

Модель сигма-пи нейронной сети для кластеризации данных

Р. А. Жилов

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

Аннотация. Селевые потоки представляют собой один из наиболее разрушительных геологических процессов, сложность прогнозирования которых обусловлена их многофакторной природой и сильными нелинейными зависимостями между определяющими факторами. Традиционные методы моделирования ограничены в интерпретируемости и способности учитывать взаимодействия признаков, что требует разработки новых подходов.

Цель исследования. Разработка и апробация архитектуры сигма-пи нейронной сети для кластеризации селевых потоков с учетом морфометрических и генетических характеристик, а также выявление ключевых факторов и их комбинаций, определяющих формирование различных типов селей.

Материалы и методы. В качестве исходных данных использовались кадастровые сведения о селевых процессах юга европейской части России. Для анализа применялась сигма-пи нейронная сеть, способная учитывать как линейные признаки, так и их взаимодействия второго порядка. Для выбора числа кластеров использовался коэффициент силуэта. Проведено сравнение с результатами кластеризации, полученными методом самоорганизующихся карт Кохонена (SOM).

Результаты. Модель выделила три устойчивых кластера, соответствующих грязевым, каменным и грязекаменным типам селей. Анализ значимости признаков показал, что наибольший вклад в формирование кластеров вносят площадь бассейна, уклон русла, максимальный объем отложений и их попарные комбинации. Сравнение с SOM подтвердило более высокую интерпретируемость предложенной модели и ее способность выявлять скрытые нелинейные зависимости.

Выводы. Применение сигма-пи нейронных сетей позволяет не только повысить качество кластеризации селевых потоков, но и обеспечить интерпретируемость результатов за счет анализа значимости признаков и их комбинаций. Такой подход перспективен для инженерной геологии и может быть использован в системах геоэкологического мониторинга и прогнозирования опасных процессов.

Ключевые слова: селевые потоки, кластеризация, сигма-пи нейронная сеть, машинное обучение, интерпретируемость, нелинейные зависимости, инженерная геология, геоэкологический мониторинг

Поступила 14.07.2025, одобрена после рецензирования 26.08.2025, принята к публикации 25.09.2025

Для цитирования. Жилов Р. А. Модель сигма-пи нейронной сети для кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2025. Т. 27. № 5. С. 34–42. DOI: 10.35330/1991-6639-2025-27-5-34-42

MSC: 68T09

Original article

Sigma-pi neural network model for data clustering

R.A. Zhilov

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

Abstract. Mudflows are some of the most destructive geological phenomena, and their prediction is challenging due to their complexity and the strong nonlinear relationships between the various factors that contribute to their formation. Traditional modeling methods have limitations in their ability to interpret and account for the complex interactions between different factors, and this lead to the need for the development of more advanced approaches.

Aim. The study aims to develop and test a sigma-pi neural network architecture for mudflow clustering based on morphometric and genetic characteristics as well as to identify the key factors and their combinations that contribute to the formation of different mudflow types.

Materials and methods. Cadastral data on mudflows in the southern European part of Russia is used as the initial data. A sigma-pi neural network capable of accounting for both linear features and their second-order interactions is employed for analysis. A silhouette coefficient is used to determine the number of clusters. The results are compared with those obtained using Kohonen's self-organizing maps (SOM).

Results. The model identified three stable clusters corresponding to mud, rock, and mud-rock types of mudflows. Analysis of the significance of features has revealed that the basin area, channel slope, and maximum sediment volume make the greatest contributions to cluster formation, as well as their various pairwise combinations. Comparison with the SOM (self-organizing map) confirmed the improved interpretability of the proposed model and its ability to identify hidden, nonlinear relationships.

Conclusions. The use of sigma-pi neural networks not only improves the accuracy of mudflow clustering, but also ensures the interpretability of the results by analyzing the significance of features and their combinations. This approach is promising for engineering geology and can be used in geoeological monitoring systems and forecasting of hazardous processes.

