

МЕТОДЫ ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЙ

УДК 556.535.5

ПРОГНОЗИРОВАНИЕ РЕДКИХ ГИДРОЛОГИЧЕСКИХ ЯВЛЕНИЙ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ НА ПРИМЕРЕ ЛЕДОВЫХ ЗАТОРОВ НА РЕКЕ ПЕЧОРЕ

С.М. Иглин¹, В.М. Морейдо², К.И. Головнин³¹⁻³ *Институт водных проблем РАН, лаборатория гидроинформатики*¹ *Науч. сотр., канд. геогр. наук; e-mail: iglin.sm@bk.ru*² *Ст. науч. сотр., зав. лаб., канд. геогр. наук; e-mail: vsevolod.moreydo@iwp.ru*³ *Инж.; e-mail: golovnin98@bk.ru*

Редкие гидрологические явления, как следует из названия, характеризуются малой повторяемостью, что, наряду с их зачастую катастрофическими проявлениями для человека, также говорит о малой доле измерений, сопровождающих эти события (что называется дисбалансом классов). Это, в свою очередь, мешает создавать надежные модели для прогнозирования таких процессов и особенно проявляется при построении моделей природных процессов с использованием алгоритмов машинного обучения, которым свойственна чувствительность к таким выборкам, проявляющим дисбаланс классов. В настоящем исследовании предпринята попытка обойти высказанные ограничения путем дополнения ряда для обучения моделей искусственно сгенерированными событиями. В качестве предмета и объекта исследований были выбраны, соответственно, долгосрочные прогнозы ледовых заторов, происходящих в устье Печоры – арктической реки европейской территории России. За длительный период наблюдений были собраны данные о ледовых заторах, подобраны предикторы и модели. Были использованы алгоритмы машинного обучения: *k* ближайших соседей (KNN), логистическая регрессия, градиентный бустинг (CatBoost) и многослойный перцептрон (MLP). Результаты показывают, что все модели после применения дополнения ряда искусственными событиями показали прирост качества моделирования. Это подтверждает перспективность метода дополнения ряда для обучения моделей редко повторяющихся процессов.

Ключевые слова: прогнозирование ледовых заторов, дисбаланс классов, машинное обучение, балансировка классов

DOI: 10.55959/MSU0579-9414.5.80.1.6

ВВЕДЕНИЕ

Экстремальные гидрологические события, наносящие максимальный урон, как правило, находятся в области редкой повторяемости. Освещенность таких событий данными наблюдений зачастую недостаточна, так как в процессе их прохождения оказывается экстремальное воздействие на наблюдательные системы, вплоть до выхода их из строя. По определению, такие события лежат в области низкой обеспеченности, и количество данных об интересующих нас характеристиках в ряду наблюдений может составлять не более 1%. В мировой практике такие ряды наблюдений принято называть несбалансированными (imbalanced dataset [Bourel et al., 2021; Wang et al., 2024]). Использование таких рядов данных для настройки параметров моделей машинного обучения может

приводить к неудовлетворительным результатам расчетов, поскольку выборка данных недостаточна для надежного снижения ошибки моделирования. Наиболее ярко эта особенность несбалансированных рядов может проявляться в задачах классификации, когда требуется определить вероятность возникновения экстремального события по набору входных данных. В настоящее время разработаны несколько подходов к балансировке классов в рядах для обучения моделей классификации, один из которых (SMOTE: Synthetic minority over-sampling technique [Chawla et al., 2002]) мы применили к задаче бинарной классификации в настоящей статье. К числу таких задач относится и определение вероятности возникновения ледовых заторов при прохождении весеннего половодья на реках с устойчивым ледоставом.

Ледовые заторы являются одним из наиболее опасных гидрологических явлений, в особенности характерным для рек бассейна Северного Ледовитого океана, верховьям которых свойственно более раннее начало весеннего ледохода. Возникновение ледовых заторов часто приводит к стремительному подъему уровня воды, а выход льда на берег также сопровождается повреждениями и разрушениями. Наиболее катастрофические заторы в России наблюдались в таких городах, как Великий Устюг, Котлас, Ленск, Томск, Якутск.

Прогнозирование ледовых заторов является довольно сложной задачей, что связано прежде всего с механизмом их формирования под воздействием множества факторов, таких как морфология русла, речной сток во время ледохода, осенний шугоход, температура воздуха и количество оттепелей за зимний период, сумма твердых осадков и т. д. [Бужин, 2004; Агафонова, Фролова, 2007]. Разработано множество методик оценки вероятности развития затора, учитывающих различные группы факторов [Козлов и др., 2015]. Однако на практике исследователям этих явлений часто приходится сталкиваться с неполнотой исходных данных для построения прогностических моделей. Кроме того, отсутствует четкий критерий определения факта наличия затора. Наблюдениями обычно охвачены только участки реки в непосредственной близости от гидрологического поста, а регулярные авиаразведки в последнее время являются редкостью и данные по ним, как правило, недоступны. В силу отмечающихся в последние годы климатических изменений, разработанные для других условий методики прогноза, требуют существенной переработки [Методические рекомендации..., 2004].

К настоящему времени имеется ряд довольно успешных примеров применения методов машинного обучения для прогнозирования ледовых заторов [Малыгин, 2014; Малыгин, Алешин, 2022; Сумачев и др., 2024]. Во многих из них, ввиду недостатка информации о ледовых явлениях, большое значение имеют данные метеорологических наблюдений. При этом, как правило, используются данные ближайшей метеостанции, атмосферного реанализа или их сочетание [Малыгин, Алешин, 2022]. Среди моделей машинного обучения применяются как простые методы: k ближайших соседей (*K Nearest Neighbors*), дерево решений (*Decision Tree, DT*), так и ансамблевые алгоритмы, например случайный лес (*Random Forest, RF*) [Малыгин, Алешин, 2022]. Особенно следует отметить нейронные сети, из которых наиболее распространены модели с архитектурой многослойного перцептрона (*Multi-layer perceptron, MLP*) [Агафонова и др., 2016; Guo et al., 2018; Wang et al., 2019; Graf et al., 2022]. Выбор ме-

тодики определяется как особенностями объекта исследования, так и исходными данными. Как упоминалось выше, при прогнозе ледовых заторов часто приходится сталкиваться с малой повторяемостью прогнозируемого события, что может быть связано как с экстремальным характером явления (затор, вызвавший катастрофическое наводнение), так и с редкой повторяемостью самого события (на участке поста) или недостаточной частотой наблюдений. Следовательно, для построения модели на этапе предобработки требуется приведение исходных данных к сопоставимому количеству наблюдений с наличием и отсутствием события (балансировка классов). Сочетание методов балансировки классов, подбора на кросс-валидации наиболее оптимальных гиперпараметров и вероятностного порога помогает моделям машинного обучения определять классы целевой переменной в случае сильного дисбаланса с более высокой точностью.

