

基于区块链与机器学习结合的洪水预测模型研究

赵 鹏^{1,2} 柴荣沐^{1,2} 袁富江^{1,2} 马宇群^{1,2} 高 嘉^{1,2}

ZHAO Peng CHAI Rongmu YUAN Fujiang MA Yuqun GAO Jia

摘 要

近年来,人类活动愈发频繁,深刻改变自然生态,致使极端天气与自然灾害呈高发态势,其中,洪涝灾害影响尤为突出。洪水直接威胁群众生命安全,造成溺水悲剧;洪灾过后,卫生环境恶化,又极易引发疾病传播,严重损害群众健康。这些灾害不仅给人民生命财产带来难以估量的损失,还对我国的可持续发展构成严峻挑战。文章提出一种基于 MLP 多层感知机与区块链技术结合的洪水预测模型,首先使用皮尔逊相关系数法对数据进行关联性分析,选取关联性强的指标,然后,基于 MLP 模型对洪水数据中关联性强的指标进行预测,并将数据和预测结果上传至区块链的智能合约,触发合约可靠且不可篡改的预警信息发布,及时通知相关部门和公众,同时利用区块链记录的数据对模型进行优化和更新。实验结果表明,MLP 的预测精度达到 82.47%,并且 MLP 模型输出的 MAE、MSE 和 F_1 -score 指标也较为可靠,模型性能稳定。

关键词

MLP; 区块链技术; 洪水预测; 皮尔逊相关系数

doi: 10.3969/j.issn.1672-9528.2025.04.013

0 引言

近年来,由于人类活动的影响,洪水灾害频发,这种难以预测的自然灾害,不仅可能导致溺水事故,还会导致灾后疾病的高发,严重威胁着人民生命安全和财产。2023 年 6 月 24 日至 25 日,我国 15 条河流出现超警戒水位的洪水态势,水位超警戒线意味着洪水风险加剧。这一系列洪水灾害不仅对中国造成了严重影响,而且在全球范围内也导致了数十亿美元的经济损失。洪水的频发和强度加剧,与人口快速增长、耕地扩张、围湖造田以及无节制的森林砍伐等活动密切相关。这些行为改变着地球表面的自然状态,也改变了水流的自然汇流条件,从而加剧了洪灾的程度和频率。

因此,加强洪水预警系统的建设,提高防灾减灾能力,以及实施可持续的土地管理和水资源管理策略,对于减轻洪水灾害的影响至关重要。同时,国际社会需要加强合作,共同应对这一全球性挑战,以保护我们共同的家园。

在这种数据量骤增的洪水案例事件背景下,利用机器学习^[1]、人工智能和区块链技术^[2]进行洪水的预测的重要性显

得愈加突出。机器学习模型可以通过分析历史洪水数据,识别洪水发生的模式和相关环境因素,从而提高预测的准确性。人工智能技术能够处理大量的气象、地质和水文数据,实时更新洪水风险评估,为决策者提供科学的依据。同时,能够优化警报系统的响应时间,确保信息的快速传递,这对于洪水应急响应和灾害准备至关重要。区块链技术具有不可篡改、去中心化和数据可追溯性,将其应用于存储洪水相关的数据,有利于提高洪水数据的安全性以及解决中心化服务器存储数据的单点故障问题。因此,整合机器学习、人工智能和区块链技术进入现有的洪水预警和管理框架,将极大地提升防范和减轻洪水灾害的能力。

传统的洪水预测方法主要是利用统计模型^[3]、水文模型和经验法则。这些模型和方法数据依赖性强,且局限于历史数据,在数据使用和处理上还存在一些显著问题,以至于模型的实际落地应用仍然面临着诸多挑战。近年来,随着机器学习和区块链技术的发展,基于数据驱动的洪水预测方法^[4]进入了研究人员的视野。这些方法通过大量的数据收集,利用机器学习算法对收集到的数据进行处理和分析,有效地提升了预测的准确度。但在数据透明性和共享性以及模型优化方面存在不足,影响了洪水预测模型的可靠性,如何有效地整合不同机构(如气象局、水利部门、科研机构)共享和访问洪水相关的数据,以构建更为全面的洪水预测模型,打破数据孤岛,也是一个亟待解决的问题。

