基于数据分析与机器学习的洪水灾害预测与风险评估

张 舒[1]
ZHANG Shu

摘 要 洪水是由多种自然因素引发的，包括暴雨、冰雪融化和风暴潮等，属于极具破坏性的自然灾害。这些灾害不仅会对基础设施造成严重损毁，还会对人类生命安全和生态环境产生深远的负面影响。随着全球气候变化的加剧和人类活动的日益频繁，洪灾的发生频率和严重程度显著增加，导致全球范围内的防洪压力不断加大。因此，研究和应对洪灾的成因及其影响，成为当前环境科学和灾害管理领域的重要课题。在实际应用中，获取和处理所有潜在指标的数据成本较高，因此选择少量关键指标显得尤为重要。结合数据分析与机器学习技术，预测洪水发生概率，并提供科学的预防与应对措施。采用两种方法选取关键指标：首先，采用斯皮尔曼相关系数来确定与洪灾发生高度相关的指标；其次，通过改进的 K-means 聚类方法将洪水事件风险分为高、中、低三级，并利用随机森林分类器选择重要特征，建立风险预警评价模型。基于线性回归、决策树、随机森林和多重感知机等多种机器学习模型，构建了洪水发生概率预测模型。研究结果表明，优化后的特征选择方法，尤其是通过风险分级并结合随机森林分类器，显著提高了模型的预测准确性和泛化能力，同时提升了模型的运行效率和决策的可解释性。所提出的方法为洪水风险评估和防灾减灾提供了更加高效和可靠的技术支持。

关键词 洪水灾害；数据分析；机器学习；K-means 聚类；随机森林；洪水预测模型

doi：10.3969/j.issn.1672-9528.2024.08.044

0 引言

洪水是由多种自然因素引发的现象，包括暴雨、冰雪融化、风暴潮等，是全球范围内常见且破坏性极大的自然灾害。洪水不仅会对基础设施造成严重损毁，还会对人类生命安全和生态环境产生深远的负面影响。近年来，全球气候变化加剧，导致极端天气事件频发，洪灾的发生频率和严重程度显著增加。这种趋势不仅增加了防洪减灾的难度，还对社会经济发展和生态平衡构成了巨大威胁。同时，人类活动如森林砍伐、城市化扩展和不合理的土地利用，进一步加剧了洪水风险，造成了重大经济损失和生命财产威胁。因此，对洪水灾害的预测[1]与风险评估进行研究显得尤为重要。

传统的洪水预测方法[2]主要依赖水文学和气象学模型，这些模型在处理复杂多变的洪水成因时，往往预测准确性有限。此外，获取和处理所有潜在影响指标的数据成本较高，使得模型的实际应用面临挑战。近年来，随着大数据技术[3]和机器学习[4]方法的发展，基于数据驱动的洪水预测方法[5-7]逐渐受到关注。这些方法通过利用大量历史数据和复杂算法，如随机森林、支持向量机和神经网络，显著提升了洪水预测的性能。然而，现有研究在特征选择和模型优化方面仍存在不足，影响了模型的应用效果和实际操作的可行性。

为了克服这些不足，本研究提出了一种结合改进的聚类算法和多种机器学习模型的洪水预测方法。首先，采用 Spearman 相关系数来筛选与洪灾发生高度相关的关键指标。

然后，利用改进的 K-means 聚类方法将洪水事件的风险等级划分为高、中、低三个级别，并通过随机森林分类器选择最重要的特征，构建风险预警评估模型。最后，基于线性回归、决策树、随机森林[8-10]和多重感知机[11-12]等多种机器学习模型，构建洪水发生概率预测模型，并通过均方误差和决定系数对模型进行优化。本研究的目标是通过优化特征选择方法和模型结构，提高洪水预测的准确性和效率，并增强模型的决策可解释性，从而为洪水风险评估和防灾减灾[13]提供科学依据和技术支持。

1 数据分析与处理

1.1 数据描述

本研究使用的数据集来自 Kaggle 平台，包含了超过 100 万条与洪水事件相关的记录。该数据集为本研究提供了丰富的数据源，能够支持本文利用数据分析与机器学习技术，构