В данной работе была опробована процедура балансировки обучающей выборки SMOTE [Chawla et al., 2002] для разработки метода прогнозирования возникновения ледовых заторов на устьевом участке Печоры с использованием различных моделей машинного обучения, что во многом позволило добиться удовлетворительных результатов определения вероятности возникновения затора.

В первой части статьи дана краткая характеристика заторообразования в устье Печоры, его гидрологической изученности и описаны имеющиеся данные наблюдений за ледовыми явлениями. Вторая часть посвящена используемым материалам – данным метеорологических наблюдений, методам предобработки исходных данных, построению моделей прогноза и анализу полученных результатов. В заключении приведена оценка результатов, сделаны выводы о выявленных особенностях и преимуществах рассматриваемых методов для прогнозирования редких событий, намечены перспективы дальнейших исследований.

Объект исследования. Несмотря на широкое распространение в Северном полушарии на реках, текущих с юга на север [Природные опасности России, 2001], в значительной мере ледовым заторам подвержены устьевые участки рек в арктической зоне. Хотя в целом частота заторов на крупных реках от верховий к устьям уменьшается и весеннее половодье в устьях имеет более плавный ход развития, протяженность заторов в нижнем течении может достигать существенных масштабов, а морфометрические факторы, такие как сложная структура гидрографической сети и снижение уклонов русла также способствуют накоплению льда во время весеннего ледохода [Василенко, Банщикова, 2010; Сумачев и др. 2024]. Для устьев рек, в том числе

и в Арктической зоне, как правило, характерна высокая степень хозяйственного освоения, следовательно, наводнения и другие опасные и неблагоприятные явления заторного генезиса могут приводить к значительным ущербам [Магрицкий и др., 2024]. К подобным объектам относится и устье Печоры, которое было взято в качестве тестового объекта в настоящем исследовании. Как упоминалось ранее, имеющейся информации о ледовых заторах в устье Печоры недостаточно для полного представления о закономерностях их формирования [Магрицкий и др., 2024]. Хотя, согласно результатам периодических исследований [Лупачев, 1979], ледовые заторы в устье Печоры происходят практически каждый год, данные ежегодников свидетельствуют только о менее чем 50% повторяемости заторов, что также объясняется недостаточным пространственным охватом и частотой наблюдений за ледовыми явлениями (53% у села Ермицы и 21% у села Оксина) [Сумачев, Банщикова, 2021]. Ледовые заторы в устье Печоры не всегда приводят к катастрофическим последствиям: хотя вклад заторной составляющей максимальных уровней половодья может быть достаточно высок (до 2,2 м) [Михайлов, 1997], часто превышения критических отметок наблюдались в отсутствии затора или свободный ото льда период [Михайлов, Магрицкий, 2011].

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Подготовка исходных данных. В исследовании были использованы метеорологические и гидрологические данные. Использованные метеорологические данные были взяты с метеостанции Нарьян-Мар (индекс ВМО 23205) и включали среднесуточную, минимальную и максимальную температуру воздуха, сумму осадков за сутки за 1926–2022 гг. [Архив погоды..., 2024].

Гидрологические данные о наличии или отсутствии ледовых заторов относятся к участку устьевой области Печоры от гидрологического поста Оксина (индекс 70827, 141 км от морского края дельты) до морского гидрологического поста Осколково (индекс 70841, 53 км от устья).

На основе исходных данных были сформированы признаки, или предикторы, определяющие наличие или отсутствие затора (табл. 1). Признаки формировались на один календарный месяц с шагом по времени до семи месяцев назад, а также рассчитаны средние многолетние значения признаков на один месяц вперед. Следует отметить, что признаки в настоящем исследовании не включали гидрологические характеристики, такие как уровень и расход воды, толщину льда и т. п. Это было сделано в силу двух обстоятельств: 1) наличие таких данных было ограничено как по времени, так и по

Таблица 1

Сформированные признаки для каждого месяца

Признак	Единица измерения
Средняя температура	°C
Минимальная температура	°C
Сумма минимальных температур	°C
Максимальная температура	°C
Сумма максимальных температур	°C
Сумма осадков	мм
Максимальное количество осадков	мм
Количество дней с суммой осадков более 1 мм	ед.
Количество дней с суммой осадков более 10 мм	ед.
Разница между максимальной и минимальной температурой	°C
Количество дней со среднесуточной температурой ниже 0°C	ед.
Количество дней со среднесуточной температурой выше 0°C	ед.
Сумма отрицательных температур	°C
Сумма положительных температур	°C
Количество дней с среднесуточной отрицательной температурой и наличием осадков	ед.
Сумма квадратов отрицательных температур	°C ²
Сумма квадратов положительных температур	°C ²
Сумма осадков в дни с отрицательными температурами	мм
Сумма квадратов среднесуточных температур	°C ²
Сумма квадратов суточного количества осадков	мм ²
Количество переходов через 0°C	ед.
Среднее изменение среднесуточных температур	°C
Среднее изменение суточного количества осадков	мм

составу; 2) модель создавалась только с целью проверки алгоритма балансировки имеющихся данных наблюдений, а не для рекомендаций по внедрению в оперативную практику.

Целевая переменная представляла собой закодированное бинарное значение наличия затора в следующем месяце (1 – затор наблюдался, 0 – затор не наблюдался). Отбор признаков, оказывающих влияние на целевую переменную, производился на основе корреляционного анализа с отбрасыванием тех признаков, у которых коэффициент корреляции Пирсона и ранговой корреляции Спирмена менее 0,15. Итоговый размер сформированного набора данных составляет 1162×61 (1162 события, 60 признаков, 1 целевая переменная).

