ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 519.7 Научная статья
DOI: 10.35330/1991-6639-2024-26-6-139-145
EDN: FIUPQE

Применение метода машинного обучения для анализа неполных данных

Л. А. Лютикова

Институт прикладной математики и автоматизации –
филиал Кабардино-Балкарского научного центра Российской академии наук
360000, Россия, г. Нальчик, ул. Шортанова, 89 А

Аннотация. В данной работе представлен комплексный подход к анализу неполных и неточных данных, проиллюстрированный на примере прогнозирования селей. Целью исследования является демонстрация того, как сочетание различных методов позволяет не только получать адекватные прогнозы, но и глубоко понимать логику принятия решений моделью, выявляя ключевые факторы, влияющие на прогноз. Ключевым моментом работы является использование категоризации числовых данных для повышения устойчивости моделей к выбросам и шуму, а также для учета нелинейных зависимостей. Комплексный подход основан на сочетании ассоциативного анализа данных и построения логического классификатора, который выступает в роли интерпретатора полученных решений. Такое сочетание позволило выявлять критически важные входные признаки и понимать, как модель использует информацию для формирования прогноза, выделять факторы, оказывающие наибольшее влияние на результат прогнозирования, обеспечивать точность и устойчивость прогнозов с учетом специфики и сложности данных о селевых потоках. Полученные в ходе исследования правила, являющиеся ключевыми принципами изучаемой области, способствуют более глубокому пониманию природы селей.

Ключевые слова: машинное обучение, нейронные сети, кластерный анализ, ассоциативные правила

Поступила 15.10.2024, одобрена после рецензирования 04.12.2024, принята к публикации 10.12.2024

Для цитирования. Лютикова Л. А. Применение метода машинного обучения для анализа неполных данных // Известия Кабардино-Балкарского научного центра РАН. 2024. Т. 26. № 6. С. 139–145. DOI: 10.35330/1991-6639-2024-26-6-139-145

MSC: 68T027 Original article

Application of machine learning method to analyse incomplete data

L.A. Lyutikova

Institute of Applied Mathematics and Automation –
branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360000, Russia, Nalchik, 89 A Shortanov street

Abstract. This paper presents an integrated approach to the analysis of incomplete and inaccurate data, illustrated by the example of mudflow forecasting. The aim of the study is to demonstrate how a combination of different methods allows not only to obtain adequate forecasts, but also to deeply understand the logic of decision-making by the model, identifying the key factors influencing the forecast. The key point of the work is the use of categorization of numerical data to increase the stability of models to outliers and noise, as well as to take into account nonlinear dependencies. The integrated approach

is based on a combination of associative data analysis and the construction of a logical classifier, which acts as an interpreter of the obtained decisions. This combination made it possible to identify critical input features and understand how the model uses information to form a forecast, identify factors that have the greatest impact on the forecast result, ensure the accuracy and stability of forecasts taking into account the specificity and complexity of mudflow data. The rules obtained during the study, which are the key principles of the studied area, contribute to a deeper understanding of the nature of mudflows.

Keywords: machine learning, neural networks, cluster analysis, associative rules

Submitted 15.10.2024, approved after reviewing 04.12.2024, accepted for publication 10.12.2024

For citation. Lyutikova L.A. Application of machine learning method to analyse incomplete data. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2024. Vol. 26. No. 6. Pp. 139–145. DOI: 10.35330/1991-6639-2024-26-6-139-145

Введение

Несмотря на разнообразие подходов и методов обработки данных логический анализ позволяет выявить причинно-следственные связи и зависимости между различными переменными в данных, найти скрытые закономерности и тенденции в данных, которые могут быть незаметны при другом анализе.

Объединение различных методов логического анализа позволяет учесть больше аспектов данных, снизить влияние шума и выбросов и получить более точные прогнозы. Различные методы, используемые в комплексном анализе, могут выявлять как линейные, так и нелинейные взаимосвязи между переменными, что важно для сложных систем [1].

Исследование применяет машинное обучение для анализа характеристик селевых потоков на Северном Кавказе, используя имеющийся кадастр селей [2] с данными о генезисе, типе селя, площади бассейна, уклоне русла, длине реки, высоте истока и объеме выноса (табл. 1).

