唐远志, 陈清乐, 周园, 等. 基于改进的 BiLSTM 网络能见度预测研究[J]. 智能计算机与应用,2024,14(5):241-246. DOI: 10.20169/j.issn.2095-2163.240534

基于改进的 BiLSTM 网络能见度预测研究

唐远志[1], 陈清乐[2], 周 园[3], 苏静文[1], 李 丽[3], 廖 波[4]

(1 贵州省气象台, 贵阳 550002; 2 贵州新气象科技有限责任公司, 贵阳 550002; 3 贵州省习水县气象局, 贵州 遵义 564500; 4 贵州省气象服务中心, 贵阳 550002)

摘 要: 能见度对人们日常生活、生产方面的影响越来越大,因此实现对能见度的精确预测就显得尤为重要。近年来很多研究者利用人工智能技术对能见度进行预测,但都聚焦于逐小时数据、精细化程度较低。为了提升能见度预测的精度,本文提出了 IM_BiLSTM_Attention 网络。一方面,获取大量的逐分钟气象的数据,并计算 Spearman 相关系数,衡量其与能见度的相关性;另一方面,引入稀疏注意力机制对 BiLSTM 网络进行改进,进而选择性地关注时间序列中的重要信息以减少注意力分散和噪声数据干扰,提高了能见度预测的精度。通过在数据集上的实验结果表明,IM_BiLSTM_Attention 在逐分钟能见度预测问题上效果更优。

关键词: 能见度预测; 气象因子; 稀疏注意力; IM_BiLSTM_Attention

中图分类号: TP391;P457 **文献标志码**: A **文章编号**: 2095-2163(2024)05-0241-06

Research on visibility prediction based on improved BiLSTM network

TANG Yuanzhi[1], CHEN Qingle[2], ZHOU Yuan[3], SU Jingwen[1], LI Li[3], LIAO Bo[4]

(1 Meteorological Observatory of Guizhou Province, Guiyang 550002, China; 2 Guizhou New Meteorological Technology Co., Ltd., Guiyang 550002, China; 3 Xishui Meteorological Bureau, Zunyi 564500, Guizhou, China; 4 Meteorological Service Center of Guizhou Province, Guiyang 550002, China)

Abstract: Visibility has an increasing impact on people's daily life and production, so it will be especially important to realize the accurate prediction of visibility in a more refined way. In recent years, many researchers have utilized artificial intelligence techniques for visibility prediction, but all of them focus on the hour-by-hour data with a low degree of refinement. In order to improve the accuracy of visibility prediction, this paper proposes IM_BiLSTM_Attention network. On one hand, obtaining a large amount of minute by minute meteorological data and calculating Spearman correlation coefficients to measure their correlation with visibility; on the other hand, the BiLSTM network is improved by introducing the sparse attention mechanism, which selectively focuses on the important information in the time series to reduce distraction and interference from noisy data, and improves the accuracy of visibility prediction. The experimental results on four datasets show that the IM_BiLSTM_Attention is more effective in the minute-by-minute visibility prediction problem.

Key words: visibility prediction; weather factor; sparse attention; IM_BiLSTM_Attention

0 引 言

能见度是指人类视觉系统在大气中所能识别物体的距离,在气象学中常用来描述可见光下能够从地面或水平面上方看到的最远距离,常用于评估雾、霾、沙尘暴等灾害天气的影响范围[1-2]。现今,交通出行、航空和海运的安全性和高效性在很大程度上受到天气因素的影响,尤其是雾天能见度较低的情境极易造成较为严重的交通事故,这将对公众的生命财产安全造成极为严重的损失,可见能见度与人们的生活息息相关。因此探索大雾演化规律,预测大雾变化趋势成为一个重要的课题[3-4]。

一般来说,对于能见度的预报主要分为数值模式预报和统计预报两种传统方法。对于数值模式预报来说,主要是通过使用各类气象数据和大气环境监测数据建立气象数值模型,对大气能见度进行预

