左右. 基于决策树的水灾预测模型[J]. 水利水电技术(中英文), 2025, 56(8): 19-31. DOI: 10.13928/j.cnki.wrahe.2025.08.002
ZUO You. Flood prediction model based on decision trees[J]. Water Resources and Hydropower Engineering, 2025, 56(8): 19-31. DOI: 10.13928/j.cnki.wrahe.2025.08.002

基于决策树的水灾预测模型

左 右
(杭州电子科技大学 理学院, 浙江 杭州 310018)

摘 要: 【目的】洪水是由暴雨、冰雪快速融化、风暴潮等因素引发的自然灾害, 常导致经济损失与生活不便。常规的洪水预测主要依赖于传统的水文学方法和基于经验的统计模型, 但在遇到缺乏长期、连续的水文观测数据的地区, 利用其他数据进行洪水预测的方法就至关重要。【方法】基于决策树的机器学习算法, 如随机森林、XGBoost及LightGBM, 因其直观性和强大功能, 在分类和回归任务中具有良好的预测能力, 适用于洪水预测。使用包含50 000条记录与21个变量的数据集, 评估随机森林、XGBoost和LightGBM三种算法的洪水预测能力, 通过预测效果与关键变量识别比较其性能, 并以ROC-AUC曲线衡量优劣。【结果】结果显示: 所有模型均表现出较高的预测精度, 其中XGBoost模型具有最小的均方误差0.000 186 2和最高的决定系数0.925 2, 而LightGBM模型在ROC-AUC曲线中取得了最大的AUC值0.99。随机森林模型各指标均不如以上二者。【结论】结果表明: XGBoost模型在洪水概率的预测方面效果最好, 预测误差最小; 而对于预测洪水是否发生这类二分类情况, LightGBM则是最优的选择。

关键词: 洪水预测; 决策树; 随机森林; XGBoost; LightGBM
DOI: 10.13928/j.cnki.wrahe.2025.08.002 **开放科学(资源服务)标志码(OSID):**
中图分类号: TV122; TP181 **文献标志码:** A **文章编号:** 1000-0860(2025)08-0019-13

Flood prediction model based on decision trees

ZUO You
(School of Science, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)

Abstract: [Objective] Floods are natural disasters triggered by factors such as heavy rainfall, rapid snow and ice melt, and storm surges, often resulting in significant economic losses and severe disruption to daily life. Conventional flood prediction primarily relies on traditional hydrological method and experience-based statistical models. However, in areas lacking long-term and continuous hydrological monitoring data, alternative data-driven method for flood prediction are essential. [Methods] Machine learning algorithms based on decision trees, including Random Forest, XGBoost, and LightGBM, demonstrated excellent performance in classification and regression tasks due to their interpretability and strong functions, making them suitable for flood prediction. A dataset containing 50 000 records and 21 variables was used to evaluate the flood prediction performance of these three algorithms, namely Random Forest, XGBoost, and LightGBM. Their performance was assessed based on prediction

收稿日期: 2024-08-06; **修回日期:** 2024-09-06; **录用日期:** 2024-09-12; **网络出版日期:** 2024-10-15
基金项目: 国家社会科学基金项目(21BTJ071)
作者简介: 左 右(1999—), 男, 硕士研究生, 主要从事统计与机器学习研究。E-mail: 221070107@hdu.edu.cn

accuracy and key variable identification, with the ROC-AUC curve used for comparative analysis. [Results] The result showed that all three models achieved high prediction accuracy. Among them, the XGBoost model exhibited the lowest mean squared error (0.000 186 2) and the highest coefficient of determination (0.925 2). Moreover, the LightGBM model achieved the highest AUC value (0.99) in the ROC-AUC curve. The Random Forest model underperformed the other two across all indicators. [Conclusion] The findings indicate that XGBoost delivers optimal performance for flood probability prediction with lowest prediction errors, while LightGBM is the optimal choice for binary classification tasks, such as predicting flood occurrence.

Keywords: flood prediction; decision trees; Random Forest; XGBoost; LightGBM

0 引　言

地球的气候系统正经历着前所未有的变化，极端气候事件的发生频率和强度显著增加，其中最引人注目的现象之一便是洪水事件的增多。近年来，中国南方地区洪涝灾害不断，许多地区因为基础设施老化、排水系统不足和水土流失等问题，大大加剧了洪水的风险和破坏力，损害了人们的生命和财产安全，因此，提升洪水概率预测的准确性对灾害管理和减灾具有重要意义[1]。过去，洪水预测主要依赖于传统的水文学方法和基于经验的统计模型，利用过去或现在的水文气象数据，包括流域、地区或特定的水文站，对未来的水文条件进行定性或定量预测[2]。这些方法在已监测流域中的表现尚可，但对于未测量流域(即缺乏长期、连续观测数据的地区)或数据稀少的地区，传统的预测模型由于缺乏历史数据，难以建立有效的预测机制。而近年来，研究者们热衷于利用人工神经网络(ANN)、支持向量机(SVM)等新方法来进行洪水预测。

机器学习方法的出现和发展为解决洪水预测问题提供了新的可能性[3]。尤其是深度学习算法，由于其强大的模式识别能力和对大量非结构化数据的处理能力，在处理复杂系统预测问题上具有巨大潜力。近年来，有研究基于神经网络的方法，捕捉到时空数据中的复杂关联[4]，从而在没有直接观测数据的情况下也能进行准确预测；基于卷积神经网络和长短期记忆网络的方法能捕捉到复杂的气候模式和时空关联，显著提高预测精度[5]。因此，研究者们也尝试将深度学习方法运用于洪水预测。如支持向量机(SVM)[6]、人工神经网络(ANN)[7]、决策树等[8]。SAHOO 等[9]验证了使用混合模型(包括 SVM 和其他机器学习算法)来预测城市洪水的有效性，展示了 SVM 模型如何与数值模型结合，提供快速的预测；APARAJITA 等[10]探讨了不同类型的 ANN 模型在河流系统洪水预测中的应用，包括具有和不具有记忆功能的 ANN，其中大于 0.90 的效率系数和较低的均方根误差说明模型在这类问题中的适用性；LAWAL 等[11]在研究后提出了决策树类方法在洪水预测中具有很好的效果，具有很高的准确率。SYEED 等[12]用决策树分类器进行洪水预测，取得了 0.787 9 的准确率和 0.583 3 的召回率。GRZESIAK 等[13]验证了随机森林模型在洪水预测上具有优秀的精度。尽管这些方法在许多情况下表现良好，但是都建立在拥有长期水文数据的情况下，而对于那些缺乏此类数据的区域，如何构建可靠的洪水预测模型仍然是一个开放性问题。此外，如何有效地整合多源异构数据，包括遥感数据、气象预报、地形信息等，以构建更全面的洪水预测模型，也是一个亟待解决的问题。

考虑到有些地区技术落后、数据不足或缺失，没有包含时间的某个地域的水文数据集，无法进行洪水的有效预测或防范。本文尝试结合其他可以影响洪水的相关特征，整合多源异构数据，尝试使用适合回归和分类问题的基于决策树的深度学习预测模型，解决洪水预测问题。随机森林是一种集成学习技术[14]，性能优良，在常规的洪水预测领域已经有了广泛地应用[15]，是水文学中最成熟的基于树的机器学习方法之一，具有很高的可操作性和计算速度。XGBoost 是基于梯度提升的机器学习算法[16]，具备高效的优化及并行计算功能。LightGBM 同样基于梯度提升且速度更快，内存使用率更低[17]，但它在分裂与生长策略上与 XGBoost 有所区别。本文选择了一个包含诸多特征的数据集，将使用以上三种模型，实现对未来洪水事件的预测，并判断出在不同情况下，该选用何种方法才能取得最优效果，从而为全球洪水风险管理提供有力的支持。

