计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):215−219 [doi: 10.15888/j.cnki.csa.006889]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 LSTM 网络的大雾临近预报模型及应用①

苗开超[1], 韩婷婷[2], 王传辉[1], 章　军[2], 姚叶青[1], 周建平[1]

[1](安徽省气象局, 合肥 230031)
[2](安徽大学, 合肥 230039)
通讯作者: 苗开超, E-mail: mkc2005@126.com

摘　要: 长短期记忆网络 (LSTM) 是一种时间递归神经网络, 适合于预测时间序列延续性相对较长的事件. 本文基于 LSTM 网络构建了一个全新的大雾临近预报框架, 首先将地面气象要素观测资料转化成时间序列数据, 并基于此序列进行建模. 为了验证提出的模型的准确性, 将安徽省 81 个国家站近 2 年地面气象要素数据转换为序列数据, 基于该数据集对未来 1–4 小时进行逐小时大雾预报实验, 实验结果显示本文提出的模型其 TS-Score 分别为 61%、55%、36% 和 31%, 明显优于卷积神经网络 (CNN) 以及传统机器学习算法如支持向量机 (SVM) 和 K-近邻算法 (KNN) 的预测结果, 是大雾临近预报的一种有效预报方法.

关键词: LSTM; 气象要素时间序列; 大雾; 临近预报

引用格式: 苗开超,韩婷婷,王传辉,章军,姚叶青,周建平.基于 LSTM 网络的大雾临近预报模型及应用.计算机系统应用,2019,28(5):215–219. http://www.c-s-a.org.cn/1003-3254/6889.html

Fog Nowcasting Model Based on LSTM Network and Its Application

MIAO Kai-Chao[1], HAN Ting-Ting[2], WANG Chuan-Hui[1], ZHANG Jun[2], YAO Ye-Qing[1], ZHOU Jian-Ping[1]

[1](Meteorological Bureau, Anhui Province, Hefei 230031, China)
[2](Anhui University, Hefei 230039, China)

Abstract: Long-Term and Short-Term Memory (LSTM) network is a time recursive neural network, which is suitable for predicting events with relatively long delay in time series. In this study, a new fog proximity prediction framework based on LSTM network is constructed, which can transform meteorological observation data into time series data and model them based on time series data. In order to validate the proposed model effectively, this study transforms the surface meteorological data of 81 national stations in Anhui Province from October 2015 to June 2017 into sequence data and constructs a validation data set. Based on this data set, the future 1–4 hourly fog forecasting experiments are carried out. The experimental results show that the proposed model's TS-Scores are 61%, 55%, 36%, and 31%, respectively, which are obviously better than CNN and those of traditional machine learning algorithms such as SVM and KNN. It is an effective method for fog prediction.

Key words: LSTM; time series of meteorological elements; fog; nowcasting

浓雾作为一种灾害性天气现象, 近年来受到越来越广泛的关注. 雾导致的视程障碍对交通的安全带来严重影响[1–3]. 近年来, 随着经济的发展和全球气候变暖加剧, 中国区域大气能见度整体呈下降趋势, 其中东部地区下降趋势最为明显[4–7]. 低能见度天气的增多成为诱发交通事故的主要气象因素, 如我国 2017 年道路

① 基金项目: 江苏省气象科学研究所北极阁基金 (BJG201707)
Foundation item: Beijige Fund of Jiangsu Institute of Meteorological Sciences (BJG201707)
收稿时间: 2018-11-14; 修改时间: 2018-12-10; 采用时间: 2018-12-17; csa 在线出版时间: 2019-05-01

