

陈锦鹏, 黄奕丹, 朱婧, 等, 2024. 集成学习和动态融合算法在福建省短时强降水预报中的应用[J]. 气象, 50(1): 48-58. Chen J P, Huang Y D, Zhu J, et al, 2024. Application of ensemble learning and dynamic fusion for short-time severe rainfall forecasting in Fujian Province[J]. Meteor Mon, 50(1): 48-58(in Chinese).

集成学习和动态融合算法在福建省短时强降水预报中的应用*

陈锦鹏^{1,2,3} 黄奕丹³ 朱婧^{1,4} 林辉³ 程晶晶^{1,3} 杨德南^{2,3}

1 厦门市海峡气象开放重点实验室, 厦门 361012

2 福建省灾害天气重点实验室, 福州 350001

3 福建省漳州市气象局, 漳州 363005

4 集美区气象局, 厦门 361021

提 要: 为了提高短时强降水预报准确性, 在 2019—2020 年 4—9 月福建省逐时降水实况观测资料与中国气象局广东快速更新同化数值预报系统(CMA-GD)模式预报产品的基础上, 应用 LightGBM 集成学习算法框架, 建立以 $30 \text{ mm} \cdot \text{h}^{-1}$ 为阈值的逐时降水预报模型。通过特征处理、自助聚合及超参数搜索等技术对模型进行优化, 结合 AUC、AUPR 与传统分类指标, 设计了包括业务模拟测试在内的多项试验, 通过对比各建模方案验证了模型对于较长时效的短时强降水预报的适用性。结果表明: 模式预报本身的命中率和空报率均较高, 各建模方案具有不同程度的改善作用。自助聚合可以增强模型预测稳定性, 轻微不平衡子训练集能降低模型预测空报率而取得更高的综合评分, 在验证集中最佳 TS 评分可达 17.5%; 对分类信息增益贡献最大的特征变量为 K 指数, 其次为 500 hPa 露点温度和时间参数特征; 试验指标从优到劣依次为: 随机交叉验证、小时划分的随机交叉验证、业务模拟测试, 可见模型有效性主要来自相同或相邻时刻的样本信息; 设计基于逻辑回归的异质模型动态融合方案以改善静态同质模型表现, 各项指标均有小幅提升, 在命中率接近 50% 时削减空报样本超过 52 万个。

关键词: LightGBM, 短时强降水预报, 样本不平衡, 动态融合

中图分类号: P456

文献标志码: A

DOI: 10.7519/j.issn.1000-0526.2023.073001

Application of Ensemble Learning and Dynamic Fusion for Short-Time Severe Rainfall Forecasting in Fujian Province

CHEN Jinpeng^{1,2,3} HUANG Yidan³ ZHU Jing^{1,4} LIN Hui³ CHENG Jingjing^{1,3} YANG Denan^{2,3}

1 Xiamen Key Laboratory of Straits Meteorology, Xiamen 361012

2 Fujian Key Laboratory of Severe Weather, Fuzhou 350001

3 Zhangzhou Meteorological Office of Fujian Province, Zhangzhou 363005

4 Jimei Meteorological Station of Xiamen, Xiamen 361021

Abstract: In order to improve the accuracy of short-time severe rainfall forecasts, the LightGBM algorithm is applied to build the hourly precipitation forecasting model based on the precipitation observation data and CMA-GD model forecast products of Fujian Province from April to September in 2019 and 2020. Correction models are optimized by the feature processing, Bagging (bootstrap aggregating) and hyperparameter searching. Combined with AUC, AUPR and traditional classification indices, a series of experiments are

* 厦门市气象局科技专项(3502Z20214ZD4014)、中国气象局复盘总结专项(FPZJ2023-065)和中国气象局/广东省区域数值天气预报重点实验室开放基金课题(J202005)共同资助

2022 年 6 月 21 日收稿; 2023 年 7 月 30 日收修定稿

第一作者: 陈锦鹏, 主要从事模式应用与机器学习研究. E-mail: chenjp9010@163.com

designed to evaluate different modeling schemes and verify the applicability in short-time severe rainfall forecasting. The results show that all modeling schemes can improve the original numerical model forecast representing the high POD and FAR in varying degrees. Bagging can enhance the stability of model prediction, and the slightly unbalanced sub-training set contributes to the higher TS scores by reducing the FAR with the best TS score of validation set being about 17.5%. The largest contribution of feature variable to the classification information gain is K index, followed by 500 hPa dew point and time parameters. The ranking of experiment indices in good to bad order is random cross-validation, random hourly cross-validation and operational simulation test which indicates that the validity of correction models mainly result from the sample information at the same or adjacent moments. The dynamic fusion scheme of heterogeneous models based on logistic regression increases indices of static homogeneous models, which decreases at least 520 000 false alarm samples with approximately 50% POD.

Key words: LightGBM, short-time severe rainfall forecasting, sample imbalance, dynamic fusion

引 言

短时强降水是福建汛期较常发生的强对流天气之一,这种极端降水天气主要由超级单体和中尺度对流系统(MCS)造成(章丽娜等,2014)。随着高分辨率中尺度数值模式在短时临近预报预警业务中的广泛应用,对于短时强降水预报的客观订正技术需求也日益迫切。结合福建省业务标准,将小时雨强超过 $30 \text{ mm} \cdot \text{h}^{-1}$ 的降水事件定为短时强降水。从全年的逐时数据来看,短时强降水的样本比例极低。尽管过去大多数研究是基于逐日降水资料开展的,但是逐时降水数据能反映出更加精细的特征(李德帅,2016)。另外,在降水事件时间尺度为 1 h 的情况下其非线性和随机性特点将进一步放大,加大了预报订正的难度。

与降水客观预报订正技术相关的研究工作逐年增多。唐健等(2018)实现了主客观融合定量降水预报(QPF)平台,其中包括多模式 QPF 集成技术与 QPF 预报场调整和订正技术等,对预报业务起到了良好的支撑作用;唐冶等(2021)采用消空订正方案对新疆区域数值天气预报系统 DOGRAFS 的降水预报产品进行订正,小幅提高了晴雨准确率和 TS 评分;张华龙等(2021)基于因子分析法构建分期、分区短时强降水逐 6 h 格点概率预报模型,在汛期业务试验中相对于模式预报提升较大,尤其对于容易漏报的暖区短时强降水具有明显优势;张武龙等(2021)通过时间滞后集合预报方法构建多个不同权重系数的集合成员进行逐小时降水预报,发现晴雨 TS 评分提升了 10% 左右,且有效减小了模式空报率;洪伟和郑玉兰(2018)在分析福建前汛期短时强