Основной задачей в настоящей работе являлась разработка прогноза наличия (1) или отсутствия (0) ледового затора на рассматриваемом участке устья Печоры. Исходные данные по целевой переменной имеют выраженный дисбаланс классов: число событий, когда заторы отсутствовали (1141 событие), значительно превышает количество самих заторов (23 события). В таком случае первый класс (также называемый мажоритарным) имеет существенно больший вес, и обучение модели для выявления событий второго класса (миноритарного) практически невозможно. Существуют различные способы решения такой проблемы [Bourel et al., 2021; Chawla et al., 2002], одним из них является балансировка классов: уменьшение числа объектов преобладающего класса (*downsampling*) или, напротив, увеличение числа меньшего (*oversampling*). В нашем случае ввиду небольшого объема исходных данных был выбран второй способ, в ходе выполнения которого в пространстве признаков генерируются новые объекты класса редких событий (в данном случае, ледовых заторов). Эта процедура была основана на алгоритме синтетического увеличения числа миноритарного класса (Synthetic minority over-sampling technique – SMOTE [Chawla et al., 2002]). Суть его заключается в искусственном генерировании случайного события из миноритарного класса в пространстве признаков. С этой целью для каждого из имеющихся событий методом ближайшего соседа выбираются k ближайших событий, анализируются расстояния (в пространстве признаков) до них, после чего каждое из этих расстояний умножается на случайную величину в диапазоне от 0 до 1, чтобы определить положение дополнительных сгенерированных событий. Количество ближайших соседних событий k является одним из параметров оптимизации. Такой подход позволяет генерировать дополнительные события внутри выборки уже имеющихся. Сгенерированные синтетические события добавляются к исходному набору данных, пока соотношение миноритарного и мажоритарного классов

не станет 1:1. Пример балансировки представлен на рис. 1. Процедура *oversampling* (SMOTE) применялась только для части данных, используемой для обучения модели.

Модели. Были опробованы различные алгоритмы машинного обучения, среди которых классификация методом ближайших соседей (*K-Nearest Neighbors* – *KNN*), логистическая регрессия (*Logistic Regression*), градиентный бустинг (*Catboost*) и многослойный перцептрон (*Multi-Layer perceptron*, *MLP*). Для построения моделей машинного обучения использовались библиотеки *sklearn* [Pedregosa et al., 2011] и *Catboost* [Catboost..., 2024] языка программирования *Python*, *imblearn* [Imblearn..., 2024] для балансировки классов, *optuna* [Optuna..., 2024] для подбора гиперпараметров моделей. Обучение моделей, анализ и визуализация результатов выполнялись в среде разработки *Jupyter Notebook*. Полученные алгоритмы доступны в публичном репозитории [Github..., 2024].

Для всех моделей на обучение отводилось 80% исходных данных, на тестирование – оставшиеся 20%. Для определения параметров процесса обучения моделей (гиперпараметров моделей) применялась процедура кросс-валидации обучающей выборки также на 80 и 20% (всего пять блоков). В обучающую выборку попало 18 событий с ледовыми заторами, в тестовую выборку пять событий.

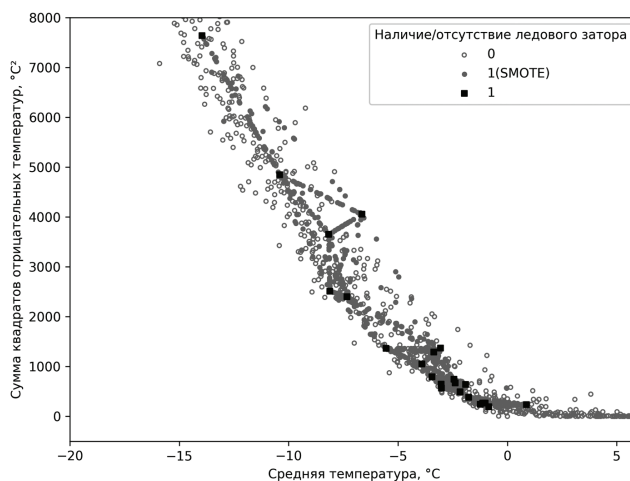


Рис. 1. Пример балансировки набора данных с помощью алгоритма SMOTE в пространстве двух признаков:

0 – отсутствие затора; 1 – заторное событие по данным наблюдений; 1(SMOTE) – искусственно сгенерированные алгоритмом SMOTE события наличия заторов

Fig. 1. Illustration of SMOTE dataset balancing within two-dimensional feature space:
0 – no congestion; 1 – congestion event according to observations;
1(SMOTE) – congestion events artificially generated by the SMOTE algorithm

В зависимости от используемой модели машинного обучения в ходе кросс-валидации определялись следующие гиперпараметры. Для всех моделей, кроме модели градиентного бустинга деревьев решений (Catboost), проводилось предварительное масштабирование данных. Для этого использовались алгоритмы StandardScaler (стандартизация, вычитание выборочного среднего и деление на выборочное стандартное отклонение), RobustScaler (вычитание выборочного медианного значения и деление на выборочный межквартильный интервал), MinMaxScaler (вычитание минимального значения выборки и деление на выборочный размах). В качестве гиперпараметра использовалось количество k ближайших соседей (от 4 до 10), до которых в алгоритме SMOTE оценивалось расстояние, на котором генерировались новые события. Для каждого алгоритма классификации также подбирались значения характерных для них гиперпараметров. Например, для модели KNN также подбиралось число ближайших соседей (1–15), тип определения значимости (весов) соседей: равнозначно (uniform) или в зависимости от расстояния (distance), а также метрика расстояний до них: евклидово (euclidean), манхеттенское (manhattan), минковского (minkowski).

Для модели Logistic Regression подбирались метод (L1 или регуляризация Тихонова L2) и коэффициент регуляризации (0,001–1000). Для модели деревьев решений Catboost подбирались количество деревьев (100–1000), глубина дерева (4–10), минимальный размер выборки для формирования листа (1–100) и скорость обучения (0,001–0,5). Для искусственной нейронной сети MLP подбирались размер скрытого слоя (32–512 нейронов), функция активации (identity, relu, logistic, tanh), скорость обучения (0,0001–0,01), коэффициент регуляризации (0,00001–0,01).

Одним из гиперпараметров моделей был вероятностный порог – это минимальное значение вероятности, при превышении которого объект относится к классу «затор». Пороговое значение также подбиралось на кросс-валидации обучающей выборки с помощью библиотеки Optuna. Слишком высокий порог может привести к игнорированию значимых положительных примеров, в то время как слишком низкий – к увеличению числа ложных срабатываний.

Итоговая последовательность процедур по подбору гиперпараметров, обучению, валидации и тестированию моделей представлена на рис. 2.