基金项目: 贵州省科技厅科技支撑计划项目(黔科合支撑[2022]一般 286)。

作者简介: 唐远志(1994-),男,助理工程师,主要研究方向:公共气象服务和预警研究。

通讯作者: 苏静文(1976-),女,高级工程师,主要研究方向:公共气象服务和预警研究。Email: 554910503@qq.com

收稿日期: 2024-01-31

测[5-6]。而统计预报则是以气象要素对能见度的影响关系为基础进行大气能见度的预测[7-8]。近年来,随着以人工智能技术的飞速发展,许多学者使用 KNN、GBRT、SVM 和 XGBoost 等机器学习算法进行预测研究且取得优异结果,这是因为时间序列数据可能存在高度复杂的非线性关系,相比传统的方法、机器学习技术在处理时序数据时有着强大的非线性建模能力,这可以很好地提高预测准确性[9-10]。比如,余东昌等学者[11]通过 LightGBM 算法对逐小时能见度进行预测。邓拓[12]通过 LSTM 神经网络进行机场能见度预测。方楠等学者[13]通过长短期记忆神经网络(LSTM)模型对低能见度天气做预报研究。陆冰鉴等学者[14]基于相关性分析和数据均衡构建能见度分层预测模型。但是这些研究聚焦逐小时数据甚、至逐日数据,精细化程度较差。

因此,本文以 BiLSTM 为主干网络,引入稀疏注意力机制并进行改进,构建 IM_BiLSTM_Attention 模型。并以低能见度天气频发的冬、春季为对象,选取能见度、温度、相对湿度、风速、风向、水汽压、气压、露点温度、蒸发和降水等 10 项气象因子的逐分钟数据制作数据集,对能见度进行预测实验。通过实验,分析发现本文改进的 IM_BiLSTM_Attention 模型可以显著提高习水能见度预测的准确率,降低能见度预测的误差。

1 资料与方法

能见度作为一个复杂的气象概念,气象因子和环境因子都会对其变化造成影响。通过查阅习水县 2022 年生态环境政务信息,发现习水全年监测空气质量优良率为 98.9%,可见当地大气受环境污染程度极低,因此本研究仅聚焦气象因子对能见度进行预测研究。对于气象因子数据的选取,考虑到在西南山区能见度变化速度非常快,常常几分钟到几十分钟就能变化几百米、甚至上千米,当在极端天气情况下甚至能变化数千米。因此,本文采用逐分钟气象数据对能见度数据展开研究。最终获取到的数据为习水县国家基本气象站的实际观测数据,包含能见度(*Vis*)、温度(*Tem*)、相对湿度(*Re_hum*)、风速(*wnd_spd*)、风向(*wnd_dir*)、水汽压(*Wat - v_pre*)、气压(*Air_ pre*)、露点温度(*Rew_ p*)、蒸发(*Evaporate*)和降水(*Rain*)这 10 个气象因子,总共有 892 800 条数据。

在众多的研究中,通过计算相关系数去衡量各个因子和能见度之间的相关性,筛选出对能见度影响较高的影响因子,进而能够更好地开展能见度预测研究[15-16]。对于气象因子数据相关系数的构建,Pearson 相关系数常常需要确定数据间的线性关系或正态分布特性,但是这些气象数据因子与能见度之间并一定完全是呈现线性关系和正态分布的,如若简单地按照线性模型计算相关系数,可能导致最终的计算结果误差较大。所以,本文采用 Spearman 相关系数计算相关性,计算公式见式(1):