交通事故万车死亡人数为 2.06 人[8], 成为危害人身安全的重要因素之一. 准确预报出大雾的生成、发展和消亡能有效减少交通事故的发生. 为此, 交通和气象部门开展了广泛合作, 以减少低能见度因素导致交通事故发生. 周须文等[9]应用天气学原理和数理统计方法对低能见度雾的生消机理进行研究, 建立能见度与气象因子的回归方程, 从而对雾的等级进行预报; 吴彬贵等[10]、黄政等[11]基于数值预报模式数据, 结合逆向传播 (BP) 神经网络或具体的要素阈值来判别雾是否出现, 在雾的预报方面做出了积极的探索. 也有学者在对大雾中平流雾气象要素特征分析的基础上, 给出了预报思路, 在实际应用中取得较好的效果[12]. 在已有的大雾预报研究中, 大多为短期预报 (未来 24 小时大雾是否发生), 空报和漏报率较高. 因此, 大雾临近预报显得尤为重要, 目前关于大雾的临近预报的研究相对较少.

随着人工智能的兴起, 一些深度学习算法相继出现[13–15], 基于深度学习的短期天气预测已成为一种新的趋势[16]. 深度学习算法中循环神经网络 (RNN) 是一种适合序列数据的模型[17], 能够提取时间序列数据中的有效信息, 已广泛应用于股票预测、语音识别等领域[18,19]. 长短期记忆网络 (LSTM)[20]是一种特殊的循环神经网络 (RNN), 适合处理和预测时间序列相对较长的重要事件. 与 RNN 相比, 它解决了 RNN 训练过程中梯度爆炸和梯度消失的问题, 可以学习长期的依赖信息. 近年来 LSTM 在各行业中得到较为广泛的应用[21–23]. 本文提出了一种基于 LSTM 的网络模型, 该模型通过自动提取气象要素历史数据中的相关信息, 预报未来 4 小时内逐小时能见度的变化, 是对以往利用经验预报的有效补充.

1 LSTM 大雾预报模型的设计

建模数据选取安徽省 81 个国家站 2015 年 10 月 1 日到 2017 年 6 月 1 日逐时地面气象数据. 要素包括气压、气温、露点温度、降水量、风速、风向和能见度.

1.1 预报模型的建立

LSTM 是一种 RNN 神经网络, 每个 LSTM 单元增加了三个门, 即输入门, 忘记门和输出门. 与 RNN 相比, 解决了 RNN 训练过程中梯度爆炸和梯度消失的问题, 可以学习长期的依赖信息. 对于时间序列 x_t (t= 1, 2, 3, 4, 5, ⋯), 前一个 LSTM 单元的输出结合当前时间点数据作为输入, 每个时间步都有一个输出. 同时, 存储器单元产生当前时间步长的状态向量. 图 1 中展示出了一个 LSTM 单元的存储块, 其中灰色圆圈 (中间大圆圈) 表示存储器单元, 存储了 LSTM 的状态信息, 图 1 中黑色实心圆 (小圆圈) 表示乘法, 空心圆圈表示激活函数. LSTM 工作的具体过程遵循以下公式:

$$i_t = \sigma(W^{(i)}H + b_i) \tag{1}$$

$$f_t = \sigma(W^{(f)}H_{(f)} + b_f) \tag{2}$$

$$o_t = \sigma(W^{(o)}H_{(o)} + b_o) \tag{3}$$

$$C_t = \tanh(W^{(c)}H_{(c)} + b_c) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