1. 济南浚达信息技术有限公司 山东济南 250101

建洪水发生概率预测模型，并为洪水风险评估和预防措施的制定提供科学依据。

1.2 数据预处理

在数据预处理中，对数据集进行了完整性检查，确认不存在缺失值，因此无需进行填补或删除操作。接着，通过箱线图识别并剔除少量异常值，以确保数据的质量和一致性。图 1 展示了各特征的箱型图和异常值分布。如图 1 所示，箱型图显示了每个属性的五个统计量：最小值、第一四分位数（Q1）、中位数（Q2）、第三四分位数（Q3）和最大值，同时也展示了异常值的分布情况。

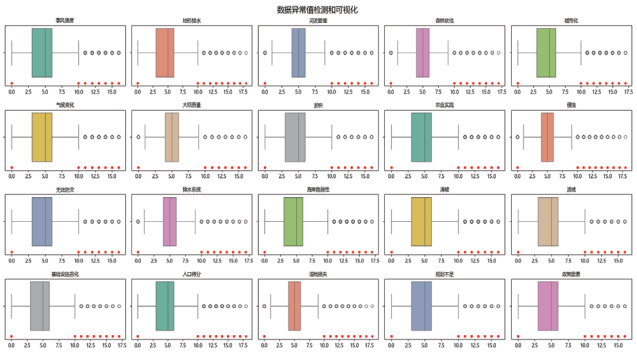

图 1 箱型图

从图 2 中可以看到，多数属性的中位数（箱体中的线）较低，表明数据倾向于较小的值，而一些属性的箱体较宽，例如河流管理和森林砍伐，表明数据在这些属性上的分散度较大。同时，注意到指标值大于 10 的数据点占比非常小（见图 2），它们的占比约等于 0，因此将这些数据点视为异常值并删除掉。

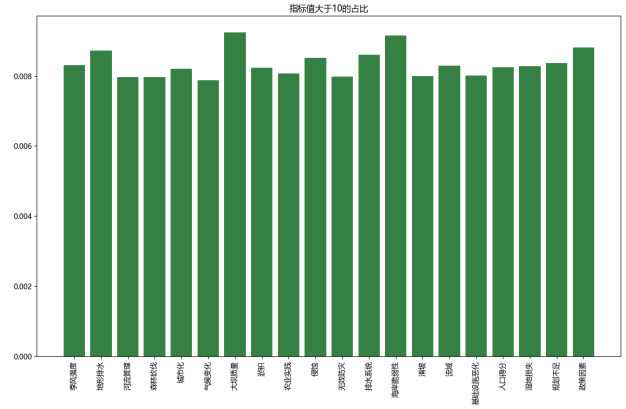

图 2 异常值数据占比统计

2 Spearman 相关系数选择特征

为了确定与洪水发生密切相关的关键指标，首先使用 Spearman 相关系数进行了相关性分析。作为一种非参数统计方法，Spearman 相关系数用于衡量两个变量之间的单调关系，适用于变量间即便非线性关系也能进行有效测量。相比于皮尔逊相关系数，Spearman 相关系数对异常值和非正态分布的数据更为鲁棒，因此在处理复杂环境数据时具有优势。通过计算每个指标与洪水发生概率之间的相关系数，发现季风强度、地形排水、基础设施恶化、河流管理、气候变化和大坝质量等指标与洪水发生存在显著相关性。可以选取这些相关性较强的指标作为模型的预测输入，从而提高模型的准确性和可靠性。Spearman 相关系数的矩阵图如图 3 所示。

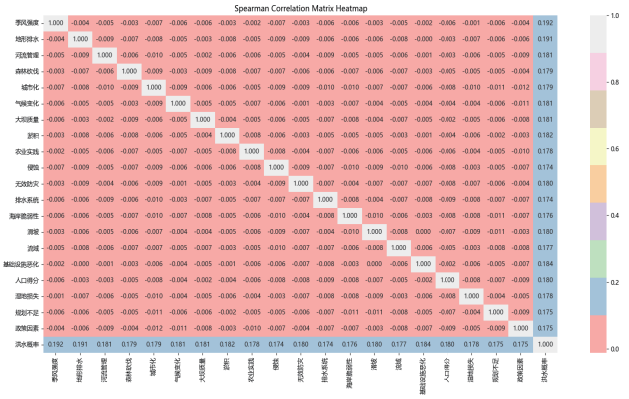

图 3 Spearman 相关系数的矩阵图

3 改进的 K-means 聚类方法和随机森林分类器选择特征

在风险预警采用改进的 K-means 聚类方法对洪水事件进行风险分类，并结合随机森林算法建立了风险预警评价模型，通过精细化的风险分类和高效的预测模型，提升防洪管理的科学性和应对效率。改进的 K-means 方法能够更准确地划分高、中、低风险指标，从而优化资源配置，确保高风险区域得到优先保护。随机森林算法则提供了强大的预测能力和稳健性，使得预警系统能够更准确地预测洪水发生概率，及时发布预警信息。这一组合方法不仅提升了风险评估的准确性和可靠性，还增强了应急响应的效率和效果，有助于减少洪水灾害对公共安全和财产的影响，促进社会的可持续发展和安全保障。