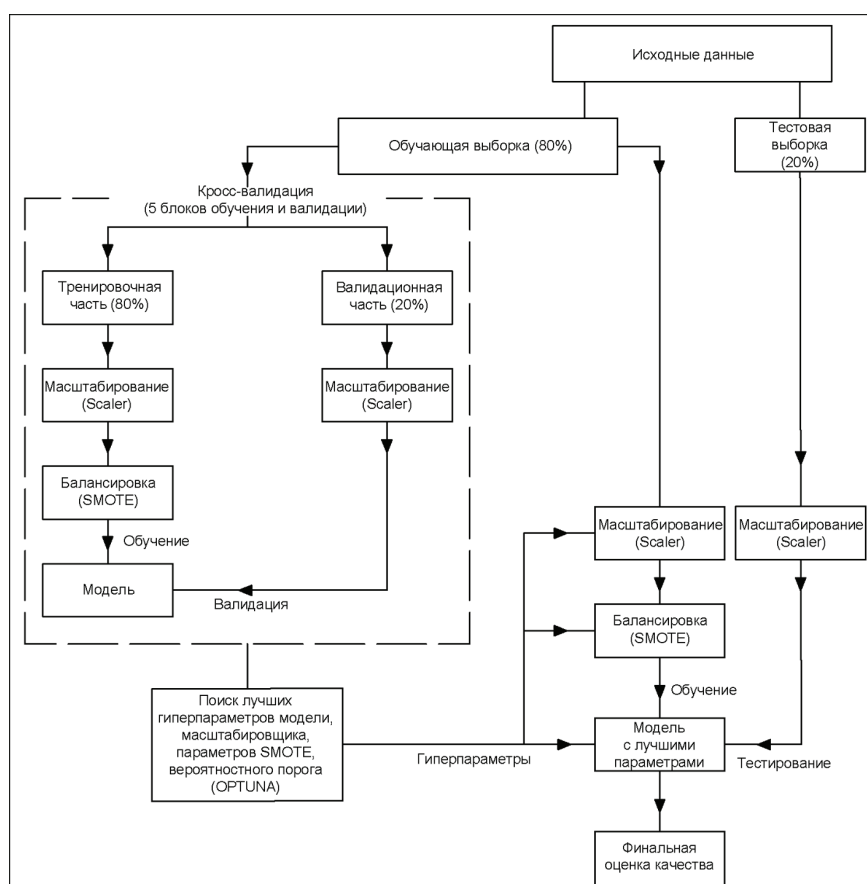


Рис. 2. Принципиальная схема по подбору лучших гиперпараметров, обучению, валидации и тестированию моделей

Fig. 2. Flowchart of hyperparameter optimization, model fit, validation and testing

Метрики оценки качества. Первичной оценкой в задачах классификации является вычисление матрицы ошибок (табл. 2). На ее основе рассчитываются показатели условной вероятности правильных прогнозов событий, которые показывают различные аспекты качества создаваемой модели. Так, по матрице ошибок рассчитывается показатель точности (Precision) (1), определяемый как доля правильных прогнозов среди всех прогнозов наступления события, и показатель полноты определения (Recall) (2), определяемый как отношение правильных прогнозов заторов ко всем наблюдавшимся заторам. Значения обоих показателей определены на отрезке [0, 1], где 0 – наихудший показатель, 1 – наилучший. В свою очередь, точность и полнота могут быть выражены в единой метрике F1-score (3), представляющей собой их среднее гармоническое. Для настоящего исследо-

вании F1-score является наиболее оптимальной метрикой оценки качества, так как в ней совмещаются задачи максимизирования истинных положительных предсказаний (TP) и минимизирования значений ложных положительных срабатываний (FP) и ложных отрицательных предсказаний (FN). Значения F1-score находятся на отрезке от 0 до 1, причем 1 – наилучшее качество классификации, 0 – наихудшее.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

$$F1\text{-score} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Таблица 2

Матрица ошибок прогноза заторов

		Наблюдения	
		Затор (1)	Нет затора (0)
Прогноз	Затор (1)	Верный прогноз затора (true positive – TP)	Ошибочный прогноз затора (false positive – FP)
	Нет затора (0)	Ошибочный прогноз отсутствия затора (false negative – FN)	Верный прогноз отсутствия затора (true negative – TN)

Оценка важности признаков моделей. Оценка важности признаков выполнена для модели логистической регрессии через анализ коэффициентов перед переменными *X*. Каждый коэффициент указывает на величину и направление влияния соответствующего признака на вероятность отнесения объекта к одному из классов. Если коэффициент положителен, то увеличение значения признака повышает вероятность отнесения объекта к положительному классу; если отрицателен – вероятность снижается.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Табл. 3 отражает результаты оценки F1-score для четырех моделей: CatBoost, KNN, MLP и логистической регрессии на валидационной (validation) и тестовой (test) выборках. Оценка производилась на различных подходах: 1) модели с параметрами по умолчанию по исходному ряду; 2) применение метода балансировки выборки (SMOTE); 3) использование SMOTE с оптимизацией гиперпараметров и вероятностного порога.

Без применения методов балансировки данных все модели, за исключением MLP и KNN, показали нулевые значения F1-score как на валидационной, так и на тестовой выборке. Это обусловлено сильной несбалансированностью исходного набора данных, в результате чего модели не смогли эффективно обучиться для прогноза редкого класса – наличия затора. MLP и KNN, однако, дали F1-score 0,258 и 0,080 на валидационной выборке соответственно.

Применение метода SMOTE для балансировки выборки улучшило результаты всех моделей, за исключением Catboost, которая продолжила показывать нулевые значения F1-score как на валидационной, так и на тестовой выборках. Результаты по моделям KNN и MLP значительно улучшились после применения SMOTE, F1-score достигла на валидационной выборке 0,333 и 0,258 соответственно, хотя на тестовой выборке KNN показал нулевое значение F1-score, тогда как MLP сохранил значение 0,333. Это указывает на ограниченную обобщающую способность KNN на тестовой выборке при использовании только SMOTE. Результаты по логистической регрессии также улучшились, F1-score достигла 0,250

на валидационной выборке и 0,000 на тестовой, что указывает на сложность обобщения модели с синтетически сбалансированными данными.

Наиболее значительные улучшения были достигнуты при использовании SMOTE в сочетании с оптимизацией гиперпараметров и вероятностного порога. По всем моделям произошел рост показателей: Catboost, ранее показывавший нулевые результаты, достиг F1-score 0,403 на валидационной и 0,286 на тестовой выборке. Хотя результат на тестовой выборке остался сравнительно низким, это указывает на опре-

деленное улучшение способности модели к обобщению после оптимизации. Результаты KNN также улучшились, F1-score вырос до 0,459 на валидации и 0,444 на тесте, что подтверждает положительный эффект от оптимизации. Результаты использования логистической регрессии также улучшились, F1-score принял значение 0,336 на валидационной и 0,444 на тестовой выборке. Значительное улучшение отмечено для MLP: F1-score составил 0,430 на валидации и 0,500 на тесте, что подтверждает пользу от подбора гиперпараметров для обобщающей способности модели.