Таблица 1. Характеристики селевых потоков

Table 1. Mudflow characteristics

№	Название водотока	Генезис селя	Тип селя	Площадь бассейна, S, км^2	Средний уклон русла, α, ‰	Длина реки, L, км	Высота истока, H, м абс.	Объем максимального единовременного выноса, W, м^3	Максимальный объем твердых отложений селя, W, м^3 (аналитическим методом)	Повторяемость 1 раз в n лет/ даты схода
1	Кичмалка	Д*	ВК	152,7	30	36	1909	10000	147240	1–2/ 31.05.93
2	Рхыкол	Д	ВК	9,8	52	10	1440	100000	81946	1–2/–
3	Кызылкол	Д	ВК	14,1	220	4,5	1520	50000	52140	1–5/–
4	Тазакол	Д	ВК	11,3	150	5	1525	50000	43200	1–5/–
5	Лахран	Д	ВК	22,2	102	5	1629	20000	35712	1–5/–
6	Большой Лахран	Д	ВК	21,8	190	6	1642	50000	53400	1–5/–

Примечание: Д – дождевой, ВК – водокаменный

Работа направлена на демонстрацию способности машинного обучения выявлять закономерности и создавать эффективные модели для классификации и прогнозирования селей. Анализ позволит углубить понимание процессов формирования селей, определить ключевые факторы риска и в конечном счете создать прогнозные модели для оценки последствий и управления селеопасными территориями. Полученные результаты имеют практическую ценность для инженерной и научной деятельности [3, 4].

Анализ данных

Задача состоит в разработке модели прогнозирования и классификации селей, основанной на логическом анализе данных. Цель анализа – выявление общих правил, порождающих эти зависимости, отбор наиболее информативных переменных и классификация типов селей.

Анализ кластеризации выявил слабую структуру в данных, разделив их на три группы, характеризующиеся различными физическими свойствами и типами селевых потоков.

Эти группы демонстрируют некоторые интересные закономерности.

Группа 1 отличается большими бассейнами и низким уклоном, что нетипично для селевых потоков. Это может свидетельствовать о более медленных и постепенных процессах формирования селей в этой группе.

Группы 0 и 2 различаются по высоте источника селей и объему селевых масс, но имеют схожий генезис и тип селей. Возможно, эти группы связаны с определенными географическими условиями, например, с особым рельефом или климатом.

Модель многопараметрической регрессии, построенная для прогнозирования целевой переменной, оказалась неэффективной. Высокое значение MSE (92477727488,7331) свидетельствует о значительных ошибках прогнозирования, а низкое значение R-квадрата (0,1235) указывает на крайне низкую объясняющую способность модели [5, 6].

Линейная модель неадекватно описывает нелинейные взаимосвязи между предикторами и целевой переменной, что является основной причиной неудовлетворительных результатов. В качестве меры по преодолению этой проблемы была применена категоризация числовых данных [2, 7, 8].

Категориальные данные

Преобразование непрерывных данных в категориальные позволяет учитывать нелинейные зависимости путем разбиения данных на интервалы, в которых взаимосвязи могут быть аппроксимированы линейными. Такой подход повышает устойчивость моделей к выбросам и шуму, упрощая интерпретацию результатов (табл. 2). Вместо анализа непрерывного спектра значений модель оперирует более компактным набором дискретных категорий, что упрощает сравнение и анализ.

Таблица 2. Диапазон значений для дискретизации

Table 2. Range of values for discretization

Группа	Площадь бассейна, S, км^2	Средний уклон русла, α, ‰	Длина реки, L, км	Высота истока, H, м абс.	M1, м^3	M2, м^3
Малый (0)	0 – 12,64	0 – 44,52	0 – 1492,8	0 – 1492,8	0 – 8300	0 – 71811,96
Средний (1)	12,64 – 58,45	44,52 – 105,76	1492,80 – 1644,48	1492,80 – 1644,48	8300 – 38800	71811,96 – 102840,08
Большой (2)	58,45 – +∞	105,76 – +∞	1644,48 –+∞	1644,48 – +∞	38800 – +∞	102840,08 –+∞