本文先对数据进行相关性分析，确定目标变量与特征变量，划分训练集与测试集。之后通过随机森林、XGBoost、LightGBM 等三种基于决策树的深度学习方法，可视化特征重要性。比较均方误差(*MSE*)、决定系数(R^2)、准确率、精确率、召回率、*F*1 值和

ROC-AUC 曲线，来评判洪水风险量化预测效果和洪水事件二分类预测效果，得出在不同情况和需求下模型的选择。

1　数据描述

本研究使用的数据集来源于 Kaggle 平台的 Flood Prediction Dataset，网址为 https：//www. kaggle. com/datasets/naiyakhalid/flood-prediction-dataset，由 Naiya Khalid 提供，发布于 2024 年 6 月。此数据集不包含长期连续的历史观测数据，不包含时间、地域等特征，只包含全球范围内不同城市和地区的多个与洪水预测相关的特征，包括环境因素和社会经济指标。数据集有 50 000 行和 21 列；其中，21 列分别是通过量化分级得来的数字变量，数据集中所有特征变量及其代表含义、计算方法、单位和时空分辨率的特征变量如表 1 所列。

这些数据中指标的选择都是基于以往的研究和历史数据，这些因素被证明与洪水发生有显著相关性。这些变量的数据相对容易获取，且具有较高的准确性和可靠性。对于季风强度[18]，GOSWAMI 等[19]研究发现强降雨会导致河流和水库水位迅速上升，增加洪水风险；FLEISCHMANN 等[20]研究表明排水不良的地区更容易积水，导致洪水；良好的河流管理可以有效减少洪水风险[21]，JACOBSON[22]分析了不透水面积增加对洪水频率的影响，城市化导致不透水表面的增加，减少了雨水的自然渗透，增加了洪水风险；气候变暖和异常气候事件的增加导致极端天气事件的频率增加[23]，包括暴雨和洪水；淤积会减少河道和水库的蓄水能力，增加洪水风险[24]；不当的农业实践会增加地表径流和土壤侵蚀，增加洪水风险[25]；侵占河道、洪泛区等自然防洪空间会减少自然蓄水空间，增加洪水风险；TIERNEY[26]讨论政策变化如何对自然灾害和其他极端事件产生影响并且米胤瑜等[27]还就伦敦、纽约、郑州等国内外城市研究了政策对洪水破坏的影响程度；WONG[28]研究表明，沿海地区更容易受到洪水的影响，尤其是在极端气候事件下；流域特征影响降雨径流和洪水的形成与传播[29]；张海凤等[30]对由暴雨引起的洪涝灾害进行了具体分析，发现基础设施老旧恶化对洪水的发生也有着巨大影响；MCGRANAHAN[31]提及高人口密度区域在洪水发生时面临更大风险，湿地可以吸收洪水和缓冲洪峰，但湿地损失会增加洪水风险[32]。在了解各特征之后，为了清晰地看出这 21 个变量的数据分布，绘制变量分布图，如图 1 所示。从分布情况可以看出，前 20 个变量取值范围都在 0~15 之间，而大量的数据处在中小取值处，各个变量的取值都呈现从中间向两边减少的趋势。第 21 个特征洪水概率主要在 0. 3~0. 7 区间，符合概率条件。因此，从图 1 可以看出数据合理有效。

表 1　特征变量

Table 1　Feature variables

特　征	代表含义及取值方法	时空分辨率
季风强度	取风速和降水量加权平均值	多年平均
地形排水	使用地形坡度、土壤渗透率和地表覆盖物的综合指数	当前状况
河流管理	堤坝建设、河道疏浚等河流治理设施质量和数量的评估	当前状况
森林砍伐	该地区的森林砍伐程度，取森林覆盖率的变化率	年变化率
城市化	城市化的情况，取城市化率(城市人口占总人口的比率)	当前状况
气候变化	取用温度变异系数	多年平均
水坝质量	水坝的结构强度和维护状况评估	当前状况
淤　积	河床或水库中的沉积物量	年淤积量
农业实践	农田管理方式(如轮作、灌溉方式)的综合评估	当前状况
侵　占	自然保护区或洪泛区被开发的比例	年变化率
无效灾备	灾害应急计划的有效性评估	当前状况
排水系统	城市或区域排水系统的覆盖率	当前状况
沿海脆弱性	沿海区域对风暴潮、海平面上升等自然灾害的脆弱性评估	当前状况
山体滑坡	区域内滑坡发生的频率和潜在风险评估	多年平均
流　域	流域内水文和土地利用状况的综合评估	当前状况
基础设施恶化	基础设施的老化程度和维护状况评估	当前状况
人口评分	人口密度及其对洪水风险的影响评估	当前状况
湿地损失	湿地面积减少的比例	年变化率
规划不足	城市或区域规划对洪水风险管理的不足之处评估	当前状况
政治因素	政府政策对洪水管理的影响评估	当前状况
洪水概率	对历史洪水事件发生频率进行统计分析，以此计算洪水概率	当前状况

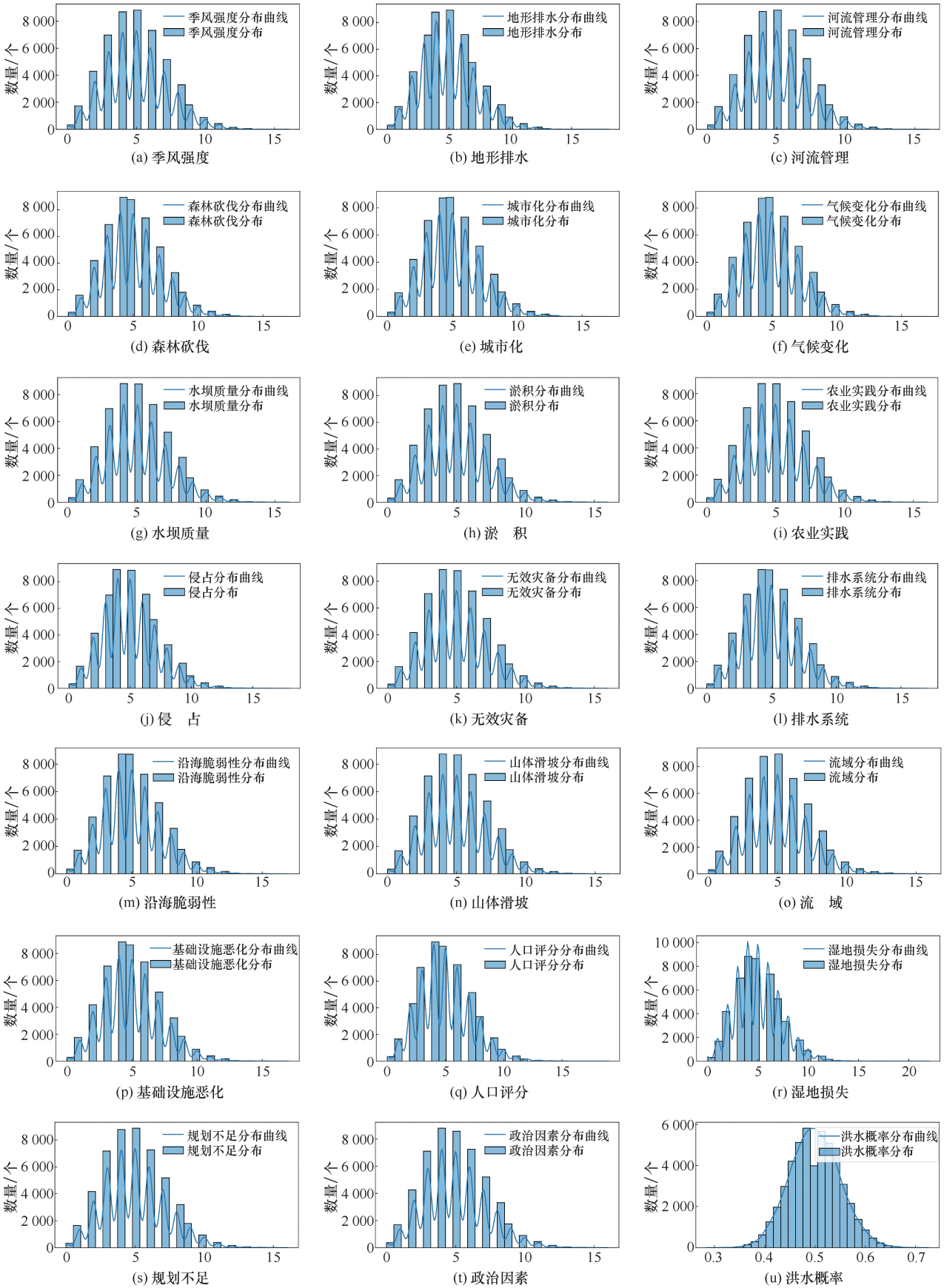


图 1　数据集变量分布直方图

Fig. 1　Distribution of dataset variables

2　数据处理及分析

首先读取数据，用平均值填充缺失值并用 Z-score 标准化方法对所有变量进行标准化。在原数据中，直接将所有的缺失值替换为相应列的平均值，然后进行相关性分析，绘制洪水相关变量热力图如图 2 所示。