1. 太原师范学院计算机科学与技术学院 山西晋中 030619

2. 太原师范学院智能优化计算与区块链技术山西省重点实验室 山西晋中 030619

[基金项目] 山西省科技战略研究专项重点项目“山西省信创产业科教融汇、产教融合的模式与实施路径研究”(2023040314 01011)

1 相关工作

高小雲等人^[5]将水动力模型与机器学习技术相结合，以改善城市洪水预报与灾害分析效果，并通过参数确定后的一维、二维耦合水动力模型，实现较高的洪水预测精度，并且运用高分辨率激光雷达影像与精密模型，模拟英国伦敦西部泰晤士河的洪水场景。付贵^[6]研究的三岔河普定电站以上流域水情测报系统，是一种结合光纤通信组网方式和遥测站GPRS（GSM）先进洪水预报系统，采用改进的API模型（降雨径流预报模型）、改进的马斯京根流量演算及有自动实时校正功能的马斯京根模型等对洪水进行实时预报，在流域雨水情监测中发挥了及时有效的作用。刘明锦等人^[7]针对山洪预警问题，在基于机器学习中的算法以及历史雨量数据和实时采集雨量数据组成的大数据的支撑之下，搭建了山洪预测模型，并应用到阿坝州小金川河流域和阿坝州龙溪沟流域进行模拟和验证。张越^[8]首先提出了一种新的洪水数据处理方式，在河道流域内筛选数据相关性最大的站点作为预测输入，降低数据冗余、去除无效数据，使模型能够更快达到最优。李静^[9]通过对当地气象和水文现状的分析，构建了洪涝灾害预警的气象水文耦合预测模型。同时，针对耦合模型在暴雨天气中径流量预测响应慢的问题，采用长短期记忆网络对模型进行优化。

为解决洪水预测模型中的数据安全问题，提高预测精确性，本文提出了一种基于区块链技术和多层感知机的洪水预测架构。首先，对异构数据集进行上传区块链，所有上传的数据需要多方机构达成共识才可上链，并且可以被所有参与者查看，这种透明性有助于建立信任，专注于分析20个指标的分布情况；其次，对上链数据进行相关性分析，以量化20个指标与洪水发生概率之间的联系，总结出哪些指标对洪水发生的影响最为关键，据此提出预防措施，构建模型并进行洪水预测，预测洪水发生概率。最后，将预测结果和相关数据集上传至区块链，通过智能合约自动通知相关部门和公众，确保及时响应，对比预测结果与实际洪水事件，评估模型的准确性。将新的洪水事件数据和预测结果上传至区块链，作为模型优化的基础。

2 数据画像和特征分析

本文数据集是一个关于洪水的数据集，涵盖超过20万条与洪水发生的多个特征和相关指标，该数据集为模型研究提供了数据支持，并且能精准分析洪水的发生规律、影响因素和预测模型。

首先，需要对train.csv数据集进行初步的探索性数据分析。这一过程包括数据的统计描述、数据清洗（处理缺失值和异常值）以及确认变量类型等基本任务，为后续的

分析工作奠定坚实的基础。将专注于分析20个指标的分布情况，以识别可能的离群点或异常模式。可以通过散点图等可视化工具，研究指标与洪水发生概率之间的关系。接下来，进行相关性分析，以量化指标与发生概率之间的联系。使用皮尔逊相关系数^[10]来衡量变量之间的线性相关性。最后，得出哪些指标对洪水发生的影响最关键，并提出预防措施。

在不确定是否有缺失值或者异常值的情况下，本文使用K-S检验判断方法对附件中的train.csv分别进行判断，因为train.csv中的数据量较大，文中选取其中50条数据对季风气候和排水进行K-S检验，具体检验结果如图1、图2所示。