降水发生背景下模式预报物理量分布特征的基础上,建立基于阈值判定方法的短时强降水预报模型,对福建西部关键区的 TS 评分白天可达 0.5、夜间约为 0.3;危国飞等(2020)设计了全球模式与区域模式相结合的降水分级最优化权重集成预报算法,24 h 累计降水预报 TS 评分比主观预报高出 0.9%~2.3%;赵渊明和漆梁波(2021)基于上游降水实况与模式预报的匹配程度,设计动态权重多模式短时临近定量降水概率预报方法,有效提高定量降水概率预报的准确性;潘留杰等(2022)基于卡尔曼滤波方法动态匹配预报和观测降水频率进行降水预报订正,发现能够改善模式对小量级降水预报偏大、大量级降水预报偏小的现象。但总体上,精细到逐小时降水的客观预报方法研究仍相对较少。

近年来,以神经网络为代表的机器学习方法逐渐展现出在强对流预测方面的潜力。金子琪等(2021)采用卷积神经网络算法构建爬线识别模型,揭示了该算法能够学习并识别爬线和非爬线回波的图像特征;黄骄文等(2021)构建基于深度学习网络的降水相态判识模型,针对两年数据以及一次大范围雨雪天气过程进行检验,判识准确率为 98.2%,雨、雪的 TS 评分分别为 97.4% 和 94.4%,较传统指标阈值法的判识准确率有较大提高;张焯方等(2021)以卷积神经网络模型为基础,结合多个时间序列的雷达产品与闪电数据研发雷电临近预报方法,与常规采用雷达、闪电阈值控制的雷电预警算法相比准确率有所提高;韩丰等(2021)以大气层结和对流参数作为特征参数,基于 XGBoost 集成学习方法建立短时强降水预报模型,同时使用分段权重损

失函数进行模型调优,命中率为 0.65、空报率为 0.37、TS 评分为 0.47,表明模型对短时强降水天气具有一定预报能力。

集成学习(ensemble learning)是机器学习领域最热门的研究方向之一,基本思想是通过提升(boosting)、自助聚合(Bagging)和推叠(stacking)等方式将多个弱学习器进行组合以获得比单一模型更好的表现和更小的误差(余东昌等,2021),其中梯度提升决策树(gradient boosting decision tree, GBDT)算法在分类、回归、排序等问题上取得了优异的性能,在学术界和工业界中被广泛使用(江佳伟等,2019)。轻量梯度提升机(light gradient boosting machine, LightGBM)则是在 GBDT 的基础上引入多项优化技术进行改进和提升的一种算法框架,主要目的在于解决高维度大样本数据运行耗及时可拓展性差的问题(刘新伟等,2021),具备内存占用少、并行化学习和准确率较高的优点。在几乎相同的精度上,LightGBM 可以使传统 GBDT 的训练过程加速 20 倍以上(Ke et al,2017),自 2016 年开源以来逐渐成为 Kaggle 等数据挖掘竞赛中的夺冠热门算法。

本研究应用 LightGBM 集成学习算法框架建立逐小时的短时强降水预报模型,在特征处理、自助聚合以及超参数搜索等模型优化技术的基础上,通过多项对比试验对不同方案下的模型预测能力进行评估,并针对业务实践中的应用难点设计了基于逻辑回归的异质模型动态融合方案,在一定程度上提高了长时效、精细化的短时强降水预报能力,为集成学习与模型融合方法在数值预报订正中的应用提供了有益参考。

1 资料

1.1 实况资料

选取 2019—2020 年 4—9 月福建地区的逐小时自动站观测数据作为实况资料,全省自动站数量约为 2200 个。2019 年和 2020 年收集到的有效样本数分别为 4491604 个和 4368497 个。其中,2019 年作为建模所用的已知的训练集和验证集,而 2020 年作为未知的测试集,旨在最大限度模拟实际业务情

况下模型的表现。以 $30 \text{ mm} \cdot \text{h}^{-1}$ 为阈值将实况资料划分为正负类样本,其中 $\geq 30 \text{ mm} \cdot \text{h}^{-1}$ 为正样本,具体分布如表 1。从样本比例来看,实况资料具有极端不平衡的特点,短时强降水属于极小概率事件,后续试验将会对此问题进行探讨。

表 1 2019 年和 2020 年的正负类样本数量

Table 1 Positive and negative sample sizes in 2019 and 2020

年份	降水区间/ $(\text{mm} \cdot \text{h}^{-1})$	样本数/个	样本比例/%
2019 年	$[0, 30)$	4488151	99.923
	$[30, \infty)$	3453	0.077
2020 年	$[0, 30)$	4366462	99.953
	$[30, \infty)$	2035	0.047

1.2 预报资料

在中国气象局广东快速更新同化数值预报系统(CMA-GD)模式预报的基础上开展订正试验。该模式预报产品的空间分辨率约为 $0.03^\circ \times 0.03^\circ$,预报间隔为 1 h,每日起报时次为 08 时和 20 时(北京时,下同)。以当日白天的短时强降水潜势预报为例,由于计算耗时、传输延迟等因素,一般只能参考前日 20 时起报的 CMA-GD 模式产品。因此选择 2019 年和 2020 年 4—9 月、预报时效为 18~23 h(预报时段只涵盖了下午至傍晚及后半夜)的预报产品,其预报变量主要有各等压面层的温度、位势高度、相对湿度、露点温度、水平风场、垂直速度、沙氏指数和 K 指数等。以观测站点为中心,应用最邻近 16 点平均插值计算对应于该站点的特征变量。如此得到的“点”特征称为结构化特征,便于集成学习模型训练。另外,由于数值模式升级或气候背景变化,2019 年和 2020 年的数据分布规律可能会存在差异,称为数据漂移,因此订正模型需要具备一定的泛化能力以适应新的数据。

2 方法

2.1 LightGBM 集成学习算法框架与建模思路

集成学习是指通过训练与集成多个弱分类器来提高最终学习效果的一种技术(李勇等,2014),本研究具体采用 LightGBM 算法框架进行建模。LightGBM 作为集成学习代表性算法之一,其基本原理是通过梯度下降法不断拟合残差(真实值与预测值的偏差)来迭代学习的决策树,再将所有决策树的单独预测进行相加即得最终结果,因此可以由如下公

式表示:

$$f_n(\mathbf{X}) = \sum_{n=1}^N T(\mathbf{X}; \Theta_n) \quad (1)$$

式中: \mathbf{X} 代表输入特征变量, T 代表决策树, Θ_n 为决策树的超参数(如最大深度、叶子节点数等), N 为决策树的个数。LightGBM 应用了两种创新性采样技术:基于梯度的单侧采样(gradient-based one-side sampling)和互斥特征捆绑(exclusive feature bundling)。前者可以缩减用于计算信息增益的样本数量,后者则能减少用于参与分裂点筛选计算的特征数量(任师攀和彭一宁,2020)。