3.1 改进的 K-means 聚类方法

改进的 K-means 聚类方法通过选择尽可能远的聚类质心点来优化初始聚类中心的选择。这种方法增加了算法的稳定性和聚类结果的一致性，更适合处理复杂多维数据，从而提高分类准确性，减少迭代次数和计算成本，提升效率，并能更精准地识别不同风险等级特征，从而制定更有效的防洪措施和资源配置策略，具体算法原理如图 4 所示。使用改进 K-means 聚类将洪水事件聚为高风险、中风险、低风险 3 类，基于聚类结果，分析具有高、中、低风险的洪水事件的各个指标特征，通过计算每个风险类别的各指标均值，绘制各指标在不同风险的均值条形图，可以直观地体现出不同风险类别的特征差异。

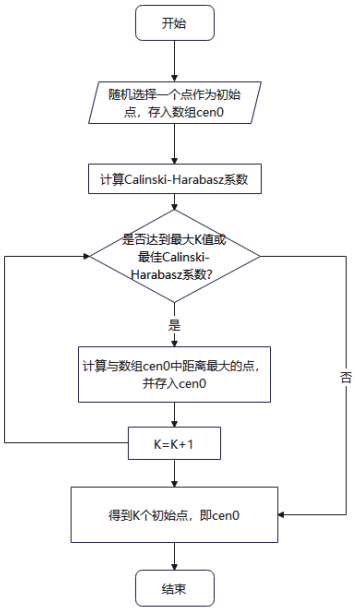

图 4 改进 K-Means 聚类流程图

图 5 展示了不同洪水风险类别在各特征上的均值，每个子图分别代表一个特征，横轴为洪水风险类别，纵轴为各个特征的均值，有助于识别哪些特征对洪水风险评估具有较高的重要性，这些信息可以作为重要特征选择的依据，从图中可以观察到不同风险类别的指标特征。

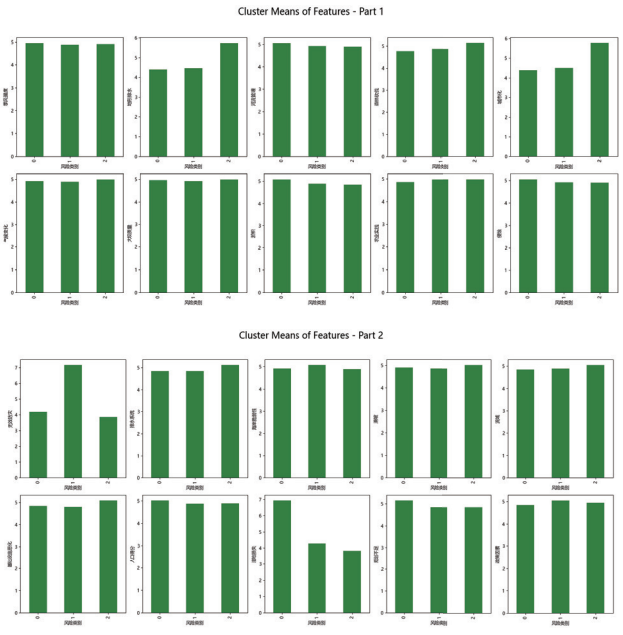

图 5 各指标在不同风险的均值条形图

3.2 随机森林分类器

随机森林是一种集成学习方法，通过构建多棵独立的决策树来完成分类或回归任务，具有降低过拟合风险和提高模型稳定性的优势。每棵树通过从原始数据中随机抽取子样本（Bootstrap 抽样）和随机选择特征进行训练。最终的预测结果通过所有树的投票（分类任务）或取平均（回归任务）得到。主要步骤包括：首先，进行 Bootstrap 抽样构建多个训练集；然后，在每个节点上基于随机选择的特征进行最优划分，直至达到预定的停止条件；接着，重复上述过程构建多棵独立的决策树；最后，集成所有树的预测结果。随机森林分流器流程图如图 6 所示。

