

李春涵,孙存永,毕陟,等. 基于机器学习的淄博市浓雾预报研究[J]. 陕西气象,2024(4):32-38.

文章编号:1006-4354(2024)04-0032-07

基于机器学习的淄博市浓雾预报研究

李春涵^{1,2},孙存永³,毕陟²,李君²,韩国泳²

(1. 山东省气象防灾减灾重点实验室,济南 252006;

2. 淄博市气象局,山东淄博 255048;3. 郑州市气象局,郑州 450005)

摘要:雾是诱发交通事故的主要气象因素,具有很强的局地性。统计分析表明,淄博市的浓雾多发生在秋冬季节和后半夜,弱的东风、北风分量和高相对湿度时有利于浓雾发生。选取 2015—2023 年淄博市 8 个国家级气象观测站逐小时气象观测资料,以 AdaBoost 算法为基础建立能见度诊断预报模型,针对单站进行浓雾诊断测试。试验结果表明:各站点机器学习模型预报的精确率基本在 90% 以上,空报率、漏报率基本在 10% 以下,大多数空报和漏报出现在能见度阈值附近;模型对较强浓雾的预报较为准确,而对持续轻雾的空报率较高;对模型预报贡献最大的气象要素为相对湿度,其次是风向风速;模型有一定的局地性,存在优化空间。

关键词:浓雾;能见度;机器学习;AdaBoost;预报模型

中图分类号:P457.7: TP181

文献标识码:A

雾是一种局地性很强的低能见度天气现象,在华东地区四季均可见^[1]。近年来,随着经济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^[2]。淄博市处于鲁中山地和华北平原的过渡地带,地形多山、多丘陵,整体地形南高北低。淄博北部是山东大雾天气易发区,而南部山区的大雾天气则明显少于周边地区^[3]。研究表明,山东内陆地区多局地性和小范围的辐射雾,主要发生在秋冬季节^[4]。辐射雾的发生主要由于下垫面夜间辐射冷却,使空气中水汽凝结而成雾,伴随近地面逆温、微风以及水汽辐合,有很强的局地性^[5],即相似的气象条件在不同的地形、下垫面作用下亦有可能产生强度差异较大的大雾天气。雾的增多是诱发交通事故的主要气象因素之一,如何准确及时地对大雾天气进行预报,是气象领域重要的研究方向。当前对于雾的预报,大部分是通过发生大雾时的天气特征进行总结归纳,建立经验

预报模型,具有较强的主观性^[6]。或是借助数值模式进行预报,但目前数值预报产品对于雾的预测效果并不能满足社会发展需求^[7]。在已有的大雾预报研究中,大多为短期预报(即未来 24 小时大雾是否发生),其空报率和漏报率均较高^[8];因此预报人员进行雾的预报时,往往采用的是主观经验为主、模式预报为辅的经验预报模型。

机器学习方法是研究在计算机上从数据中产生“模型”的算法,即学习算法^[9]。机器学习在气象领域的应用从 20 世纪 80 年代开始,目前已在短期雷暴、强对流天气、降水、海雾等方面的预报中取得了较好的应用成果^[10-15]。相比传统预报方法,机器学习方法能够抓取传统预报方法难以描述的细节,寻找预报因子与预报量之间的非线性关系,同时略过大量复杂计算过程,提升计算效率。目前,对于算法效果探究和实际业务应用分析仍然较少。当前我国气象预报向着智能化、精细化的趋势发展,先进预报技术需要融入优于传

收稿日期:2023-12-07

作者简介:李春涵(1997—),男,汉族,山东淄博人,硕士,助工,主要研究方向为机器学习、气溶胶参数化方案设计。

通信作者:孙存永(1995—),男,汉族,江苏连云港人,硕士,助工,主要研究方向为机器学习集成算法。

基金项目:山东省气象局 2023 年青年项目(2023SDQN12);淄博市气象局 2023 年科学技术研究项目(2023zbqx01)

统预报方式的前沿方法。本文针对于目前预报预警业务中出现的难点,以 AdaBoost 算法模型为例研究淄博市的浓雾天气,探究机器学习方法在大雾预报应用中的可行性,以期为大雾精细化预报预警提供有益参考。