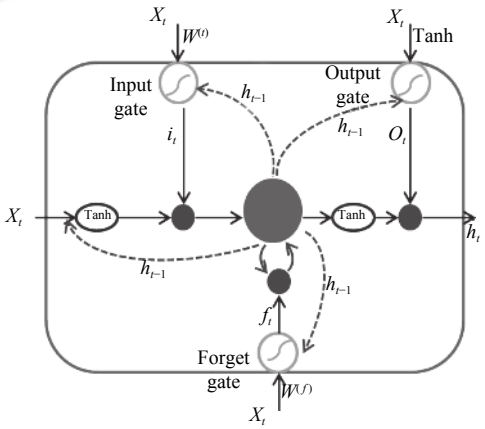

图 1 LSTM 记忆单元示意图

其中, i_t, f_t, o_t分别表示输入门, 忘记门和输出门的输出, C_t表示当前 Cell 的状态, h_t是 Cell 的输出, w_i, w_f和 w_o是输入门, 遗忘门和输出门的权重, 在每个时间步骤共享. H 表示当前输入向量x_t和前一时刻单位的输出向量h_{t-1}的叠加$H = [x_t, h_{t-1}]$. 输入门用于控制保留信息, 防止无用信息进入存储器单元. 忘记门用于决定从上一步骤中的单元状态丢弃信息. 忘记门和输入门一起来更新存储器单元的状态.

1.2 基于 LSTM 的大雾临近预报模型建立

本文基于 LSTM 模型提出了一种大雾预报框架, 基本原理如图 2 所示. 每小时返回的气象要素数据根据需要被转化成不同长度的时间序列, 并将其作为网络的输入时间序列X, 每小时返回的气象要素数据被视为一个时间步, 表示为x_t, 网络的输出是下一个时间序列中雾的类别标签. LSTM 工作的具体过程可归纳为以下步骤:

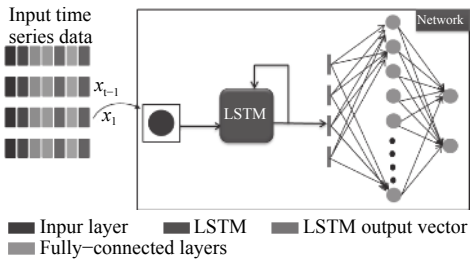

图 2 基于 LSTM 的大雾预报模型

1) 首先, 将气象要素时间序列 X 作为输入进入输入层.

2)LSTM 接收输入向量, 结合 Cell 在上一时间点的输出, 当 t=1 时, 隐藏层状态为 0. LSTM 的输入门和忘记门分别由 Sigmoid 函数值决定要输入的信息, 进入存储单元时存储单元应该忘记哪些信息. 输入门和隐藏门的输出更新隐藏层的状态,根据单元状态决定输出信息, 最终 Cell 输出当前单元.

3) 然后, 将下一时间点的气象要素 x_t 和前一时间点的单元的输出 h_{t-1} 输入 LSTM, 并重复上述过程.

4) 最后, 将 LSTM 的最后时间步的输出输入到全连接层. 全连接层进一步提取气象要素时间序列的特征, 最终输出预测大雾的类别的标签.

2 实验与分析

通过将逐时气象要素数据处理成时间序列以构建数据集, 从 2015 年 10 月到 2016 年 12 年的数据中随机选出 1500 个正样本和 1500 个负样本作为验证集, 使用剩余数据构建训练集.利用 2017 年 1 月到 2017 年 6 月的数据构建测试集. 具体制作时间序列的方式为: 预测未来小时雾是否存在, 先选择前一段时间的气象要素数据构建时间序列, 使用下一小时的能见度作为训练标签. 通常大家关心的是有雾状态, 因此标记时间序列的标准是当能见度值小于 1000 米时, 标记为有雾, 否则, 则标记为无雾. 例如, 第 1 至第 4 个小时为一个时间序列, 第 5 个小时的能见度值则为标签. 第 2 至第 5 个小时作为时间序列, 用第 6 各小时能见度值构建标签. 在预测未来 1–2 小时能见度时, 选择的时间序列长度为 2, 预测 3–4 小时能见度时选择的时间序列长度为 4. 训练样本和测试样本的数量见表 1. 在样本中, 有雾的样本相对较少, 这就造成了正负样本比例严重不平衡, 其比例约为 1:20. 本文使用随机过采样的方法对训练正样本进行数据扩充, 使最终正负样本比例为 1:2 (在本文中, 正样本为有雾, 负样本为无雾.). 测试集按照真实情况下取出的正负样本的数量. 当预测未来 1 到 2 小时是否有雾时我们选择的时间序列长度 2. 预测 3 到 4 小时选择时间序列的长度为 4.