Keywords: mudflows, clustering, sigma-pi neural network, machine learning, interpretability, nonlinear dependencies, engineering geology, geo-environmental monitoring

Submitted on 14.07.2025,

approved after reviewing on 26.08.2025,

accepted for publication on 25.09.2025

For citation. Zhilov R.A. Sigma-pi neural network model for data clustering. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2025. Vol. 27. No. 5. Pp. 34–42. DOI: 10.35330/1991-6639-2025-27-5-34-42

ВВЕДЕНИЕ

Селевые потоки относятся к числу наиболее разрушительных и трудно прогнозируемых геологических явлений. Их формирование происходит стремительно, а воздействие сопровождается высокой энергией, масштабными разрушениями и угрозой для жизни людей, инфраструктуры и природных комплексов горных и предгорных территорий. Сложность изучения этих процессов связана с их многофакторной природой: на динамику потоков одновременно влияют геологические условия, морфометрия водосборных бассейнов, гидрометеорологические факторы, а также сложное взаимодействие твердой и жидкой составляющих. Вследствие этого наблюдается значительная изменчивость селевых потоков по морфологическим и реологическим параметрам, что затрудняет создание универсальных классификационных и прогностических моделей [1].

Для анализа селевых процессов на протяжении последних десятилетий применяются различные типы моделей. Физико-математические (реологические) модели позволяют описывать движение и трансформацию селевой массы с учетом ее вязкости, плотности и гранулометрического состава. Несмотря на их физическую обоснованность, такие модели требуют большого количества исходных параметров, которые сложно измерить в полевых условиях, и потому их использование ограничено. Статистические подходы, напротив, базируются на выявлении эмпирических зависимостей и отличаются относительной простотой и интерпретируемостью, но плохо справляются с описанием нелинейных эффектов и ситуаций с перекрывающимися классами, что характерно для природных потоков [2, 3].

Развитие машинного обучения открыло новые возможности в исследовании сложных геопроцессов. Методы этой группы позволяют работать с многомерными наборами данных, выявлять скрытые закономерности и обеспечивать высокую предсказательную точность при решении задач классификации и кластеризации [4, 5]. При этом в задачах анализа природных систем часто используются алгоритмы кластеризации на основе нейронных сетей – такие как карты Кохонена (SOM), автоэнкодеры и их вариации, гибридные модели с элементами вероятностных подходов. Эти методы показали высокую эффективность в выделении скрытых структур в данных, но в то же время страдают от ограниченной интерпретируемости, что снижает их применимость в инженерной геологии, где важно не только получить прогноз, но и объяснить ключевые факторы его формирования.

В этом контексте особый интерес представляют архитектуры, способные сочетать возможности нейросетевых моделей с большей прозрачностью анализа. Одним из перспективных направлений являются сигма-пи нейронные сети, которые развивают классическую модель, вводя мультипликативные связи между признаками. Такая архитектура позволяет учитывать как линейные зависимости, так и нелинейные взаимодействия второго и более высоких порядков, что критически важно для изучения сложных природных процессов. Применение сигма-пи сетей в задачах кластеризации селевых потоков открывает возможность более точного выявления типов селей и одновременно обеспечивает интерпретируемость получаемых результатов.

В настоящей работе исследуется применение сигма-пи нейронной сети для кластеризации селевых потоков на основе их морфометрических и генетических характеристик. Отдельное внимание уделено анализу важности факторов и их комбинаций, что позволяет выделить ключевые детерминанты формирования различных типов селевых процессов. Результаты исследования вносят вклад в понимание закономерностей возникновения селей и могут найти практическое применение в прогнозировании, инженерной защите и мониторинге опасных процессов.

АЛГОРИТМ КЛАСТЕРИЗАЦИИ ДАННЫХ СИГМА-ПИ НЕЙРОННОЙ СЕТЬЮ

Классические многослойные персептроны (MLP) выполняют суммативную агрегацию входных признаков: каждый нейрон скрытого слоя вычисляет линейную комбинацию входных переменных с соответствующими весами, после чего результат проходит через нелинейную функцию активации. Такой подход хорошо работает при наличии сложных нелинейных зависимостей, однако он не всегда позволяет явно учитывать взаимодействия между признаками, особенно когда речь идет о геологических процессах, где комбинация факторов зачастую важнее их индивидуального вклада.

Сигма-пи нейронные сети (σ - π сети) расширяют классическую архитектуру, добавляя к суммативной (σ) части нейрона мультипликативную (π) часть. Это позволяет учитывать не только линейные эффекты отдельных признаков, но и их попарные, тройные и более сложные взаимодействия. Таким образом, нейрон данной архитектуры фактически моделирует не только сумму взвешенных признаков, но и произведения переменных, что делает модель особенно полезной в задачах, где решающую роль играют нелинейные комбинации факторов.