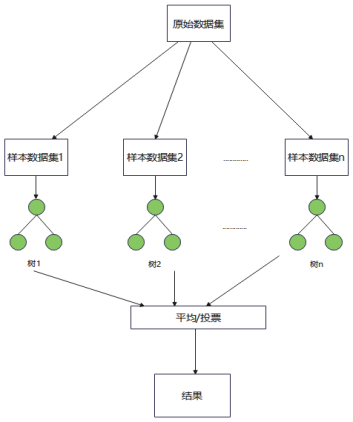

图 6 随机森林分流器流程图

随机森林是一种强大的集成学习方法，通过构建多棵决策树并对它们的结果进行平均或投票来提高预测的准确性和稳定性。特征重要性是通过评估每个特征在决策树节点分裂时对不纯度减少的贡献来计算的。具体步骤包括：首先构建随机森林模型，然后在每个节点选择最佳特征进行分裂并计算相应的不纯度减少量，最后将所有树中的不纯度减少量进行平均，得出每个特征的重要性得分。最终得到的指标权重如表 1，并绘制指标重要性的条形图，可以直观地看到特征重要性之间的差异。

表 1 指标权重

指标	权重	指标	权重
季风强度	0.010 6	无效防灾	0.335 8
地形排水	0.081 6	排水系统	0.012 7
河流管理	0.011 8	海岸脆弱性	0.012 0
森林砍伐	0.013 6	滑坡	0.011 2
城市化	0.084 9	流域	0.011 5
气候变化	0.010 9	基础设施恶化	0.012 8
大坝质量	0.010 7	人口得分	0.011 6
淤积	0.012 5	湿地损失	0.306 1
农业实践	0.011 2	规划不足	0.013 8
侵蚀	0.011 6	政策因素	0.012 1

表 1 展示了使用随机森林分类器计算得到的各特征的重要性，图 7 可以直观地看出不同指标在模型中的相对重要性。结果显示，无效防灾、湿地损失和城市化是最重要的影响指标。无效防灾的重要性明显高于其他指标，对模型的预测性能起到关键作用。

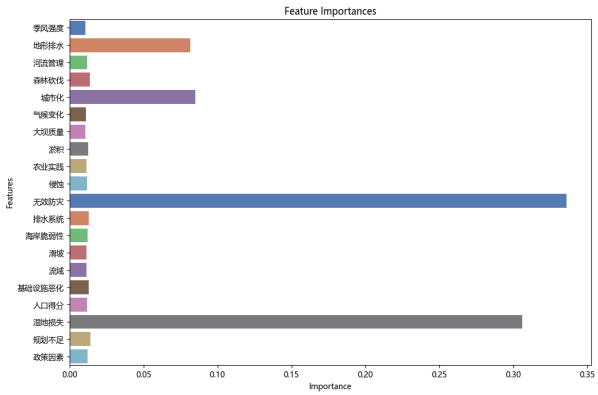

图 7 指标权重条形图

从表 1 中可以得出，无效防灾是最重要的影响指标，权重为 0.335 8；其次影响程度较大的为湿地损失，权重为 0.306 1；其他重要性较强的指标为城市化、地形排水和规划不足，权重依次为 0.084 9、0.081 6、0.013 8。上述五个指标在提高洪水风险预测准确性方面具有重要作用。从图 7 可以更加直观地看出各指标的重要性差异，无效防灾的重要性明显高于其他指标，它对风险预测评价模型起到重要作用，其次是湿地损失的重要性较为突出，然后是城市化、地形排水和规划不足的权重较为突出，其他指标权重较小。

最终选取重要性排名前 5 的指标，作为模型的输入指标，建立基于随机分林分类器的洪水风险预警评价模型。

为分析选取的五个指标分别对识别哪一类风险具有突出贡献，绘制了小提琴图，展示了重要特征在不同洪水风险类别上的分布情况。

由图 8 可以看出，无效防灾在不同风险分布上有明显的差异，风险类别 1 的总体分布要明显高于风险 0 和 2，说明规划不足这个指标能够更好地识别中风险洪水事件，同样，湿地损失和规划不足在识别低风险洪水事件（风险类别 0）上具有优势，而城市化和地形排水在识别高风险的洪水事件上具有优势。