1 资料与方法

1.1 数据集来源

样本数据的质量限制着机器学习模型的预报效果,因此,选取自 2015 年 9 月—2023 年 4 月(去除 2016—2022 年 5—8 月)共计 64 个月,淄博市 8 个国家级气象观测站(淄博、沂源、博山、淄川、周村、临淄、桓台、高青)的逐小时自动观测资料作为试验数据。因为夏天汛期多雨,降雨后因高湿度形成的雾并不是业务的关注点,因此去除了 5—8 月观测数据,减少非典型样本数,保证样本数据的质量。结合预报经验并参考已有的研究,数据集选取气温、相对湿度、12 h 降水量、风向、风速、地面温度和能见度等要素,所有数据均进行了质量控制。

机器学习需要划分训练集与测试集,训练集与测试集的比例最好大于 4 : 1,因此选用 56 个月的数据作为训练集,8 个月的数据作为测试集。由于样本量较少,并且浓雾天气为小概率事件,如果单纯选用某一年份的观测数据作为测试集,必然会造成测试样本存在单一性,影响预报效果。因此,将 64 个月的观测数据分为 16 组,每组 4 个月;循环使用其中的 2 组作为测试集,其他作为训练集;总共进行 120 次无重复选取试验,再将所有试验结果进行平均。此举能够尽量减少单一测试集对于模型预报效果的干扰,使试验结果更具可信性。

1.2 AdaBoost 算法及使用原理

Boosting,也称为增强学习或梯度提升法,能够将预测精度仅比随机猜度略高的弱学习器增强为预测精度高的强学习器,在过去十几年间取得了巨大的成功,已经发展为一种理论完善的机器学习技术^[16],而 AdaBoost 是其中最成功的代表之一。选择当下较为流行的 AdaBoost 算法构建机器学习模型:一是因为该算法构建思路简便,仅需要指定迭代次数,运行过程中所有参数由算法

自适应调整,易于理解和操控;二是因为该算法包容性强,添加其他算法改进模块较为方便。因此,基于 AdaBoost 算法设计一个能够自动将能见度分类的机器学习模型。

算法核心的数据分类流程可调用 Python 中的 AdaBoostClassifier 模块实现。训练流程(图 1)可分为以下几步。

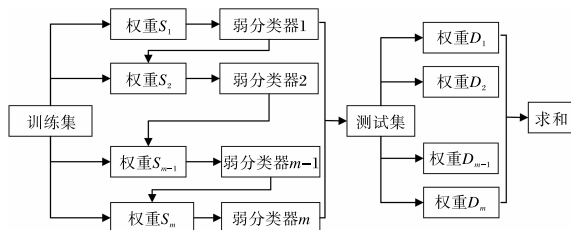


图 1 AdaBoost 算法构建模型流程示意图

第一步:初始化训练样本的权值分布,使每个样本的权重都为 $1/m$ (m 为样本数)。

第二步:依次使用带有权重的训练集构造并训练弱分类器,通过学习误差率 e 得到该弱分类器的权重系数 α ;误差率与权重系数的计算公式为

$$e = \frac{\text{错误数}}{\text{样本总数}}, \quad (1)$$

$$\alpha = \frac{1}{2} \ln\left(\frac{1-e}{e}\right). \quad (2)$$

第三步:通过前一个弱分类器的权重系数,重新调整每个样本的权重 S 。如果第 m 个样本被正确分类 ($i=1, 2, \dots, m$),则该样本权重更新为

$$S_m = \frac{S_{m-1} e^{-\alpha}}{\sum_{i=1}^{m-1} S_i}, \quad (3)$$

其权重将会降低。如果第 m 个样本被分错,则该样本权重更新为

$$S_m = \frac{S_{m-1} e^{\alpha}}{\sum_{i=1}^{m-1} S_i}, \quad (4)$$

其权重将会提高。

第四步:重复第二、三步,直到训练错误率为 0,或者弱分类器的数目达到指定值。此时 AdaBoost 算法为每个弱分类器分配了基于其学习误差率的权重值 D ,若干个弱分类器最后通过简单的加法加权集成为一个强分类器,其中误差率小的弱分类器会在最终的强分类器里占据更大的权

重,否则较小。最后,将测试集输入强分类器后即可获取预报数据集。

可以看出,AdaBoost 的关键在于其样本权重更新(样本权重模拟了概率分布)的方法以及弱分类器加权组合,适合小样本的气象数据集,但分类器的数目并非越多越好,过多的分类器会浪费有限的计算资源;因此,选择适当数量的分类器是一个权衡问题,需要根据具体问题和数据来确定。本文设置的分类器数目为 40 个时,已经能够较好地完成预报任务,在实际应用时应根据训练数据的多少进行调整。另外,在模型训练前,还需加入一步随机过采样来处理数据集不均衡的问题,可通过 Python 中的 RandomOverSampler 模块实现。