表 1 训练集、验证集和测试集正负样本数量

	选择的时间序列长度	预测时长	训练集	验证集	测试集
正样本	2	1,2	361500	1500	10246
负样本	2	1,2	764 341	1500	2 282 373
正样本	4	3,4	396 400	1500	9399
负样本	4	3,4	761 605	1500	269 430

由于各气象要素因子数值区间的差异性较大, 在数据使用前对原始数据做归一化处理, 将各个因子缩放到一个尺度, 公式如下:

$$\overset{*}{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

式中, $\overset{*}{x}$ 为某类气象因子归一化后的数据. x 为每一类气象因子原始数据. $x_{\max}$ 为每一类气象因子的最大值, $x_{\min}$ 为每一类气象因子的最小值.

为了评估训练模型的性能, 本文选用 *precision*, *F*1-score, *accuracy* 和 *TS-Score* 作为评价指标[24,25], 指标公式如下:

$$precision = TP/(TP + FP) \tag{8}$$

$$F1 = 2 * precision * recall/(precision + recall) \tag{9}$$

$$recall = TP/(TP + FN) \tag{10}$$

$$accuracy = (TP + TN)/(TP + TF + NP + NF) \tag{11}$$

$$TS = TP/(TP + FP + FN) \tag{12}$$

其中, *TP* 是正样本被正确分类的样本数, *TN* 是负样本分类正确的样本数量; *recall* 是指正样本被正确分类数量与总正样本的比率; *FN* 指正样本被分类为负样本的数量; *FP* 是负样本被分类为正样本数量; *TS-Score* 则是一种气象部门广泛用来评价预测效果的指标.

本文使用了交叉熵做为雾预测的目标函数, 目标函数如下:

$$loss = -\sum_{i}^{K} y_i \ln f_i(x) \tag{13}$$

式中, y_i 代表真实标签, $f_i(x)$ 代表全连接层的输出值, i 代表全连接层输出向量的第 i 个值. 本文利用长度为 2 的时间序列做未来 1–2 小时大雾预报过程中, 使用 LSTM 的层数为 3 层, *Cell* 的数目为 50 个, 全连接层数为 2 层, 2 层全连接的卷积核数分别为 100. 当使用

长度为 4 的时间序列预测未来 3–4 小时大雾是否存在时, LSTM 的层数选择 1 层, *Cell* 的数目选择 70 个, 全连接层为 1 层, 节点数目为 200 个. 利用上述数据集对模型进行训练, *loss* 曲线如图 3 所示.