图 8 指标风险类别小提琴图

3.3 模型评价和灵敏度分析

对建立的洪水风险预警评价模型进行训练，并使用精确率（precision）、召回率（recall）和 F_1 分数来衡量模型在每个类别上的分类准确性。此外，采用准确率（Accuracy）、宏平均（Macro Avg）和加权平均（Weighted Avg）对模型的整体效果和性能进行评估。

进一步对模型进行灵敏度分析，通过对选取的 5 个重要指标依次移除，重新训练模型，得到模型的评估指标值，与未移除特征的评估结果进行对比，分析各个指标对模型性能的影响。

由表 2 可以看出，全部引入 5 个重要指标初始模型的准确率高达 94%，且对每种风险类别的识别准确率都高于 94%，说明模型的分类效果和泛化能力很好。

表 2 模型评价与灵敏度分析

未移除指标	初始模型			移除指标	规划不足		
类别	Precision	Recall	F_1-score	类别	Precision	Recall	F_1-score
0	0.93	0.93	0.93	0	0.91	0.92	0.92
1	0.95	0.94	0.94	1	0.94	0.92	0.93
2	0.93	0.94	0.94	2	0.92	0.93	0.93
Accuracy			0.94	Accuracy			0.92
Macro Avg	0.94	0.94	0.94	Macro Avg	0.92	0.92	0.92
Weighted Avg	0.94	0.94	0.94	Weighted Avg	0.92	0.92	0.92
移除指标	地形排水			移除指标	城市化		
类别	Precision	Recall	F_1-score	类别	Precision	Recall	F_1-score
0	0.87	0.87	0.87	0	0.86	0.87	0.87
1	0.88	0.89	0.89	1	0.88	0.90	0.89
2	0.84	0.83	0.83	2	0.84	0.82	0.83
Accuracy			0.86	Accuracy			0.86
Macro Avg	0.86	0.86	0.86	Macro Avg	0.86	0.86	0.86
Weighted Avg	0.86	0.86	0.86	Weighted Avg	0.86	0.86	0.86
移除指标	湿地损失			移除指标	无效防灾		
类别	Precision	Recall	F_1-score	类别	Precision	Recall	F_1-score
0	0.60	0.45	0.51	0	0.76	0.89	0.82
1	0.81	0.88	0.84	1	0.57	0.32	0.41
2	0.71	0.81	0.76	2	0.70	0.84	0.77
Accuracy			0.72	Accuracy			0.70
Macro Avg	0.71	0.71	0.71	Macro Avg	0.68	0.69	0.67
Weighted Avg	0.71	0.71	0.71	Weighted Avg	0.68	0.70	0.68

对于灵敏度分析结果，可以从表 2 中看出，在去除无效防灾指标后的模型准确率最低，值为 0.70，说明无效防灾对于模型分类效果的影响程度最大；风险类别 1 的分类准确性明显更低，说明无效防灾对于识别类别 1 具有重要作用，与前面分析的结果一致。

4 模型求解

为准确预测洪水发生的概率，将数据集分割为训练集和测试集，其中 30% 的数据作为测试集，70% 的数据作为训练集。选择线性回归、多层感知机、决策树、随机森林以及线性回归和多层感知机集成的模型对洪水发生的概率进行预测。