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{1}$$

对于能见度与温度、相对湿度、风速、风向、温度、相对湿度、水汽压、气压、露点温度和降水这 10 个气象因子数据之间的相关性计算结果如图 1 所示。

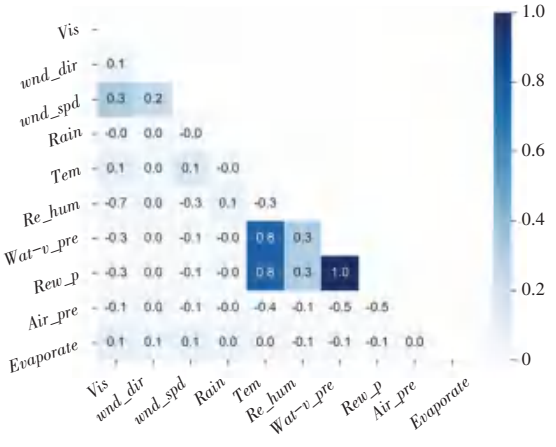


图 1 各气象因子相关性分析热力图

Fig. 1 Heat map of correlation analysis of each meteorological factor

通过 Spearman 相关系数的计算结果可以看出,相对湿度、降水、露点温度、水气压和气压与能见度呈负相关,温度、风速、风向和蒸发与能见度呈正相关,这一结论也和其他研究者类似[14-15]。

2 能见度预测的模型

2.1 LSTM 网络模型及其演进

LSTM 网络模型的基本成分是一个单元(Cell),利用遗忘门、输入门和输出门等机制来管理历史信息、当前输入和输出状态,并对单元状态进行调整;当 *Sigmoid* 函数输出为 0,信息完全舍去;当 *Sigmoid* 函数输出为 1,则信息完全保留[17-18]。研究给出阐释分述如下。

(1)遗忘门。用于计算信息的丢弃或保留状

态，通过 *Sigmoid* 处理后为 0 到 1 的值；可由式(2)进行描述：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}\ x_t] + b_f) \tag{2}$$

(2)输入门。计算哪些信息保存到单元状态。主要有 2 部分信息。一部分是 i_t，当前有多少信息需要保存到单元状态中，可由式(3)表示为：

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

另一部分是 $C_t^{'}$，用于将当前输入所产生的信息加入到单元状态中，并以这 2 部分产生新的记忆状态；可由式(4)表示为：

$$C_t^{'} = \tanh(\boldsymbol{W}_c \cdot [h_{t-1}\ x_t] + b_c) \tag{4}$$

此刻，当前单元状态是由遗忘门控制的历史信息和上一时刻状态的乘积，加上输入门两部分积，即：

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t^{'} \tag{5}$$

(3)输出门。用来计算当前时刻信息被输出程度。可由式(6)、式(7)进行描述：

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}\ x_t] + b_o) \tag{6}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{7}$$

其中，t 表示时刻；C_t 表示记忆单元；h_t 表示状态；x_t 表示输入；f_t 表示遗忘门；i_t 表示输入门；o_t 表示输出门；b_f、b_i、b_c、b_o 分别表示各自的偏差。

但在实际的时间序列预测任务中，不仅仅需要考虑长序列历史数据之间的依赖性，还需要关注当前位置上下文信息的双向依赖性，因此有研究者提出了 BiLSTM 模型。一般来说，BiLSTM 模型是在输入序列中引入了前向和后向两个方向的 LSTM 层，再将状态做求和操作，得到最后的隐状态序列，以捕获不同时间步的上下文信息[16]。状态求和的数学公式可写为：

$$h_t = [\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{8}$$

其中，“$\oplus$”表示向量逐元素求和，$\overrightarrow{h}_t$ 和 $\overleftarrow{h}_t$ 分别表示每个 t 时刻前向、后向的隐状态。可见，BiLSTM 能够同时利用当前位置之前和之后的信息，从而在时间序列预测任务中获得更全面的上下文信息。

除此之外，对于长序列的任务，LSTM 和 BiLSTM 可能会面临长期依赖问题，即较早时刻的输入对当前时刻的预测或输出产生较小的影响；并且 BiLSTM 捕捉上下文信息并不总能够确定哪些上下文是最相关的。因此有研究者试图对 LSTM 和 BiLSTM 模型加入注意力机制，进而解决上述问题，并进一步提高模型的可解释性和基于输入位置的重要性给予不同的权重分配。

2.2　IM_BiLSTM_Attention 模型

一般 BiLSTM_Attention 对信息不敏感、且计算复杂度高，因此通过引入稀疏注意力机制并进行必要的改进生成 IM_BiLSTM_Attention 模型。IM_BiLSTM_Attention 模型结构如图 2 所示。

