СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ, СТАТИСТИКА

УДК 519.7 Научная статья
DOI: 10.35330/1991-6639-2024-26-5-129-137
EDN: MKXTYO

Построение самоорганизующейся карты Кохонена (SOM) для прогнозирования типов селевых потоков

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. В работе построена самоорганизующаяся карта Кохонена (SOM), которая производит анализ типа сели. Обучение SOM производится на реальных данных кадастра селевой опасности Юга европейской части России. Цель работы – получить прогнозы типов селевых потоков. Результаты исследования показывают, что SOM дает хорошую точность в предсказании типов селей. Основной задачей будет кластеризация данных, связанных с геологическими и метеорологическими факторами, с целью выявления закономерностей, которые могут быть использованы для прогнозирования риска возникновения различных типов селевых потоков. Ожидается, что результаты данной работы смогут способствовать более точному и своевременному прогнозированию селевых потоков, что в свою очередь поможет минимизировать ущерб от этих природных явлений.

Ключевые слова: кластеризация данных, метод кластеризации SOM, модель SOM, самоорганизующиеся карты Кохонена, классификация типа сели

Поступила 12.08.2024, одобрена после рецензирования 16.09.2024, принята к публикации 23.09.2024

Для цитирования. Жилов Р. А. Построение самоорганизующейся карты Кохонена (SOM) для прогнозирования типов селевых потоков // Известия Кабардино-Балкарского научного центра РАН. 2024. Т. 26. № 5. С. 129–137. DOI: 10.35330/1991-6639-2024-26-5-129-137

MSC: 68T09 Original article

Construction of Kohonen self-organizing map (SOM) for prediction of mudflow types

R.A. Zhilov

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. The paper describes a self-organizing Kohonen map (SOM) that analyzes the mudflow type. SOM is trained on real cadastre data of mudflow danger in the south of the European part of Russia. The purpose of the work is to obtain forecasts of mudflow types. The results of the work show that SOM provides good accuracy in predicting mudflow types. The main task will be to cluster data related to geological and meteorological factors in order to identify patterns that can be used to predict the risk of occurrence of various mudflow types. It is expected that the results of this work will contribute to more accurate and on time forecasting of mudflows, which, in turn, will help minimize damage from these natural phenomena.

© Жилов Р. А., 2024

Keywords: data clustering, SOM clustering method, SOM model, Kohonen self-organizing maps, mudflow type classification

Submitted 12.08.2024, approved after reviewing 16.09.2024, accepted for publication 23.09.2024

For citation. Zhilov R.A. Construction of Kohonen self-organizing map (SOM) for prediction of mudflow types. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2024. Vol. 26. No. 5. Pp. 129–137. DOI: 10.35330/1991-6639-2024-26-5-129-137

Введение

Селевые потоки представляют собой опасное природное явление, вызывающее разрушение инфраструктуры, экологические катастрофы и угрожающее жизни людей. В связи с увеличением частоты экстремальных погодных условий и изменением климата прогнозирование и своевременное предупреждение о селевых потоках становится важной задачей для ученых и инженеров [1]. Одним из перспективных подходов к решению этой задачи является использование методов машинного обучения, в частности самоорганизующихся карт Кохонена (Self-Organizing Maps, SOM).

Предобработка данных для определения типа селевых потоков

Изучением селевых явлений на Кавказе занимались и занимаются многие советские и российские ученые, что привело к значительному количеству научных трудов по этой тематике. Однако существующая ведомственная разрозненность издающихся научных публикаций и отсутствие единого методического центра в стране по изучению селепроявлений создают серьезные препятствия для ознакомления и использования результатов исследований в практических целях, а также при проведении новых научно-исследовательских работ.

Кадастр, из которого берутся данные для обучения и тестирования данной модели, является справочным изданием, в котором в систематизированном виде представлена обобщенная информация о пространственном распределении основных параметров и режиме селевых процессов на территории Юга европейской части России [2].

Предобработка данных является важным шагом перед построением самоорганизующихся карт Кохонена. Она помогает улучшить качество модели, обеспечивая точность и надежность кластеризации и прогнозирования. Внимание к деталям на этапе предобработки данных способно существенно повысить эффективность применения SOM для анализа и прогнозирования.