4.1 采用 Spearman 选择的特征的模型预测性能

多模型预测比较图见图 9，集成模型的预测性能见图 10，各模型均方误差见图 11。

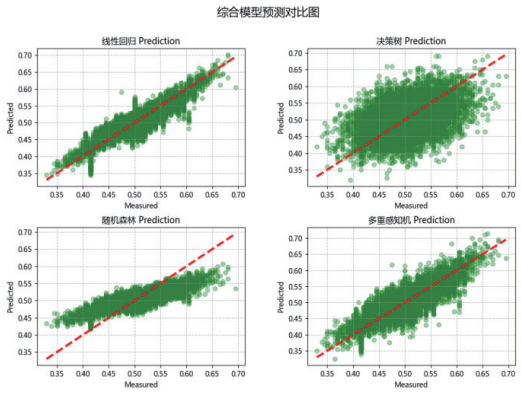

图 9 多模型预测比较图

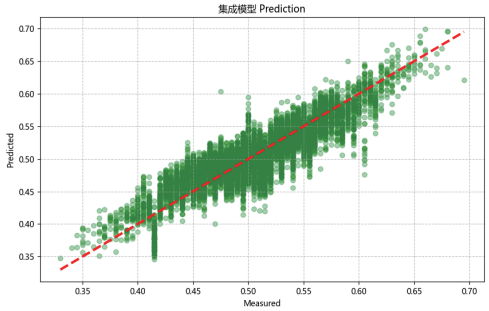

图 10 集成模型的预测性能

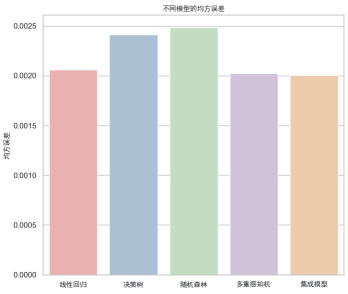

图 11 各模型均方误差

4.2 采用聚类方法和随机森林分类器选择特征的模型预测性能

多模型预测比较图见图 12，集成模型的预测性能见图 13，各模型均方误差见图 14。

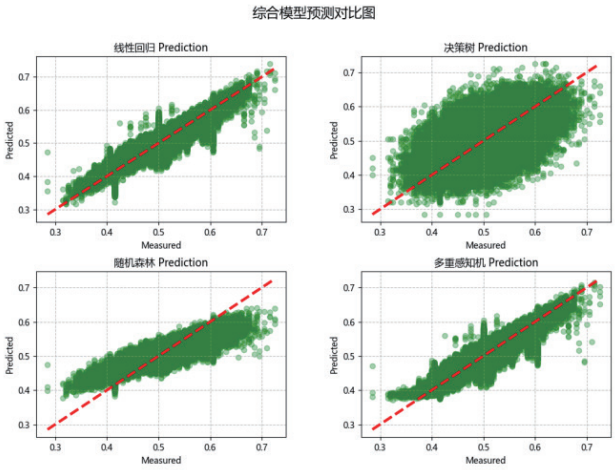

图 12 多模型预测比较图

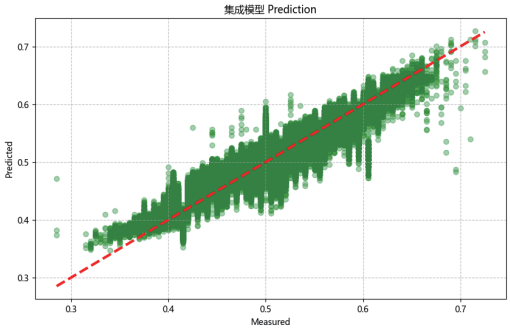

图 13 集成模型的预测性能

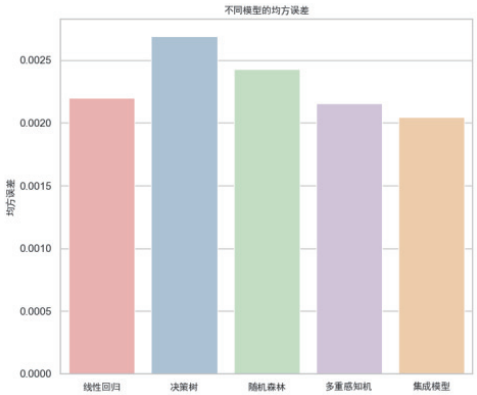

图 14 各模型均方误差

4.3 实验结果讨论

本文详细分析了不同机器学习模型在洪水灾害预测中的表现。决策树模型在数据捕捉的局部性上表现良好，但整体趋势捕捉不佳，预测点分布较为分散。随机森林模型通过集成多个决策树，增强了稳健性，整体趋势捕捉较好，但仍存

在一些偏差。多层感知器（MLP）[14-15] 模型在处理复杂非线性关系方面表现出色，预测点沿对角线高度集中，显示出较高的准确性。线性回归模型在捕捉线性关系方面表现良好，但对于复杂关系的处理需要进一步改进。集成模型结合多个模型的优势，显著提升了预测的准确性和稳健性，在处理复杂数据集和非线性关系时表现出色。

根据实验结果，采用改进的 K-means 聚类方法和随机森林分类器选择特征，使各模型的均方误差（MSE）显著降低，显示出在捕捉洪水风险关键因素方面的优势。集成线性回归和多重感知机模型进一步提高了预测的准确性，通过多模型协同作用，减少了单一模型的局限性和过拟合风险。这种方法不仅优化了模型的预测能力，还提升了结果的稳定性和可靠性，为洪水风险评估和防灾减灾提供了更为高效和可靠的技术支持。