Теперь задача регрессии, описанная в предыдущем разделе, сводится к задаче классификации, поскольку целевая переменная становится категориальной. И задача может быть описана следующим образом [2, 9]:

$$X = \{x_1, x_2, ..., x_n\},\ \ x_i \in \{0, 1, ..., k_i - 1\}.$$

В нашей системе входными данными будут являться n=6, а выходными m=387:

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & \dots & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & \dots & x_n(y_2) \\ \dots & \dots & \dots & \dots \\ x_1(y_m) & x_2(y_m) & \dots & x_n(y_m) \end{pmatrix} \rightarrow \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix} \quad (1)$$

Методы решения

Вместо того, чтобы предсказывать непрерывное значение объема максимального единовременного выноса «M1», мы теперь предсказываем, к какой из трех категорий (50, 51 или 52) относится «M1». После построения модели классификации с использованием дерева решений мы получили впечатляющие результаты, представленные в табл. 3.

Таблица 3. Результат классификации объема максимального единовременного выноса

Table 3. Result of classification of the maximum one-time removal volume

Метрика / Объем выноса ($м^3$)	Recall (чувствительность)	Precision (точность)	Accuracy (правильность)	F1-мера
Малый (50)	1	1	1	1
Средний (51)	1	1	1	1
Большой (52)	1	1	1	1

Логические методы анализа – это построение ассоциативных правил и логического классификатора [10, 11].

Метод построения ассоциативных правил – это метод обнаружения скрытых взаимосвязей и закономерностей в больших объемах данных. Он фокусируется на поиске наборов элементов, которые часто встречаются вместе в данных. Этот метод обычно используется для анализа данных о транзакциях, где каждая запись представляет собой отдельную транзакцию.

В данной работе использовался алгоритм FP-Growth – это эффективный алгоритм для поиска ассоциативных правил в больших объемах данных. Он основан на построении специального дерева (FP-дерева), которое содержит часто встречающиеся элементы и их взаимосвязи. Алгоритм FP-Growth обходит это дерево, а не все данные целиком, что значительно ускоряет процесс поиска правил с высокой поддержкой. В итоге он находит все правила, удовлетворяющие заданным порогам поддержки и достоверности, значительно быстрее и эффективнее, чем другие алгоритмы, такие как Apriori [12].

После работы алгоритма самые важные ассоциативные правила представлены в табл. 4.

Таблица 4. Самые важные ассоциативные правила

Table 4. The most important association rules

№	Antecedents (причина)	Consequents (следствие)
232619	(D, GK, 11, 62)	(40, 32, 51, 22)
200538	(40, 11, 62)	(D, 32, 22, GK)
200510	(40, GK, 11, 62)	(D, 32, 22)
230187	(40, 32, 61, 51, 11, VK)	(C-D, 20)
230216	(40, 32, 61, 11, VK)	(C-D, 51, 20)

Анализ пяти выявленных закономерностей, используя методы булевой алгебры [9], показывает, что грязекаменные сели, даже при средней площади бассейна, отличаются значительным объемом максимального единовременного выноса и высоким содержанием твердых отложений.

Построение логического классификатора

Каждая строка (1) является зависимостью и может быть представлена следующим правилом:

$$\&_{j=1}^{m} x_j\left(y_i,\right) \to y_i. \tag{2}$$

Эти правила описывают зависимость конкретного выноса твердых отложений от остальных параметров данного селевого потока [8].

Представим их в следующей дизъюнктивной форме:

$$\bigvee_{j=1}^{m} \overline{x_j(y_i)} \vee y_i, \tag{3}$$

а зависимость всех исследуемых селевых потоков от своих параметров как

$$f(x,y) = \&_{i=1}^{n} \bigvee_{j=1}^{m} \overline{x_j(y_i)} \vee y_i. \tag{4}$$

В нашем случае $f(X) = \&_{j=1}^{387}\left(\&_{i=1}^{7} x_i \to P(y_j)\right)$

$$x_1 \in \{D,L,C-D,\}; \quad x_2 \in \{VK,GK\}; \quad x_3 \in \{10,,11,12\}; x_4 \in \{20,21,22\};$$
$$x_5 \in \{30,31,32\}; x_6 \in \{40,41,42\}; x_7 \in \{50,51,52\}.$$

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

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

Из огромного количества полученных правил (93 237) были отобраны наиболее значимые, в основном правила, содержащие категории 60, 61 или 62. Для упрощения и обобщения информации схожие правила были объединены, что сократило общее количество правил без потери ключевой информации [9].