Таблица 3

Результаты оценки моделей по метрике F1-score на валидационной и тестовой выборке

Модель	С параметрами по умолчанию по исходному ряду		SMOTE		Подбор гиперпараметров + SMOTE + вероятностный порог	
	valid	test	valid	test	valid	test
Catboost	0,000	0,000	0,000	0,000	0,403	0,286
KNN	0,080	0,000	0,333	0,000	0,459	0,444
MLP	0,258	0,333	0,258	0,333	0,430	0,500
Logistic Regression	0,183	0,000	0,250	0,000	0,336	0,444

Таблица 4

Дополнительные метрики оценки качества моделей на тестовой выборке со SMOTE, подбором лучших параметров и вероятностного порога

Модель	Recall	Precision	TN	FP	FN	TP
CatBoost	0,400	0,222	221	7	3	2
KNN	0,400	0,500	226	2	3	2
MLP	0,400	0,667	227	1	3	2
Logistic Regression	0,800	0,308	219	9	1	4

Табл. 4 содержит результаты четырех моделей (Catboost, KNN, MLP, Logistic Regression) по основным метрикам классификации: Recall (полнота), Precision (точность), а также по количеству истинных отрицаний (TN), ложных положительных срабатываний (FP), ложных отрицательных предсказаний (FN) и истинных положительных предсказаний (TP). Эти метрики позволяют оценить качество работы моделей в задаче классификации ледовых заторов.

Модели Catboost, KNN и MLP сумели предсказать каждая по два затора из пяти, представленных в тестовой выборке, однако у них наблюдается разное количество ложноположительных срабатываний (FP). У модели MLP наблюдается наименьшее количество ложноположительных срабатываний на тестовой выборке (FP = 1, Precision = 0,667).

Модель логистической регрессии показывает сбалансированные результаты, достигая высокой полноты (Recall = 0,800), но достаточно низкой точ-

ности (Precision = 0,308). Модель правильно классифицировала 219 примеров без заторов (TN), совершив девять ложных положительных ошибок (FP), и допустила одну ложную отрицательную ошибку (FN). Она правильно распознала четыре из пяти заторов (TP).

График результата подбора наилучшего вероятностного порога на валидационной и тестовой выборках с балансировкой SMOTE и лучшими гиперпараметрами по модели логистической регрессии представлен на рис. 3. Наилучший F1-score на валидационной выборке был достигнут при значении вероятностного порога в 77–79%. Библиотека Optuna подобрала лучший порог в 78% отношения объекта классификации к классу «1». Подбор порога позволяет балансировать между точностью (precision) и полнотой (recall), что особенно важно для миноритарного класса. Лучший порог в 78% также позволяет добиться наилучших результатов и на тестовой выборке.

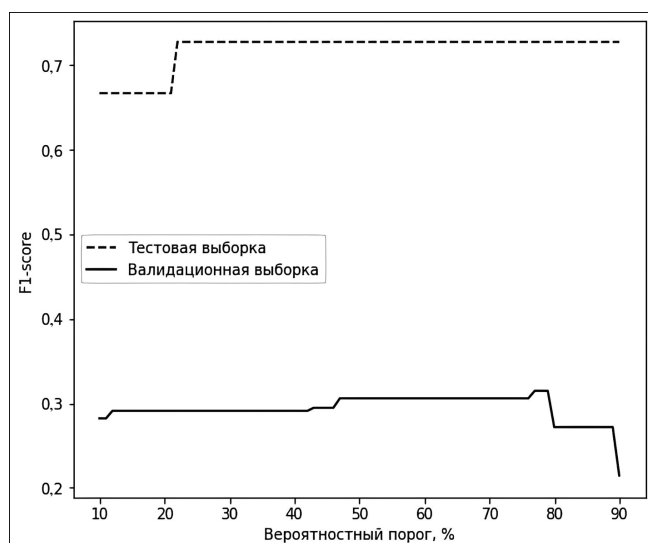


Рис. 3. График зависимости F1-score от вероятностного порога на валидационной и тестовой выборке по модели логистической регрессии со SMOTE и лучшими параметрами

Fig. 3. F1-score dependency from probability threshold on validation and test subsets for logistic regression model with SMOTE preprocessing and best parameter set

Рис. 4 содержит весовые коэффициенты 20 самых важных признаков, полученных с помощью логистической регрессии для задачи прогнозирования ледовых заторов на тестовой выборке со SMOTE, лучшими гиперпараметрами и значением вероятностного порога. Признаки с положительными весовыми коэффициентами влияют на вероятность отношения к классу «1», т. е. вносят больший вклад в наличие ледового затора в следующем месяце. В свою очередь признаки с отрицательными коэффициентами влияют на вероятность отношения к

классу «0», т. е. вносят больший вклад в отсутствие ледового затора в следующем месяце. Наиболее важными являются признаки, определяющие температуру воздуха в текущем и следующем месяцах, а также шесть месяцев назад, т. е. относятся к характеристикам периода ледостава. Также согласно рис. 4 отмечается важность среднего изменения суточного количества осадков шесть месяцев назад, т. е. в период ледостава. Признаки, сформированные на текущий месяц, имеют отрицательные весовые коэффициенты, а за следующий месяц – положительные (см. рис. 4). Признаки, представленные на рис. 3, соответствуют факторам, определяющим наличие ледовых заторов на различных изучаемых объектах [Massie et al., 2003; Wang et al., 2019; Семенова и др., 2020]. Следует отметить, что набор признаков включает не все возможные переменные, в том числе отсутствуют признаки, ответственные за речной сток и состояние ледового покрова, наблюдаемые на гидрологических постах. Это обстоятельство иллюстрирует тот факт, что примененные в исследовании алгоритмы позволяют добиться удовлетворительных результатов прогнозов на используемом материале, а также открывает дополнительные перспективы для продолжения исследований на более обширном материале.

ВЫВОДЫ

В статье продемонстрирована целесообразность использования методов предварительной балансировки данных и оптимизации параметров при работе с несбалансированными выборками. Без применения процедуры SMOTE большинство моделей не смогли справиться с задачей классификации ледовых заторов, показав в основном нулевые значения метрики F1-score.