由图 2 可以很直观地看出，除了洪水概率之外，其他的变量之间相互关系极小，几乎没有关联性。而洪水概率与其余的 20 个变量的关系程度相似。因此，以洪水概率作为目标变量，其余 20 个变量作为特征变量来进行预测比较合理，并且无须剔除任何特征变量。

接下来，将数据划分为目标变量和特征变量。从数据中删除表示洪水概率的列，得到特征矩阵 $\boldsymbol{X}$，并将该列作为目标变量 $\boldsymbol{y}$。由于要进行机器学习，因此将数据按 70% 的比例划分为训练集，剩余 30% 作为测试集。在对特征矩阵 $\boldsymbol{X}$ 进行标准化处理后，再将其与目标变量 $\boldsymbol{y}$ 一同拆分为训练集和测试集，之后便开始进行模型的训练和预测。

3　基于决策树的预测方法介绍

本研究按照图 3 所示流程进行，在完成数据的分析和处理之后选择了三种基于决策树的模型来进行训练。

决策树是一种树状模型，用于对数据进行分类或回归。随机森林模型由多个决策树组成，XGBoost 模型和 LightGBM 模型也是基于决策树的集成方法。

3.1　随机森林模型

研究的数据集包含多个特征，这些特征可能存在复杂的非线性关系。随机森林通过构建多棵决策树，可以捕捉这些复杂关系。随机森林能够很好地处理数

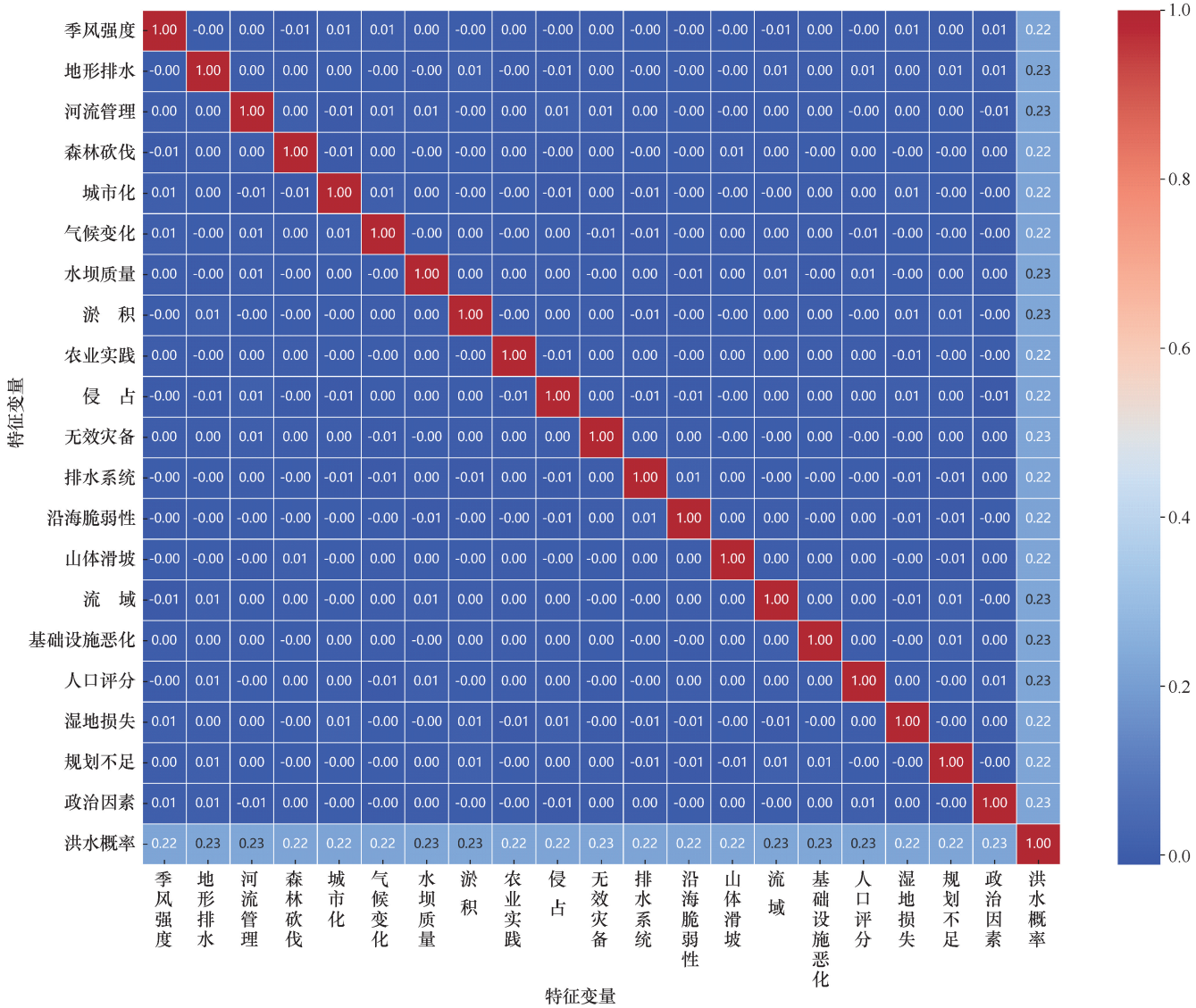


图 2　洪水相关变量热力图

Fig. 2　Heat map of flood-related variables

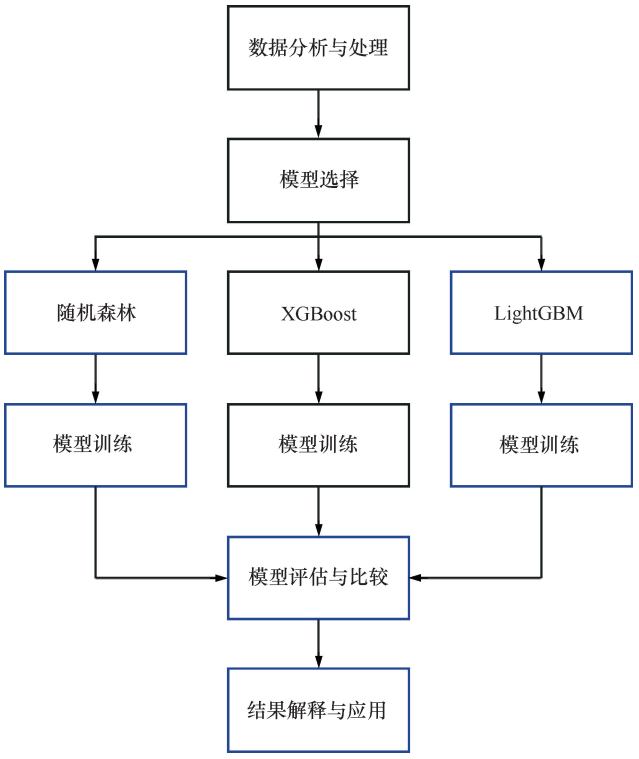


图 3　研究技术流程

Fig. 3　Technical workflow

据集中 21 种特征，并且自动识别较为重要的特征进行预测；同时通过对多棵树取平均，减少了过拟合的问题。结合其在回归和分类问题中的优秀表现，选用随机森林模型研究缺乏长期连续水文观测信息的洪水数据是合理的。