В результате часть картины полученных правил изображена на рис. 1.

```
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
```

Рис. 1. *Результирующие правила (здесь обозначения: «|»-«∨»; «,»- «&»)*

Fig. 1. *Resulting rules (here the notations are: «|»-« »; «,»- «&»)*

Данные можно проинтерпретировать следующим образом: сели с малым объемом твердых отложений (60) – это преимущественно небольшие грязекаменные дождевые потоки с низкой интенсивностью. Сели со средним объемом (61) характеризуются преобладанием дождевого генезиса, но включают как грязекаменные, так и водокаменные сели преимущественно со средними и крупными бассейнами. Сели с большим объемом (62) чаще всего вызваны ливневыми дождями (L-D) и связаны с крупными бассейнами и руслами.

Заключение

В результате можно утверждать, что логический анализ данных позволяет выделить набор фундаментальных правил, которые объясняют основные закономерности и взаимосвязи в данных. Эти правила являются основой исследуемой области, способствуют более глубокому пониманию ее природы и оптимизируют поиск решений.

Результаты исследования показывают, что даже неполные и неточные данные могут стать основой для создания эффективных моделей прогнозирования, что дает возможности в области управления рисками и повышения безопасности в зонах, подверженных селевым потокам. Это подчеркивает потенциал интеллектуальных аналитических систем для эффективного управления рисками и минимизации негативных последствий селевых процессов.

СПИСОК ЛИТЕРАТУРЫ

1. *Кондратьева Н. В.* Предварительная оценка максимального объема твердых отложений селя методами математической статистики для Центрального Кавказа // Современные проблемы науки и образования. 2014. № 4. С. 50–56. URL: http://www.science-education.ru/118-13897

2. *Кондратьева Н. В., Аджиев А. Х., Беккиев М. Ю. и др.* Кадастр селевой опасности Юга европейской части России. М., Нальчик: Феория, 2015. 148 с.

3. *Caiafa C. F., Jordi Solé-Casals J.S.-C., Marti-Puig P. et al.* Decomposition methods for machine learning with small, incomplete or noisy datasets // Applied Sciences. 2020. Vol. 10. No. 23. P. 8481. DOI: 10.3390/APP10238481

4. *Kainthura P., Sharma N.* Hybrid machine learning approach for landslide prediction, Uttarakhand, India // Scientific reports. 2022. Vol. 12. No. 1. P. 20101. DOI: 10.1038/s41598-022-22814-9

5. *Hadi F. A. A., Sidek L. M., Salih G. H. A. et al.* Machine learning techniques for flood forecasting // Journal of Hydroinformatics. 2024. Vol. 26. No. 4. Pp. 779–799. DOI: 10.2166/hydro.2024.208

6. *Lombardo L., Mai P. M.* Presenting logistic regression-based landslide susceptibility results // Engineering Geology. 2018. Vol. 244. Pp. 14–24. DOI: 10.1016/j.enggeo.2018.07.019

7. *Rahmati O., Kornejady A., Samadi M. et al.* PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches // The Science of the total environment. 2019. Vol. 664. Pp. 296–311. DOI: 10.1016/j.scitotenv.2019.02.017

8. *Кюль Е. В., Езаов А. К., Канкулова Л. И.* Теоретические основы геоэкологического мониторинга горных геосистем // Устойчивое развитие горных территорий. 2019. Т. 11. № 1. С. 36–43. DOI: 10.21177/1998-4502-2019-11-1-36-43

9. *Lyutikova L. A.* Methods for Improving the Efficiency of Neural Network Decision-Making // Advances in Automation IV. RusAutoCon 2022. Lecture Notes in Electrical Engineering – 2023. Vol. 986. Pp. 294–303. DOI: 10.1007/978-3-031-22311-2_29

10. *Радеев Н. А.* Предсказание лавинной опасности методами машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19. № 2. С. 92–101. DOI: 10.25205/1818-7900-2021-19-2-92-101

11. *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5–68.

12. *Флах П.* Машинное обучение: наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015.