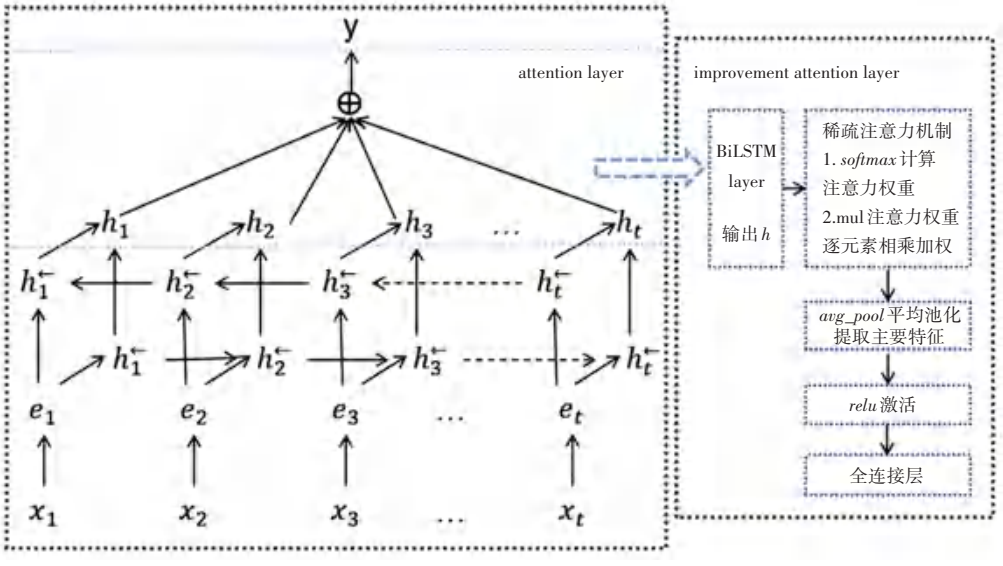


图 2　IM_BiLSTM_Attention 结构

Fig. 2　IM_BiLSTM_Attention structure

对于稀疏注意力层，见图 2 右侧所示。首先，研究令 H 为 LSTM 层前后向输出拼接的隐状态矩阵 $[h_1\ h_2, \cdots, h_t]$。采用函数的形式表示为：

$$H = LSTM(h) \tag{9}$$

再通过 *softmax* 对输入向量进行非线性变换，将原始的实数向量转化为概率分布，得到注意力权重：

$$att\_weights = softmax(H) \tag{10}$$

接着将 LSTM 层的输出与注意力权重 *att_weights* 进行矩阵乘法，得到加权后的注意力，这样可以使模型更关注重要的时间步并提供更准确的预测结果。对此可以表示为：

$$soft\_att = H \times att\_weights^{\mathrm{T}} \tag{11}$$

接着，对序列特征进行平均池化操作，并通过 *Relu* 函数进行激活、对负值数据进行过滤，进一步剔除不重要的信息降低计算的复杂度。这一计算方法如式(12)所示：

$$act = Relu(avg\_pool(soft\_att)) \tag{12}$$

最后，通过全连接层对激活后的张量进行线性变换，得到与目标维度匹配的新张量，再输出最终结果，即：

$$out = fc(linear(act)) \tag{13}$$

总地来说，IM_BiLSTM_Attention 能够利用稀疏注意力机制关注与当前时间步相关的重要信息，忽略了与当前时间步不相关的时间步，极大地减少了长序列的依赖性和计算量，提高计算效率，并且允许模型选择性地关注时间序列中的重要信息以提供更好的解释性，从而提高模型的预测精度和鲁棒性。

3 实验部分

3.1 数据集介绍和数据预处理

本研究采用习水县国家基本气象站的逐分钟气象数据资料，分别取冬、春两个低能见度天气频发的季节中各一个月的逐分钟观测数据，其训练集为各月前 28 天，测试集为后 3 天，数据集的具体划分参见表 1。