使用随机森林模型，首先要构建决策树。对于当前节点，选择一个特征 X_j 和一个阈值 t，根据 X_j 将数据集划分为两个子集

$$D_{\text{left}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid x_{ij} \leqslant t\} \tag{1}$$

$$D_{\text{right}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid x_{ij} > t\} \tag{2}$$

式中，$\boldsymbol{x}_i$ 为特征向量，表示单个数据点的输入特征，假设数据集中有 p 个特征，那么 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$；$\boldsymbol{y}_i$ 为目标变量，表示单个数据点的输出或结果；x_{ij} 表示特征向量 $\boldsymbol{x}_i$ 中的第 j 个特征，假设第 j 个特征是我们用来进行节点划分的特征。

左子节点 D_{left}：包含特征 x_{ij} 小于或等于阈值 t 的数据点，右子节点 D_{right} 包含特征 x_{ij} 大于阈值 t 的数据点。而选择的标准是使得分裂后的节点纯度最大化，这边用到均方误差 *MSE* 作为纯度度量的解释，分裂后的节点均方误差为

$$MSE_{\text{split}} = \frac{1}{N_{\text{left}}} \sum_{i \in D_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \frac{1}{N_{\text{right}}} \sum_{i \in D_{\text{right}}} (y_i - \bar{y}_{\text{right}})^2 \tag{3}$$

式中，N_{left} 和 N_{right} 分别为左子节点和右子节点的数据点数量；$\bar{y}_{\text{left}}$ 和 $\bar{y}_{\text{right}}$ 分别为左子节点和右子节点的目标变量均值。

左右子节点的目标变量均值分别为

$$\bar{y}_{\text{left}} = \frac{1}{N_{\text{left}}} \sum_{i \in D_{\text{left}}} y_i \tag{4}$$

$$\bar{y}_{\text{right}} = \frac{1}{N_{\text{right}}} \sum_{i \in D_{\text{right}}} y_i \tag{5}$$

通过上述公式和解释，可以帮助我们理解特征向量 $\boldsymbol{x}_i$、特征 x_{ij}、目标变量 y_i 以及节点划分的过程，在构建决策树时，算法会递归的分割数据集，选择最佳的特征和阈值来进行分裂。当达到预设的最大树深度或是当前节点的数据量小于预设的最小样本数，又或是节点的不纯度低于预设的阈值时，这种分裂停止。

随机森林通过构建多个决策树并结合其预测结果来提高模型的性能，具体步骤依次为[33]：先从训练集中有放回地随机抽取 n 个样本，形成一个新的子集，称为 bootstrap 样本。然后在每个节点划分时，随机选择 m 个特征（通常 $m = \sqrt{p}$，p 为总特征数），在这些特征中选择最优特征和阈值进行节点划分[34]；接着使用 bootstrap 样本和随机选择的特征，训练决策树；最后对于回归问题，随机森林模型的预测结果是所有决策树预测结果的平均值；对于分类问题，预测结果是所有决策树的投票结果。

其中对于回归模型，给定训练数据集 $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1, \cdots, n\}$，随机森林回归模型的预测公式为

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{6}$$

式中，$\hat{y}$ 为预测值；T 为决策树的数量；$h_t(x)$ 为第 t 棵决策树对输入 x 的预测结果。

本项研究中，选用平均减少不纯度来作为随机森林模型的重要性度量[33]，公式为

$$MDI(X_j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{m \in \text{nodes}} I_m \cdot 1(x_m = X_j) \tag{7}$$

式中，I_m 为节点 m 的不纯度减少量；$1(x_m = X_j)$ 表示节点 m 是由特征 X_j 分裂的指示函数。

3.2　XGBoost 模型和 LightGBM 模型

XGBoost 模型和 LightGBM 模型都使用了梯度提升框架，它们通过构建一个决策树的序列，每棵新树都试图纠正前一棵树的错误预测，从而逐步提高模型的预测性能；并且它们都能够有效地捕捉洪水预测中的非线性特征组合。由于数据集较大，XGBoost 和 LightGBM 在处理大规模数据集时表现尤为突出，它

们也提供了正则化手段进一步增强了模型的鲁棒性，防止在多特征的复杂数据集中过拟合。结合两者也都适用于回归问题与二分类问题，因此它们也成为研究缺乏长期连续水文观测信息的洪水数据的合理选择。

XGBoost 模型和 LightGBM 模型的差异体现在，XGBoost 通过逐步扫描特征来寻找最佳分裂点，使用基于样本的分裂方式，并且采用了层级分裂(Level-wise)的生长策略，这意味着它会在每个节点上考虑所有样本，并寻找最佳分裂点；而 LightGBM 通过直方图方法将特征值离散化[34]，极大地提高了训练速度，使用基于特征值的分裂方式，并且采用了叶子节点分裂(Leaf-wise)的生长策略，这意味着它会先分裂增益最大的叶子节点。不过两者使用的目标函数和预测公式相同[35]。

构建 XGBoost 模型和 LightGBM 模型，首先是给出目标函数(Objective Function)，它是模型要优化的整体函数。在模型中，目标函数通常由两部分组成，损失函数和正则化项[36]，即

$$Obj = \sum_{i=1}^{n} l(\hat{y}_i,\ y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{8}$$

目标函数用于衡量模型的预测误差和复杂度，训练损失函数训练的样本数量为 n，$l(\hat{y}_i,\ y_i)$ 为损失函数，用于度量预测值 $\hat{y}_i$ 和真实值 y_i 之间的误差；$\Omega(f_k)$ 为正则化项，用于控制模型的复杂度，防止过拟合；K 是树的数量。

对于损失函数，在回归任务中，常用均方误差(*MSE*)作为损失函数

$$l(\hat{y}_i,\ y_i) = (\hat{y}_i - y_i)^2 \tag{9}$$

在分类任务中，则常用对数损失(log loss)

$$l(\hat{y}_i,\ y_i) = -\left[y_i \log(\widehat{y_i}) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{10}$$

而正则化项包括树的复杂度和叶子节点权重的惩罚项

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2 \tag{11}$$

式中，γ 为控制树的叶子节点数量的参数；T_k 为 k 棵树的叶子节点数量；λ 为控制叶子节点权重的 L2 正则化参数；w_j 为第 j 个叶子节点的权重。

通过机器学习训练模型之后，代入预测公式得出预测结果

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{12}$$

式中，f_k 为第 k 棵树；x_i 为输入特征。

本项研究中 XGBoost 模型选用增益(*Gain*)作为重要性度量[36]，即

$$Gain(f) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} G_{f,\ i,\ j}}{\sum_{i=1}^{N} C_{f,\ i}} \tag{13}$$

式中，N 为树的总数；M_i 为第 i 棵树中的分裂点总数；$G_{f,\ i,\ j}$ 为特征 f 在第 i 棵树的第 j 个分裂点的增益；$C_{f,\ i}$ 为特征 f 在第 i 棵树中的分裂点次数。

