柯余洋，梁肖，et al. 基于 LSTM 的发电机组污染物排放预测研究 [J]. 电气自动化，2016，38（5）.