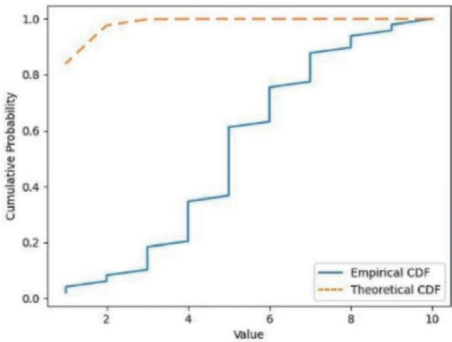


图 1 K-S 检验季风强度图

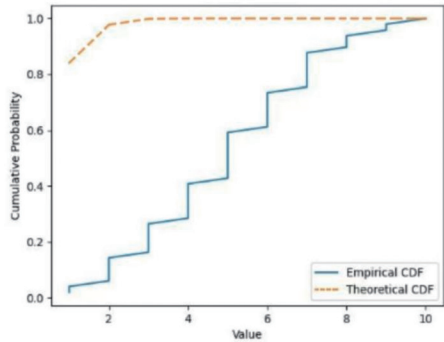


图 2 K-S 检验地形排水图

在对训练（train）数据集进行异常点检验时，可以发现季风强度、地形排水坐标的变化呈现出较好的线性规律，数据中并未出现离散点，也未发现异常值。如表1所示，本次仅对训练数据集中异常点的检验结果进行展示。

表 1 季风和排水 K-S 检验表

类别	<i>P</i> 值	是否拒绝原假设
季风强度	4.562 527	是
地形排水	4.1625 26	是

这里，先对数据进行了预处理之后，在预处理过的数据基础上，对train.csv数据中的一部分进行可视化的展示，方便更加清晰地分析数据。可视化数据如图3所示。

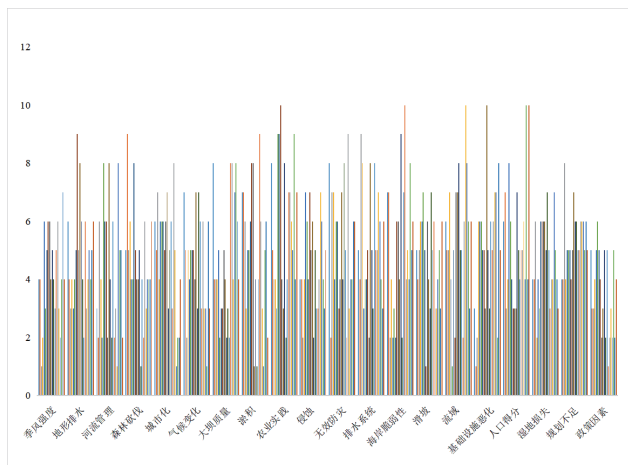


图3 50条数据中的20参数可视化分析图

本文使用了Pandas库中的corr()方法来计算相关值。corr()方法用于计算DataFrame中所有列之间的相关性，返回一个相关性的矩阵。在此只关心与洪水概率列的相关性，具体各个影响因素与之相关性如表2和图4所示。通过表2可以看出，湿地损失对洪水概率发生的相关性最密切，其次是政策因素、排水系统和规划不足等，然后季风强度和农业实践对洪水发生的概率影响性最小，仅有0.104 085和0.131 002。

表2 各个参数与洪水发生概率的相关性

影响洪水发生概率的因素类别	影响因子
湿地损失	0.256 100
政策因素	0.244 400
排水系统	0.209 701
规划不足	0.207 369
大坝质量	0.202 084
河流管理	0.200 474
人口得分	0.195 899
城市化	0.191 512
无效防灾	0.190 619
侵蚀	0.188 992
海岸脆弱性	0.182 665
滑坡	0.181 650
地形排水	0.181 223
基础设施恶化	0.181 094
淤积	0.177 818
森林砍伐	0.158 018
气候变化	0.138 528
流域	0.138 248
农业实践	0.131 002
季风强度	0.104 085

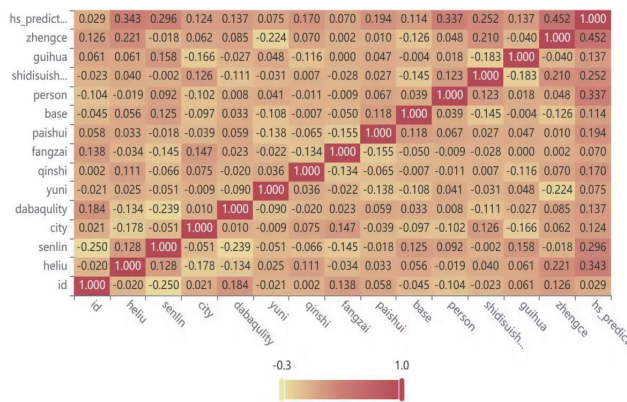


图4 各参数与洪水发生概率热力图

皮尔逊相关性系数（Pearson correlation coefficient）是衡量两个变量之间线性关系强度和方向的统计指标，其值范围为-1~1。具体来说，1表示完全正相关；-1表示完全负相关；0表示没有线性相关性。