LightGBM 模型在研究中使用次数(*Split*)来计算贡献度，代表特征 f 在所有树中作为分裂点出现的总次数，即

$$Split(f) = \sum_{i=1}^{N} C_{f,\ i} \tag{14}$$

3.3　模型评估指标

三种模型根据其在分类和回归方面的优势，对洪水进行风险量化预测和二分类预测，以应对不同需求。在进行洪水风险量化预测时，选择使用均方误差(*MSE*)和决定系数 R^2 来评估模型效果。*MSE* 越小，说明模型的预测结果越接近真实值。决定系数 R^2 为 1 表示模型完美地拟合了数据；0 表示模型的表现与只使用平均值预测一样好；负数则表示模型表现更差。当把目标设置为判断是否发生洪水时，洪水风险量化预测问题就转化为了洪水事件二分类预测问题。设定一个阈值 0.5，概率大于等于 0.5 视为“有洪水”，小于 0.5 视为“无洪水”。对于二分类问题，常用准确率、精确率、召回率、F1 值来评估模型性能，现决定用来评估模型[37]，计算公式为

$$\boldsymbol{Confusion\ Matrix} = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \tag{15}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

式中，*TP* 为实际为正类的样本被模型正确地预测为正类的数量；*TN* 为实际为负类的样本被模型正确地预测为负类的数量；*FP* 为实际为负类的样本被模型错误地预测为正类的数量；*FN* 为实际为正类的样本被模型错误地预测为负类的数量。

4　模型的评估与比较

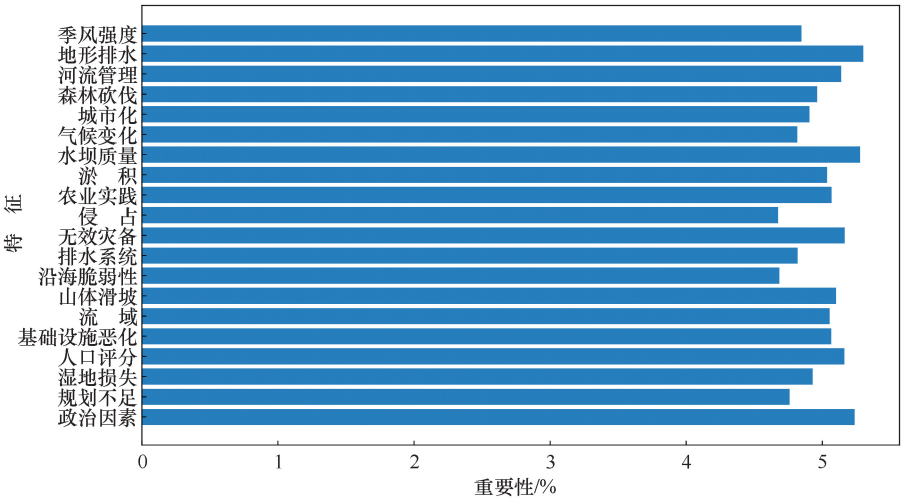
为了评估以上三种模型在面对缺乏长期连续水文

观测信息的数据时的洪水预测情况，本文进行了详尽的试验分析。可视化了使用不同模型进行预测时的特征重要性，用来了解哪些特征对模型的预测最为关键；计算了决定系数和均方误差，用来评估和比较不同模型在预测洪水发生的具体概率时的效果和误差；计算了如精确率、准确率、召回率、*F*1 分数和混淆矩阵等指标，用来评估模型关于预测是否发生洪水这种二分类问题中的模型性能，同样也对模型绘制了 ROC-AOC 曲线图横向对比了不同方法在这种情况下的优劣。

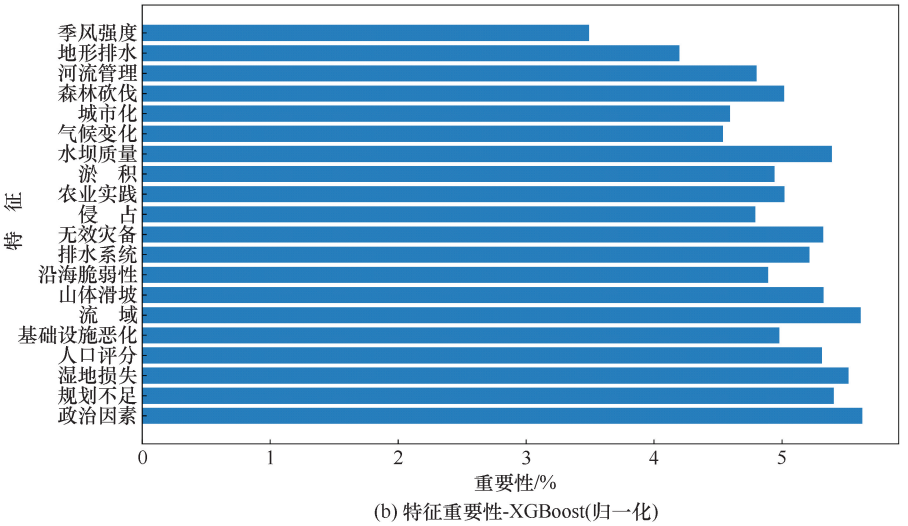
4.1　特征重要性的评估与比较

使用随机森林模型、XGBoost 模型和 LightGBM 模型进行训练和预测，通过使用 Python 软件，根据不同模型各自的特征重要性指标，输出相应的变量重要性柱状图，用来评估各模型的重要性。变量重要性如图 4 所示，能够很明确地看出不同变量对于洪水概率预测的重要性。

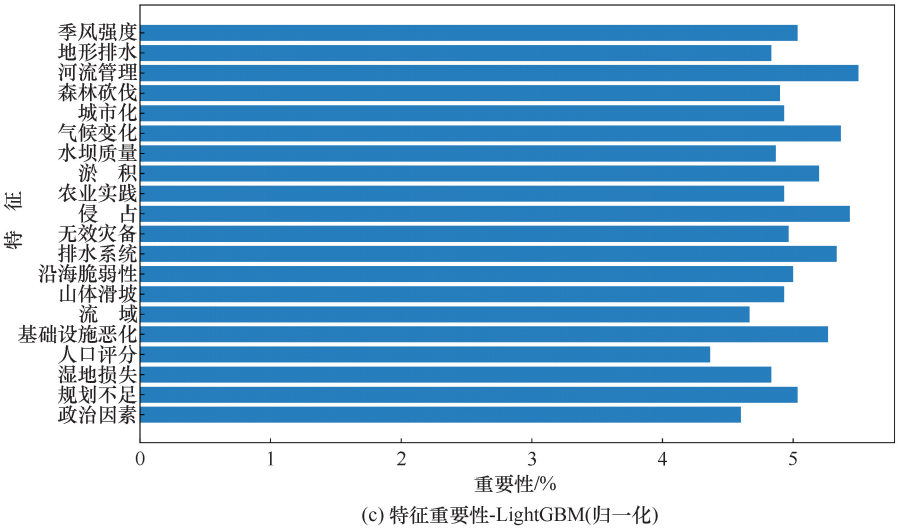
从图 4 中可以看到，每个特征都有一个对应的条形，条形的长度代表了该特征的重要性。条形越长，表示该特征对模型的预测结果影响越大。在随机森林模型的预测中，因为数据集中的特征具有较高的多样性，即每个特征都对目标变量有一定的影响，同时随机森林模型在构建决策树时，会随机选择特征和样本进行训练，这种随机性可能会导致各个特征的重要性较为接近。而其中“地形排水”“水坝质量”和“政治因素”等特征对于预测目标变量洪水概率最为关键。对于 XGBoost 模型预测来说，最重要的指标为“政治因素”“流域”“湿地损失”。而对于 LightGBM 模型，“河流管理”“气候变化”“侵占”等因素重要性更高。比较发现，使用不同模型时，发挥关键作用的特征并不相同，这有助于了解在使用不同模型时哪些因素最显著地影响了模型的预测结果。



(a) 特征重要性-随机森林(归一化)



(b) 特征重要性-XGBoost(归一化)



(c) 特征重要性-LightGBM(归一化)