2 试验与分析

在进行试验前,有必要分析淄博市浓雾发生时的时间和气象要素特征,以便有针对性地进行试验设计,并且能够为机器学习的特征因子选取提供指导。气象灾害预警信号及防御指南中将雾的预警信号分三级^[17]:以黄色表示可能出现能见度小于 500 m 的雾或者已经出现能见度小于 500 m、大于等于 200 m 的雾并将持续;以橙色表示可能出现能见度小于 200 m 的雾,或者已经出现能见度小于 200 m、大于等于 50 m 的雾并将持续;以红色表示可能出现能见度小于 50 m 的雾,或者已经出现能见度小于 50 m 的雾并将持续。在实际业务中,预测或观测到实际能见度出现低于 500 m 并可能持续的趋势时才会考虑发布预警,因此将能见度低于 500 m 的观测数据标签设置为“能够发布预警”、能见度大于等于 500 m 的观测数据标签设置为“不能发布预警”。

2.1 浓雾的时间特征和气象要素特征

淄博市的浓雾天气有显著的季节特征,主要出现在 1—2 月和 11—12 月。其原因可能是,秋冬季节淄博市多静稳天气,辐射降温明显。浓雾天气的日变化特征是,后半夜和上午多、前半夜和下午少。傍晚前后随着辐射降温,浓雾时次逐渐增加,02—08 时为浓雾较为集中的时段,08 时前后浓雾时次达到峰值,此后逐渐减小(图略)。这是因为辐射雾的日变化与太阳辐射和气温密切相

关,日落后气温开始下降,到日出前温度降到一天中最低,日出前相对较低的气温使得地面空气很容易达到饱和,从而形成辐射雾。日出以后温度逐渐升高,雾开始消散,直至中午前后大部分雾消散。因此,浓雾发生的时间段相对集中,很可能只持续数个小时乃至更短,且消散较为迅速,导致业务中难以做到能见度的精确预报,需要引入新的客观预报方法加以补充。

分析发现,淄博市浓雾次数基本与相对湿度的变化呈正相关(图略),相对湿度越大则越有可能发生浓雾。淄博地区地形较为复杂,北部多平原而南部多山,因此还需对浓雾形成时的风场特征进行统计分析(图 2)。可以看出浓雾的出现依赖于 1~2 级(0.2~3.3 m/s)的微风,过大或者过小的风速都不利于大雾形成。这是因为静风时不利于雾气扩散,而过大的风速加速了水汽流通,会抑制局地的水汽饱和。浓雾出现时北风的频率最大,东北风和东风次之,其他风向的频率相差不大。这可能是因为秋冬季节北风、东北风和东风一般给淄博市带来海上的水汽和小幅降温,而淄博市北部多平原无遮挡,更能满足浓雾生成的条件。因此,高环境湿度与合适的风场极有可能成为影响机器学习的重要因素。

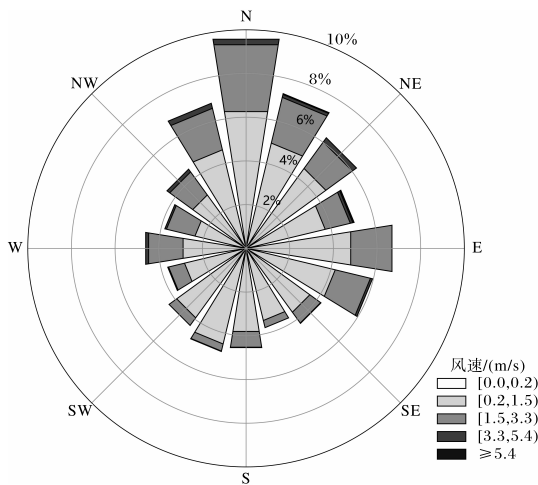


图 2 2015—2023 年淄博市浓雾发生时的风场特征

在实际业务中,预报员大多数情况下遇到的天气状态并不具备浓雾形成的全部有利条件。比如,在初春季节,预报的近地面相对湿度在 95%左右、出现风力为 2~3 级的西北风时,能否确定

今夜不需要发布大雾预警,此时依靠预报经验并不能保证能见度预报的正确性。也就是说,业务上需要定量而非定性地对大雾的产生进行诊断。而数值模式预报被分辨率、参数化方案误差和时效性所限,很难对某些区县、村镇、街道进行能见度的精确预报。因此,本文希望借助非传统手段中的机器学习算法对已有的观测数据进行梳理和统计,总结出一套适用于淄博的预报模型,单独地对某个地点进行能见度的预报。