3 模型架构

系统架构图如图5所示。

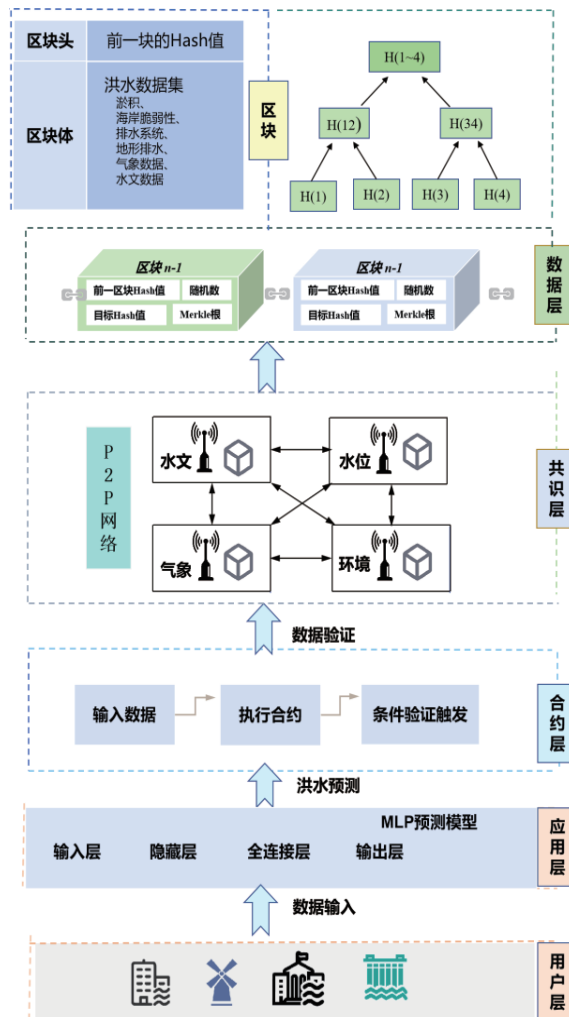


图5 系统架构图

传统洪水预测中,往往涉及多个部门和机构的数据共享。中心化存储容易遭受恶意攻击和单点故障,一旦服务器被攻击,数据就会被伪造、非法篡改甚至造成整体丢失,中心化存储出现该问题的时候,想要追回数据极其困难。并且洪水隐私数据安全性也不能被保障,也容易被攻击者非法窃取。但是通过将区块链技术应用于洪水预测模型,将区块链中的每一个区块使用 SHA256 方法进行链接,当其中一个信息被非法篡改,其他区块会同时受到影响。以保证数据不可篡改。

3.1 用户层

在基于区块链的洪水预测模型中,各种检测站、气象部门和水文部门作为共识节点,承担着重要的“角色”,负责上传与洪水相关的数据。这些节点的协作不仅提高了数据的可靠性和透明度,还增强了洪水管理的整体效率。各个共识节点会定期收集洪水相关数据,并上传至区块链。区块链网络的共识机制可以确保所有节点对上传数据的有效性达成一致。这意味着,只有经过验证的数据才会被记录在链上,提高数据的可信度。

3.2 应用层

该层主要使用 MLP 多层感知机模型,用于洪水数据分析和预测,主要的优点如下:

