УДК 004.085
DOI: 10.26102/2310-6018/2024.47.4.029

Анализ характеристик селевых потоков при ограниченных данных с использованием моделей машинного обучения

Л.А. Лютикова

Институт прикладной математики и автоматизации Кабардино-Балкарского научного центра Российской академии наук, Нальчик, Российская Федерация

Резюме. В работе предложен комбинированный метод анализа неполной и искаженной информации, продемонстрированный на примере прогнозирования селей. Основная цель исследования заключается в демонстрации возможности не только создавать точные прогнозы, но и разбирать механизмы принятия решений модели, идентифицируя значимые параметры, влияющие на предсказания. Для представления выявленных комплексов параметров, влияющих на объем селевого потока, в виде логических правил потребовалось использование категоризации данных. Это позволило повысить надежность моделей при наличии выбросов и шума, а также учесть нелинейности. Для формирования логических правил применялись два подхода: метод ассоциативного анализа и оригинальная методика построения логического классификатора. В результате ассоциативного анализа были выявлены правила, отражающие определенные закономерности в данных, которые, как оказалось, нуждались в значительной коррекции. Использование логического классификатора позволило уточнить и скорректировать закономерности, обеспечив определение комплекса факторов, влияющих на объем селевого потока. Такой подход позволил выявить наиболее существенные входные переменные и понять, каким образом модель обрабатывает данные для генерации прогноза, определить факторы, играющие ключевую роль в результатах прогнозирования, обеспечить адекватную точность и стабильность прогнозов с учетом особенностей и сложности данных о селевых явлениях. Выведенные в результате исследования закономерности, отражающие скрытые принципы исследуемой предметной области, методы логического анализа, использованные в исследовании, помогли установить возможные причины формирования разных объемов выносимых твердых отложений. Полученные результаты могут быть использованы для совершенствования систем мониторинга и предотвращения негативных последствий селевых явлений.

Ключевые слова: машинное обучение, нейронные сети, кластерный анализ, ассоциативные правила, селевые потоки, модель.

Для цитирования: Лютикова Л.А. Анализ характеристик селевых потоков при ограниченных данных с использованием моделей машинного обучения. *Моделирование, оптимизация и информационные технологии.* 2024;12(4). URL: https://moitvivt.ru/ru/journal/pdf?id=1747 DOI: 10.26102/2310-6018/2024.47.4.029

Analysis of mudflow characteristics with limited data using machine learning models

L.A. Lyutikova

Institute of Applied Mathematics and Automation of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, Nalchik, the Russian Federation

Abstract. In paper, a combined method for analyzing incomplete and distorted information is proposed, demonstrated by the example of mudflow forecasting. The main purpose of the study is to demonstrate the ability not only to create accurate forecasts, but also to analyze the decision-making mechanisms of the model, identifying significant parameters that affect predictions. To represent the identified sets of

parameters affecting the volume of the mudflow in the form of logical rules, it was necessary to use data categorization. This made it possible to increase the reliability of models in the presence of emissions and noise, as well as to take into account non-linearities. Two approaches were used to form logical rules: the method of associative analysis and the original method of constructing a logical classifier. As a result of associative analysis, rules were identified that reflect certain patterns in the data, which, as it turned out, required significant correction. The use of a logical classifier made it possible to clarify and correct the patterns, ensuring the determination of a set of factors influencing the volume of mudflow. This approach made it possible to identify the most significant input variables and understand how the model processes data to generate a forecast, identify factors that play a key role in forecasting results, and ensure adequate accuracy and stability of forecasts, taking into account the specifics and complexity of mudflow data. The patterns deduced as a result of the study, reflecting the hidden principles of the subject area under study, and the methods of logical analysis used in the study helped to identify possible causes of the formation of different volumes of carried-out solid deposits. The results obtained can be used to improve monitoring systems and prevent the negative consequences of mudslides.

Keywords: machine learning, neural networks, cluster analysis, associative rules, mudflows, model.