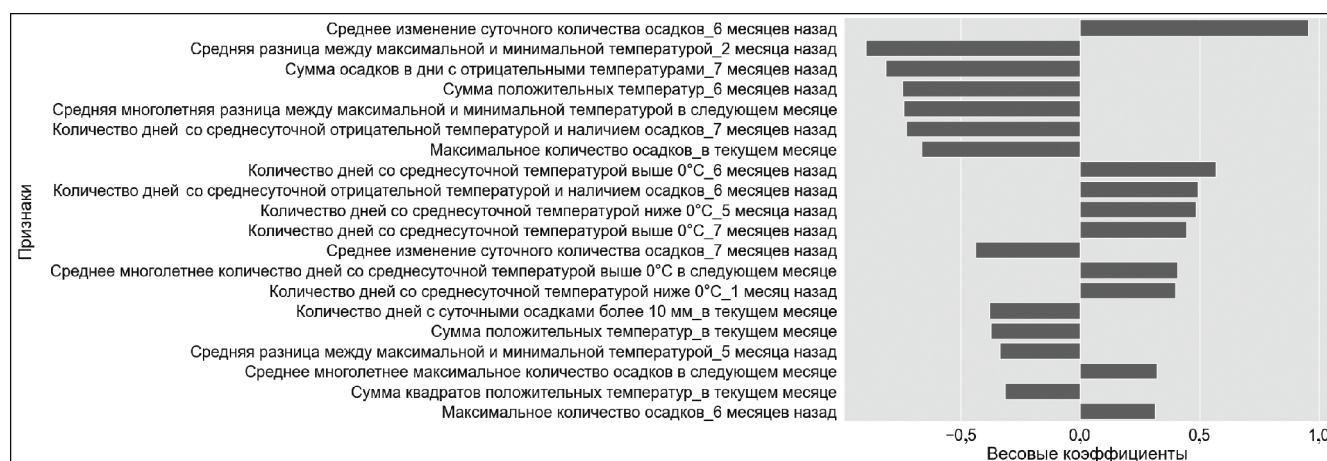


Рис. 4. Весовые коэффициенты признаков модели логистической регрессии на тестовой выборке со SMOTE, лучшими параметрами и вероятностным порогом

Fig. 4. Weight coefficients of logistic regression model features for a test subset with SMOTE preprocessing, best parameter set and probability threshold

Применение SMOTE совместно с оптимизацией гиперпараметров моделей и подбором вероятностного порога выделения события ледового затора значительно улучшило результаты прогнозирования его возникновения с заблаговременностью 1 месяц.

Перспективами развития этой задачи будет являться добавление новых признаков о водности реки, ледовых явлениях, расширения географии поиска событий заторов на весь устьевой участок. Дополнительными перспективами исследова-

ний может быть также применение других методов машинного обучения, таких как LSTM и Transformer.

Использование алгоритма балансировки классов SMOTE важно для задач с несбалансированными данными, так как оно помогает увеличить количество примеров редкого класса за счет создания синтетических образцов и улучшить обучение моделей машинного обучения. Это может быть полезно в задачах классификации и предсказания редких гидрологических явлений.

Благодарность. Работа выполнена в рамках темы № FMWZ-2022-0003 государственного задания ИВП РАН.

СПИСОК ЛИТЕРАТУРЫ

- Агафонова С.А., Василенко А.Н., Фролова Н.Л. Факторы образования ледовых заторов на реках бассейна Северной Двины в современных условиях // Вестн. Моск. ун-та. Сер. 5. Геогр. 2016. № 2. С. 82–90.
- Агафонова С.А., Фролова Н.Л. Особенности ледового режима рек бассейна Северной Двины // Водные ресурсы. 2007. № 2. С. 141–149.
- Бузин В.А. Заторы льда и заторные наводнения на реках. СПб.: Гидрометеоздат, 2004. 203 с.
- Василенко Н.Г., Банищикова Л.С. Оперативная оценка участков образования заторов льда и их параметров // Лед и снег. 2010. № 2(110). С. 61–65.
- Козлов Д.В., Бузин В.А., Фролова Н.Л. и др. Опасные ледовые явления на реках и водохранилищах России: монография / под ред. Д.В. Козлова. М.: Изд-во РГАУ-МСХА им. К.А. Тимирязева. 2015. 348 с.
- Лупачев В.Г. Ледовые заторы на реке Печоре и их прогнозирование // Метеорология и гидрология. 1979. № 4. С. 45–51.
- Магрицкий Д.В., Агафонова С.А., Банищикова Л.С. и др. Гидрологические опасности в устье Печоры // Проблемы Арктики и Антарктики. 2024. Т. 70. № 2. С. 185–209. DOI: 10.30758/0555-2648-2024-70-2-185-209.
- Малыгин И.В. Методика прогноза образования ледовых заторов на реках на основе теории распознавания образов // Вестн. Моск. ун-та. Сер. 5. Геогр. 2014. № 3. С. 43–47.
- Малыгин И.В., Алешин И.М. Прогнозирование заторов льда на р. Лене методами машинного обучения // Геофизические процессы и биосфера. 2022. Т. 21. № 3. С. 18–26. DOI: 10.21455/GPB2022.3-3.
- Методические рекомендации по предотвращению образования ледовых заторов на реках Российской Федерации и борьбе с ними. М.: ФЦ ВНИИ ГОЧС, 2004. 234 с.
- Михайлов В.Н., Магрицкий Д.В. Новое в исследовании экстремальных гидрологических процессов // Вестн. Моск. ун-та. Сер. 5. Геогр. 2011. № 6. С. 108–109.
- Михайлов В.Н. Заторные явления на реках России // Водные ресурсы. 1997. Т. 24. № 3. С. 345–353.
- Природные опасности России. Т. 5. Гидрометеорологические опасности / под ред. Г.С. Голицына, А.А. Васильева. М.: КРУК, 2001. 480 с.
- Семенова Н.К., Сазонов А.А., Крыленко И.Н. Прогнозирование возможности образования заторов льда с помощью методов машинного обучения // Четвертые Виноградовские чтения. Гидрология от познания к мировоззрению: сб. докладов Междунар. науч. конф. Санкт-Петербург, 23–31 октября 2020 г. СПб.: Изд-во ВВМ, 2020. С. 358–361.
- Сумачев А.Э., Банищикова Л.С. Ледовый режим реки Печоры в современных климатических условиях и принципы прогнозирования высшего уровня воды за период весеннего ледохода // Успехи современного естествознания. 2021. № 10. С. 75–80. DOI: 10.17513/use.37701.
- Сумачев А.Э., Банищикова Л.С., Грига С.А. Применение методов обучения искусственных нейронных сетей при прогнозировании высших уровней воды на примере рек Двинско-Печорского бассейнового округа // Метеорология и гидрология. 2024. № 4. С. 104–115. DOI: 10.52002/0130-2906-2024-4-104-115.
- Bourel M., Segura A.M., Crisci C. et al. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters, *Water Research*, 2021, vol. 202, DOI: 10.1016/j.watres.2021.117450.
- Chawla N.V., Bowyer K.W., Hall L.O. et al. SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, 2002, vol. 16, p. 321–357, DOI: 10.1613/jair.953.
- Graf R., Kolarski T., Zhu S. Predicting ice phenomena in a river using the artificial neural network and extreme gradient boosting, *Resources*, 2022, vol. 11, no. 12, 29 p.
- Guo X., Wang T., Fu H. et al. Ice-jam forecasting during river breakup based on neural network theory, *Journal of Cold Regions Engineering*, 2018, vol. 32, iss. 3, p. 04018010, DOI: 10.1061/(ASCE)CR.1943-5495.0000168.
- Massie D.D., White K.D., Daly S.F. Application of neural networks to predict ice jam occurrence, *Cold Reg. Eng.*, 2002, vol. 35, no. 2, p. 0–122, DOI: 10.1016/s0165-232x(02)00056-3.
- Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 2011, vol. 12, no. 85, p. 2825–2830.
- Wang T., Guo X., Fu H. et al. Breakup ice jam forecasting based on neural network theory and formation factor, *E-proceedings of the 38th IAHR World Congress*, Pan-