5 结语

本文提出的洪水预测方法在多个方面提升了洪水风险管理的效能。首先，通过改进的 K-means 聚类方法和随机森林分类器，能够准确识别和分类洪水风险，不仅提高了预测的精度，还减少了误报和漏报的概率，为城市防洪决策提供了更可靠的依据。其次，分级预警系统的应用，使得资源配置更加合理高效。高风险区域可以得到更多的防护资源和应急措施，确保在洪水发生时能够迅速有效地应对，降低了洪水带来的损失和影响；中低风险区域则可以进行常规监测和维护，保持警惕，在风险增加时能够及时响应。此外，本文结合多种机器学习模型，如随机森林和多层感知器（MLP），显著提升了预测的准确性和稳健性。这种综合方法能够处理复杂的非线性关系，更好地捕捉洪水发生的潜在因素，为城市规划和应急管理提供了科学依据。总的来说，本文的研究成果不仅在洪水预测的精度和可靠性方面取得了显著进展，还在实际应用中提升了洪水风险管理的效率和效果。通过优化特征选择和模型结构，本文为洪水风险评估和防灾减灾提供了更加高效和可靠的技术支持，有助于减少洪水灾害对城市的影响，保障公共安全和财产安全。

参考文献：

[1] 杨禄记 . 涨率分析法在广西中小河流洪水预测预警中的应用——以东安江沙头镇水文站洪水为例 [J]. 广西水利水电 , 2024(2):52-57.

[2] 刘志雨 . 洪水预测预报关键技术研究与实践 [J]. 中国水利 , 2020(17):7-10.

[3] 张雄灵 , 杨贯中 . 数据挖掘在河道洪水准确预测中的应用研究 [J]. 计算机仿真 ,2013,30(1):401-403+414.

[4] 赵松 . 基于机器学习与空时特征融合的洪水预测 [D]. 西安 : 西安电子科技大学 ,2021.

[5] 崔雅博 , 罗清元 , 刘丽娜 . 基于改进非线性自回归网络的洪水预测算法 [J]. 沈阳工业大学学报 , 2023, 45(1): 84-89.

[6] 杨晨 . 基于流域异质性的洪水特征分析与分类预测 [D]. 成都 : 电子科技大学 ,2022.

[7] 刘扬 , 王立虎 . 基于改进 EMD-LSTM 的洪水预测方法研究 [J]. 水利水电技术 (中英文),2022,53(1):35-44.

[8] 黄天星 , 臧兆祥 , 陈露露 , 等 . 基于 ResNet 和随机森林的海洋鱼类分类方法 [J]. 工业控制计算机 ,2023,36(10): 78-80+83.

[9] 张锟滨 , 陈玉明 , 吴克寿 , 等 . 粒向量驱动的随机森林分类算法研究 [J]. 计算机工程与应用 ,2024,60(3):148-156.

[10] 汤圣君 , 张韵婕 , 李晓明 , 等 . 超体素随机森林与 LSTM 神经网络联合优化的室内点云高精度分类方法 [J]. 武汉大学学报 (信息科学版), 2023,48(4): 525-533.

[11] 蒋东浩 , 赵洪华 , 王真 . 基于 EWT 和 NeuralProphet-MLP 的蜂窝网络流量长期预测方法 [J]. 现代信息科技 , 2024, 8(6):52-57.

[12] 魏璐露 , 程楠楠 . 基于 SVM-DT-MLP 模型的 Web 日志异常流量检测研究 [J]. 现代信息科技 , 2024,8(4): 171-174+179.

[13] 王智 . 洪水预测模型对抗灾减灾的相关研究 [J]. 低碳世界 , 2020, 10(11):63-64.

[14] 刘璐瑶 , 陈志刚 , 沈欣炜 , 等 . 基于 EMD-MLP 组合模型的用电负荷日前预测 [J]. 南方能源建设 , 2024, 11(1): 143-156.

[15] 刘俊文 , 谢劭峰 , 钟雁琴 , 等 . 基于 MLP 神经网络的中国南方地区多因子 PWV 预测模型 [J]. 中国科技论文 , 2024, 19(1): 99-107+122.

【作者简介】

张舒（1987—），男，山东菏泽人，硕士，研究方向：人工智能。

（收稿日期：2024-07-24）