For citation: Lyutikova L.A. Analysis of mudflow characteristics with limited data using machine learning models. *Modeling, Optimization and Information Technology*. 2024;12(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1747 DOI: 10.26102/2310-6018/2024.47.4.029

Введение

Данное исследование было предложено Географическим центром КБНЦ РАН для выявления совокупностей наиболее существенных параметров оценки объема селевых потоков с целью включения этих параметров при формировании нового кадастра.

Основными составляющими селевого кадастра выступают такие параметры, как происхождение селя (категориальная характеристика), тип селя (категориальная характеристика), площадь водосборного бассейна, выраженная в квадратных километрах (количественная величина), средний уклон русла (α, количественный показатель), длина реки в километрах (L, количественное значение), высота истока (количественный показатель), объем наибольшего разового выноса (W, кубические метры, количественная величина) и максимальный объем твердых отложений селя, также измеряемый в кубометрах (W).

Для анализа взаимосвязи между различными переменными был проведен комплексный анализ данных. Была построена корреляционная матрица. Коэффициенты корреляции между переменными в данной матрице варьируются от [-0,36, 0,91]. Высокая корреляция наблюдается только между площадью бассейна и длиной реки – 0,91, остальные данные показывали значения от [-0,36, 0,19].

Для построения прогнозной модели объема твердых отложений селевых потоков были использованы методы машинного обучения, такие как метод опорных векторов (SVM), многопараметрическая регрессия, случайный лес, нейронные сети. Даже после дополнительной обработки данных, включая удаление выбросов, наибольшую точность на тестовом наборе данных продемонстрировала нейронная сеть – 42 %. Модель случайного леса показала точность 29 %, многопараметрическая регрессия – 13 %, а SVM – 35 %.

С учетом специфики предметной области и целей исследования, направленных на выявление совокупности признаков, влияющих на процессы селевого переноса, были дополнительно применены методы выявления ассоциативных правил и метод логического классификатора. Эти подходы позволили определить влияние ключевых факторов на селевые процессы, что существенно углубило понимание механизмов формирования различных объемов выноса твердого материала [2, 3].

Материалы и методы

Кластерный анализ, проведенный для начального анализа данных, разделил данные на три группы, отличающиеся друг от друга по физическим свойствам и типу селевых потоков.

Каждая группа демонстрирует определенные особенности:

Группа 1: Характеризуется крупными бассейнами и малым уклоном, что необычно для селевых явлений. Это может указывать на более медленное и постепенное формирование селей в рамках данной группы.

Группы 0 и 2: Отличаются высотой источников селей и объемом переносимых масс, однако обладают сходным генезисом и типом селей. Вероятно, эти группы обусловлены специфическими географическими факторами, такими как особый рельеф местности или климатические условия [4].

Исследование данных моделями, указанными во введении, не дали хороших результатов, так как связи между данными были слишком сложными для этих моделей. Чтобы решить эту проблему, мы применили метод категоризации числовых данных. [5].

Категоризация данных. Преобразование непрерывных данных в категориальные позволяет применять методы ассоциативного анализа и логической классификации, что обеспечивает выявление скрытых закономерностей и структурных зависимостей в данных. Такой подход улучшает робастность моделей к наличию выбросов и шумов, одновременно облегчая интерпретацию полученных результатов (Таблицы 1, 2). Вместо рассмотрения всего диапазона непрерывных значений, модель работает с ограниченным числом дискретных категорий, что делает сравнительный анализ и выводы более простыми и понятными [6].

