基于 KNN 和决策树算法的苏霍纳河冰塞预测

Yuxuan Cui

Lomonosov Moscow State University, Moscow, Russian Federation

【摘要】冰塞预测对于寒冷地区减少和预防冰塞洪水具有重要意义。本文主要对冰塞预测在寒冷地区应用的可能性进行评估。Sukhona River 基于 Russia 选定的最重要水文和气象特征，对冰塞进行预测。冰流期间的最高水位和冰塞引起的水位上升是主要决定因素。基于决策树算法的 KNN 算法开发了最佳预测模型。研究发现，由 KNN 算法建立的模型表现最佳，并准确地预测了所有堵塞年份。本文的研究为该 Veliky Ustyug 地区的冰塞预测提供了帮助。knn 方法对所 1 in 研究的河段具有回忆性，比其他预测方法更准确地预测了冰塞的发生。这意味着所选的预测因子具有高度可靠性。

【关键词】冰塞；KNN；决策树方法；多元线性回归模式

【收稿日期】2024 年 10 月 22 日　　【出刊日期】2024 年 11 月 20 日　　【DOI】10.12208/j.aiml.20240003

Ice Jam Prediction for Sukhona River Based on KNN and Decision Tree Algorithm

Yuxuan Cui

Lomonosov Moscow State University, Moscow, Russian Federation

【**Abstract**】Prediction of ice jam is very important for reduction and prevention of ice jam floods in cold regions. This article focuses on the assessment of the possibility of predicting ice jam on the Sukhona River in Russia based on selected most significant hydrological and meteorological features. The maximum water level during the ice drift and ice-jam induced water level rising are the main determinants. The optimal prediction model is developed based on KNN algorithm with decision tree algorithm. The model built by the KNN algorithm was found to perform best and accurately found all blockage years. The research in this paper provides help to establish ice jam prediction in the Veliky Ustyug region. The knn method has a recall of 1 in the studied river segment, which predicts the occurrence of ice jam more accurately than other prediction methods. This implies that the chosen forecast factor is highly reliable.

【**Keywords**】Ice jam; KNN; Decision Tree Method; Multiple linear regression mode

1 引言

苏霍纳河是俄罗斯欧洲部分北德维纳河的一条支流，长 558 公里，发源于沃洛格达州的库别纳湖，向东北流动。河流上游河道平缓，河岸宽阔，下游河谷陡峭狭窄。苏霍纳河春季洪水持续三个月左右，随后水位继续下降，7 月进入平水期。下游严冬冰厚平均 1m，春季流冰期易发生冰塞洪涝灾害[1]。研究河段位于苏霍纳河下游，卡利基诺市以北、科特拉斯市以南。

该河段夏季温暖短暂，冬季漫长严寒，积雪不断，1 月平均气温-14℃[2,3]。冬季受北极冷空气侵袭，气温急剧下降，最低气温可达-46℃。同时苏霍纳河下游河道变窄，河床坡度加大，在大乌斯秋格市，河道宽度达 500 米。河道狭窄，气温急剧下降，造成该段河流严重冰塞[4]。

2 方法

本研究选择度量 KNN 方法作为预测冰塞出现的算法，该方法基于相似物体的紧密性和接近性假设。该方法的优点是抗异常值、易于实施、可解释性强、能够处理小数据。预测结果还与决策树算法模型进行了比较[5,6]。

2.1 预后因素分析

注：本文于 2022 年发表在 Advances in Computer and Communications 期刊 3 卷 2 期，为其授权翻译版本。

影响冰塞的因素包括冰塞发生前的最高水位、自 9 月 1 日至气温低于零度的累计天数、冰塞持续时间、最大冰厚以及春季洪水期间的河流流量。本文收集了 1936 年至 2020 年研究河段 3 个气象站和 5 个水文站的数据。所得特征列于表 1 中。

在机器学习中，通常将数据分为训练集和测试集来评估模型的预测效果。本文将 2000 年至 2018 年的数据指定为测试集，将 1960 年至 1999 年的数据设置为训练集。目标变量分别设置为 0（无过度冲刷）和 1（过度冲刷）。

2.2 特征变换

度量分类算法对数据的大小非常敏感。初始特征可以属于不同的范围并在不同的范围内变化，对度量的贡献也不同[7]。特征值之间的不平衡会导致不稳定并降低模型的质量[8]。为了避免这种情况，必须对特征进行规范化。在此任务中，MaxAbsScaler 工具用于此目的。根据这种规范化，每个特征都按其最大绝对值缩放，从而将每个特征的变化范围转换为（-1；1）的范围。