在具体算法应用中,对于 t 时刻至 $t+1$ 时刻的多个站点累计雨量 P , 依托 LightGBM 建模工具包来挖掘与 t 时刻模式预报的结构化特征变量 \mathbf{X} 之间的隐含关系,即假设存在 $P=f(\mathbf{X})+\epsilon$, 其中 ϵ 代表预测误差。这样的建模思路不仅可以大幅增加样本数据量,而且不会受到外推时效的制约,能够在数值模式的基础上进行较长时效的预报。

2.2 特征处理

对 2019 年特征变量进行标准化处理,再利用其均值和标准差对 2020 年特征变量做相同变换。由于短时强降水在不同的季节具有不同的日变化规律(付超等,2019),为了让模型尽可能地捕捉到这种规律以增强待挖掘的映射关系,在原始特征的基础上添加时间参数特征,具体包括距离当年 1 月 1 日的天数(日期参数)和降水所在的时刻(时刻参数)。

最后根据皮尔逊相关系数大小进行特征过滤。分析特征变量两两之间的相关系数,以 0.95 为判断阈值,选择两个强线性相关特征变量中的一个进行剔除,从而减少建模过程中的冗余数据与过拟合现象对模型的不利影响。

2.3 分层 K 折交叉验证与自助聚合

数据采样和模型训练过程中均不可避免地存在随机性。为了增加试验结果的可信度以及提高样本数据利用率,有关 2019 年数据的试验中采用随机分层 4 折交叉验证:将数据集进行随机采样而划分为 4 份,且保持每份子集中的正负样本比例与全集基本一致,每次试验取其中 3 份作为训练集,剩余 1 份作为验证集以跟踪模型性能变化,最后取 4 次试验结果的平均值进行分析。

针对逐小时降水样本极端不平衡的问题,可以采用自助聚合(bootstrap aggregating, 简称 Bagging)方法缓解其不利影响。在训练集中,对大数量

负样本随机采样 N 次,每次采样数与正样本数的比例为 P , 然后与同一份正样本组成子训练集,且通过调节 N 和 P 可在一定程度上提高样本利用率与削弱样本不平衡程度。相比于单纯的平衡欠采样,经过 Bagging 处理后的数据利用率至少增加了 $(N-1)/2$ 倍。基于 N 份子训练集可训练出 N 个超参数不同的同质基模型,再对每个基模型的预测结果进行平均。

2.4 模型跟踪指标

利用模型输出的概率预测结果对逐时降水进行分类预报,会涉及到分类阈值的问题,传统的检验指标(如 TS 评分、命中率 POD 和空报率 FAR 等)会随分类阈值变化而变化。当阈值较低时,命中率提高,空报率也随之增加,反之则都会降低,因此传统的检验指标不利于实时跟踪和评估模型性能随训练代数的变化。引入受试者工作特征曲线(receiver operating characteristic curve, ROC) 曲线下面积(area under ROC curve, AUC)、精准率和召回率曲线(precision-recall curve, PR)曲线下面积(area under PR curve, AUPR)来解决该问题。ROC 曲线和 PR 曲线均是指遍历所有分类概率阈值后由相应坐标系中的散点所连成的曲线。其中,ROC 曲线以伪阳性率(false positive rate, FPR)为横坐标,以真阳性率(true positive rate, TPR)为纵坐标;而 PR 曲线是以召回率 Recall 为横坐标,以精准率 Precision 为纵坐标(黄苏琦,2020)。其中,TPR 与召回率是一致的,等同于气象业务检验中常用的命中率 POD,而精准率与空报率之和为 1。当 AUC 为 0.5 时说明预测完全是随机的;AUC 在 0.5 以上才能说明模型具有正向预测价值;AUC 和 AUPR 越接近 1,则模型整体的预测效果越趋于完美。在正负样本不平衡的分类问题中,PR 曲线与 ROC 曲线的评价结果并不一致,PR 曲线更能突出不同算法之间的性能差异(Davis and Goadrich, 2006),其纵轴也更加贴近气象业务检验习惯。

3 分析

3.1 模式预报检验

分别对 2019 年和 2020 年 CMA-GD 模式预报进行以 $30 \text{ mm} \cdot \text{h}^{-1}$ 为阈值的二分类检验。2020 年的 AUC 为 0.7577, 高于 2019 年的 0.7022; 但从更

具参考意义的 AUPR 来看,2019 年的 0.2354 略高于 2020 年的 0.2178。

传统的分类检验指标与分类阈值密切相关,从命中率、空报率、TS 评分和 FPR 随分类阈值变化(图 1)的角度来对比两者表现。可以发现,2019 年和 2020 年模式预报的命中率变化曲线基本重合,但在分类阈值接近 $30 \text{ mm} \cdot \text{h}^{-1}$ 时后者仍保持在 47% 以上,略高于 2019 年;对于空报率和 TS 评分而言,2020 年模式预报均显著优于 2019 年,侧面反映了在极端不平衡数据中空报情况对 TS 评分影响很大;2019 年模式预报的 FPR 略优于 2020 年。整体来看模式对逐小时强降水预报仍不够理想,业务应用存在困难。

3.2 基于 LightGBM 的建模方案与超参数优化

在 2019 年数据集上训练模型和调整超参数。通过随机 4 折交叉验证将 2019 年数据集划分为 75% 的训练集和 25% 的验证集。验证集作为模型未曾学习过的陌生数据,跟踪其指标变化能够在一定程度上反映模型的过拟合程度,还可以为模型的超参数调优提供必不可少的参考依据。

在训练集的全部样本数据直接用于建模的情况下,由于过度拟合极端不平衡数据,模型性能变得极不稳定,验证集变化曲线大幅波动,即模型失效(图略)。因此需应用 Bagging 方法进行处理,设定子模型数量为 3 个,采样的正负样本比例为 1。不同的超参数组合对模型表现影响极大,最大深度和叶子节点数从小到大意味着模型拟合能力越来越强,但泛化能力可能下降。如图 2 所示,Bagging 采样融合方法带来的最大改变是模型在验证集上的稳定性大大增加,AUC 与 AUPR 均随着训练次数增加而大致趋于收敛。同时注意到 AUC 和 AUPR 变化曲线均呈“V”形。当超参数较小时(图 2a,2b),前 5 代以内的模型反而取得了较高的 AUC 和 AUPR,这主要是因为此时模型拟合能力偏弱、泛化能力偏强而具有高命中率的特点导致的。随着训练代数的增加,模型拟合能力逐渐增强,AUC 和 AUPR 也有所回调,隐含了模型拟合能力与泛化能力相互制约的关系。