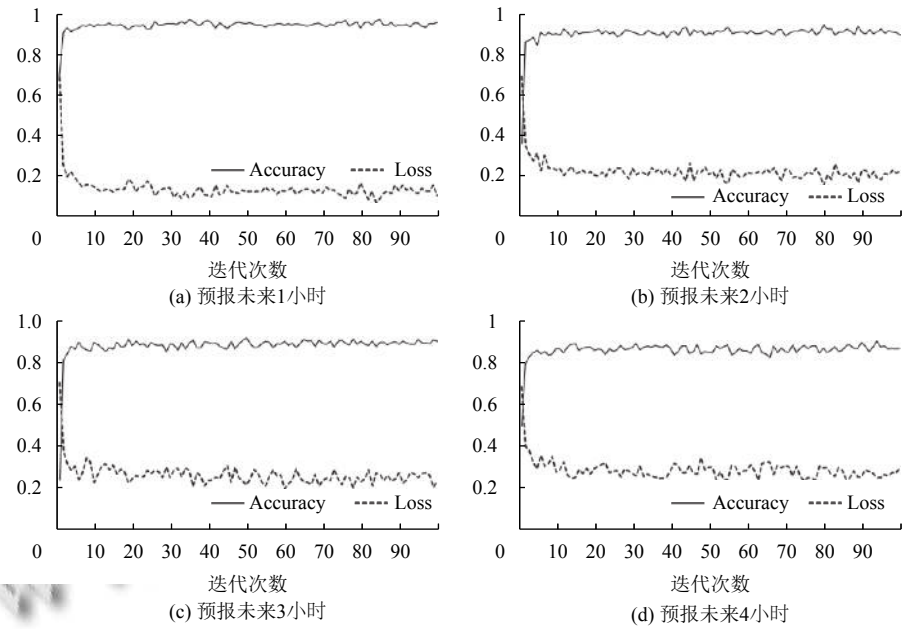

图 3 训练 *loss* 与 *accuracy* 曲线 (横坐标单位: 100)

设计的实验应用于 Intel i7 3.4 Ghz*8, 16 GB 内存和英伟达 GTX1080. Ubuntu16.04 操作系统, 网络框架是基于 Tensorflow1.4.0[26].

表 2 给出了 LSTM 与 CNN 以及传统的 SVM、KNN 对未来 1–4 小时的大雾预测结果的对比. 从表 2 中可以看出在对未来 1–4 小时的预测, LSTM 在四种评估标准的结果优于 SVM、KNN 以及 CNN 方法的预测结果. 用该方法对未来 4 小时预测的 TS-score 分别为 61%, 55%, 36% 和 31%. 由此可以发现基于 LSTM 的预测模型相比于 CNN 以及传统的分类模型 SVM、KNN, 在预测精度上有明显优势.

表 2 LSTM、SVM、KNN 和 CNN 对未来 4 小时大雾预测结果的各项指标对比

预报时长	1 小时				2 小时				3 小时				4 小时			
模型类型	LSTM	SVM	KNN	CNN	LSTM	SVM	KNN	CNN	LSTM	SVM	KNN	CNN	LSTM	SVM	KNN	CNN
F1-score	0.76	0.60	0.60	0.51	0.66	0.51	0.50	0.49	0.53	0.43	0.40	0.44	0.47	0.40	0.36	0.33
PRE	0.64	0.71	0.48	0.38	0.51	0.42	0.38	0.35	0.39	0.29	0.29	0.34	0.34	0.26	0.73	0.21
TS-score	0.61	0.42	0.43	0.34	0.55	0.33	0.33	0.32	0.36	0.25	0.25	0.28	0.31	0.25	0.22	0.20

3 结语

本文基于 LSTM 网络提出了一个新的临近大雾预报框架, 与传统的大雾预报方法不同, 该框架基于气象要素时间序列数据进行建模. 利用安徽省国家地面气象观测站气象要素转换的时间序列数据, 该框架能够有效地预测未来 1–4 小时的大雾生成情况. 对比分析发现, 和当前比较常见的 CNN、SVM、KNN 等其他机器学习方法相比较, 本文提出的预测框架能够达到更好的预测效果.

参考文献

1 吴兑, 邓雪娇, 毛节泰, 等. 南岭大瑶山高速公路浓雾的宏微观结构与能见度研究. 气象学报, 2007, 65(3): 406–415. [doi: 10.3321/j.issn:0577-6619.2007.03.009]

2 李秀连, 陈克军, 王科, 等. 首都机场大雾的分类特征和统计分析. 气象科技, 2008, 36(6): 717–723. [doi: 10.3969/j.issn.1671-6345.2008.06.008]

3 崔新强, 周小兰, 付佳, 等. 高速铁路安全运行高影响天气条件等级标准研究. 灾害学, 2016, 31(3): 26–30. [doi: 10.3969/j.issn.1000-811X.2016.03.005]