Подготовка данных

Для обучения модели используется таблица признаков, описывающих каждый селевой поток. В нашем случае набор данных включает семь переменных, среди которых могут быть: количество осадков, уклон склона, влажность почвы, площадь водосборного бассейна и другие морфометрические или гидрометеорологические характеристики.

Поскольку значения признаков могут отличаться на порядки, перед подачей в сеть проводится стандартизация. Каждый параметр преобразуется по формуле

$$x' = \frac{x - \mu}{\sigma},$$

где μ – среднее значение признака, а σ – его стандартное отклонение. Это необходимо для того, чтобы параметры с большими абсолютными значениями (например, площадь бассейна) не начали доминировать над другими признаками и не искажали процесс обучения.

Определение числа кластеров

Поскольку заранее не известно, сколько типов селевых потоков можно выделить в данных, используется критерий естественности кластеризации. Одним из наиболее распространенных является коэффициент силуэта. Он показывает, насколько хорошо объект принадлежит своему кластеру и насколько он отделен от соседних. Значения, близкие к 1, указывают на четкую кластеризацию, близкие к 0 — на неопределенность. Это позволяет подобрать оптимальное число кластеров без привлечения априорной информации.

Архитектура нейрона

Каждый кластер в сети соответствует отдельному нейрону. Такой нейрон состоит из двух частей:

- **σ-часть** вычисляет линейную комбинацию признаков аналогично классическому персептрону:

$$z_{\sigma} = \sum_i w_i x_i,$$

где w_i – веса, а x_i – значения признаков.

- **π-часть** формирует попарные произведения признаков и также агрегирует их с весами:

$$z_{\pi} = \sum_{i,j} v_{ij} (x_i * x_j).$$

Итоговое значение нейрона вычисляется как сумма этих двух компонент с учетом порогового параметра:

$$z = z_{\sigma} + z_{\pi} + b.$$

После этого результат передается через функцию активации (обычно используется сигмоида), которая переводит его в значение от 0 до 1. Чем ближе это значение к 1, тем выше вероятность принадлежности данного объекта к рассматриваемому кластеру.

Процесс обучения

Обучение сети направлено на то, чтобы минимизировать разброс данных внутри каждого кластера и увеличить различия между кластерами. На каждом шаге веса обновляются по следующему принципу:

- если признаки в кластере имеют устойчивое среднее значение, веса σ-части смещаются в сторону этого среднего;
- если в кластере часто встречаются определенные комбинации признаков (например, осадки × влажность почвы или уклон × площадь бассейна), π-часть усиливает их вклад.

Таким образом, сеть «обучается» выделять не только отдельные важные факторы, но и их взаимодействия, которые могут быть решающими для образования определенного типа селевого потока.

Преимущества метода

Селевые процессы отличаются высокой сложностью: один и тот же параметр может иметь различное значение в зависимости от сочетания с другими факторами. Например, высокая влажность почвы не всегда приводит к формированию селя, но в сочетании с большим количеством осадков и крутым склоном вероятность резко возрастает. Линейные методы кластеризации, такие как алгоритм К-средних, не способны учитывать такие эффекты, так как работают только с расстояниями в пространстве признаков.

Сигма-пи нейронная сеть, напротив, благодаря своей π -части выявляет скрытые нелинейные зависимости и позволяет моделировать сложные взаимодействия параметров. Это обеспечивает более высокое качество кластеризации и при этом сохраняет интерпретируемость, что делает метод удобным для применения в инженерной геологии и мониторинге процессов.

СРАВНЕНИЕ С САМООРГАНИЗУЮЩИМИСЯ КАРТАМИ КОХОНЕНА

Одними из наиболее распространенных нейросетевых алгоритмов кластеризации являются самоорганизующиеся карты Кохонена (SOM). Они представляют собой двумерное проективное отображение многомерного пространства признаков, где объекты, обладающие сходными характеристиками, проецируются в соседние области карты. Такой подход позволяет визуализировать многомерные данные и выявлять группы объектов без предварительного знания числа кластеров.