2.2 测试结果分析及讨论

机器学习模型最后会返回一个混淆矩阵(confusion matrix),其内容包含:模型正确预报为“不能发布预警”(实际能见度大于等于 500 m)的样本数(true negative number, N_{TN})、模型错误预报为“不能发布预警”(实际能见度低于 500 m)的样本数(false negative number, N_{FN})、模型错

误预报为“能够发布预警”(实际能见度大于等于 500 m)的样本数(false positive number, N_{FP})、模型正确预报为“能够发布预警”(实际能见度低于 500 m)的样本数(true positive number, N_{TP})。

基于机器学习模型输出的混淆矩阵,使用气象上常用的精确率(accurate rate, R_A)、漏报率(false negative rate, R_{FN})、空报率(false positive rate, R_{FP})三个指标评估模型的预报效果,其公式如下:

$$R_A = (N_{TP} + N_{TN}) / (N_{TP} + N_{TN} + N_{FP} + N_{FN}) \times 100\%, \quad (7)$$

$$R_{FN} = N_{FN} / (N_{TP} + N_{FN}) \times 100\%, \quad (8)$$

$$R_{FP} = N_{FP} / (N_{TN} + N_{FP}) \times 100\%。 \quad (9)$$

分别对淄博市 8 个国家级气象观测站的测试数据集使用机器学习模型测试其预报结果,试验结果如表 1。

表 1 机器学习模型对淄博各站点多年浓雾预报的测试结果

站名	淄博	沂源	博山	淄川	周村	临淄	桓台	高青
精确率/%	94.4	94.6	92.6	94.2	92.5	95.0	94.6	88.4
空报率/%	5.5	5.3	7.3	5.7	7.4	4.9	5.3	11.8
漏报率/%	7.9	12.4	13.5	8.6	7.4	10.2	9.7	5.2

从机器学习模型对淄博 8 站的预报结果(表 1)来看:除高青站外,预报精确率均在 90%以上,空报率在 10%以下,其中临淄的预报精确率最高达 95%,同时空报率最低为 4.9%;沂源、博山与临淄的漏报率大于 10%,其中博山最高达 13.5%,其他站点的漏报率都低于 10%,漏报率最低的站点为高青,只有 5.2%。淄博南部区县(沂源、博山、淄川)漏报率较高的原因可能是地形多山,产生浓雾的次数比平原地区少,浓雾消散速度也比平原地区快,给机器学习增加了一定难度;而高青站的空报率较高可能是由于站点附近多水域和树木,环境湿度较高,风速较小,同样的环境条件下模型更倾向于预报“能够发布预警”。因此,若要提升预报效果还需考虑站点附近的地形、下垫面等客观因素,将机器学习模型真正投入业务使用时,必须根据每个地区各自的特点加入更多训练要素以优化模型。

另外,大多数空报和漏报出现在阈值(500 m)附近,因此,本模型的主要技术难点在于阈值附近能见度的判断。这符合常理,例如实际能见度为 490 m 和 510 m 时,环境条件的差别相当小,导致模型容易出现判断失误。如果某年出现的浓雾强度均较小,能见度均接近 500 m,就会对机器学习的训练和测试造成一定的干扰;因此,解决预报误差的最简单方法是提高阈值至 600~700 m,尽可能减小漏报率,并且需要循环划分训练集与测试集,否则预报效果会被极端样本稀释。

经过测试,机器学习模型的空报、漏报率基本在 10%以下,精确率在 90%以上,其预报效果接近目前几种主流的数值预报模式。换言之,如果模式预报的相对湿度、气温、风速等气象要素足够精准,那么将模型的测试集换为模式预报数据集后,其预报结果应比模式预报结果更优。当前主流数值预报模式对精细至区县乃至村镇的能见度

还无法做到靶向预报,且预报时间分辨率很难精确至小时内,机器学习方法的引入可能为解决这一难点提供一种新的思路。

2.3 大雾个例预报效果分析

下面通过一个淄博地区的大雾个例来检验机器学习模型的预报效果。2023 年 12 月 27—31 日,淄博市发生一次较为严重的低能见度天气过

程,数个国家级气象观测站的最低能见度降至 50 m 以下并持续数小时。使用各自站点训练好的预报模型,并将 26 日 20 时—31 日 20 时共计 121 个时次的地面观测要素(气温、相对湿度、12 小时降水量、风向、风速、地面温度)输入机器学习模型后,试验结果见表 2。