图 4　变量重要性

Fig. 4　Bar charts of variable importance

4.2　洪水风险量化预测的评估与比较

对随机森林、XGBoost、LightGBM 三种模型分别进行训练，再分别用特征矩阵 $\boldsymbol{X}$ 的测试数据集进行预测，生成目标变量 y 的预测值，这些预测值将会用来与实际的目标变量 y 进行比较，以评估模型性能。最终得出来随机森林模型预测的均方误差 MSE = 0.000 686 1，R^2 = 0.728 4；XGBoost 模型预测的均方误差 MSE = 0.000 186 3，R^2 = 0.925 2；LightGBM 模型预测的均方误差 MSE = 0.000 191 9，R^2 = 0.922 9。可见三个模型均方误差的值都非常小，模型的预测结果与真实值十分接近，尤其是 XGBoost 模型误差最小；决定系数的值也是 XGBoost 最大，拟合效果较好，模型具有很好的解释能力，能够很好地捕捉洪水发生概率的变化趋势。而随机森林模型在数值上与另外两者有着较大的差距。

分别根据这三种模型的训练和预测结果，输出散点图(见图 5)。此图用来比较实际与预测洪水概率，横轴表示实际洪水概率，纵轴表示预测洪水概率，对角的红线是理想预测线，意味着如果预测完全准确，所有的点都应该落在这条线上。结果显示，第二个子图中 XGBoost 模型预测效果最为优良，基本所有点都在完美预测线上下；第三个子图中 LightGBM 模型有些许偏移，但总体效果依旧良好；而第一个子图中随机森林模型的预测结果与真实值相差较大，效果并不理想，可能因为随机森林在每棵树的分裂过程中都会随机选择一部分特征，这可能导致某些重要的特征未能在某些树中得到充分利用，从而影响整体模型的性能，而 XGBoost 模型在训练过程中会根据特征的重要性动态调整特征的选择，从而更好地利用重要的特征。

4.3　洪水事件二分类预测的评估与比较

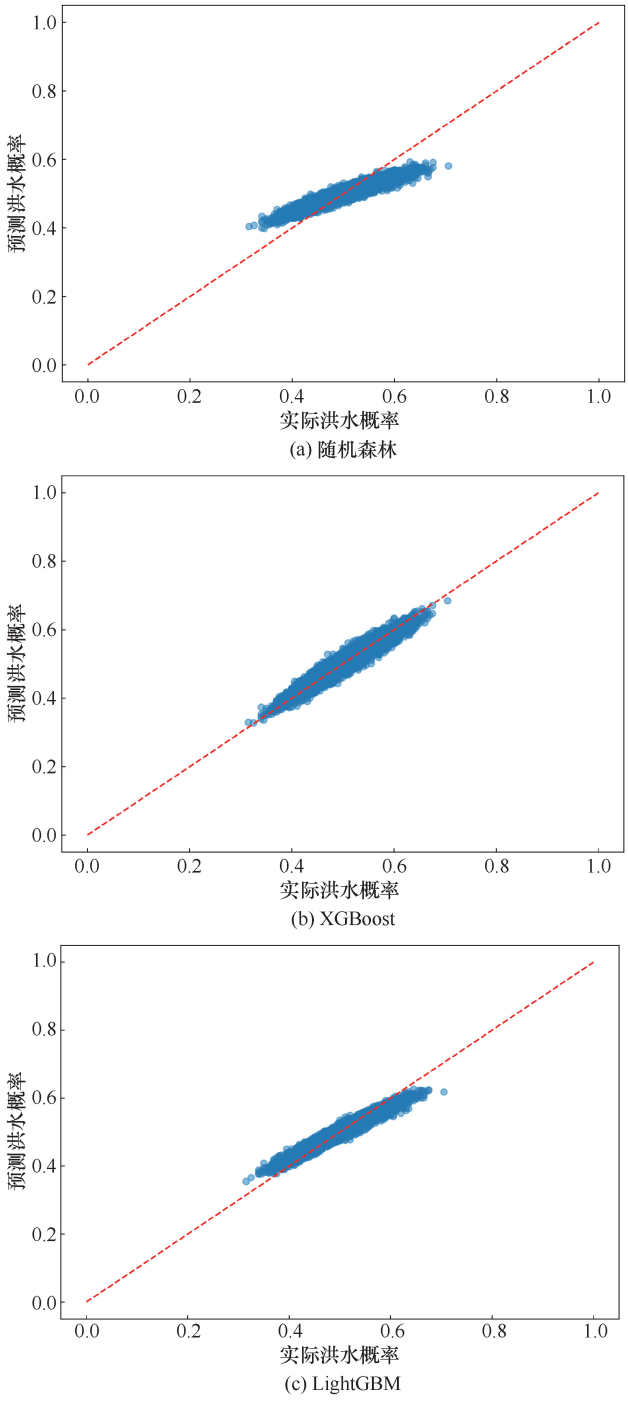
使用三种模型分别通过 Python 软件得到各自的混淆矩阵，如表 2 所列。

表 2　三种模型的混淆矩阵

Table 2　Confusion matrices of three models

混淆矩阵	随机森林模型		XGBoost 模型		LightGBM 模型	
	预测为正	预测为负	预测为正	预测为负	预测为正	预测为负
实际为正	6 757	577	4 508	313	4 591	230
实际为负	926	6 740	552	4 627	413	4 766

通过混淆矩阵中的这些数值，可据此计算出诸如准确率、召回率、精确率和 $F1$ 分数等各种性能指标，以全面评估分类模型的表现。各项性能指标输出结果如表 3 所列。



(a) 随机森林

(b) XGBoost

(c) LightGBM

图 5　预测结果与实际结果散点图

Fig. 5　Scatter plots of predicted and actual results

表 3　模型评估参数

Table 3　Model evaluation parameters

模　型	准确率	精确率	召回率	$F1$ 值
随机森林模型	0.899 8	0.921 1	0.879 2	0.899 7
XGBoost 模型	0.913 5	0.936 6	0.893 4	0.914 5
LightGBM 模型	0.935 7	0.954 0	0.920 3	0.936 8

准确率衡量的是所有预测结果中正确分类的比

例。较高的准确率意味着模型在整体上表现较好；精确率衡量的是预测为正类的样本中实际为正类的比例，精确率越高，表示模型在预测正类时的错误越少；召回率衡量的是实际为正类的样本中被正确预测为正类的比例，召回率越高，表示模型在捕获正类时的能力越强；*F*1 值是精确率和召回率的调和平均数，用于平衡两者之间的关系，*F*1 值越高，表示模型在精确率和召回率之间取得了较好的平衡。表 3 显示 LightGBM 模型的各项指标在三者中均为最好，准确率为 0.935 7，精确率为 0.954 0，召回率为 0.920 3，*F*1 值为 0.936 8；尤其在召回率上，LightGBM 模型对比其他两种模型都具有显著的优势，这意味着它在捕获洪水事件方面更为有效。随机森林模型在所有指标上的表现都相对较弱。

在分别对随机森林，XGBoost 和 LightGBM 三种模型进行评估之后，再选择运用 *ROC-AUC* 曲线对它们进行横向比较[38]。*ROC* 曲线常用于评价二分类模型的预测性能，利用真阳性率（*TPR*）和假阳性率（*FPR*）之间的关系曲线评估模型性能。真阳性率是在所有实际为正的样本中被正确预测为正样本的比例，即之前提到的召回率；假阳性率为所有实际为负的样本中，被错误预测为正样本的比例。*AUC* 是 *ROC* 曲线下的面积，*AUC* 值越接近 1，模型的性能越好。利用此曲线在 Python 帮助下，绘制 *ROC-AUC* 曲线（见图 6）。