Таблица 1 – Диапазон значений для дискретизации
Table 1 – Range of values for discretization

Группа	Площадь бассейна реки ($км^2$) 1	Средний уклон русла реки (%) 2	Длина реки (км) 3	Высота истока (м. абс.) 4
Малый (0)	0 – 12,64	0 – 44,52	0 – 1492,8	0 – 1492,8
Средний (1)	12,64 – 58,45	44,52 – 105,76	1492,80 – 1644,48	1492,80 – 1644,48
Большой (2)	58,45 – +∞	105,76 – +∞	1644,48 – +∞	1644,48 – +∞

Таблица 2 – Диапазон значений для дискретизации
Table 2 – Range of values for discretization

Группа	M1, ($м^3$) 5	M2, ($м^3$) 6
Малый (0)	0 – 8300	0 – 71811,96
Средний (1)	8300 – 38800	71811,96 – 102840,08
Большой (2)	38800 – +∞	102840,08 – +∞

При категоризации переменных использовался двухзначный код, в котором первая цифра обозначала номер поля, а вторая – категорию внутри этого поля.

Нумерация полей следующая: 1 – площадь бассейна реки, 2 – уклон русла реки, 3 – длина реки, 4 – высота истока, 5 – M1 (объем максимального единовременного выноса), 6 – M2 (максимальный объем твердых отложений селя).

Категории внутри каждого поля классифицировались следующим образом: малая – 0, средняя – 1, большая – 2.

Например, большой уклон русла реки будет представлен значением 22.

После категоризации наши данные будут иметь следующие значения:

x_1 *(генезис селя)* $\in \{D, L, L-D\}$;

x_2 *(тип селя)* $\in \{VK, G\,K\}$;

x_3 *(площадь бассейна реки)* $\in \{10,11,12\}$;

x_4 *(уклон русла реки)* $\in \{20,21,22\}$;

x_5 *(длина реки)* $\in \{30,31,32\}$;

x_6 *(высота истока)* $\in \{40,41,42\}$;

x_7 *(объем единовременного выноса)* $\in \{50,51,52\}$;

x_8 *(объем твердых отложений селя)* $\in \{60,61,62\}$.

Теперь задача регрессии, рассмотренная ранее, трансформируется в задачу классификации, так как целевая переменная принимает категориальное значение.

Вместо того чтобы прогнозировать непрерывное значение объема максимального единовременного выноса «M1», наша цель теперь – определить, к какой из трех категорий (50, 51 или 52) принадлежит «M1». После создания модели классификации на основе дерева решений [7, 8] мы достигли следующих результатов (Рисунок 1).

Точность: 1.0

	precision	recall	f1-score	support
60.0	1.00	1.00	1.00	16
61.0	1.00	1.00	1.00	17
62.0	1.00	1.00	1.00	13
accuracy			1.00	46
macro avg	1.00	1.00	1.00	46
weighted avg	1.00	1.00	1.00	46

Рисунок 1 – Результат классификации объема максимального единовременного выноса
Figure 1 – Result of the maximum one-time removal volume classification

Если модель машинного обучения демонстрирует высокую точность при решении задачи, то это свидетельствует о том, что модель находит определенные закономерности и зависимости в данных. Эти зависимости могут быть выявлены и интерпретированы с помощью методов логического анализа.

В работе используются такие методы логического анализа, как построение ассоциативных правил и создание логических классификаторов.

Один из эффективных подходов к обнаружению таких зависимостей – использование метода построения ассоциативных правил, который направлен на поиск регулярных сочетаний элементов в наборах данных. В рамках данного исследования был применен алгоритм FP-Growth, который отличается высокой скоростью обработки больших объемов информации за счет использования специализированной структуры данных – FP-дерева. Алгоритм FP-Growth позволяет эффективно находить наборы элементов, которые часто встречаются совместно, основываясь на критериях поддержки и достоверности [9]. Это делает его предпочтительным выбором перед другими методами, такими как Apriori, благодаря значительной экономии времени и ресурсов. Результаты применения этого подхода представлены в виде наиболее значимых ассоциативных правил в Таблице 3.