相比于模式预报在验证集上的 AUC 为 0.706,不同超参数组合的订正模型在该项指标上均有显著提升,且表现出随着模型超参数增大而小幅增加的趋势。另外,模式预报的 AUPR 为 0.114,当最大深度和叶子节点数较大时模型的该项指标有所下

滑,逐渐劣于数值模式。

为了进一步改善模型表现,尝试对 Bagging 采样的样本比例进行调节。以最大深度为 8、叶子节点数为 22 的订正模型为研究对象,当负类样本数与正类样本数之比 P 逐渐增大时,AUC 与 AUPR 均有不同程度提高,AUPR 尤其显著(图 3)。当 $P=6$ 时 AUC 最大值超过 0.97, $P=10$ 时 AUPR 最大值超过 0.2,且注意到此时模型还存在继续训练的潜力,表明了模型能够从轻微不平衡训练集中学习到比平衡训练集更多的数据信息。

分析基于轻微不平衡与平衡采样训练集的建模方案在传统分类指标上的差异。从图 4 可以看到,平衡方案的优势在于高命中率,除此之外的其他指标在全部概率分类阈值上均劣于轻微不平衡方案。以 $P=10$ 为例,当分类阈值达到 0.8 以上时,命中率快速下滑至 30% 左右,同时空报率也降低至约 70%,TS 评分则快速上升至超过 17.5%,FPR 则变化不大,整体保持在 10% 以下。由此可见,面对极端不平衡的数据集,提高预测 TS 评分的关键和难点在于降低空报率。

基于 LightGBM 的订正模型在训练完成后可从累计信息增益的角度统计特征重要性。信息增益 g 的公式如下:

$$g(\mathbf{D}, \mathbf{A}) = H(\mathbf{D}) - H(\mathbf{D} | \mathbf{A}) \quad (2)$$

其中

$$H(\mathbf{X}) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (3)$$

式中: \mathbf{D} 和 \mathbf{A} 分别表示待分类数据和分类条件, H 表示信息熵, n 为类别数(二分类中为 2), $p(x_i)$ 为该类别事件发生概率。对轻微不平衡模型而言,排序前十的特征变量如表 2 所示,最重要的特征变量为 K 指数,且重要性远远大于之后的特征变量,可见 K 指数对于模型预测的指示意义最佳。其次为 500 hPa 露点温度,第三和第四分别为自定义添加的时间特征——时刻参数和日期参数,表明了日变化和年变化规律的挖掘对于模型预测具有较大的帮助。另外,700 hPa 以上经向风、中高层位势高度与 850 hPa 相对湿度也占据一定的重要性。

3.3 订正模型的有效性分析

上述建模与调参试验中均无视数据集中存在的时间关系而进行随机交叉验证,为了进一步分析轻微不平衡模型的有效性来源以及在具有时间关系的业务模拟中的表现,分别设计 2019 年小时划分的交叉验证试验以及 2020 年测试集试验,模型超参数设

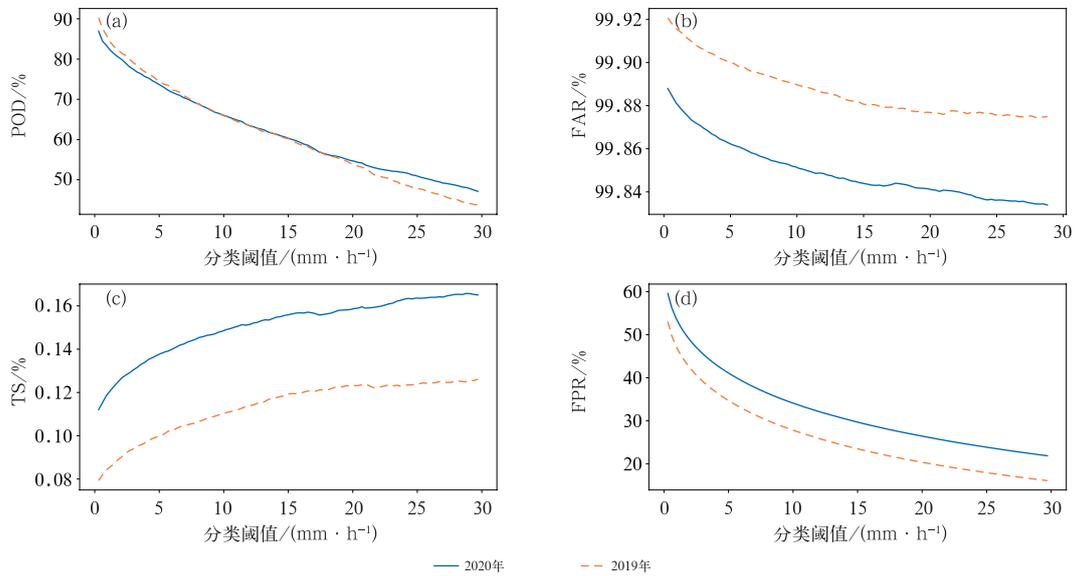


图 1 2019 年和 2020 年 CMA-GD 模式预报检验指标随分类阈值的变化

Fig. 1 Change of CMA-GD model forecast skill with classification thresholds in 2019 and 2020

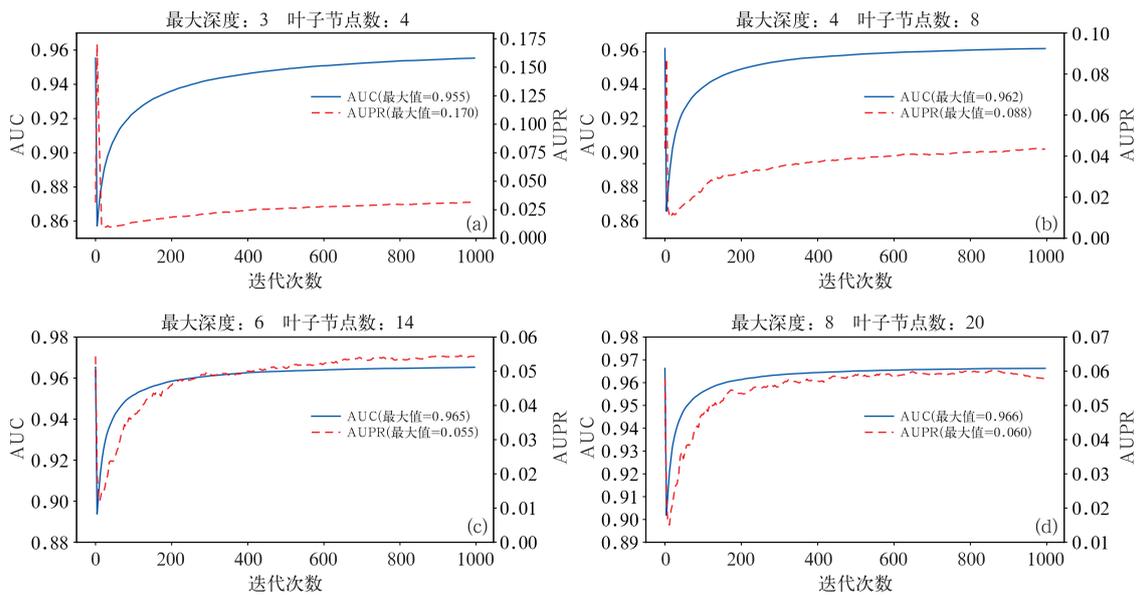


图 2 不同超参数的模型在验证集上的 AUC 与 AUPR

Fig. 2 AUC and AUPR of correction models with different hyperparameters in validation set