表 1 数据集划分

Table 1 Division of dataset

季节	时段范围	数据量
冬季	训练集：2021 年 12 月 31 日 20:01 至 2022 年 1 月 28 日 17:36	40 176 * 10
	测试集：2022 年 1 月 28 日 17:37 至 2022 年 1 月 31 日 20:00	4 464 * 10
春季	训练集：2022 年 5 月 1 日 20:01 至 2022 年 5 月 29 日 17:36	40 176 * 10
	测试集：2022 年 5 月 29 日 17:37 至 2022 年 6 月 1 日 20:00	4 464 * 10

实际情况中，各数据集中往往部分数据会存在离群值、空值和非逻辑值的情况。对于离群值、非逻辑值，研究应用了极大似然估计的概念[19]，这意味着数据值不在 $\mu \pm 2\sigma$（在正态分布假设性下，这包含了 95%以上的数据）区间内将会被标记为离群值，然后通过插值法对离群值进行替代；对于空值，利用 KNN 进行邻近均值计算和填充，进而确保了数据的完整性。并通过 min-max 做数据标准化处理，就是将实数范围数值映射到一定区间上，目前主要缩放的区间为 [0,1] 或 [-1,1]，能见度不存在负值本研究缩放区间为 [0,1]。对于序列 $x_1, x_2, \cdots, x_n$，本文 min-max 标准化的计算为：

$$x_i^{'} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{14}$$

其中，$x_i^{'}$ 表示 min-max 标准化后的结果。

3.2 评价指标设置

通过评价指标去衡量时间序列预测模型性能是比较常见的方法，为了评价本文能见度模型的预测性能，研究选用平均绝对误差（Mean Absolute Error，*MAE*）、平均相对误差（mean relative error，*MRE*）、均方根误差（Root Mean Square Error，*RMSE*）和决定系数（R^2）[20]，其计算公式具体如下：

$$MAE(y, y^{'}) = \frac{1}{m}\sum_{i=1}^{m}(|y_i - y_i^{'}|) \tag{15}$$

$$MRE(y, y^{'}) = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - y_i^{'}}{y_i}\right| \times 100\% \tag{16}$$

$$RMSE(y, y^{'}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - y_i^{'})^2} \tag{17}$$

$$R^2(y, y^{'}) = 1 - \frac{\frac{1}{m}\sum_{i=1}^{m}(y_i - y_i^{'})^2}{\frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{18}$$

其中，m 表示预测序列的时间长度；y_i 表示实际的能见度数据；$y_i^{'}$ 表示预测的能见度数。为了评估本研究模型的性能，将 IM_BiLSTM_Attention 与 BiLSTM 和 BiLSTM_Attention 三种模型预测结果进行 4 种评价指标的对比。