Предобработка данных играет ключевую роль в построении самоорганизующихся карт Кохонена, так как качество исходных данных напрямую влияет на результаты модели. Основные цели предобработки данных включают:

- устранение шума и выбросов;
- приведение данных к единому масштабу;
- устранение пропущенных значений;
- выделение наиболее значимых признаков (фичей).

Перед началом построения карты Кохонена важно убедиться, что данные очищены от аномалий и выбросов. Выбросы могут негативно повлиять на процесс обучения карты, приводя к неправильным результатам кластеризации.

Самоорганизующиеся карты Кохонена чувствительны к масштабу данных, так как они используют евклидовое расстояние для оценки сходства между объектами. Если признаки данных имеют разные масштабы, более крупные значения могут доминировать над другими признаками, что приведет к неверной кластеризации. Для повышения эффективности обучения карты Кохонена необходимо сосредоточиться на наиболее значимых признаках данных. Это позволит сократить время обучения и улучшить качество кластеризации.

Самоорганизующиеся карты Кохонена

Самоорганизующиеся карты Кохонена (СОМ, англ. Self-Organizing Maps, SOM) – это вид искусственных нейронных сетей, разработанных финским ученым Тейво Кохоненом в 1980-х годах. Эти карты предназначены для решения задач кластеризации и визуализации многомерных данных, что делает их полезными в различных областях, включая анализ данных, машинное обучение и биоинформатику [4]. В этой работе будут рассмотрены принципы работы самоорганизующихся карт Кохонена, их структура, алгоритмы обучения, а также примеры практического применения.

Самоорганизующиеся карты Кохонена относятся к нейросетям, использующим неконтролируемое обучение. Основная идея заключается в том, что карта преобразует входные данные многомерного пространства в выходное пространство с меньшей размерностью (обычно двумерное), сохраняя при этом топологические свойства данных, такие как близость и структура.

Карта Кохонена представляет собой двумерную сетку нейронов, каждый из которых ассоциирован с вектором весов, имеющим ту же размерность, что и входные данные. Входные данные подаются на карту, и каждый нейрон сравнивает свои веса с входным вектором. Нейрон, чьи веса наиболее близки к входному вектору, называется нейроном-победителем [5].

Процесс обучения карты Кохонена состоит из следующих этапов:

1. *Инициализация*. Веса каждого нейрона инициализируются случайными значениями или на основе какой-либо эвристики.

2. *Выбор случайного примера*. Один из входных векторов случайным образом выбирается из набора данных.

3. *Определение нейрона-победителя*. Для каждого нейрона вычисляется расстояние между его вектором весов и входным вектором. Нейрон, имеющий наименьшее расстояние, становится победителем.

4. *Обновление весов*. Веса нейрона-победителя и его соседей обновляются так, чтобы стать ближе к входному вектору. Это делается по формуле

$$W_i(t+1) = W_i(t) + \eta(t) * h_i(t) * \big(X(t) - W_i(t)\big),$$

где $W_i(t)$ – вектор весов i-го нейрона на итерации t,

$\eta(t)$ – коэффициент обучения,

$h_i(t)$ – функция соседства, определяющая, насколько сильно должны обновляться веса соседних нейронов,

$X(t)$ – входной вектор на итерации t.

5. *Повторение*. Шаги 2–4 повторяются для заданного количества итераций или до тех пор, пока изменения в карте не станут минимальными.

Самоорганизующиеся карты Кохонена находят применение в различных областях, где требуются кластеризация и визуализация многомерных данных.

СОМ часто используются для задач кластеризации, так как они могут эффективно выявлять кластеры в данных без предварительных меток. В отличие от других методов кластеризации, таких как k-means, СОМ не требует заранее заданного числа кластеров.

Хотя СОМ в первую очередь используются для кластеризации и визуализации, они также могут быть применены для задач прогнозирования и обнаружения аномалий. Обученная карта может быть использована для прогнозирования поведения новых данных или для выявления данных, сильно отличающихся от обучающего набора.

К преимуществам SOM можно отнести:

• *Ненадзорное обучение*. СОМ не требуют меток на данных и могут самоорганизовываться на основе входных данных.

• *Топологическая сохранность*. Карта сохраняет пространственную структуру данных, что упрощает визуализацию и интерпретацию.

• *Гибкость*. СОМ могут работать с данными высокой размерности и эффективно выявлять скрытые паттерны.