SOM хорошо зарекомендовали себя в задачах анализа геологических процессов благодаря своей способности выявлять скрытые структуры в данных и обеспечивать наглядную интерпретацию результатов. Однако данный метод имеет ряд ограничений. Во-первых, SOM в большей степени ориентированы на топологическое упорядочивание данных, нежели на моделирование взаимодействий между признаками. Во-вторых, получаемая кластеризация зависит от параметров обучения (размера карты, радиуса окрестности), что может приводить к неоднозначным результатам при разных настройках. Кроме того, SOM не предоставляет явный механизм для анализа вклада отдельных признаков и их комбинаций в формирование кластеров.

В отличие от SOM сигма-пи нейронные сети обладают способностью явно моделировать взаимодействия второго порядка за счет π -части архитектуры. Это дает возможность не только выделять кластеры, но и объяснять их природу через ключевые комбинации признаков. Например, можно напрямую выявить, что высокая влажность почвы в сочетании с большим уклоном склона существенно увеличивает вероятность формирования определенного типа селевого потока. Таким образом, сигма-пи сети обеспечивают более высокий уровень интерпретируемости и лучше подходят для задач инженерной геологии, где требуется не просто группировка данных, но и понимание причинно-следственных связей.

ПРИМЕР РАБОТЫ СЕТИ НА ДАННЫХ О СЕЛЕВЫХ ПОТОКАХ

Для демонстрации возможностей разработанного алгоритма на вход системы был подан набор данных в формате CSV, содержащий шесть признаков, описывающих морфометрические и генетические характеристики селевых потоков:

- генезис селя;
- площадь бассейна;
- средний уклон русла;
- длина реки;
- высота истока;
- максимальный объем отложений.

Задачей сети являлось разбиение выборки на несколько устойчивых кластеров без априорного знания их числа, а также выявление наиболее значимых факторов и их комбинаций, определяющих принадлежность потока к определенному типу (рис. 1).