3.3 实验与结果分析

在本研究中根据多次重复的预实验，并逐步调整优化参数，最终选取用于本文能见度预测研究的IM_BiLSTM_Attention 网络模型实验参数见表 2。

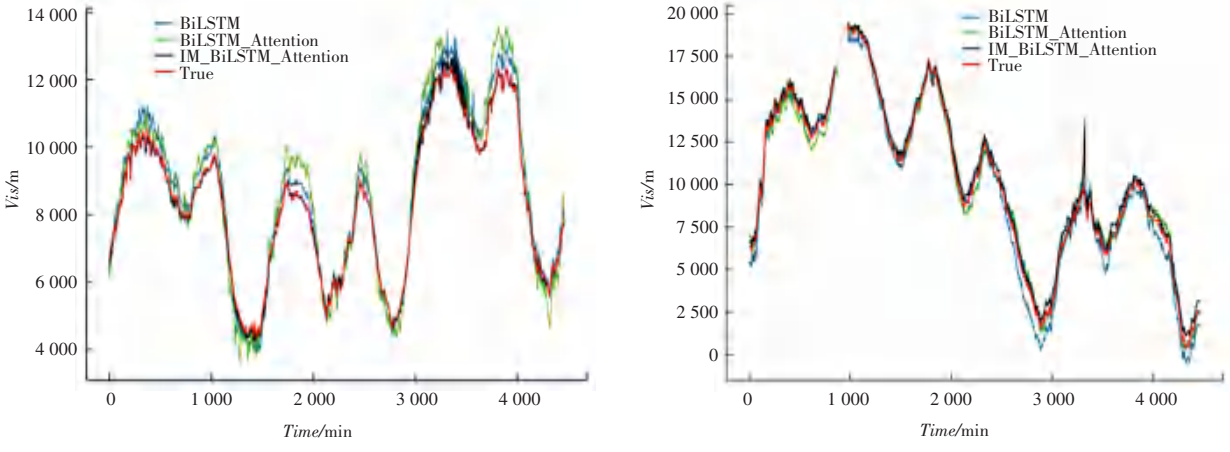
表 2　实验参数设置

Table 2　Experimental parameters settings

参数	参数描述	实验参数值
Learning_Rate	学习率	0.01
Batch_Size	批处理量	64
Epoch	周期数	10
num_layers	LSTM 的层数	7
hidden_dim	隐藏层 1 大小	64
hidden_dim2	隐藏层 2 大小	32

基于上述参数，研究通过 BiLSTM、BiLSTM_Attention 和 IM_BiLSTM_Attention 这 3 种网络模型分别对表 2 所述数据集进行实验，并与习水国家基本气象站实际观测值进行对比，进而衡量对习水能见度预测的有效性。

对于冬季能见度的预测实验，将 2021 年 12 月 31 日 20:01 至 2022 年 1 月 28 日 17:36 的能见度、相对湿度、降雨、露点温度、水气压、气压、温度、风速、风向和蒸发等 10 项气象因子数据输入模型进行模型训练；训练后进行预测，并与实际值进行对比。对于春季能见度预测实验，其数据集划分和输入输出采用和冬季同样的方式。最终预测结果如图 3 所示。各模型对冬、春季能见度的预测效果值见表 3。



(a) 冬季　　(b) 春季

图 3　各模型对冬、春季能见度的预测值与实际值对比

Fig. 3　Comparison between predicted values and actual values of winter and spring visibility for various models

由图 3 可知，3 种模型都能对习水冬、春季能见度有较好的预测，能在一定程度上反应习水冬、春季能见度的真实变化。研究可知 IM_BiLSTM_Attention 效果最优、BiLSTM 和 BiLSTM_Attention 各有优势。但是，IM_BiLSTM_Attention 模型更能反映低能见度下的情况，能够很好地捕捉低能见度的变化趋势。

表 3　各模型对冬、春季能见度的预测效果

Table 3　Prediction effects of various models on winter and spring visibility

季节	Model	*MAE*	*MRE*/%	*RMSE*	R^2
冬季	BiLSTM	366.61	4.24	418.64	0.862
	BiLSTM-Attention	493.96	5.92	601.03	0.822
	IM-BiLSTM-Attention	94.18	1.19	125.09	0.896
春季	BiLSTM	455.54	11.04	609.85	0.873
	BiLSTM-Attention	353.90	6.31	463.94	0.890
	IM-BiLSTM-Attention	349.36	5.19	441.26	0.916

对比表 3 的实验结果，可见 IM_BiLSTM_Attention 的预测效果最优，其 *MAE* 和 *RMSE* 均相较另外 2 种模型有较大降低，*MRE* 和 R^2 均为较好的数值水平。以春季能见度的预测为例，相较于另外 2 种模型，其 *MRE* 分别下降 5.85%和 1.12%；R^2 分别上升 0.43 和 0.40。毫无疑问，基于 IM_BiLSTM_Attention 的预测方法在 4 个评价指标上都优于其他方法。

4 结束语

由于能见度受多种因素的影响且随时间变化极快，因此以逐分钟数据对能见度做预测研究就显得尤为重要，本文通过习水国家气象站观测资料对不同模型做对比实验可以得出如下结论：