为:最大深度为 8、叶子节点数为 22、子模型数为 3、负类与正类样本之比为 10。

同样基于 2019 年数据集,小时随机交叉验证与 3.2 节中随机交叉验证的最大不同在于,模型无法从训练集中学习到与验证集中所属同一小时的样本数据,从而避免了时间上的信息泄露。在这种情况下,模型预测能力明显下降,尤其是 AUPR 最高仅为 0.024,几乎缩减了一个数量级(图 5a),此时最高 TS 评分只有 3.671%,命中率下降至 10.959%,空

报率也升高至 94.768%。不难发现,同一小时内的样本信息对模型能力提升的贡献极大,即便是相邻时刻的样本也难以弥补其作用。换言之,该试验验证了基于数值模式的逐时降水数据集具有时间敏感的特点,这会给模式后处理技术带来很大的困难。

2020 年测试集的条件更为苛刻,与训练集之间基本不存在时间相关性,模型甚至无法获取邻近时刻的样本信息。因此,模型的最高 AUC 和 AUPR 再度下降,分别仅有 0.883 和 0.003,表明了模型从

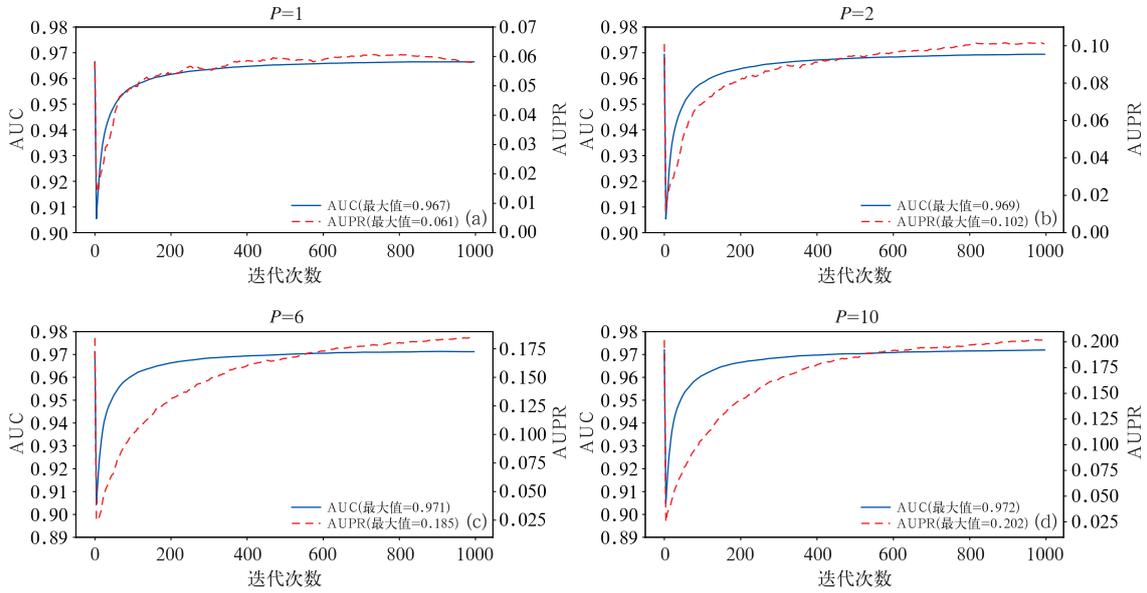


图 3 不同 P 值下轻微不平衡采样的模型在验证集上的 AUC 与 AUPR

Fig. 3 AUC and AUPR of correction models with slightly imbalanced sampling in validation set under different P values

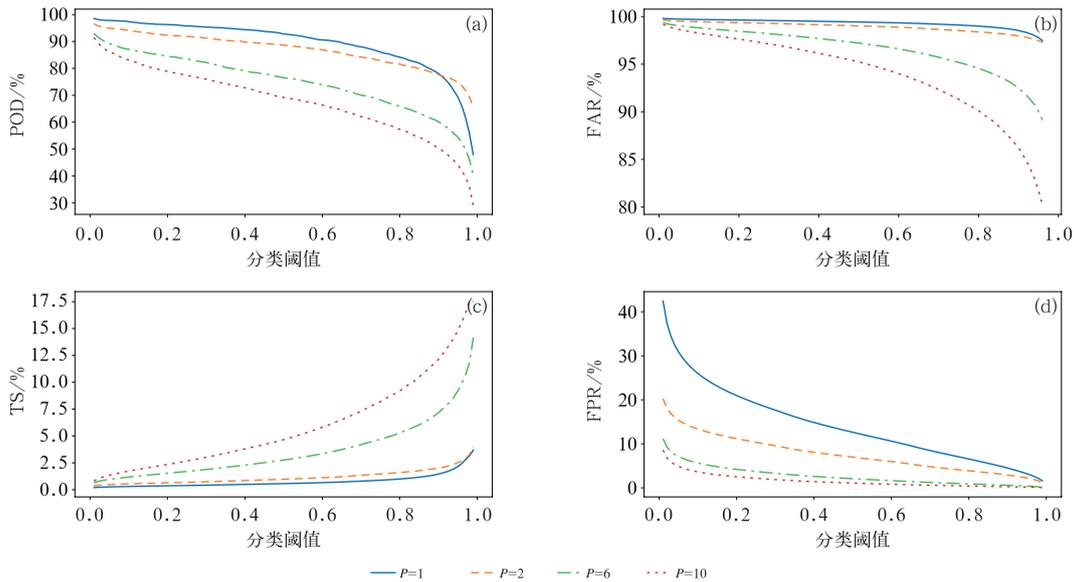


图 4 不同 Bagging 采样比例的 LightGBM 模型预报检验指标对比

Fig. 4 Comparison of LightGBM model forecast skill with different Bagging sampling scales