(1) 非线性建模能力: MLP 能够通过其多层结构和激活函数捕捉输入特征之间的复杂非线性关系。这对于洪水预测尤为重要,因为洪水的发生通常受到多种因素的影响。

(2) 自适应学习: MLP 通过反向传播算法自动调整权重,这使得模型在处理复杂的洪水数据时更加灵活和高效。

(3) 强大的预测能力: 经过适当训练的 MLP 模型通常表现出较高的准确性。通过使用历史数据训练,模型能够识别出潜在的模式,提高洪水发生概率的预测能力。总之,使用多层感知机模型进行数据分析和预测,能够有效捕捉复杂的非线性关系,并自动提取特征,提供准确性高的预测结果。这些优势使得 MLP 成为洪水管理和风险评估中一种强大的工具,帮助决策者制定更有效的应对策略。

MLPRegressor 类是 Scikit-Learn 库中用于创建多层感知器 (MLP) 回归模型^[1]的一种方法。这类模型特别适用于解决回归问题,即预测一个具体的数值,而非分类问题中的类别标签。多层感知器 (MLP) 作为一种人工神经网络,包含输入层、至少一个隐藏层和一个输出层,能够通过学习捕捉数据中的非线性关系。

构建多层感知机模型首先对权重进行初始化,将输入样本送入网络中进行前向传播。对于每个样本,计算输入层到隐藏层的加权和,输入激活函数中得到隐藏层的输出。再将其传递到下一层,直至输出层输出分类结果。将网络的

输出与样本的标签进行比较,计算损失函数。常用的损失函数包括均方误差 (Mean squared error) 和交叉熵 (Cross entropy)。反向传播算法是一种常用的优化方法,通过计算梯度并更新权重和偏置来最小化损失函数。并通过反向传播算法计算每个权重对损失的梯度。本文采用 6 层 MLP 模型,由输入层、3 个隐藏层、1 个全连接层和输出层组成。这种结构使得 MLP 能够处理复杂的数据模式。

(1) 输入层: 输入层是网络的第一层,接收原始数据作为输入。每个输入节点代表数据的一个特征。

(2) 隐藏层: 在输入层之后是隐藏层,执行大部分的计算工作。每个隐藏层完全连接到前一层的所有节点上。六层的 MLP 模型中有 3 个隐藏层,隐藏层的神经元数量和激活函数可以根据具体任务进行调整。常用的激活函数有 ReLU、sigmoid 和 tanh 等,本文采用 ReLU 函数。

(3) 输出层: 输出层是 MLP 的最后一层,产生最终的网络输出。输出层的节点数量和激活函数类型取决于特定的任务。例如,对于二分类问题,输出层可能只有一个神经元,使用 sigmoid 激活函数;对于多分类问题,输出层的神经元数量将等于类别数量,并使用 softmax 激活函数。

MLP 出色的全局优化和泛化能力使其成为在数据集中进行拟合和预测的理想选择。验证预测模型的准确性,将数据集分为训练集和测试集,使用训练集数据训练 MLP 回归模型。使用测试集数据评估模型的性能,可以使用平均绝对百分比误差 (MAPE) 作为评估指标。MAPE 的计算公式为:

$$\text{MAPE} = \sum (|y_{\text{pred}} - y_{\text{true}}| \times 100\% / n) \quad (1)$$

式中: n 是测试集样本数; y_{pred} 是预测值; y_{true} 是真实值。

使用 MLPRegressor 类创建一个多层感知器 (MLP) 回归模型。设置模型的隐藏层大小为 100 和 50,随机状态为 1,最大迭代次数为 500。使用训练集 (X_{train} 和 Y_{train}) 来训练模型。使用训练好的模型对测试集 (X_{test}) 进行预测,得到预测的洪水概率 (y_{pred})。使用 mean_absolute_percentage_error 函数计算测试集上预测洪水概率的平均绝对百分比误差 (MAPE)。计算模型的准确率,即 100% 减去 MAPE。

3.3 智能合约层

传统合约是由双方一起制定规则,需要第三方机构对合约执行进行监督和授权。智能合约规则代码公开透明、不可篡改,可以在没有第三方情况下可信交互,避免了人工错误与人为干涉。该系统中将洪水数据通过共识节点上传至区块链,区块链中的智能合约将先校验数据的规范性以及完整性。当前合约定义只有水文局、气象局、水文站等监测部门节点才有权对特定范围内的链上洪水信息数据进行检索和确认。并且共识节点按需上传洪水数据,当有新的数据上传至区块