- ama City, September 1–6, 2019, Panama, IAHR, 2019, p. 2488, DOI: 10.3850/38WC092019-0641.
- Wang H., Meng Y., Xu H. et al. Prediction of flood risk levels of urban flooded points through using machine learning with unbalanced data, *J. Hydrol.*, 2024, vol. 630, p. 130742, DOI: 10.1016/j.jhydrol.2024.130742.
- Wu Y., Ding Y., Feng J. SMOTE-Boost-based sparse Bayesian model for flood prediction, *J. Wireless Com. Network*, 2020, vol. 78, DOI: 10.1186/s13638-020-01689-2.
- Wu Y., Yukai D., Feng J. Sparse Bayesian flood forecasting model based on SMOTEBoost, *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Atlanta, GA, USA, 2019, p. 279–284, DOI: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00067.
- Электронные источники
- Архив погоды в Нарьян-Маре. URL: <http://www.pogodaiklimat.ru/weather.php?id=23205> (дата обращения 01.06.2024).
- Catboost – высокопроизводительная библиотека с открытым исходным кодом для градиентного бустинга на деревьях решений. URL: <https://catboost.ai/> (дата обращения: 02.08.2024).
- Imblearn: библиотека. URL: <https://imbalanced-learn.org/> (дата обращения 02.08.2024).
- Optuna: фреймворк оптимизации гиперпараметров. URL: <https://optuna.readthedocs.io/> (дата обращения 02.08.2024).
- Sklearn: машинное обучение на Python. URL: <https://scikit-learn.org/stable/> (дата обращения 02.08.2024).
- Github. URL: https://github.com/SergeyIglin/ML_rarehydroevents_with_smote (дата обращения 16.10.2024).

Поступила в редакцию 19.10.2024

После доработки 15.11.2024

Принята к публикации 26.11.2024

FORECASTING RARE HYDROLOGICAL EVENTS BY MACHINE LEARNING METHODS: CASE STUDY OF ICE JAMS ON THE PECHORA RIVER

S.M. Iglin¹, V.M. Moreido², K.I. Golovnin³

^{1–3} Institute of Water Problems of the Russian Academy of Sciences, Laboratory of Hydroinformatics

¹ Scientific Researcher, Ph.D. in Geography; e-mail: iglin.sm@bk.ru

² Senior Scientific Researcher, Head of the Laboratory, Ph.D. in Geography; e-mail: vsevolod.moreido@iwp.ru

³ Engineer; e-mail: golovnin98@bk.ru

Rare hydrological events, as the name suggests, occur quite infrequently, but are often catastrophic for humans. They are also inadequately provided with measurements (the so-called class imbalance). In its turn, this hinders the creation of reliable models for predicting such processes. This is especially evident when constructing models of natural processes using machine learning algorithms, which are particularly sensitive to class-imbalanced samples. The study attempts to overcome the above-mentioned limitations by supplementing a series for model training with artificially generated events. The subject and object of the study were long-term forecasts of ice jams occurring at the mouth of the Pechora River in the Arctic area of the European Russia. Data on ice jams were collected over a long period of observations, and applicable predictors and models were selected. The following machine learning algorithms were used: *k*-nearest neighbors (KNN), logistic regression, gradient boosting (CatBoost), and multilayer perceptron (MLP). As a result all models demonstrated higher quality of modeling after supplementing artificial events to a series. This confirms the prospects of the method of series supplementing for training models of rarely occurring processes.

Keywords: ice jam forecasting, class imbalance, machine learning, class balancing

Acknowledgements. The study was carried out under the state task of the RAS Institute of Water Problems, theme no. FMWZ-2022-0003.

REFERENCES

- Agafonova S.A., Frolova N.L. Specific features of ice regime in rivers of the Northern Dvina Basin], *Water Resources*, 2007, no. 2, p. 123–131.
- Agafonova S.A., Vasilenko A.N., Frolova N.L. Faktory obrazovaniya ledovykh zatorov na rekah bassejna Severnoj Dviny v sovremennykh usloviyakh [The present-day factors of ice jam formation on the rivers of the Severnaya Dvina River Basin], *Vestn. Mosk. un-ta, Ser. 5, Geogr.*, 2016, no. 2, p. 82–90. (In Russian)
- Bourel M., Segura A.M., Crisci C. et al. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters, *Water Research*, 2021, vol. 202, DOI: 10.1016/j.watres.2021.117450.
- Buzin V.A. Zatory l'da i zatornye navodneniya na rekah [Ice Jams and Ice Jam Floods on Rivers], St Petersburg, Gidrometeoizdat Publ., 2004, 203 p. (In Russian)
- Chawla N.V., Bowyer K.W., Hall L.O. et al. SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, 2002, vol. 16, p. 321–357, DOI: 10.1613/jair.953.