表 2 模型特征重要性前十排序

Table 2 The top 10 ranking of model feature importance

序号	特征变量	特征重要性
1	K 指数	12322.9
2	500 hPa 露点温度	2288.8
3	时刻参数	2133.1
4	日期参数	1822.9
5	700 hPa 经向风	1715.0
6	200 hPa 经向风	1591.7
7	500 hPa 位势高度	1567.3
8	200 hPa 位势高度	1535.1
9	850 hPa 相对湿度	1495.3
10	500 hPa 经向风	1414.2

2019 年数据集中学习到的映射关系只有很小一部分适用于 2020 年,再次验证了模型有效性主要来自于相同或相邻时刻的样本信息。

3.4 异质模型动态融合方案

从模拟实际业务的角度来讲,尽管无法得到超出当前时刻的样本信息,但可以将已发生的最新样本信息用来更新模型,使模型逐渐适应新的数据分布。上述 LightGBM 模型在本质上是基于决策树的加性模型,导致很难将最新样本信息引入其中。

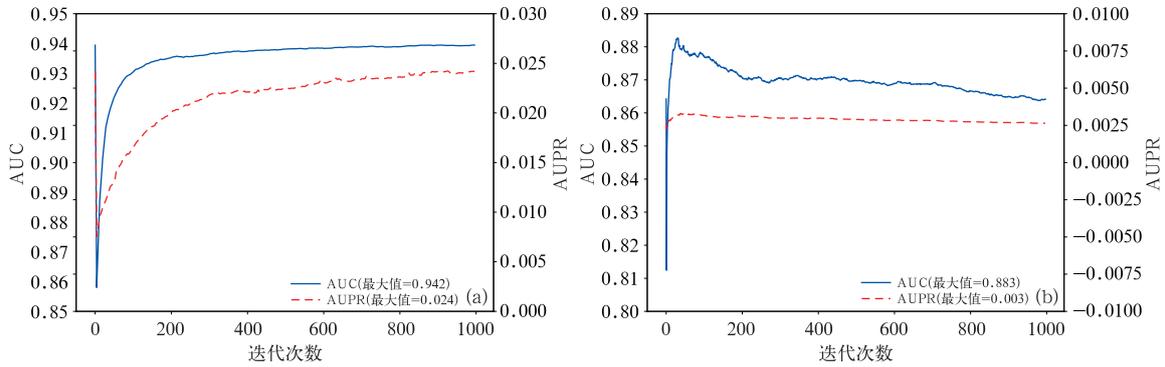


图 5 轻微不平衡采样的模型在(a)2019 年小时划分的交叉验证和(b)2020 年测试集的 AUC 与 AUPR

Fig. 5 AUC and AUPR of correction models with slightly imbalanced sampling in (a) 2019 hourly cross-validation and (b) 2020 test set

因此,选择逻辑回归(logistic regression, LR)算法作为次级模型以达到对初级静态模型“再订正”的目的。LR 是一种相对简单的分类算法,在具备非线性拟合能力的同时又可以避免在小样本的情况下发生严重的过拟合。此外,仅把 LightGBM 模型的输出概率作为次级模型的输入特征,会存在变量过少、信息过于单调的问题。深度神经网络(deep neural net, DNN)模型作为目前受到广泛应用的机器学习模型,可以提供不同的统计视角来作为信息补充(陈锦鹏等,2021)。

具体的异质模型动态融合方案为:分别用 2019 年训练集提前训练好 LightGBM 模型和 DNN 模型,其中 LightGBM 模型超参数与 3.3 节中一致, DNN 模型则采用 3 层全连接层结构来匹配点特征的输入,网络层的连接顺序为输入层、64 个神经元的全连接层、32 个神经元的全连接层、16 个神经元的全连接层、失活比例为 0.1 的随机失活层(用来减轻模型过拟合)以及实现概率归一化的 Softmax 函数输出层。在对 2020 年测试集的预测过程中,发现由于过拟合问题的存在,训练代数越多反而会削弱泛化能力,故分别取第 50 代 LightGBM 模型和第 10 代 DNN 模型进行预测。另外在进行动态融合前,需要先判断过去 5 d 的最新数据中正样本数量是否充足,当正样本数 ≥ 10 时采用 LR 进行融合建模,正样本数 < 10 时只对输出概率求平均。具体的流程示意图如图 6 所示。

对比试验中包含了数值模式、LightGBM 模型、DNN 模型与 LR 融合模型共四种预测。在命中率方面(图 7a),模式预报表现较稳定,区间大致为 42%~90%;LightGBM 模型和 DNN 模型均有随阈

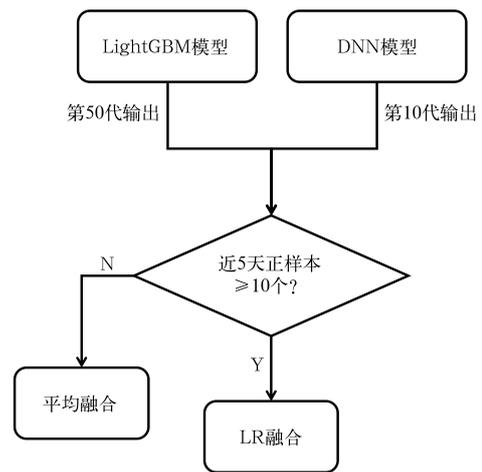


图 6 异质模型动态融合流程示意图

Fig. 6 Flow chart of dynamic fusion scheme of the heterogeneous models

值升高而快速下降的特点,实际应用中阈值设定不宜过高;LR 融合模型的下降速率则表现出“先高后低”趋势,分类阈值 < 0.4 时命中率维持在 90% 以上,在 0.4~0.6 时则为快速下滑阶段,之后与单模型趋于一致。在空报率方面(图 7b),三种订正模型均低于模式预报,其中 LightGBM 模型和 LR 融合模型在高阈值情况下会出现反弹。TS 评分来看(图 7c),LR 融合模型在 LightGBM 模型和 DNN 模型的基础上能够小幅提升,分类阈值达 0.78 时可得最高 TS 评分为 0.568%。三种订正模型的 FPR 在分类阈值 ≥ 0.5 时均优于模式预报,其中 LR 融合模型的 FPR 变化趋势与命中率相似。