Таблица 3 – Самые значимые ассоциативные правила
Table 3 – The most important association rules

№	Antecedents (причина)	Consequents (следствие)
232619	(51, D, 20, VK, 40)	(32, 62, 12)
200538	(51, 32, 20, VK, 12, 40)	(D, 62)
200510	(51, 11, 21)	(61)
230187	(D, 11)	(61)
230216	(D, GK, 10, 30)	(60, 50)

Так, правило в таблице номер 230187 и формулируется следующим образом: «Если генезис селя является дождевым (D) и объем бассейна реки средний (11), то объем выноса твердых веществ будет средним (61)».

При минимальной поддержке 0,2, минимальной достоверностью 0,6.

Результаты анализа выявленных правил позволяют получить представление о взаимосвязях в данных.

Для повышения достоверности полученных результатов путем ассоциативных правил был применен метод анализа данных, основанный на построении логического классификатора. Этот подход, выходящий за рамки стандартных алгоритмов машинного обучения, был разработан для коррекции результатов, полученных с помощью методов машинного обучения.

Применение логического классификатора для анализа исходных данных.

Метод построения логического классификатора, доказательство его свойств, а также применение в качестве корректора результатов нейронной сети подробно изложены в работе [10]. В данной работе метод адаптирован под рассматриваемые данные.

Исследуемые нами данные – это совокупность строк, которых в данных 387, каждая из которых может быть представлена в следующем виде:

$$\&_{j=1}^{m} x_j(y_i) \to P(y_i), i = 1, \dots, l; x_j(y_i) \in \{0,1\}.$$

Тогда функция, описывающая совокупность всех заданных объектов и их признаков, будет следующая:

$$f(X) = \&_{m}^{j=1} \left(\&_{n}^{i=1} x_i \to P(y_j) \right).$$

Для рассматриваемого случая: $f(X) = \&_{387}^{j=1} \left(\&_{7}^{i=1} x_i \to P(y_j) \right)$.

$$x_1 \in \{D, L, L-D\};\ x_2 \in \{VK, G\,K\};\ x_3 \in \{10,11,12\};\ x_4 \in \{20,21,22\};$$

$$x_5 \in \{30,31,32\};\ x_6 \in \{40,41,42\};\ x_7 \in \{50,51,52\}; y_i \in \{60,61,62\}.$$

$$P(60) = \begin{cases} 0\ при\ y_i = 61\ или\ 62 \\ 1\ при\ y_i = 60 \end{cases}; \qquad P(61) = \begin{cases} 0\ при\ y_i = 60\ или\ 62 \\ 1\ при\ y_i = 61 \end{cases};$$

$$P(62) = \begin{cases} 0\ при\ y_i = 60\ или\ 61 \\ 1\ при\ y_i = 62 \end{cases}.$$

Такая функция будет принимать значение «ложь» на наборах $\left(x_1(y_i),\ldots,x_n(y_i),P^{\sigma}(y_1),\ldots,\overline{P^{\sigma}(y_i)},\ldots,P^{\sigma}(y_n)\right)$ там, где есть признаки объекта y_j, *но* отрицается сам объект и «истина» в остальных случаях.

Учитывая большое количество правил (93 237), применялась процедура фильтрации и сокращения, включающая отбор правил с интересующими нас свойствами, такими как присутствие в них категорий 60, 61 или 62. Объединение схожих правил позволило уменьшить общее число правил за счет генерализации.

В результате часть картины полученных правил изображен на Рисунке 2.

```
D ,GK, 12, 50, 60 |D, GK, 11,20,60|D, GK, 10, 30, 60
D, VK, 11, 50, 61| D, GK, 10,51, 61| D, GK, 31, 61
L-D, GK, 12, 32, 52, 62| L-D, VK, 11, 31 ,52, 62
```

Рисунок 2 – Результирующие правила (здесь обозначения: «|» – «V»; «,» – «&»)
Figure 2 – Resulting rules (here the notations: «|» – «V»; «,» – «&»)

Результаты

Полученные результаты демонстрируют существенную коррекцию первоначально выявленных закономерностей, что позволяет уточнить взаимосвязи между характеристиками и выявить новые закономерности, не обнаруженные на этапе первичного анализа ассоциативных правил.