2.3 KNN 算法模型与决策树算法模型比较

KNN 算法和决策树算法是针对分类问题最基本、最简单的机器学习算法之一。

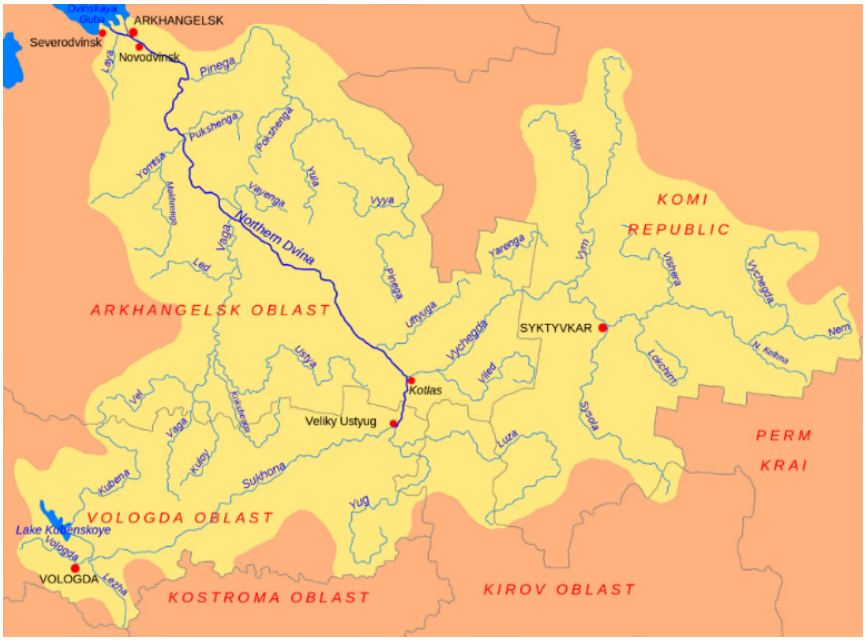

图 1 苏霍纳河流域和研究河段的位置

表 1 水文气象特征列表

特征编号	特征名称	城市	特性类型、测量单位
1	结冰前最高水位	别列佐瓦亚斯洛博德卡	水文特征，厘米
2		卡利基诺	
3	流冰持续时间	大乌斯秋格	水文特征，白天
4		托特马	
5	冰盖持续时间	别列佐瓦亚斯洛博德卡	水文特征，白天
6	最大冰厚	大乌斯秋格	水文特征，厘米
7		科特拉斯	
8	冻结期温度变化特征	大乌斯秋格	气象标志，温度过渡至 0°C，从 9 月 1 日起的天数，天
9		纽克谢尼察	
10		尼科尔斯克	

KNN 算法寻找与测试样本最相似的训练样本，并将其分类。为了选择合适的参数值，我们使用交叉验证方法来寻找 k 值。交叉验证方法通过将训练集分成 n 个相等的子集，用 n-1 个子集训练模型，用剩余的子集测试模型，从而防止过度学习。该过程重复 n 次，取 n 个结果的平均值。在 python 中调用 sklearn 库中的 GridSearchCV 函数来寻找最佳参数。knn 算法中 GridSearchCV 得到的最佳模型是 k=5，采用均匀的权重函数。在此模型下样本准确率达到 81%。

决策树算法将逻辑条件组合成树形结构，具有由根节点、决策节点和终端节点组成的分层结构。使用相同的 GridSearchCV 函数来寻找最佳参数，对于决策树算法，我们得到了树的最大深度为 3，特征数的最大数量为 3 的最佳模型参数。但该模型的准确率仅为 43.8%。

3 结果与讨论

对于机器学习方法，我们用更多的指标来评估得到的最佳模型，除了最常见的准确度指标外，精度和召回率也很重要。KNN 模型指标是 Minkowski 距离，对于具有均匀权重的模型，k 值为 5，召回率为 1，这意味着该模型正确地发现了所有年份的冰堵情况。另一方面，决策树模型表现不佳，这是因为较少的观察年份改变了模型对异常值的敏感性。

4 结论

（1）根据史书数据分析，大乌斯克河段超标的水位 SukhonaRiver 与 700cm 冰塞洪水的形成有关。1936 年至 2018 年冰塞洪水发生率为 64%，每三至四年发生一次。令人担忧的是，受全球气候影响，该河段冰层厚度呈下降趋势。

（2）本文对 KNN 模型与决策树模型进行了比较，KNN 算法取得了较好的效果，通过交叉验证得到最适合本地区的模型，其中 K=5，权重均匀，距离度量为 Minkowski 距离，该模型的准确率为 81%，召回率为 1，即该模型正确地找到了所有的冰堵点。

参考文献

[1] Buzin V.A. Ice jams and jamming floods on rivers, 2004.

[2] Buzin V.A. Dangerous hydrological phenomena, 2008.

[3] Sazonov A.A. Probable scenarios and calculation of the characteristics of flooding of residential areas based on a set of mathematical models, 2021.

[4] Du, W. Application of support vector regression in prediction model using genetic algorithm optimized. Journal of Physics: Conference Series. IOP Publishing, 2021.

[5] Beltaos, S. Assessing Ice-Jam Flood Risk: Methodology and Limitations, 2021.

[6] Vanantwerp R.L. Ice jam flooding: causes and possible solutions, 1994.

[7] Yang, Z. The research on the RS Dynamie Monitoring Mode Of The Yellow River Icicle Hazard And The Ice Regime Information Extraction Model, 2006.

[8] Hu, W. Research on product recommendation system based on deep learning. Фундаментальные и прикладные научные исследования: актуальные вопросы, достижения и инновации. 2022.