表 2 机器学习模型对 2023-12-27—31 淄博各站点浓雾预报的空报率、漏报率和精确率

站名	淄博	沂源	博山	淄川	周村	临淄	桓台	高青
精确率/%	81.8	96.6	62.8	82.6	68.6	85.1	76.8	86.0
空报率/%	33.3	7.7	44.1	23.9	50.0	23.7	37.3	24.0
漏报率/%	3.6	0.0	0.0	3.0	0.0	0.0	4.4	4.0

试验结果表明,机器学习模型对此次大雾过程的预报效果较好。除博山和周村外,其他站点的空报率均低于 40%,预报精确率均在 75% 以上,8 个站点的漏报率均接近 0。由于此次大雾过程影响时段长,范围广,各个站点的能见度普遍较低,出现浓雾的样本数与未出现浓雾的样本数较为接近;因此,模型对能见度接近 500 m 的样本普遍倾向于预报“能够发布预警”,导致空报率较高,与上面的分析结论一致。其中,沂源受本次大雾过程影响最小,能见度偏高,模型空报率最低;周村受本次大雾过程影响最大,整体能见度在

1 500 m 以下,空报率最高。此次大雾过程模型的漏报率几乎为 0,说明模型对能见度低于 500 m 的样本预报较准确,即模型对较强浓雾的预报效果较好;而对大范围持续性轻雾的空报率较高,空报率较高的问题可通过测试模型参数和设定预报因子阈值的方法来实现优化。

2.4 预报要素对模型的贡献

根据前面的分析可知,影响大雾产生的主要要素可能是相对湿度和风场,则通过减少预报要素的方式测试要素对模型的贡献,试验结果见表 3。

表 3 机器学习模型在减少各预报要素时对淄博站多年浓雾预报的空报率、漏报率和精确率

缺少的预报要素	温度	相对湿度	过去 12 h 降水量	风速风向	地面温度
精确率/%	94.3	77.6	94.2	88.2	94.4
空报率/%	5.7	22.3	5.7	12.0	5.6
漏报率/%	7.7	26.7	6.6	5.3	7.3

与表 1 对比可以发现,当缺少相对湿度时,模型的精确率大幅下降,空报率、漏报率大幅上升;当缺少风场要素时,精确率、漏报率小幅下降,空报率小幅上升;当减少其他预报要素时,机器学习模型的精确率、漏报率略微下降,空报率略有增加。综上,相对湿度对于机器学习模型的贡献最大,其次为风速风向,其他预报要素对于机器学习的贡献很小,这与前面的分析结论一致。同时根

据减少气象要素后漏报率下降、空报率上升这一现象能够看出,同样的环境条件下机器学习预报更倾向于“能够发布预警”。因此,建议在使用机器学习模型时预报要素应以相对湿度和风场为主,其他预报要素根据本地区的实际情况测试后添加,以此为基础找寻空报率与漏报率之间的平衡点。

2.5 机器学习模型的局地性

虽然机器学习模型的预报效果较好,但仍然

存在一些问题和优化空间。机器学习模型的测试集和训练集均来自同一站点,只能用于该站的预报,若在业务中使用,则必须满足一定的泛化能力。表4展示了当训练与测试集出自不同站点时模型的空报率、漏报率和精确率。与高青、沂源站的试验结果(表1)对比可以看出,当训练集出自其他站点时,空报率略微下降,而漏报率明显上升。由于浓雾的样本量较少,空报率下降、漏报率上升反而导致预报精确率上升。目前国内大部分

对于使用机器学习预报大雾的研究更注重预报精确率,而忽视了空报率和漏报率。由于雾的出现本身就是小概率事件,有无雾的样本数相差悬殊,预报精确率大部分来源于无雾的样本,而不均衡数据构建的模型更偏向于大比重样本的标签,因此其精确率不具有太大的参考价值。实际预报业务中更关注出现大雾时预报是否准确,则漏报率应为三个预报指标中最值得关注的。

表4 机器学习模型在训练集与测试集出自不同站点时对多年浓雾预报的空报率、漏报率和精确率

训练站名(测试站名)	淄博(高青)	淄博(沂源)	高青(沂源)	沂源(高青)
精确率/%	90.6	95.9	94.5	90.8
空报率/%	9.3	3.9	5.4	8.9
漏报率/%	12.7	19.5	19.6	17.4