Например, для категории максимального объема твердых отложений (62) логический классификатор указывает на преобладание генезиса типа «L-D», в то время как ассоциативные правила выявили преимущественно генезис типа «D», представленный с меньшей частотой в данных, чем «L-D». Аналогичные расхождения наблюдаются и в распределении типов селей: равномерное соотношение типов «GK» и «VK», выявленное логическим классификатором, дополняет данные, полученные с помощью ассоциативных правил.

В случае малых объемов твердых отложений (60) результаты обоих методов согласуются, указывая на преобладание селей дождевого генезиса и грязекаменных селей с низкой интенсивностью.

Однако для средних объемов (61) логический классификатор выявил преобладание дождевого генезиса и равномерное распределение между грязекаменным и водокаменным типами, что не может быть извлечено из исходных ассоциативных правил. Кроме того, логический классификатор выявил значительную долю средних и крупных водосборных бассейнов (категории «11» и «12») и средние значения объемов стока (категория «51»), что также отсутствует в результатах анализа ассоциативных правил.

Таким образом, применение логического классификатора привело к существенному уточнению и коррекции закономерностей, выявленных с помощью ассоциативных правил. Это позволило выявить комплекс факторов, определяющих объем селевого потока.

Обсуждение

Для более полного понимания процессов образования и эволюции селевых потоков необходимо расширить набор данных, включив в него параметры, характеризующие количество осадков, состояние растительного покрова, антропогенное воздействие, геоморфологические особенности и геологическое строение водосборного бассейна. Несмотря на ограниченность исходных данных в настоящем исследовании,

предложенные методы анализа позволили выявить ключевые закономерности и взаимосвязи, определяющие образование и динамику селей. Полученные результаты указывают на зависимости объемов твердых отложений от комплекса факторов. Важно отметить, что традиционные методы построения ассоциативных правил могут не обеспечивать исчерпывающего анализа предметной области, в то время как комплексный логический анализ данных, использованный в данной работе, демонстрирует более высокую эффективность в выявлении причинно-следственных связей и позволяет получить более достоверные выводы о факторах, влияющих на развитие селевых потоков. Дальнейшие исследования с расширенным набором данных позволят уточнить полученные результаты и разработать более точные и надежные модели прогнозирования селей.

Заключение

Селевые потоки представляют собой опасное природное явление, наносящее значительный ущерб инфраструктуре, населенным пунктам и сельскому хозяйству. В условиях изменения климата и растущей частоты экстремальных погодных явлений изучение факторов, влияющих на формирование и характеристики селей, приобретает особую актуальность. Понимание этих факторов критически важно для разработки эффективных стратегий предотвращения и снижения рисков, связанных с селевыми потоками.

Однако анализ селевых процессов осложняется неполнотой, неточностью и плохой структурированностью имеющихся данных. Часто отсутствуют систематические наблюдения за ключевыми параметрами, такими как интенсивность осадков, состояние почвы, характеристики растительного покрова, геоморфологические особенности рельефа и антропогенное воздействие на водосборные бассейны. Эта неполнота данных существенно затрудняет выявление причинно-следственных связей и построение точных прогнозных моделей.

В настоящем исследовании, несмотря на эти ограничения, был проведен анализ доступных данных, в результате которого удалось выделить ряд правил, описывающих ключевые закономерности и взаимосвязи, определяющие формирование и характеристики селевых потоков. Эти правила не только описывают наблюдаемые корреляции, но и позволяют глубже понять механизмы образования селей, выявляя наиболее значимые факторы риска. Полученные результаты служат основой для дальнейшего углубленного изучения предметной области и способствуют более эффективному поиску решений по минимизации рисков.