Но карты Кохонена также имеют свои недостатки:

• *Чувствительность к параметрам*. Результаты СОМ сильно зависят от параметров, таких как размер карты, скорость обучения и функция соседства.

• *Трудности в интерпретации*. Хотя карта Кохонена сохраняет топологию данных, интерпретация результатов может быть сложной, особенно для неквалифицированных пользователей.

• *Масштабируемость*. СОМ могут быть вычислительно затратными при работе с большими наборами данных, особенно если требуется высокая разрешающая способность карты.

Самоорганизующиеся карты Кохонена представляют собой мощный инструмент для анализа, кластеризации и визуализации многомерных данных. Благодаря своей способности сохранять топологическую структуру данных они находят широкое применение в различных областях – от маркетинга до биоинформатики. Однако для достижения оптимальных результатов важно учитывать ограничения этого метода и тщательно настраивать параметры модели. В будущем, с развитием вычислительных технологий и методик обработки данных, применение СОМ будет только расширяться, что откроет новые возможности для анализа данных и машинного обучения.

Интеллектуальные методы прогнозирования селевых потоков

Селевые потоки, или оползни, представляют собой одно из самых опасных природных явлений, которое может привести к значительным разрушениям и человеческим жертвам. Прогнозирование селевых потоков является важной задачей для защиты населения и инфраструктуры, особенно в горных районах. В последние годы большое внимание уделяется разработке и применению интеллектуальных методов для прогнозирования селевых потоков. Эти методы включают машинное обучение, нейронные сети, анализ больших данных и другие современные подходы, которые позволяют эффективно оценивать риск возникновения оползней и предупреждать о возможных катастрофах [6].

Прогнозирование селевых потоков – задача сложная, так как зависит от множества факторов, среди которых:

- Геологические: тип грунта, наличие трещин и разломов.
- Геоморфологические: крутизна и структура склонов, эрозионные процессы.
- Гидрометеорологические: интенсивность и продолжительность осадков, таяние снега.
- Антропогенные: строительная деятельность, вырубка лесов.

Эти факторы взаимодействуют друг с другом, и их совокупное воздействие может приводить к возникновению оползней. Задача интеллектуальных методов — анализировать эти факторы и прогнозировать вероятность селевых потоков на основе исторических данных и текущих наблюдений.

Современные интеллектуальные методы прогнозирования селевых потоков включают широкий спектр подходов – от классических статистических моделей до современных методов машинного обучения и анализа данных.

Машинное обучение (ML) включает методы, которые позволяют моделям обучаться на данных и делать прогнозы без необходимости явного программирования правил [7]. В контексте прогнозирования селевых потоков наиболее часто используются следующие методы:

• Решающие деревья и случайные леса. Эти алгоритмы позволяют выявлять взаимосвязи между различными факторами и определять наиболее значимые переменные для прогнозирования.

• Поддерживающие векторные машины (SVM). SVM хорошо справляются с задачами классификации и могут использоваться для определения условий, при которых возникает высокий риск оползней.

• Нейронные сети и глубокое обучение. Нейронные сети могут моделировать сложные нелинейные зависимости и использоваться для прогнозирования на основе больших массивов данных, включая временные ряды и пространственные данные.

ГИС – мощный инструмент для анализа пространственных данных, который используется для моделирования и прогнозирования природных явлений, таких как селевые потоки. Интеграция ГИС и методов машинного обучения позволяет создавать детализированные карты рисков, учитывать топографические и геологические характеристики местности.

Большие данные (Big Data) играют важную роль в прогнозировании природных катастроф. Данные, полученные с различных сенсоров, спутников, метеостанций и других источников, могут использоваться для тренировки моделей прогнозирования. Методы обработки больших данных позволяют эффективно анализировать эти данные в реальном времени и выявлять закономерности, которые могут указывать на повышенный риск селевых потоков.

Гибридные модели сочетают различные интеллектуальные методы для повышения точности прогнозирования. Например, можно комбинировать методы машинного обучения с физическими моделями оползней, что позволяет учитывать как статистические закономерности, так и физические процессы, происходящие в грунте.