在逐小时强降水预报检验中,命中率尤为重要,为此需要分析在较高命中率的情况下各类方案的表现。通过调节分类阈值将四种预测的命中率控制在

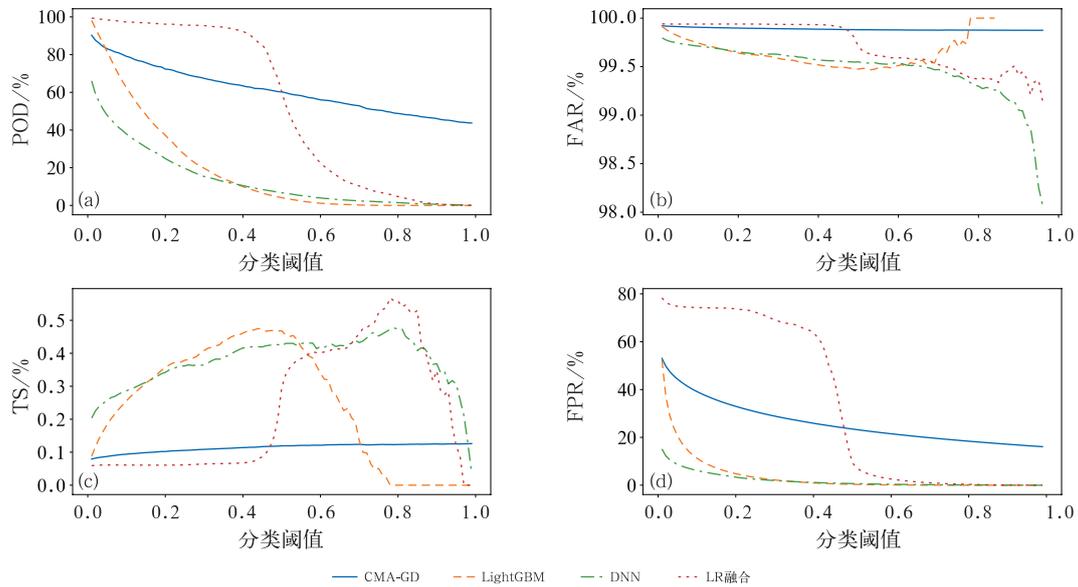


图 7 不同模型预报检验指标随分类阈值的变化

Fig. 7 Changes of different model forecast skills with classification thresholds

实际业务能够接受的 50% 左右,便于对比其他指标。如表 3 所示,此时三种订正模型对于模式的空报情况均有所改善,从而在不同程度上提高了 TS 评分。在空报数方面,LightGBM 模型减少了约 49.7 万个样本,而 LR 融合模型在命中率提高了

1.665% 的前提下减少了约 52.4 万个样本。在极端不平衡数据中空报数对 TS 评分影响极大,由于 LR 融合模型能够有效降低空报数,故 TS 评分为最优,对比模式预报提高了将近 3 倍,同时分类阈值也比较合理。

表 3 不同模型在命中率约为 50% 时的对比

Table 3 Comparison of different models at approximately 50% POD

不同模型	命中率/%	空报率/%	空报数/个	TS 评分/%	分类阈值
数值预报 (CMA-GD)	50.369	99.876	827532	0.124	0.76
	49.779	99.877	821419	0.123	0.77
LightGBM 模型	50.553	99.700	341585	0.299	0.14
	48.133	99.689	313800	0.310	0.15
DNN 模型	50.835	99.751	416986	0.249	0.04
	47.813	99.743	380552	0.256	0.05
LR 融合模型	53.661	99.665	324920	0.334	0.51
	48.354	99.644	275150	0.368	0.52

4 结 论

本文在逐时降水观测资料与数值模式预报产品的基础上开展短时强降水预报订正试验,应用 LightGBM 集成学习算法框架与多项模型优化技术进行建模和训练,通过对比不同建模方案在验证集和测试集上的表现,得到如下结论。

(1)CMA-GD 模式在短时强降水预报方面具有较低 AUC 和较高 AUPR 的特点,从传统指标来看

命中率和空报率均较高,各类建模方案的订正模型对此具有不同程度的改善作用。其中 Bagging 处理能够增强模型预测稳定性,而基于轻微不平衡子训练集的模型在验证集上表现最佳,主要优势在于显著降低了预测空报率而取得更高的综合评分,在验证集中最佳 TS 评分可达 17.5%。

(2)LightGBM 模型具有良好的可解释性,通过合理的特征处理可以进一步增强变量间的映射关系,对分类信息增益贡献最大的特征变量为 K 指数,其次是 500 hPa 露点温度和自定义添加的时间

参数特征。

(3)从随机交叉验证、根据小时划分的随机交叉验证、业务模拟测试等三类时间相关性依次减弱的试验来看,随机交叉验证的指标评分最高,而业务模拟测试为最低,验证了 LightGBM 模型在短时强降水分类数据集上的有效性主要来自相同或相邻时刻的样本信息。

(4)在业务模拟测试中 LightGBM 模型略优于 3 层全连接层架构的 DNN 模型。针对实际预报业务中客观存在的时间敏感性及数据漂移等问题,基于逻辑回归的异质模型动态融合方案能够实时学习来自不同统计视角的最新样本信息而改善静态同质模型的表现,融合模型在命中率、空报率和 TS 评分等指标上均有小幅提升,在命中率接近 50% 时削减空报样本超过 52 万个。