试验结果表明,当前的机器学习模型局地性较强,仅适用于单站训练并测试,还不具有较好的泛化能力。由于站点之间的地形、下垫面类型、植被密度、周边建筑物密度等客观因素存在差异,简单将所有样本混合后再训练是不科学的。因此,机器学习模型的优化还需要更多数据、资料和理论支撑。除此之外,机器学习模型在分类器数目与样本量的自适应、所用数据集不均衡等方面还有优化空间。

3 结语

从实际业务需求出发,基于目前主流的 Ada-Boost 算法,使用实况观测数据,结合淄博市浓雾的环境特征,构建了一套适用于淄博市单站预报能见度的机器学习模型。该模型的空报率、漏报率基本在 10% 以下,精确率基本在 90% 以上。大多数空报和漏报出现在能见度阈值附近,建议通过提高阈值的方式保证预报预警服务的准确性。各预报要素对于预报效果的贡献存在差异,使用时需要根据实际情况进行调整。

尽管预报效果较好,但机器学习模型仍然存在优化空间。模型对较强浓雾天气的预报效果较好,但对大范围轻雾天气的空报率较高;模型的泛化能力不够优秀,适合单站预报使用,并且预报效果受样本质量、算法复杂度、各站点观测的客观因

素等方面限制。尽管如此,模型也具有较大的潜力。在模式预报的相对湿度、风场等气象要素足够精准的前提下,在针对区县或村镇的能见度靶向预报方面,模型的预报效果能够接近甚至有希望超过目前几种主流模式。随着技术进步和理论更新,机器学习方法在大雾预报上的应用将会更加成熟。

参考文献:

[1] 张礼春,朱彬,耿慧,等. 南京一次持续性浓雾天气过程的边界层特征及水汽来源分析[J]. 气象, 2013, 39(10):1284-1292.

[2] 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[J]. 气候与环境研究, 2016, 21(1):47-55.

[3] 郭俊建,韩永清,赵勇. 山东区域性辐射雾时空分布及地面气象要素特征分析[J]. 海洋气象学报, 2020, 40(2):96-102.

[4] 刘畅,高留喜,王西磊,等. 山东省大雾的气候特征分析[J]. 海洋气象学报, 2014, 34(2):26-31.

[5] 吴彬贵,马翠平,蔡子颖,等. 辐射雾局地爆发性增强原因探讨[J]. 高原气象, 2014, 33(5):1393-1402.

[6] 袁成松,梁敬东,焦圣明,等. 低能见度浓雾监测、临近预报的实例分析与认识[J]. 气象科学, 2007,

- 27(6):661-665.
- [7] 史达伟,李超,史逸民,等. 基于机器学习的大雾天气背景下特强浓雾本地化诊断研究[J]. 灾害学, 2018, 33(2):193-199.
- [8] 苗开超,韩婷婷,王传辉,等. 基于 LSTM 网络的大雾临近预报模型及应用[J]. 计算机系统应用, 2019, 28(5):215-219.
- [9] HARRINGTON P. 机器学习实战[M]. 李锐,李鹏,曲亚东,译. 北京:人民邮电出版社, 2013: 13-14.
- [10] 朱玲,吴心玥. 人工智能在气象领域的应用述评[J]. 广东气象, 2019, 41(1): 35-39.
- [11] 高松,徐江玲,刘桂艳,等. 基于机器学习的青岛市区近岸海雾集成预报方法[J]. 海洋科学, 2021, 35(3):33-42.
- [12] 许立兵,孔扬,周峥,等. 基于机器学习的风场预报订正方法研究[J]. 陕西气象, 2023(1):15-20.
- [13] 张煦庭,刘慧,梁绵,等. 基于不同机器学习算法的汾渭平原 $PM_{2.5}$ 质量浓度预报与分析[J]. 陕西气象, 2023(3):8-16.
- [14] 高荣珍,李欣,任兆鹏,等. 青岛沿海海雾决策树预报模型研究[J]. 海洋预报, 2018, 33(4):80-87.
- [15] 江海英,张韧,刘科峰,等. 青岛海雾的自适应神经模糊推理系统建模[J]. 海洋科学, 2005, 29(2):26-31.
- [16] 曹莹,苗启广,刘家辰,等. AdaBoost 算法研究进展与展望[J]. 自动化学报, 2013, 39(6):745-758.
- [17] 中国气象局. 气象灾害预警信号及防御指南[M]. 北京:气象出版社, 2007: 48-51.