- Goroshkova N.I., Zavadskij A.S., Krylenko I.N. et al. Opasnye ledovye javlenija na rekah i vodohranilishchah Rossii: Monografija [Dangerous Ice Phenomena on Rivers and Reservoirs of Russia: A Monography], D.V. Kozlov (ed.), Moscow, RGAU-MSHA im. K.A. Timirjazeva Publ., 2015, 348 p. (In Russian)
- Graf R., Kolarski T., Zhu S. Predicting ice phenomena in a river using the artificial neural network and extreme gradient boosting, *Resources*, 2022, vol. 11, no. 12, 29 p.
- Guo X., Wang T., Fu H. et al. Ice-jam forecasting during river breakup based on neural network theory, *Journal of Cold Regions Engineering*, 2018, vol. 32, iss. 3, p. 04018010, DOI: 10.1061/(ASCE)CR.1943-5495.0000168.
- Lupachev V.G. Ledovye zatory na reke Pechore i ih prognozirovanie [Ice Jams on the Pechora River and Their Forecasting], *Meteorologija i gidrologija*, 1979, no. 4, p. 45–51. (In Russian)
- Magrickij D.V., Agafonova S.A., Bانشhikova L.S. et al. Gidrologicheskie opasnosti v ust'e Pechory [Hydrological hazards at the mouth of the Pechora River], *Problemy Arktiki i Antarktiki*, 2024, vol. 70, no. 2, p. 185–209, DOI: 10.30758/0555-2648-2024-70-2-185-209. (In Russian)
- Malygin I.V. Metodika prognoza obrazovaniya ledovyh zatorov na rekah na osnove teorii raspoznavaniya obrazov [Forecasting of ice clogging in rivers using the theory of image recognition], *Vestn. Mosk. un-ta, Ser. 5, Geogr.*, 2014, no. 3, p. 43–47. (In Russian)
- Malygin I.V., Aleshin I.M. Prognozirovanie zatorov l'da na r. Lene metodami mashinnogo obucheniya [Forecasting ice jams on the Lena River using machine learning methods], *Geofizicheskie processy i biosfera*, 2022, vol. 21, no. 3, p. 18–26, DOI: 10.21455/GPB2022.3-3. (In Russian)
- Massie D.D., White K.D., Daly S.F. Application of neural networks to predict ice jam occurrence, *Cold Reg. Eng.*, 2002, vol. 35, no. 2, p. 0–122, DOI: 10.1016/s0165-232x(02)00056-3.
- Metodicheskie rekomendacii po predotvrashheniju obrazovaniya ledovyh zatorov na rekah Rossijskoj Federacii i bor'be s nimi [Methodological recommendations for preventing and mitigating ice jams formation on the rivers of the Russian Federation and its counteraction], Moscow, FC VNII GOChS Publ., 2004, 234 p. (In Russian)
- Mihajlov V.N. Zatornye javlenija na rekah Rossii [Ice jam phenomena on the rivers of Russia], *Vodnye resursy*, 1997, vol. 24, no. 3, p. 345–353. (In Russian)
- Mihajlov V.N., Magrickij D.V. Novoe v issledovanii ekstremal'nyh gidrologicheskikh protsessov [New approaches in the investigation of extreme hydrological processes], *Vestn. Mosk. un-ta, Ser. 5, Geogr.*, 2011, no. 6, p. 108–109. (In Russian)
- Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 2011, vol. 12, no. 85, p. 2825–2830.
- Prirodnye opasnosti Rossii, t. 5, Gidrometeorologicheskie opasnosti [Natural hazards of Russia, vol. 5, Hydrometeorological hazards], G.S. Golitsyn, A.A. Vasil'ev (eds.), Moscow, KRUK Publ., 2001, 480 p. (In Russian)
- Semenova N.K., Sazonov A.A., Krylenko I.N. [Forecasting the possibility of ice jam formation using machine learning methods], *Chetvertye Vinogradovskie chtenija. Gidrologija ot poznaniya k mirovozzreniju* [Hydrology from understanding to thinking], St Peterburg, 23–31 oktjabrja 2020 g. St Peterburg, Izdatel'stvo VVM Publ., 2020, p. 358–361. (In Russian)
- Sumachev A.Je., Bانشhikova L.S. Ledovyj rezhim reki Pechory v sovremennyh klimaticheskikh uslovijah i principy prognozirovaniya vysshego urovnja vody za period vesennego ledohoda [Ice regime of the Pechora River under modern climatic conditions and principles of forecasting the highest water level during the spring ice drift], *Uspehi sovremennogo estestvoznaniya*, 2021, no. 10, p. 75–80, DOI: 10.17513/use.37701. (In Russian)
- Sumachev A.Je., Bانشhikova L.S., Griga S.A. Primenenie metodov obucheniya iskusstvennyh neyronnyh setej pri prognozirovanii vysshih urovnej vody na primere rek Dvinsko-Pechorskogo bassejnogo okruga [Using neural network methods for peak water level prediction: a case study for the Dvina-Pechora basin rivers], *Meteorologija i gidrologija*, 2024, no. 4, p. 104–115, DOI: 10.52002/0130-2906-2024-4-104-115. (In Russian)
- Vasilenko N.G., Bانشhikova L.S. Operativnaja ocenka uchastkov obrazovaniya zatorov l'da i ih parametrov [Operational Evaluation of Ice Jams Origin Sites and Their Parameters], *Led i sneg.*, 2010, no. 2(110), p. 61–65. (In Russian)
- Wang H., Meng Y., Xu H. et al. Prediction of flood risk levels of urban flooded points through using machine learning with unbalanced data, *J. Hydrol.*, 2024, vol. 630, p. 130742, DOI: 10.1016/j.jhydrol.2024.130742.
- Wang T., Guo X., Fu H. et al. Breakup ice jam forecasting based on neural network theory and formation factor, *E-proceedings of the 38th IAHR World Congress*, Panama City, September 1–6, 2019, Panama, IAHR, 2019, p. 2488, DOI: 10.3850/38WC092019-0641.
- Wu Y., Ding Y., Feng J. SMOTE-Boost-based sparse Bayesian model for flood prediction, *J. Wireless Com. Network*, 2020, vol. 78, DOI: 10.1186/s13638-020-01689-2.
- Wu Y., Yukai D., Feng J. Sparse Bayesian flood forecasting model based on SMOTEBoost, *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Atlanta, GA, USA, 2019, p. 279–284, DOI: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00067.

Web sources

- Arhiv pogody v Narian-Mare [Weather archive in Narian-Mar] (In Russian), URL: <http://www.pogodaiklimat.ru/weather.php?id=23205> (access date 01.06.2024).
- Catboost – high-performance open source library for gradient boosting on decision trees, URL: <https://catboost.ai/> (access date: 02.08.2024).
- Imblearn: library, URL: <https://imbalanced-learn.org/> (access date 02.08.2024).
- Optuna: a hyperparameter optimization framework, URL: <https://optuna.readthedocs.io/> (access date 02.08.2024).
- Sklearn: Machine Learning in Python, URL: <https://scikit-learn.org/stable/> (access date 02.08.2024).
- Github, URL: https://github.com/SergeyIlglin/ML_rarehydroevents_with_smote (access date 16.10.2024).

Received 19.10.2024

Revised 15.11.2024

Accepted 26.11.2024