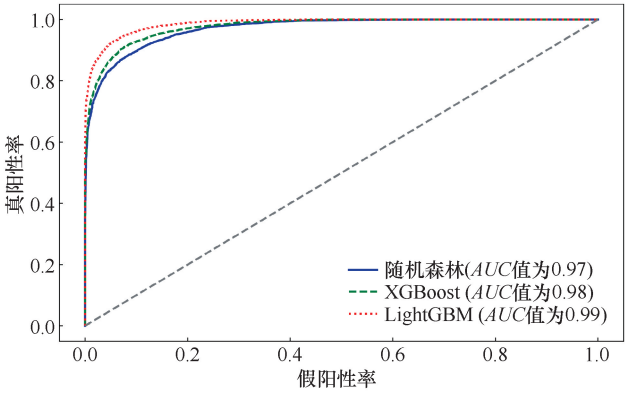


图 6　*ROC-AOC* 曲线

Fig. 6　*ROC-AOC* curves

图 6 可以很直观地看出决策树框架下的三种机器学习算法针对此类洪水预测都具有很好的预测效果，都非常接近于 1，其中 LightGBM 模型的 *AUC* 值更是高达 0.99，相比较而言随机森林模型的 *AUC* 值为 0.97 是这三种模型中效果略微差些的。

5　结果讨论

本研究旨在探索用随机森林模型、XGBoost 模型、LightGBM 模型等基于决策树的机器学习方法，在缺乏长期连续水文观测数据的情况下进行洪水预测的效果。通过对三种模型的应用，得到了一系列重要的结果。首先，在三种不同的模型预测洪水概率时，对结果影响最大的特征并不相同。随机森林模型中“地形排水”“水坝质量”和“政治因素”等特征最为重要；XGBoost 模型中“政治因素”“流域”“湿地损失”等特征最为重要；LightGBM 模型中“河流管理”“气候变化”“侵占”等特征最为重要。其次，当预测着眼于洪水风险量化预测时，XGBoost 模型取得的预测效果是最好的，LightGBM 略微差于 XGBoost 模型，而随机森林模型效果最差。XGBoost 模型的均方误差 $MSE=0.000\,186\,3$，决定系数 $R^2=0.925\,2$，分别说明了它在洪水预测结果上的低误差和在捕获洪水概率变化趋势上的高效果。最后，当预测着眼于洪水事件二分类预测，即判断洪水是否会发生的问题时，LightGBM 模型的效果是最好的，接下来是 XGBoost 模型，随机森林模型同样效果没有那么理想。LightGBM 模型的准确率为 0.935 7，精确率为 0.954 0，召回率为 0.920 3，*F*1 值为 0.936 8，AUC 值为 0.99，意味着它在捕获洪水事件上具有很好的效果，模型的性能很好，能够准确地判别洪水是否会发生这一问题。

研究结果表明随机森林、XGBoost、LightGBM 在预测中发挥重要作用的特征并不相同，在某些缺乏长期连续水文观测数据的地域或情况下，数据往往存在缺失或者不足的情况，只能通过有限的相关影响特征来进行洪水预测，此时便可以按照使用的方法，优先收集更为重要的变量和数据，尽可能多地提高预测的准确性，也可以针对性的对影响最大的特征进行改善来减少洪水的概率和风险。在运用中，如果想要预测某地域发生洪水的具体概率，优先考虑使用 XGBoost 模型，它在洪水风险量化预测中效果最好；而如果想要知道某地是否会发生洪水，将可能性划分为可能和不可能时，则优先选用 LightGBM 模型，它不仅在捕获洪水事件上效果最好，还因为独特的直方图算法有很快的训练速度。当前我国水灾多发，因此，对各地做好相关预测，掌握各地水灾风险，对影响水灾较多的特征优先针对性处理改善，可以在一定程度上减少水灾发生的可能。

研究过程中发现，此类基于决策树的模型在洪水

预测问题中的应用具有很强的解释性，它们的决策过程非常直观，每个分裂节点代表一个明确的规则，易于理解和解释。相比 SVM 的决策边界通常是难以解释的[39]，尤其是当使用复杂的核函数时；神经网络通常被视为“黑盒”模型，其内部决策过程难以解释[40]。除此之外，决策树方法易于处理缺失值，通过在分裂节点时使用替代策略来处理缺失特征值，它还可以并行化训练，适合处理大规模数据集。而 SVM 不直接支持处理缺失值，需要对缺失值进行填充或删除，且在大规模数据集上的训练和预测较慢，尤其是使用核技巧时。

尽管本文提供了诸多有价值的信息，但依然存在局限性。决策树的构建是一个贪心算法，每次分裂都试图找到最佳分割点，这可能导致局部最优而不是全局最优。而 ANN 虽然也可能陷入局部最小值，但通过适当的初始化和优化算法（如 Adam、RMSprop 等），可以更好地寻找全局最优解[41]；SVM 通过优化问题求解，也能保证找到全局最优解[42]。已有研究通过人工神经网络（ANN）和支持向量机（SVM）对巴基斯坦北部潘吉科拉河流域洪水事件进行预测[43]，决定系数 R^2 分别为 0.75 和 0.60，从此结果上显示本文的 XGBoost 模型和 LightGBM 模型预测效果准确性是高于 ANN 和 SVM 的，但是过高的决定系数也同时存在着因为特征过多导致的过拟合问题。最后在洪水事件二分类预测问题中，本文考虑将洪水事件分为“发生”和“不发生”两类，虽然直观并且取得了很好的效果，但是可能不足以捕捉洪水发生的复杂性和多样性，当分出更多的可能和类别，也许效果也会相应变差。

为了进一步巩固本研究的成果并扩展其应用范围，未来的研究考虑将洪水事件分为多个类别（多分类）可能会更加合理和实用。如轻微、中等、严重等，可以更准确地反映洪水的程度和影响。多分类模型可以提供更细粒度的预测结果，有助于制定更具体的应急响应计划。届时使用多分类评估指标，如多类准确率、宏平均精确率、宏平均召回率、宏平均 $F1$ 分数等来进一步衡量预测的效果。

6　结　论

在本研究中，我们使用了三种基于决策树的机器学习技术——随机森林（RF）、XGBoost 和 LightGBM，在缺乏长期连续水文观测数据的情况下进行洪水预测。研究结果表明如下。

（1）当要具体的预测洪水概率时，选用 XGBoost 模型为最佳选择。模型的均方误差 $MSE = 0.000\ 186\ 3$，决定系数 $R^2 = 0.925\ 2$ 均优于其余两种模型。

（2）当要判断洪水是否会发生时，选用 LightGBM 模型为最佳选择。模型的准确率为 0.935 7，精确率为 0.954 0，召回率为 0.920 3，$F1$ 值为 0.936 8，AUC 值为 0.99，均意味着模型在捕获洪水事件上优良的性能。

（3）当选用不同方法进行预测时，起关键作用的特征变量并不相同，均要根据所选方法而定。

研究的结果使得在许多没有长期观测数据、信息不全、数据缺失的地区可以通过无关时间、地域的相关特征进行合理有效的洪水预测，了解引发洪水的风险程度和欠缺的相关措施，并且有针对性地、有效地进行预防。尽管本研究还存在一定的局限性，但为洪水预测领域提供了新的视角，并为未来的研究奠定了基础。

参考文献（References）：

[1] 李国英. 为以中国式现代化全面推进强国建设、民族复兴伟业提供有力的水安全保障：在 2024 年全国水利工作会议上的讲话[J]. 水利发展研究，2024，24(1)：1-10.
LI G Y. Improved water security for China's efforts to build itself into a stronger country and rejuvenate the Chinese nation on all fronts by pursuing Chinese modernization：Speech at the 2024 National Water Conservancy Work Conference[J]. Water Resources Development Research，2024，24(1)：1-10.

[2] EL-BAGOURY H，GAD A. Integrated hydrological modeling for watershed analysis，flood prediction，and mitigation using meteorological and morphometric data，SCS-CN，HEC-HMS/RAS，and QGIS[J]. Water，2024，16(2)：356.

[3] MURPHY K P. Probabilistic machine learning：An introduction[M]. Cambridge：MIT press，2022.

[4] UTOMO D，HU L C，HSIUNG P A. Deep neural network-based spatiotemporal heterogeneous data reconstruction for landslide detection[J]. International Journal of Data Science and Analytics，2024，17(1)：93-109.