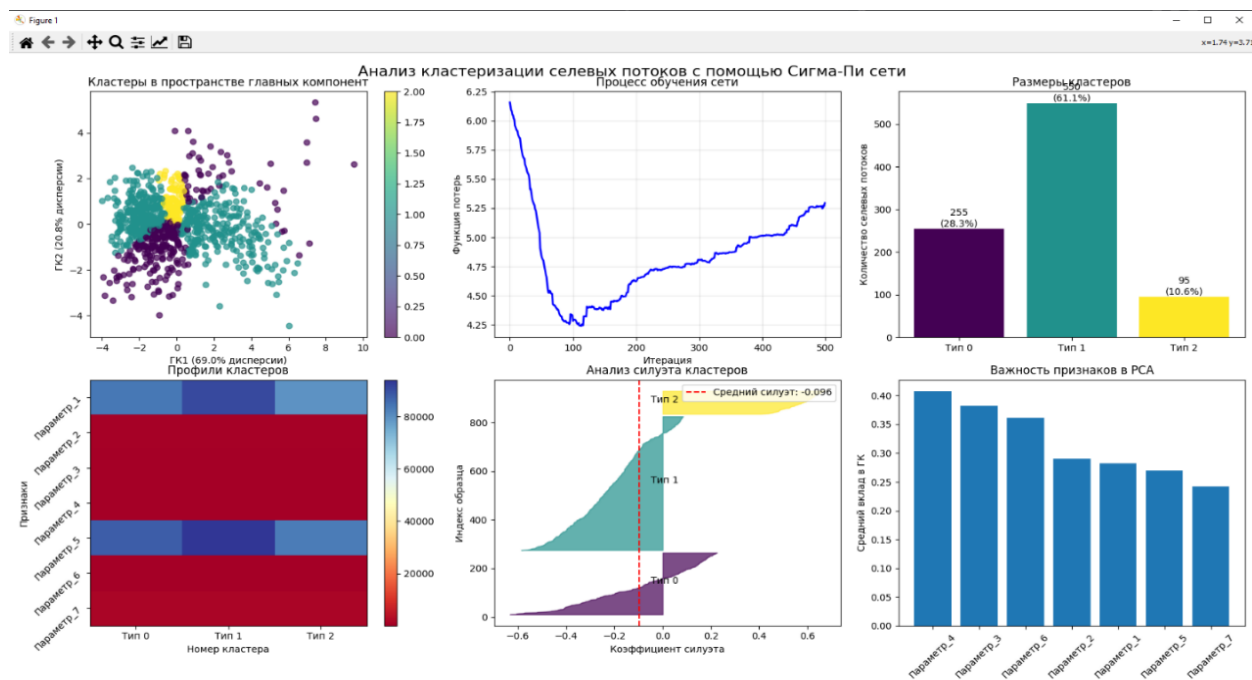


Рис. 1. Анализ кластеризации селевых потоков с помощью сигма-пи сети

Fig. 1. Analysis of mudflow clustering using the sigma-pi network

В результате работы модели было выделено три устойчивых кластера (типы 0, 1 и 2), которые соответствуют трем различным группам селевых потоков с характерными морфологическими и генетическими особенностями.

Кластер 1 (Тип 1) оказался доминирующим. Он включает 61,1 % всей выборки, что соответствует примерно 540 образцам. Данный тип представляет собой наиболее распространенную группу селей, обладающих типичными морфометрическими характеристиками.

Кластер 0 (Тип 0) занимает промежуточное положение и составляет 28,3 % выборки (около 255 образцов). Можно предположить, что данная группа соответствует среднераспространенным разновидностям селей, формирующимся при специфическом сочетании факторов.

Кластер 2 (Тип 2) является наименьшим и включает 10,6 % данных (около 95 образцов). Он, вероятно, отражает редкий или специфический генетический тип селевых процессов, формирующихся при особых условиях.

Таким образом, выявленные группы обладают как количественными, так и качественными различиями, что подтверждает способность модели выделять структурные закономерности в многомерных данных.

Процесс обучения сигма-пи сети сопровождался характерным снижением функции потерь: на первых итерациях наблюдалось ее быстрое падение, после чего кривая стабилизировалась, указывая на достижение устойчивого решения. Для выхода на плато понадобилось около 450 итераций, что свидетельствует о сходимости алгоритма и достаточной адаптивности весовых коэффициентов.

Для наглядной визуализации результатов была проведена проекция данных в пространство главных компонент. В этом пространстве кластеры демонстрируют четкое пространственное разделение:

Тип 0 (фиолетовый) формирует компактную область в левой части графика, что указывает на устойчивую однородность данной группы;

Тип 1 (бирюзовый) занимает центральную и правую части, проявляя более широкое распределение, что отражает внутреннее разнообразие наиболее массового класса;

Тип 2 (желтый) образует отдельную компактную область в верхней части, что подтверждает его уникальность и обособленность от остальных типов.

Такое разделение подчеркивает корректность кластеризации и способность модели дифференцировать селевые потоки не только количественно, но и структурно (рис. 2).