Данное исследование подтверждает важность применения интеллектуальных аналитических систем для эффективного управления рисками и снижения негативного воздействия опасных природных явлений, таких как селевые потоки. Дальнейшие исследования предполагают расширение набора данных и уточнение разработанных моделей, что позволит повысить точность прогнозов и создать более надежные инструменты для принятия управленческих решений.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Caiafa C.F., et al. Decomposition Methods for Machine Learning with Small, Incomplete or Noisy Datasets. *Applied Sciences.* 2020;10(23). https://doi.org/10.3390/app10238481
2. Kainthura P., Sharma N. Hybrid machine learning approach for landslide prediction, Uttarakhand, India. *Scientific Reports*. 2022;12(1). https://doi.org/10.1038/s41598-022-22814-9

3. Hadi F.A.A., et al. Machine learning techniques for flood forecasting. *Journal of Hydroinformatics*. 2024;26(4):779–799. https://doi.org/10.2166/hydro.2024.208
4. Lombardo L., Mai P.M. Presenting logistic regression-based landslide susceptibility results. *Engineering Geology*. 2018;244:14–24. https://doi.org/10.1016/j.enggeo.2018.07.019
5. Rahmati O., Kornejady A., Samadi M., et al. PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches. *Science of The Total Environment.* 2019;664:296–311. https://doi.org/10.1016/j.scitotenv.2019.02.017
6. Кондратьева Н.В., Аджиев А.Х., Беккиев М.Ю., Гедуев (Гяургиева) М.М., Перов В.Ф., Разумов В.В., Сейнова И.Б., Хучунаева Л.В. *Кадастр селевой опасности юга европейской части России*. Москва; Нальчик: Феория; 2015. 148 с.
7. Кюль Е.В., Езаов А.К., Канкулова Л.И. Теоретические основы геоэкологического мониторинга горных геосистем. *Устойчивое развитие горных территорий.* 2019;11(1):36–43. https://doi.org/10.21177/1998-4502-2019-11-1-36-43
Kyul E.V., Ezaov A.K., Kankulova A.K. The theoretical basis of geo-environmental monitoring of the mountain geosystems. *Sustainable Development of Mountain Territories.* 2019;11(1):36–43. (In Russ.). https://doi.org/10.21177/1998-4502-2019-11-1-36-43
8. Радеев Н.А. Предсказание лавинной опасности методами машинного обучения. *Вестник НГУ. Серия: Информационные технологии.* 2021;19(2):92–101. https://doi.org/10.25205/1818-7900-2021-19-2-92-101
Radeev N.A. Avalanches Forecasting Using Machine Learning Methods. *Vestnik NSU. Series: Information Technologies.* 2021;19(2):92–101. (In Russ.). https://doi.org/10.25205/1818-7900-2021-19-2-92-101
9. Haoxiang Wang S.S. Big Data Analysis and Perturbation using Data Mining Algorithm. *Journal of Soft Computing Paradigm*. 2021;3(1):19–28.
10. Lyutikova L.A. Construction of a Logical-Algebraic Corrector to Increase the Adaptive Properties of the ΣΠ-Neuron. *Journal of Mathematical Sciences.* 2021;253(4):539–546. https://doi.org/10.1007/s10958-021-05251-3

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Лютикова Лариса Адольфовна, кандидат физико-математических наук, заведующий отделом нейроинформатики и машинного обучения, Институт прикладной математики и автоматизации Кабардино-Балкарского научного центра Российской академии наук, Нальчик, Российская Федерация.
e-mail: lylarisa@yandex.ru
ORCID: 0000-0003-4941-7854

Larisa A. Lyutikova, Ph.D. of Physico-Mathematical Sciences, Head of the Department of Neuroinformatics and Machine Learning, Institute of Applied Mathematics and Automation of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, Nalchik, the Russian Federation.

Статья поступила в редакцию 17.11.2024; одобрена после рецензирования 10.12.2024; принята к публикации 12.12.2024.

The article was submitted 17.11.2024; approved after reviewing 10.12.2024; accepted for publication 12.12.2024.