链后，智能合约^[12]将自动触发执行洪水数据的上传和检索。

3.4 共识层

智能合约上传数据之后会通过共识机制对需要上传的数据做进一步的共识承认，最后存储到区块链上，该系统架构采用联盟链，选用联盟链常用算法——PBFT 共识算法^[13]。该区块链分布式网络中的节点为相关监管部门（包括气象监测站、水位监测站和水文监测站）。该共识算法规范了区块的传播形式，每个加入联盟链中的共识节点可以相互进行通信（包括接收数据和发送数据）以达成共识确保数据的准确性和完整性。共识完成后，将上传的数据集分别存放在各个节点的账本中，保证数据分布式存储，防止中心化故障引发系统问题。

3.5 数据层

数据存储层由区块链数据库构成，当共识机制 PBFT 算法对数据进行确认后，会将估算的洪水信息（包括海岸脆弱性、滑坡、地形排水、基础设施恶化），洪水模型的每次验证过程和更新情况进行整合区块链上链，将数据存放在各个节点的账本中。数据层核心任务是保护区块链数据库信息，确保信息能够安全保存。为了优化数据的不可篡改性和提高区块链的整体安全性，每个区块中还包含了一个 Merkle 根。这个根哈希值代表了整个区块的内容，任何对区块内数据微小的改动都会导致不同的 Merkle 根产生，从而使得任何未授权的修改立即被网络中的其他节点识别出来从而保证系统安全。

4 实验分析

本文使用 MLP 模型和随机森林模型对上述问题得出的结果进行预测。首先对洪水数据进行预处理，使用 K-S 检验判断方法进行判断，选取其中 50 条数据对季风气候和排水进行 K-S 检验，使用 StandardScaler 技术，使其具有零均值和单位方差。其次，用皮尔逊相关性系数确定了与洪水发生概率高度相关的关键指标。在本文中，选择了城镇化建设、基础设施恶化、大坝质量、地形排水、气候变化作为特征，洪水概率作为目标变量。该程序中从 df_train 中选择关键指标对应的列作为特征数据 (X_{train})，不包括洪水概率列，从 train.csv 数据表中选取 176 760 条数据作为训练集数据。将 df_train 中洪水概率列作为目标变量 (y_{train})。从 df_test 中选择与训练数据相同的特征列作为测试集的特征数据 (X_{test})。使用 StandardScaler 对训练稽核测试集的特征数据进行标准化处理。

具体实验结果为选取如参数指标：“季风强度”“地形

排水”“河流管理”“森林砍伐”“城市化”“气候变化”“大坝质量”“淤积”“农业实践”“侵蚀”“无效防灾”“排水系统”“海岸脆弱性”“滑坡”“流域”“基础设施恶化”“人口得分”“湿地损失”“规划不足”“政策因素”，对洪水发生的概率进行模型预测如表 3 所示。其中，设置模型迭代轮次为 200 轮，设置 176 760 条数据为训练集，剩下为数据位测试集。该模型随机选取了 6 组数据集进行模型训练与预测，准确率在 82% 左右。其中 MLP 模型预测准确度为 82.47%。

表 3 20 个指标的洪水概率预测模型表

序号	原始洪水概率	预测洪水概率
0	0.470	0.517 187
1	0.475	0.558 295
2	0.450	0.491 979
3	0.615	0.535 038
4	0.490	0.478 545
...	...	...
95	0.490	0.518 182
96	0.465	0.508 426
97	0.425	0.500 487
98	0.485	0.489 724
99	0.525	0.494 763

表 4 为随机选取“城市化”“基础设施恶化”“大坝质量”“地形排水”“气候变化”“洪水概率”5 个指标对洪水发生概率模型进行预测生成的预测概率与原始概率对比。

表 4 5 个指标的洪水概率预测模型表

序号	原始洪水概率	预测洪水概率
0	0.455	0.562 161
1	0.455	0.473 471
2	0.480	0.603 891
3	0.505	0.543 146
4	0.370	0.473 843
...	...	...
95	0.500	0.583 464
96	0.560	0.516 393
97	0.440	0.502 491
98	0.455	0.486 680
99	0.425	0.453 481