4 黄健, 吴兑, 黄敏辉, 等. 1954—2004 年珠江三角洲大气能见度变化趋势. 应用气象学报, 2008, 19(1): 61–70. [doi: 10.3969/j.issn.1001-7313.2008.01.009]
5 范引琪, 李春强. 1980—2003 年京、津、冀地区大气能见度变化趋势研究. 高原气象, 2008, 27(6): 1392–1400.
6 刘晓舟, 许潇锋, 杨军. 华东三市能见度、气溶胶和太阳辐射变化特征. 气象科技, 2013, 41(2): 352–359. [doi: 10.3969/j.issn.1671-6345.2013.02.027]
7 刘骞, 盛立芳, 王园香, 等. 气象要素对中国大气能见度长期变化影响的定量研究. 气候与环境研究, 2016, 21(1): 47–55.
8 中华人民共和国国家统计局. 中华人民共和国 2017 年国民经济和社会发展统计公报. 人民日报, 2018-03-01(10).
9 周须文, 时青格, 贾俊妹, 等. 低能见度雾的分级预报方法研究. 热带气象学报, 2014, 30(1): 161–166. [doi: 10.3969/j.issn.1004-4965.2014.01.018]
10 吴彬贵, 张建春, 李英华, 等. 天津港秋冬季低能见度数值释用预报研究. 气象, 2017, 43(7): 863–871.
11 黄政, 袁成松, 包云轩, 等. 基于不同参数化方案的高速公路大雾过程的数值模拟试验. 气象, 2016, 42(8): 944–953.
12 许爱华, 陈翔翔, 肖安, 等. 江西省区域性平流雾气象要素特征分析及预报思路. 气象, 2016, 42(3): 372–381.
13 崔广新, 李殿奎. 基于自编码算法的深度学习综述. 计算机系统应用, 2018, 27(9): 47–51. [doi: 10.15888/j.cnki.csa.006542]
14 童基均, 常晓龙, 赵英杰, 等. 基于深度学习的运动目标实时识别与定位. 计算机系统应用, 2018, 27(8): 28–34. [doi: 10.15888/j.cnki.csa.006525]
15 张骥, 余娟, 汪金礼, 等. 基于深度学习的输电线路外破图像识别技术. 计算机系统应用, 2018, 27(8): 176–179. [doi: 10.15888/j.cnki.csa.006458]
16 许小峰. 从物理模型到智能分析——降低天气预报不确定性的新探索. 气象, 2018, 44(3): 341–350.
17 Graves A. Generating sequences with recurrent neural networks. arXiv:1308.0850, 2013.
18 Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. Proceedings of the 31st International Conference on Machine Learning. Beijing, China. 2014. 1764–1772.
19 Sundermeyer M, Ney H, Schlüter R. From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(3): 517–529. [doi: 10.1109/TASLP.2015.2400218]
20 Graves A. Long short-term memory. Graves A. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Heidelberg: Springer, 2012. 37–45.
21 於雯, 周武能. 基于 LSTM 的商品评论情感分析. 计算机系统应用, 2018, 27(8): 159–163. [doi: 10.15888/j.cnki.csa.006483]
22 史梦飞, 杨燕, 贺樑, 等. 基于 Bi-LSTM 和 CNN 并包含注意力机制的社区问答问句分类方法. 计算机系统应用, 2018, 27(9): 157–162. [doi: 10.15888/j.cnki.csa.006536]
23 曹国清, 张晓明, 陈亚峰. 基于 PCA-LSTM 的多变量矿山排土场滑坡预警研究. 计算机系统应用, 2018, 27(11): 252–258. [doi: 10.15888/j.cnki.csa.006646]
24 李胜宇, 高俊波, 许莉莉. 面向酒店评论的情感分析模型. 计算机系统应用, 2017, 26(1): 227–231. [doi: 10.15888/j.cnki.csa.005511]
25 赵熙, 李京萌, 童红梅. 济南机场低能见度和低跑道视程对比分析. 干旱气象, 2017, 35(5): 847–856.
26 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467, 2016.