REFERENCES

1. Kondrat'eva N.V. Preliminary assessment of the maximum volume of solid mudflow deposits using mathematical statistics methods for the Central Caucasus. *Sovremennye problemy nauki i obrazovaniya* [Modern problems of science and education]. 2014. No. 4. Pp. 50–56. URL: http://www.science-education.ru/118-13897. (In Russian)

2. Kondrat'eva N.V., Adzhiev A.Kh., Bekkiev M.Yu. et al. *Kadastr selevoy opasnosti Yuga evropeyskoy chasti Rossii* [Mudflow hazard cadastre of the South of the European part of Russia]. M., Nal'chik: Feoriya, 2015. 148 p. (In Russian)

3. Caiafa C.F., ,Jordi Solé-Casals J.S.-C., Marti-Puig P. et al. Decomposition methods for machine learning with small, incomplete or noisy datasets. *Applied Sciences*. 2020. Vol. 10. No. 23. P. 8481. DOI: 10.3390/APP10238481

4. Kainthura P., Sharma N. Hybrid machine learning approach for landslide prediction, Uttarakhand, India. *Scientific reports*. 2022. Vol. 12. No. 1. P. 20101. DOI: 10.1038/s41598-022-22814-9

5. Hadi F.A.A., Sidek L.M., Salih G.H.A. et al. Machine learning techniques for flood forecasting. *Journal of Hydroinformatics*. 2024. Vol. 26. No. 4. Pp. 779–799. DOI: 10.2166/hydro.2024.208

6. Lombardo L., Mai P.M. Presenting logistic regression-based landslide susceptibility results. *Engineering Geology*. 2018. Vol. 244. Pp. 14–24. DOI: 10.1016/j.enggeo.2018.07.019

7. Rahmati O., Kornejady A., Samadi M. et al. PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches. *The Science of the Total Environment.* 2019. Vol. 664. Pp. 296–311. DOI: 10.1016/j.scitotenv.2019.02.017

8. Kyul' E.V., Ezaov A.K., Kankulova L.I. Theoretical foundations of geoecological monitoring of mountain ecosystems. *Ustoychivoe razvitie gornykh territoriy* [Sustainable development of mountain areas]. 2019. Vol. 11. No 1. Pp. 36–43. DOI: 10.21177/1998-4502-2019-11-1-36-43. (In Russian)

9. Lyutikova L.A. Methods for Improving the Efficiency of Neural Network Decision-Making. *Advances in Automation IV. RusAutoCon 2022. Lecture Notes in Electrical Engineering*. 2023. Vol. 986. Pp. 294–303. DOI: 10.1007/978-3-031-22311-2_29

10. Radeev N.A. Predicting Avalanche Hazard Using Machine Learning Methods. *Vestnik NGU. Seriya: Informacionnye tekhnologii* [Bulletin of NSU. Series: Information technology]. 2021. Vol. 19, No 2. Pp. 92–101. DOI: 10.25205/1818-7900-2021-19-2-92-101. (In Russian)

11. Zhuravlyov Yu.I. On an algebraic approach to solving recognition or classification problems. *Problemy kibernetiki* [Problems of cybernetics]. 1978. Vol. 33. Pp. 5–68. (In Russian)

12. Flakh P. *Mashinnoe obuchenie: nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh* [Machine Learning: The Art and Science of Algorithms that Make Sense of Data]. Moscow: DMK Press, 2015. (In Russian)

Финансирование. Исследование проведено без спонсорской поддержки.

Funding. The study was performed without external funding.

Информация об авторе

Лютикова Лариса Адольфовна, канд. ф.-м. наук, зав. отделом нейроинформатики и машинного обучения, Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, г. Нальчик, ул. Шортанова, 89 А;

lylarisa@yandex.ru, ORCID: https://orcid.org/0000-0002-5819-9396, SPIN-код: 1679-7460

Information about the author

Larisa A. Lyutikova, Candidate of Physical and Mathematical Sciences, Head of the Department of Neural Networks and Machine Learning, Institute of Applied Mathematics and Automation – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 89 A Shortanov street;

lylarisa@yandex.ru, ORCID: https://orcid.org/0000-0002-5819-9396, SPIN-code: 1679-7460