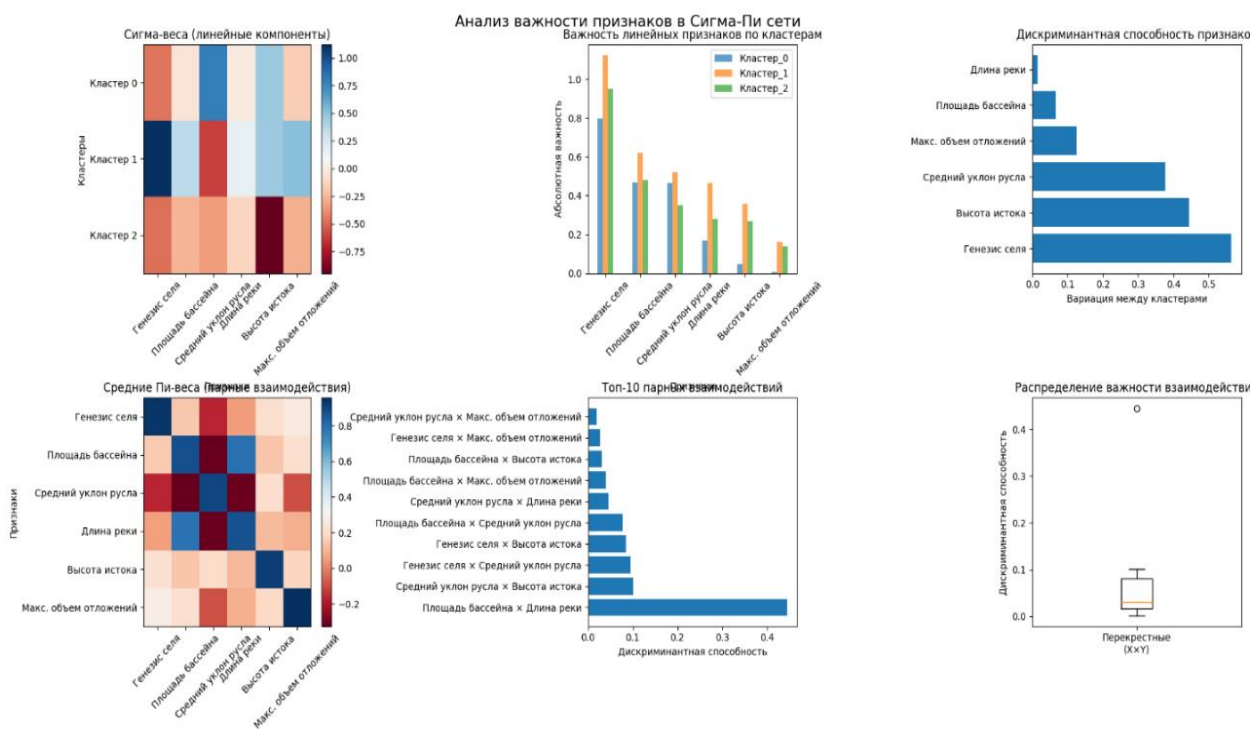


Рис. 2. Влияние попарной комбинации признаков на определение кластера

Fig. 2. The influence of pairwise combination of features on cluster definition

Интерпретация результатов осуществлялась с использованием тепловых карт влияния признаков. При этом:

синий цвет отражает отрицательный вклад признака в принадлежность объекта к определенному кластеру;

красный цвет – положительное влияние;

белый цвет – нейтральный вклад.

Такое представление позволяет наглядно выявлять не только роль отдельных параметров (например, площади бассейна или уклона русла), но и значимость их комбинаций. Например, сеть показала, что высокая влажность почвы в сочетании с крутым уклоном русла усиливает вероятность принадлежности потока к определенному типу. Аналогичным образом комбинация «максимальный объем отложений × площадь бассейна» оказывает значительное влияние на формирование другого типа селея.

Таким образом, сигма-пи сеть демонстрирует не только высокое качество кластеризации, но и позволяет проводить интерпретируемый анализ, выявляя ключевые детерминанты формирования различных типов селевых потоков.

ЗАКЛЮЧЕНИЕ

На основе проведенного исследования была выполнена интерпретация трех устойчивых кластеров, соответствующих основным типам селевых потоков – грязевому, каменному и грязекаменному. Кластеризация данных осуществлялась с использованием сигма-пи нейронной сети, которая учитывает как линейные характеристики (σ -компонент), так и попарные взаимодействия признаков (π -компонент). Такой подход позволил выявить ключевые морфометрические и генетические детерминанты, определяющие формирование различных типов селей.

Анализ важности признаков и их комбинаций показал высокую степень интерпретируемости модели. В отличие от классических методов кластеризации сигма-пи архитектура обеспечивает возможность выявления скрытых нелинейных зависимостей между параметрами, что существенно повышает качество классификации геологических процессов.

Полученные результаты демонстрируют перспективность применения гибридных нейросетевых методов в инженерной геологии. В частности, они могут быть использованы для более точной диагностики и прогнозирования селевых процессов, а также для разработки систем геоэкологического мониторинга и мер инженерной защиты на селеопасных территориях.