参考文献

- 陈锦鹏,冯业荣,蒙伟光,等,2021. 基于卷积神经网络的逐时降水预报订正方法研究[J]. 气象,47(1):60-70. Chen J P, Feng Y R, Meng W G, et al, 2021. A correction method of hourly precipitation forecast based on convolutional neural network[J]. Meteor Mon, 47(1):60-70(in Chinese).
- 付超,湛芸,朱克云,等,2019. 2010—2016 年江西省暖季短时强降水特征分析[J]. 气象,45(9):1238-1247. Fu C, Chen Y, Zhu K Y, et al, 2019. Characteristics of flash heavy rain in Jiangxi warm season from 2010 to 2016[J]. Meteor Mon, 45(9):1238-1247(in Chinese).
- 韩丰,杨璐,周楚炫,等,2021. 基于探空数据集成学习的短时强降水预报试验[J]. 应用气象学报,32(2):188-199. Han F, Yang L, Zhou C X, et al, 2021. An experimental study of the short-time heavy rainfall event forecast based on ensemble learning and sounding data[J]. J Appl Meteor Sci, 32(2):188-199(in Chinese).
- 洪伟,郑玉兰,2018. 基于 ECMWF 产品福建省前汛期短时强降水预报方法[J]. 应用气象学报,29(5):584-595. Hong W, Zheng Y L, 2018. A method of short-time strong rainfall forecasting during pre-rainy season in Fujian based on ECMWF productions [J]. J Appl Meteor Sci, 29(5):584-595(in Chinese).
- 黄骅文,蔡荣辉,姚蓉,等,2021. 深度学习网络在降水相态判识和预报中的应用[J]. 气象,47(3):317-326. Huang J W, Cai R H, Yao R, et al, 2021. Application of deep learning method to discrimination and forecasting of precipitation type[J]. Meteor Mon, 47(3):317-326(in Chinese).
- 黄苏琦,2020. 时空谱多特征联合红外弱小目标检测方法研究[D]. 成都:电子科技大学. Huang S Q, 2020. Infrared dim small target detection based on joint temporal-spatial-spectral features [D]. Chengdu: University of Electronic Science and Technology of China(in Chinese).
- 江佳伟,符芳诚,邵莹侠,等,2019. 面向高维特征和多分类的分布式梯度提升树[J]. 软件学报,30(3):784-798. Jiang J W, Fu F C, Shao Y X, et al, 2019. Distributed gradient boosting decision tree algorithm for high-dimensional and multi-classification problems [J]. J Sof, 30(3):784-798(in Chinese).
- 金子琪,王新敏,鲍艳松,等,2021. 基于卷积神经网络的飑线识别算法[J]. 应用气象学报,32(5):580-591. Jin Z Q, Wang X M, Bao Y S, et al, 2021. Squall line identification method based on convolution neural network[J]. J Appl Meteor Sci, 32(5):580-591(in Chinese).
- 李德帅,2016. 基于逐小时资料的华南地区汛期降水时空变化特征及其成因研究[D]. 兰州:兰州大学. Li D S, 2016. Research on the spatiotemporal characteristics of hourly precipitation over South China during the warm season and its possible causes [D]. Lanzhou: Lanzhou University(in Chinese).
- 李勇,刘战东,张海军,2014. 不平衡数据的集成分类算法综述[J]. 计算机应用研究,31(5):1287-1291. Li Y, Liu Z D, Zhang H J, 2014. Review on ensemble algorithms for imbalanced data classification[J]. Appl Res Comput, 31(5):1287-1291(in Chinese).
- 刘新伟,黄武斌,蒋盈沙,等,2021. 基于 LightGBM 算法的强对流天气分类识别研究[J]. 高原气象,40(4):909-918. Liu X W, Huang W B, Jiang Y S, et al, 2021. Study of the classified identification of the strong convective weathers based on the LightGBM algorithm[J]. Plateau Meteor, 40(4):909-918(in Chinese).
- 潘留杰,薛春芳,张宏芳,等,2022. 基于卡尔曼动态频率的 ECMWF 降水预报订正[J]. 气象,48(1):73-83. Pan L J, Xue C F, Zhang H F, et al, 2022. ECMWF precipitation calibration based on the Kalman dynamic frequency matching method[J]. Meteor Mon, 48(1):73-83(in Chinese).
- 任师攀,彭一宁,2020. 基于软投票融合模型的消费信贷违约风险评估研究[J]. 金融理论与实践, (4):77-83. Ren S P, Peng Y N, 2020. Default risk assessment of consumer credit based on soft voting fusion model[J]. Financial Theory Pract, (4):77-83(in Chinese).
- 唐健,代刊,宗志平,等,2018. 主客观融合定量降水预报方法及平台实现[J]. 气象,44(8):1020-1032. Tang J, Dai K, Zong Z P, et al, 2018. Methods and platform realization of the national QPF master blender[J]. Meteor Mon, 44(8):1020-1032(in Chinese).
- 唐治,李如琦,马玉芬,等,2021. DOGRAFS 逐小时气温和降水预报的释用[J]. 沙漠与绿洲气象,15(4):100-106. Tang Z, Li R Q, Ma Y F, et al, 2021. Study on the interpretation and application method of hourly temperature and precipitation forecast based on DOGRAFS[J]. Desert Oasis Meteor, 15(4):100-106(in Chinese).
- 危国飞,刘会军,吴启树,等,2020. 多模式降水分级最优化权重集成预报技术[J]. 应用气象学报,31(6):668-680. Wei G F, Liu H J, Wu Q S, et al, 2020. Multi-model consensus forecasting technology with optimal weight for precipitation intensity levels[J]. J Appl Meteor Sci, 31(6):668-680(in Chinese).
- 余东昌,赵文芳,聂凯,等,2021. 基于 LightGBM 算法的能见度预测

- 模型[J]. 计算机应用, 41(4):1035-1041. Yu D C, Zhao W F, Nie K, et al, 2021. Visibility forecast model based on LightGBM algorithm[J]. J Comput Appl, 41(4):1035-1041(in Chinese).
- 张华龙, 伍志方, 肖柳斯, 等, 2021. 基于因子分析的广东省短时强降水预报模型及其业务试验[J]. 气象学报, 79(1):15-30. Zhang H L, Wu Z F, Xiao L S, et al, 2021. A probabilistic forecast model of short-time heavy rainfall in Guangdong Province based on factor analysis and its operational experiments[J]. Acta Meteor Sin, 79(1):15-30(in Chinese).
- 章丽娜, 王秀明, 熊秋芬, 等, 2014. “6. 23”北京对流暴雨中尺度环境时空演变特征及影响因子分析[J]. 暴雨灾害, 33(1):1-9. Zhang L N, Wang X M, Xiong Q F, et al, 2014. On the evolution of mesoscale environment and influential factors of the heavy rainfall in Beijing on 23 June 2011[J]. Torr Rain Dis, 33(1):1-9(in Chinese).
- 张武龙, 王彬雁, 陈朝平, 等, 2021. SWCWARMS 模式小时降水量时间滞后集合预报研究[J]. 气象科技, 49(4):561-568. Zhang W L, Wang B Y, Chen C P, et al, 2021. Time-lagged ensemble forecast of hourly precipitation based on SWCWARMS model [J]. Meteor Sci Technol, 49(4):561-568(in Chinese).
- 张烨方, 冯真祯, 刘冰, 2021. 基于卷积神经网络的雷电临近预警模型[J]. 气象, 47(3):373-380. Zhang Y F, Feng Z Z, Liu B, 2021. Lightning nowcasting early warning model based on convolutional neural network[J]. Meteor Mon, 47(3):373-380(in Chinese).
- 赵渊明, 漆梁波, 2021. 短时强降水概率预报的多模式集成技术研究[J]. 气象, 47(5):529-538. Zhao Y M, Qi L B, 2021. Multi-model integration technology for probabilistic forecasting of short-time severe rainfall[J]. Meteor Mon, 47(5):529-538(in Chinese).
- Davis J, Goadrich M, 2006. The relationship between Precision-Recall and ROC curves[C]//Proceedings of the International Conference on Machine Learning. Pennsylvania: ACM:233-240.
- Ke G L, Meng Q, Finely T, et al, 2017. LightGBM: a highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS 2017). Long Beach: ACM:3149-3157.

(本文责编:戴洋 王婷波)