通过表 4 可知，预测值与原始值出入比较大，说明当参数指标较少时，模型预测准确度较低，准确度为 77.78%。

对该实验结果的预测值与真实值做均方误差（MSE）和平均绝对误差（MAE）误差分析，可以看出该模型有一定的可靠性和拟合性。模型的均方误差（MSE）和平均绝对误差（MAE）如表 5 所示。

表 5 两模型 MSE、MAE 和 F1 Score 对比表

序号	MSE	MAE	Test Score
MLP 模型			
1	0.000 402	0.014 934	0.844 586
2	0.001 034	0.026 411	0.600 304
3	0.001 245	0.019 634	0.836 951
4	0.000 973	0.023 924	0.696 415
5	0.001 093	0.022 869	0.839 524
6	0.001 052	0.014 963	0.848 521
随机森林模型			
1	0.000 393	0.014 982	0.848 016
2	0.001 022	0.025 864	0.600 307
3	0.000 398	0.015 693	0.844 739
4	0.000 385	0.014 398	0.851 515
5	0.000 413	0.014 852	0.843 533
6	0.000 401	0.014 947	0.843 129

5 结语

本文深入分析了洪水灾害的数据分析与预测问题，通过详细考察指标关联性与预防措施的确 定、洪水风险评估模型的构建、洪水发生概率的精确预测、大规模洪水风险预测及分布分析等关键问题。采用了区块链技术对洪水数据进行了存储，利用皮尔逊相关系数进行关联性分析，并利用多层感知器回归模型对洪水发生的概率进行预测，通过 Standard-Scaler 技术对数据进行标准化处理以提高模型性能。最终，建立了一个能够准确预测洪水概率的模型，并提出了有效的防范措施，以减轻洪水带来的损失。

参考文献:

[1] 胡义明,陈钰,周瑛,等.基于水文水动力模型和机器学习模型耦合的河道水位预报方法[J].水电能源科学,2024,42(10): 29-32

[2] 江静,吴翼.基于区块链技术与人工智能的数据共享模型研究[J].自动化与仪器仪表,2024(11):83-87.

[3] 王永庆,杨安玉.特高拱坝变形分析与监测研究[J/OL].水利水电技术(中英文),1-7[2024-12-30].http://kns.cnki.net/kcms/detail/10.1746.TV.20241014.1717.009.html.

[4] 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.

[5] 高小云,李慧.基于水动力模型与机器学习的城市洪水风险评估[J].水利水电快报,2024,45(10):5.

[6] 付贵.基于遥感数据水雨情测报系统设计[J].黑龙江水利科技,2024,52(10):83-86.

[7] 刘明锦,张智涌,王宾,等.基于机器学习和大数据的山洪预测模型研究[J].四川水利,2022,43(6):107-113.

[8] 张越.基于机器学习的汾河水系水文分析[D].太原:太原理工大学,2022.

[9] 李静.水文预测模型在洪涝灾害预警系统中的应用研究[J].水利科学与寒区工程,2023,6(12):64-67.

[10] 徐志刚,刘风雨,陈瑞刚,等.基于洪水分级的青山殿水库洪水预报[J].南水北调与水利科技(中英文),2024,22(4): 661-671.

[11] 刘俊文,谢劲峰,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.

[12] 刘星辰,杜军平,梁美玉,等.基于智能合约和联邦存储的异步联邦学习模型[J].数据采集与处理,2024,39(6):1532-1542.

[13] 张月圆,张春远,王永利,等.基于联盟链微电网交易的 PBFT 共识算法改进[J].内蒙古电力技术,2024,42(5):52-61.

【作者简介】

赵鹏（1973—），男，山西太原人，博士，教授、硕士生导师、太原师范学院计算机科学与技术学院院长、山西省新兴产业领军人才、CCF 太原主席，研究方向：区块链。

柴荣沐（1997—），男，山西长治人，硕士研究生，研究方向：区块链，email:746237646@qq.com。

袁富江（1998—），男，四川宜宾人，硕士研究生，研究方向：区块链。

马宇群（2001—），女，山西吕梁人，硕士研究生，研究方向：区块链。

高嘉（2001—），女，山东枣庄人，硕士研究生，研究方向：医学图像分割。

（收稿日期：2025-01-02）