СПИСОК ЛИТЕРАТУРЫ

1. Татаренко Н. В., Шагин С. И., Машиуков Х. В. Географические особенности распространения селевых явлений на территории КБР // Научные известия. 2019. № 17. С. 26–30. EDN: ZJPAGU
2. Жиров П. А. Построение самоорганизующейся карты Кохонена (SOM) для прогнозирования типов селевых потоков // Известия Кабардино-Балкарского научного центра РАН. 2024. № 5. С. 129–137. DOI: 10.35330/1991-6639-2024-26-5-129-137
3. Лютикова Л. А. Анализ характеристик селевых потоков при ограниченных данных с использованием моделей машинного обучения // Моделирование, оптимизация и информационные технологии. 2024. Т. 12. № 4. ID: 36. DOI: 10.26102/2310-6018/2024.47.4.029
4. Слейман А., Козлов Д. В. Использование искусственных нейронных сетей для оценки поверхностного стока в расчетах водохозяйственного баланса бассейна реки Верхний Оронгес // Водное хозяйство России: проблемы, технологии, управление. 2024. № 3. С. 21–37. DOI: 10.35567/19994508-2024-3-21-37
5. Бодянский Е. В., Кулишова Н. Е. Многомерная искусственная нейронная сигма-пи сеть и алгоритм ее обучения // Радиоэлектроника и информатика. 2005. № 4. С. 122–125.
6. Jiao J., Su K. A new Sigma-Pi-Sigma neural network based on L1 and L2 regularization and applications // AIMS Mathematics. 2024. Vol. 9. No. 3. Pp. 5995–6012. DOI: 10.3934/math.2024293
7. Deng F., Liang S., Qian K. et al. A recurrent sigma pi sigma neural network // PMC / NCBI, 2025. DOI: 10.1038/s41598-024-84299-y
8. Жиров П. А. Интеллектуальные методы кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2023. № 6(116). С. 152–159. DOI: 10.35330/1991-6639-2023-6-116-152-159

REFERENCES

1. Tatarenko N.V., Shagin S.I., Mashukov Kh.V. Geographical features of the distribution of mudflow phenomena in the Kabardino-Balkarian Republic. *Scientific News*. 2019. No. 17. Pp. 26–30. EDN: ZJPAGU. (In Russian)

2. Zhilov R.A. Construction of a Kohonen self-organizing map (SOM) for predicting types of mudflows. *News of the Kabardino-Balkarian Scientific Center of the RAS*. 2024. No. 5. Pp. 129–137. DOI: 10.35330/1991-6639-2024-26-5-129-137. (In Russian)
3. Lyutikova L.A. Analysis of mudflow characteristics with limited data using machine learning models. *Modeling, Optimization and Information Technology*. 2024. Vol. 12. No. 4. ID: 36. DOI: 10.26102/2310-6018/2024.47.4.029. (In Russian)
4. Sleiman A., Kozlov D.V. Using artificial neural networks to assess surface runoff in water management balance calculations of the upper Orontes River Basin. *Water Sector of Russia: problems, Technologies, Management*. 2024. No. 3. Pp. 92–107. DOI: 10.35567/19994508-2024-3-21-37. (In Russian)
5. Bodianskii E.V., Kulishova N.E. Multidimensional artificial neural sigma-pi network and its training algorithm. *Radioelectronics and Informatics*. 2005. No. 4. Pp. 122–125. (In Russian)
6. Jiao J., Su K. A new Sigma-Pi-Sigma neural network based on L1 and L2 regularization and applications. *AIMS Mathematics*. 2024. Vol. 9. No. 3. Pp. 5995–6012. DOI: 10.3934/math.2024293
7. Deng F., Liang S., Qian K. et al. A recurrent sigma pi sigma neural network. *PMC / NCBI*, 2025. DOI: 10.1038/s41598-024-84299-y
8. Zhilov R.A. Intelligent methods for data clustering. *News of the Kabardino-Balkarian Scientific Center of the RAS*. 2023. No. 6(116). Pp. 152–159. DOI: 10.35330/1991-6639-2023-6-116-152-159. (In Russian)

Финансирование. Исследование проведено без спонсорской поддержки.

Funding. The study was performed without external funding.

Информация об авторе

Жилов Руслан Альбердович, мл. науч. сотр. отдела нейроинформатики и машинного обучения, Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра Российской академии наук;

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

zhilov91@gmail.com, ORCID: <https://orcid.org/0000-0002-3552-4854>, SPIN-код: 9389-6188

Information about the author

Ruslan A. Zhilov, Junior Researcher, Neuroinformatics and Machine Learning Department, Institute of Applied Mathematics and Automation – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

89 A Shortanov street, Nalchik, 360000, Russia;

zhilov91@gmail.com, ORCID: <https://orcid.org/0000-0002-3552-4854>, SPIN-code: 9389-6188