Интеллектуальные методы прогнозирования селевых потоков открывают новые возможности для повышения точности и своевременности прогнозов. Эти методы позволяют эффективно анализировать сложные многомерные данные и учитывать многочисленные факторы, влияющие на риск возникновения оползней. Однако успешное применение таких методов требует внимательного подхода к сбору и обработке данных, а также интеграции различных моделей для достижения наилучших результатов. В условиях глобального изменения климата и увеличения интенсивности экстремальных погодных явлений развитие и совершенствование интеллектуальных методов прогнозирования селевых потоков будет иметь критическое значение для обеспечения безопасности населения и инфраструктуры.

SOM ДЛЯ ПРОГНОЗИРОВАНИЯ ТИПА СЕЛЕВОГО ПОТОКА

1. Определение класса SOM.

Класс SOM представляет собой самоорганизующуюся карту Кохонена.

__init__ – Инициализация объекта SOM:

m и n: Размеры сетки нейронов (в данном случае 5 x 5).

dim: Размерность входных данных, то есть количество признаков в каждом векторе данных.

n_iterations: Количество итераций для обучения.

alpha: Начальная скорость обучения.

sigma: Начальное значение для радиуса соседства, определяющего, насколько сильно изменяются веса соседних нейронов во время обучения. Если значение не указано, оно задается как половина большего измерения сетки.

weights – Инициализация весов:

Веса нейронов инициализируются случайными значениями в диапазоне от 0 до 1. Веса представляют собой трехмерный массив размером m x n x dim, где m и n – размеры сетки, а dim – количество признаков.

2. Вспомогательные функции:

_euclidean_distance(self, x, y):

Вычисляет евклидово расстояние между двумя векторами *x* и *y*.

_neighborhood_function(self, distance, iteration, total_iterations):

Вычисляет функцию соседства, определяющую, как сильно изменяются веса нейронов в зависимости от их расстояния до нейрона-победителя. Радиус соседства уменьшается с каждой итерацией.

_learning_rate(self, iteration, total_iterations):

Вычисляет скорость обучения на текущей итерации. Скорость обучения уменьшается экспоненциально с каждой итерацией.

_best_matching_unit(self, x):

Находит лучший соответствующий нейрон (BMU) на карте для входного вектора x. Это нейрон, чьи веса наиболее близки к входному вектору.

3. Основные функции:

train(self, data):

Основной метод для обучения карты.

На каждой итерации выбирается случайный вектор из данных.

Определяется BMU для этого вектора.

Обновляются веса BMU и его соседей на основе функции соседства и скорости обучения.

get_weights(self):

Возвращает текущие веса нейронов.

find_bmu(self, x):

Находит и возвращает координаты BMU на карте для заданного входного вектора *x*. На рисунке 1 показана визуализация классов обученной сети Кохонена.

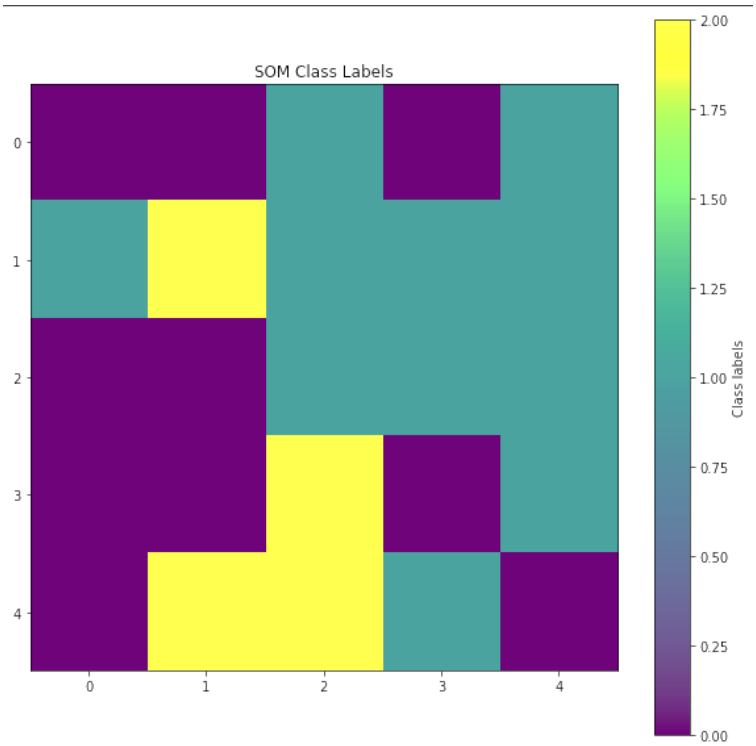

Рис. 1. *Визуализация классов обученной сети Кохонена*

Fig. 1. *Visualization of the trained Kohonen network classes*

На рисунке 2 показана визуализация SOM и нейрона-победителя BMU.