(1)IM_BiLSTM_Attention 模型，更加关注与当前时间步相关的重要信息，忽略了与当前时间步不相关的时间步，极大地减少了长序列的依赖性和计算量、提高了计算效率，增强了模型的泛化能力，防止了模型过拟合现象；

(2)使用 *RMSE*、*MRE*、*MAE* 和 R^2 作为评价指标，通过常见的时间序列预测模型及 IM_BiLSTM_Attention 模型对比实验表明，本文的方法最优，能够较好地预测不同季节的逐分钟能见度变化曲线。

该模型虽然能够有效提高各时段内能见度预测结果，但在对于能见度的极值区域的预测准确性上还有待提高，在后续的工作中，将融合地形空间特征更综合地考虑实时能见度的变化。

参考文献

[1] 田心童. 基于时空优化 LSTM-Adaboost 模型的区域能见度预测研究[D]. 南京：南京信息工程大学，2022.

[2]刘新，刘林春，尤莉. 内蒙古呼包鄂地区近 56 年来大气环境容量变化特征分析[J]. 气象与环境科学，2019，42(1)：86-92.

[3]马祖胜，李汉彬，于平. 雾对河源高速公路交通的影响[J]. 广东气象，2006(4)：61-62.

[4]刘丹枫，施佳驰，李青松. 大气能见度及影响因子特征分析[J]. 区域治理，2020(14)：158.

[5]CHEN Renjie，WANG Xi，MENG Xia，et al. Communicating air pollution-related health risks to the public：An application of the Air Quality Health Index in Shanghai，China[J]. Environment International，2013，51(1)：168-173.

[6]朱凯全，张宏伟，张兰. 大气环境多尺度数值模式系统及其应用[J]. 农业灾害研究，2014，4(8)：38-41，43.

[7]黄亮，肖鹏飞，薛梅，等. 基于多尺度融合网络的高速公路能见度估计[J]. 气象科学，2022，42(5)：668-675.

[8]康志明，桂海林，花丛，等. 国家级环境气象业务现状及发展趋势[J]. 气象科技进展，2016，6(2)：64-69.

[9]续长青，王永忠. 基于 KNN 算法的双流机场航班延误时间预测研究[J]. 信息技术与信息化，2019(2)：81-84.

[10]朱国栋. 基于 SVM 方法的乌鲁木齐国际机场多要素预测[J]. 沙漠与绿洲气象，2022，38(5)：34-41.

[11]余东昌，赵文芳，聂凯，等. 基于 LightGBM 算法的能见度预测模型[J]. 计算机应用，2021，41(4)：1035-1041.

[12]邓拓. 基于 LSTM 神经网络的机场能见度预测[D]. 济南：山东大学，2019.

[13]方楠，谢国权，阮小建，等. 长短期记忆神经网络(LSTM)模型在低能见度预报中的应用[J]. 气象与环境学报，2022，38(5)：34-41.

[14]陆冰鉴，王兴，詹少伟，等. 基于相关性分析和数据均衡的能见度分层预测模型[J]. 计算机应用与软件. 2022，39(8)：181-186.

[15]姜江，郭文利，王春玲. 2007~2015 年北京地区能见度时空变化特征[J]. 气象与环境学报，2019，35(1)：45-52.

[16]纪德洋，金锋，冬雷，等. 基于皮尔逊相关系数的光伏电站数据修复[J]. 中国电机工程学报，2022，42(4)：1514-1523.

[17]刘磊. 基于卷积神经网络和循环神经网络的时间序列分类算法研究[D]. 长春：东北师范大学，2020.

[18]赵宏，王乐，王伟杰. 基于 BiLSTM-CNN 串行混合模型的文本情感分析[J]. 计算机应用，2020，40(1)：16-22.

[19]周杰，刘三阳. 条件自回归极差模型的对数正态拟极大似然估计[J]. 西安电子科技大学学报，2007，34(5)：828-834.

[20]骆黎明. 基于机器学习树模型的 GNSS-R 海面风场反演研究[D]. 北京：中国科学院大学(中国科学院国家空间科学中心)，2019.