[5] SUN F Y，HAO W N，ZOU A，et al. A survey on spatio-temporal series prediction with deep learning：Taxonomy，applications，and future directions[J]. Neural Computing and Applications，2024，36(17)：9919-9943.

[6] HAN D，CHAN L，ZHU N. Flood forecasting using support vector machines[J]. Journal of Hydroinformatics，2007，9(4)：267-276.

[7] AL SAWAF M B，KAWANISI K，JLILATI M N，et al. Extent of detection of hidden relationships among different hydrological variables during floods using data-driven models[J]. Environmental Monitoring and Assessment，2021，193(11)：692.

[8] COSTA V G, PEDREIRA C E. Recent advances in decision trees: an updated survey[J]. Artificial Intelligence Review, 2023, 56(5): 4765-4800.

[9] SAHOO A, GHOSE D K. Flood Forecasting Using Hybrid SVM-GOA Model: A Case Study[M]. Singapore: Springer Nature Singapore, 2022: 407-416.

[10] AGARWAL S, ROY P J, CHOUDHURY P, et al. Flood forecasting and flood flow modeling in a river system using ANN[J]. Water Practice and Technology, 2021, 16(4): 1194-1205.

[11] LAWAL Z K, YASSIN H, ZAKARI R Y. Flood prediction using machine learning models: A case study of kebbi state Nigeria[C]// IEEE. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). Brisbane, Australia, 2021: 1-6.

[12] SYEED M M A, FARZANA M, NAMIR I, et al. Flood prediction using machine learning models[C]//IEEE. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). Ankara, Turkey, 2022: 1-6.

[13] GRZESIAK M, THAKKAR P. Flood prediction using classical and quantum machine learning models[EB/OL]. 2024: 2407.01001. https: //arxiv.org/abs/2407.01001v1.

[14] BOATENG E Y, OTOO J, ABAYE D A. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review[J]. Journal of Data Analysis and Information Processing, 2020, 8(4): 341-357.

[15] SCHOPPA L, DISSE M, BACHMAIR S. Evaluating the performance of random forest for large-scale flood discharge simulation[J]. Journal of Hydrology, 2020, 590: 125531.

[16] SAGI O, ROKACH L. Approximating XGBoost with an interpretable decision tree[J]. Information Sciences, 2021, 572: 522-542.

[17] AHN J M, KIM J, KIM K. Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting[J]. Toxins, 2023, 15(10): 608.

[18] SAHASTRABUDDHE R, GHAUSI S A, JOSEPH J, et al. Indian summer monsoon rainfall in a changing climate: A review[J]. Journal of Water and Climate Change, 2023, 14(4): 1061-1088.

[19] GOSWAMI B N, VENUGOPAL V, SENGUPTA D, et al. Increasing trend of extreme rain events over India in a warming environment[J]. Science, 2006, 314(5804): 1442-1445.

[20] FLEISCHMANN A, PAIVA R, COLLISCHONN W. Can regional to continental river hydrodynamic models be locally relevant? A cross-scale comparison[J]. Journal of Hydrology X, 2019, 3: 100027.

[21] POFF N L, ALLAN J D, BAIN M B, et al. A paradigm for river conservation and restoration[J]. BioScience, 1997, 47(11): 769-784.

[22] JACOBSON C R. Identification and quantification of the hydrological impacts of imperviousness in urban catchments: a review[J]. Journal of Environmental Management, 2011, 92(6): 1438-1448.

[23] Intergovernmental Panel on Climate Change (IPCC). Climate Change 2014: Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects[M]. Cambridge, UK: Cambridge University Press, 2014.

[24] WALLING D E. Human impact on land-ocean sediment transfer by the world's rivers[J]. Geomorphology, 2006, 79(3/4): 192-216.

[25] PIMENTEL D, HARVEY C, RESOSUDARMO P, et al. Environmental and economic costs of soil erosion and conservation benefits[J]. Science, 1995, 267(5201): 1117-1123.

[26] TIERNEY K J. Recent Developments in U. S. Homeland Security Policies and Their Implications for the Management of extreme Events [M]. New York, NY: Springer New York, 2007: 405-412.

[27] 米胤瑜, 孔锋. 气候变化背景下城市洪水风险管理体系国际比较与启示: 以伦敦、纽约、郑州为例[J]. 水利水电技术(中英文), 2023, 54(3): 21-34.
MI Y Y, KONG F. International comparison and enlightenment of urban flood risk management system under background of climate change: Taking London, New York and Zhengzhou as study cases [J]. Water Resources and Hydropower Engineering, 2023, 54(3): 21-34.

[28] WONG P P, LOSADA I J, GATTUSO J P, et al. Coastal systems and low-lying areas[J]. Climate Change, 2014, 2104: 361-409.

[29] WAGENER T, SIVAPALAN M, TROCH P, et al. Catchment classification and hydrologic similarity[J]. Geography Compass, 2007, 1(4): 901-931.

[30] 张海凤, 孔锋, 方建. 超常规极端暴雨洪涝灾害应对的国际比较研究: 以 2021 年中美德暴雨洪涝灾害为例[J]. 水利水电技术(中英文), 2023, 54(7): 1-13.
ZHANG Haifeng, KONG Feng, FANG Jian. International comparative study on coping with flood-waterlogging disaster from extraordinary rainstorm: Taking rainstorm flood-waterlogging disasters in China, America and Germany in 2021 as study cases[J]. Water Resources and Hydropower Engineering, 2023, 54(7): 1-13.

[31] MCGRANAHAN G, BALK D, ANDERSON B. The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones[J]. Environment and Urbanization, 2007, 19(1): 17-37.

[32] ZEDLER J B, KERCHER S. Wetland Resources: status, trends, ecosystem services, and restorability[J]. Annual Review of Environment and Resources, 2005, 30: 39-74.

[33] BREIMAN L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[34] LIAW A, WIENER M. Classification and regression by randomForest [J]. R News, 2002, 2(3): 18-22.

[35] GUOLINKE Q M, FINLEY T, WANG T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in Neural

Information Rocessing Systems, 2017, 30: 52.

[36] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system [C]//ACM. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 2016: 785-794.

[37] 李红艳, 郝景开, 刘大为, 等. 基于元启发式算法优化的洪水风险评价模型[J]. 水资源保护, 2024, 40(6): 1-15.
LI Hongyan, HAO Jingkai, LIU Dawei, et al. Flood risk assessment model optimized based on meta-heuristic algorithm [J]. Water Resources Protection, 2024, 40(6): 1-15.

[38] 桑珍珍, 崔杰, 闫寒, 等. 基于随机森林和 XGBoost 算法构建心脏骤停患者自主循环恢复后神经功能预后不良的风险预测模型[J]. 中国急救医学, 2024, 44(7): 577-585.
SANG Z Z, CUI J, YAN H, et al. Construction of prediction model for poor prognosis of neurological function in cardiac arrest patients after the return of spontaneous circulation based on Random Forest and XGBoost algorithms[J]. Chinese Journal of Critical Care Medicine, 2024, 44(7): 577-585.

[39] VINGE R, MCKELVEY T. Understanding support vector machines with polynomial kernels [C]//IEEE. 2019 27th European Signal Processing Conference (EUSIPCO). Spain: A Coruna, 2019: 1-5.

[40] BUHRMESTER V, MÜNCH D, ARENS M. Analysis of explainers of black box deep neural networks for computer vision: A survey[J]. Machine Learning and Knowledge Extraction, 2021, 3(4): 966-989.

[41] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[42] DIETRICH R, OPPER M, SOMPOLINSKY H. Statistical mechanics of support vector networks[J]. Physical Review Letters, 1999, 82(14): 2975-2978.

[43] ALI M, TAHA M, AZIZ M S, et al. Flash flood prediction of panjkora river, kpk, using artificial neural networks (ANN) and support vector machine (SVM)[J]. Technical Journal, 2024, 3(ICACEE): 758-769.

(责任编辑　王　璐)