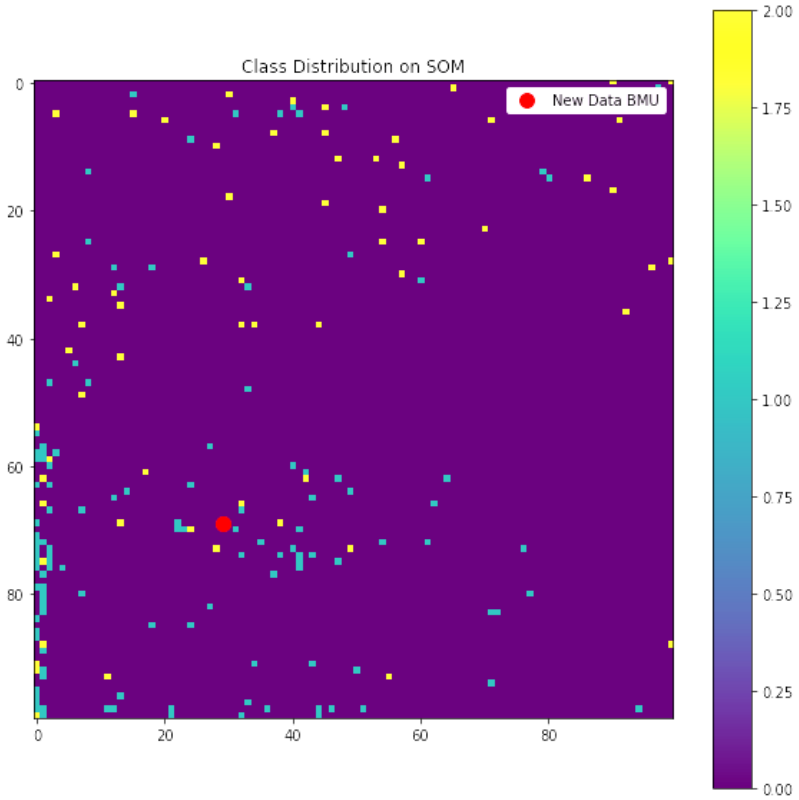

Рис. 2. *Визуализация SOM и нейрона-победителя BMU*

Fig. 2. *Visualization of SOM and BMU winner neuron*

```
Real: 2, Predicted: 1
Real: 0, Predicted: 1
Real: 2, Predicted: 2
Real: 1, Predicted: 2
Real: 1, Predicted: 1
Real: 1, Predicted: 0
Real: 1, Predicted: 1
Real: 2, Predicted: 2
Real: 2, Predicted: 2
Real: 1, Predicted: 1
Real: 1, Predicted: 1
Real: 1, Predicted: 2
Real: 1, Predicted: 1
Real: 1, Predicted: 1
Real: 1, Predicted: 1
Real: 1, Predicted: 2
Real: 1, Predicted: 1
Real: 1, Predicted: 2
Real: 1, Predicted: 2
Real: 1, Predicted: 1
Real: 2, Predicted: 2
Real: 2, Predicted: 2
Real: 2, Predicted: 2
Real: 1, Predicted: 2
Real: 1, Predicted: 1
...
Real: 2, Predicted: 2
Real: 2, Predicted: 2
Real: 2, Predicted: 2
Accuracy: 72.94%
```

Рис. 3. *Результат работы модели*

Fig. 3. *Result of the model's work*

На рисунке 3 показан результат работы модели. Как видно из рисунка, точность предсказания модели равна 72,94 %, что является достаточно неплохим показателем для заданных данных. Данные для обучения и тестирования модели были взяты из кадастра селевой опасности Юга европейской части России. Подготовленный файл данных состоял из 385 строк и 7 столбцов. Первый столбец является меткой данных или классом, к которому относится заданный объект. В этом столбце 0, 1 или 2 соответственно трем типам селевых потоков (ГК – грязекаменный, ВК – водокаменный и их комбинация ГКВК). Остальные 6 полей данных являются числовыми данными, обозначающими генезис селя, площадь бассейна реки, средний уклон реки, длину реки, высоту источника, максимальный объем твердых отложении соответственно. Для обучения модели набор данных был разделен на обучающий (300) и тестовый (85) наборы. Размер сетки нейронов в данной работе равен 5 х 5, количество итераций при обучении – 1000, начальная скорость обучения – 0.3, начальное состояние весов случайное в диапазоне от 0 до 1.

Эти значения были подобраны опытным путем. При увеличении количества нейронов и количества итераций происходит переобучение сети, что приводит к тому, что на обучающей выборке модель дает хорошие результаты, но на тестовой выдает плохие. Это связано с тем, что происходит «запоминание» моделью данных, а не обучение.

По результатам работы модели также можно сделать вывод, что данные недостаточно структурированы и объем данных не является достаточным для получения более высоких результатов.

Заключение

В работе строится самоорганизующаяся карта Кохонена для прогнозирования типа селевого потока. Обучение и тестирование SOM производится с использованием данных, взятых из кадастра селевой опасности Юга европейской части России. Точность предсказания модели – 72,94 %. Для увеличения точности в дальнейшем требуется увеличение объема данных для обучения. Основными преимуществами данной модели являются достаточно быстрая скорость обучения и практически «мгновенный» прогноз после обучения. Обучение сети требуется только один раз. В дальнейшем модель может выдавать прогноз типа селевого потока, используя настроенные веса связей.

СПИСОК ЛИТЕРАТУРЫ

1. *Хворостов В. В., Хворостов И. И.* Экстраординарные и ультраселевые потоки на территории Большого Кавказа // Материалы международной конференции «Устойчивое развитие горных территорий». 2004. С. 605.

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

3. *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.

4. *Kohonen T.* Self-Organizing Maps (Third Extended Edition). New York, 2001. 501 p.

5. *Жилов Р. А.* Применение нейронных сетей при кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2021. № 1(99). С. 15–19. DOI: 10.35330/1991-6639-2021-1-99-15-19

6. *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

7. *Флах П.* Машинное обучение: наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс. ISBN: 978-5-97060-273-7. 2015. 400 с.

REFERENCES

1. Khvorostov V.V., Khvorostov I.I. Extraordinary and ultra-seismic flows in the territory of the Greater Caucasus. *Materialy mezhdunarodnoy konferentsii «Ustoychivoye razvitiye gornykh territoriy»* [Proceedings of the international conference “Sustainable Development of Mountain Territories”]. 2004. P. 605. (In Russian)

2. Kondratyeva N.V., Adzhiev A.Kh., Bekkiev M.Yu. et al. *Kadastr selevoy opasnosti Yuga yevropeyskoy chasti Rossii* [Inventory of mudflow danger in the south of the European part of Russia]. Moscow: Feoriya, Nalchik: Pechatnuy dvor, 2015. 148 p. (In Russian)

3. Kondratieva N.V. Preliminary assessment of the maximum volume of solid mudflow deposits using mathematical statistics methods for the Central Caucasus [Electronic journal]. *Sovremennyye problemy nauki i obrazovaniya* [Modern problems of science and education]. 2014. No. 4. Pp. 50–56. (In Russian)

4. Kohonen T. Self-Organizing Maps (Third Extended Edition). New York, 2001. 501 p.

5. Zhilov R.A. Application of neural networks in data clustering. *News of the Kabardino-Balkarian Scientific Center of RAS*. 2021. No. 1(99). Pp. 15–19. DOI: 10.35330/1991-6639-2021-1-99-15-19. (In Russian)

6. Radeev N.A. Prediction of avalanche danger using machine learning methods. *Vestnik NGU. Seriya: Informatsionnyye tekhnologii* [NSU Bulletin. Series: Information Technologies]. 2021. Vol. 19. No. 2. Pp. 92–101. DOI: 10.25205/1818-7900-2021-19-2-92-101. (In Russian)

7. Flakh P. *Mashinnoye obucheniye: nauka i iskusstvo postroyeniya algoritmov, kotoryye izvlekayut znaniya iz dannykh* [Machine learning: the science and art of building algorithms that extract knowledge from data]. Moscow: DMK Press. ISBN: 978-5-97060-273-7. 2015. 400 p. (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 «Нейроинформатика и машинное обучение», Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра РАН;

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;

360000, Russia, Nalchik, 89 A Shortanov street;

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