ГИАБ. Горный информационно-аналитический бюллетень /
MIAB. Mining Informational and Analytical Bulletin, 2019;(6):27-44

УДК 622.817.4 DOI: 10.25018/0236-1493-2019-06-0-27-44

ПРИМЕНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДЛЯ ЭКСПРЕСС-АНАЛИЗА СЕЙСМИЧЕСКИХ ДАННЫХ

М.И. Абдрахманов[1], С.Э. Лапин[2], И.В. Шнайдер[2]

[1] ООО «Информационные горные технологии», Екатеринбург, Россия, e-mail: marat-ab@mail.ru
[2] Уральский государственный горный университет, Екатеринбург, Россия

Аннотация: Одной из достаточно серьезных проблем непрерывного геомониторинга являются большие объемы данных, накапливаемые при работе системы, которые необходимо анализировать и интерпретировать. Это делает важной и актуальной задачу предварительного экспресс анализа данных, заключающегося в подготовке набора характерных точек, на которые следует обратить внимание в первую очередь, в рамках исследуемой части горного массива. В качестве подхода к ее решению авторами статьи предлагается кластерный анализ, его результатом являются найденные центры кластеризации, которые могут быть использованы в качестве характерных точек. Рассмотрены три алгоритма кластеризации: k-means, Mean Shift и DBSCAN. Для оценки их работы и применимости для анализа сейсмических данных использовалось сравнение результатов, полученных с помощью алгоритмов, с точками, которые эксперт указал в качестве характерных для заданного набора данных. Оценка проведения качества самой процедуры кластеризации проводилась по индексам Calinski-Harabasz, Davies-Bouldin и силуэтному коэффициенту. В качестве сейсмических данных были использованы напряженность массива (горное давление) и вероятность флюидопроявления, численные значения которых получены в рамках локального прогноза. По результатам проведенной работы можно сделать вывод, что наиболее подходящим из исследованных алгоритмов кластеризации является алгоритм DBSCAN, его можно использовать для предварительного экспресс анализа сейсмических данных.

Ключевые слова: геомониторинг, анализ сейсмических данных, кластерный анализ, k-means, Mean Shift, DBSCAN, экспресс анализ, метрики для кластеризации.

Для цитирования: *Абдрахманов М. И., Лапин С. Э., Шнайдер И. В.* Применение алгоритмов кластеризации для экспресс-анализа сейсмических данных // Горный информационно-аналитический бюллетень. – 2019. – № 6. – С. 27–44. DOI: 10.25018/0236-1493-2019-06-0-27-44.

Clustering algorithms in express-analysis of seismic data

M.I. Abdrakhmanov[1], S.E. Lapin[2], I.V. Shnayder[2]

[1] LLC «Information mining technologies», Ekaterinburg, Russia, e-mail: marat-ab@mail.ru
[2] Ural State Mining University, Ekaterinburg, Russia

Abstract: A grand problem in continuous geomonitoring is constituted by a huge amount of accumulated data to be analyzed and interpreted. In this regard, it is relevant to carry out the preliminary express-analyses of data in order to select representative points to be in spotlight in the first place in the studied rock mass. By way of approach to this problem solution, the authors propose the clustering analysis which produces clustering centers usable as representative points. The scope

of the discussion embraces three clustering algorithms: k-means, Mean Shift and DBSCAN. Their efficiency and suitability to the seismic data analysis is assessed by comparing the results of the algorithms with the representative points found by expert for the pre-assigned set of data. The quality of the clustering procedure is evaluated by the Calinski–Harabasz and Davies–Bouldin indexes, as well as the silhouette coefficient. The set of the seismic data was composed of numerical values of stress state (rock pressure) and fluid flow potential in rock mass obtained in local prediction. The obtained results allow concluding that the best clustering algorithm is DBSCAN, and it is applicable to preliminary express-analysis of seismic data.

Key words: geomonitoring, seismic data analysis, clustering analysis, express-analysis, clustering metrics.

For citation: Abdrakhmanov M. I., Lapin S. E., Shnayder I. V. Clustering algorithms in express-analysis of seismic data. *MIAB. Mining Inf. Anal. Bull.* 2019;(6):27-44. [In Russ]. DOI: 10.25018/0236-1493-2019-06-0-27-44.

Введение

Потеря устойчивости горного массива или геологическое нарушение, которое возникает в результате техногенного воздействия на горный массив в процессе отработки месторождений подземным способом, может привести к катастрофическим последствиям. В целях предотвращения подобных событий правила промышленной безопасности [1] регламентируют применение систем контроля и прогноза гео-газодинамических явлений, в частности, — систем геомониторинга горного массива. Наиболее опасными в части обозначенной проблемы являются угольные шахты опасные по газу и пыли, поэтому правила безопасности применительно к такого рода предприятиям предусматривают применение сразу нескольких типов геомониторинга, используемых одновременно.

Наиболее важную с точки зрения обеспечения безопасности отработки месторождения информацию о состоянии и структуре горного массива на значительном расстоянии от мест ведения горных работ позволяют получить системы сейсмического геомониторинга, поэтому остановимся на нем более подробно, как представляющему для нас наибольший интерес. Данные для проведения расчетов и публикации в статье взяты с действующего объекта, где применяется серийно выпускаемая система локального и регионального прогноза «Микон-ГЕО» производства компании ООО «ИНГОРТЕХ» [2].

Для оценки зон сейсмической активности в масштабах шахтного поля добычного предприятия применяют системы регионального контроля и прогноза состояния горного массива. Они представляют собой комплекс сейсмодатчиков, распределенных по шахтному полю и непрерывно регистрирующих сейсмическую активность горного массива в автоматическом режиме. Подобные системы регистрируют энергию сейсмического события, а также определяют местоположение его эпицентра. Вычислительный комплекс на поверхности архивирует полученную информацию и, основываясь на критериях производителя системы мониторинга, делает прогноз возможности проявления и местоположения зоны газодинамического явления.

Основываясь на информации, полученной от систем регионального контроля и прогноза, технический руководитель добычного предприятия принимает решение о целесообразности применения систем локального контроля и прогноза состояния горного массива. Кроме того, применение систем локального прогноза также регламентировано действующими правилами промышленной безопасности в угольных шахтах. В отличие от

систем регионального прогноза, датчики системы локального прогноза располагаются в радиусе 100 м от интересующей области массива, регистрируют упругие волны в более высоком диапазоне частот и позволяют оценить не только состояние, но и структуру исследуемой области, то есть оценить прочность пород, напряжения в массиве, а также обнаружить геологические нарушения и трещиноватые зоны, возможно, заполненные газом и/или водой.

В итоге, главной задачей системы геомониторинга является прогноз зон возможного возникновения гео-газодинамических явлений (внезапного выброса породы, газа, вывалов), геологических нарушений и линз с флюидом (т.е. водой и/или газом), зон трещиноватости горного массива, или зон дезинтеграции.

При этом процедура прогноза состояния горного массива может быть разделена на четыре основные части:

- сбор данных (сейсмограмм);
- обработка сейсмограмм в соответствии с принятой методикой;
- визуализация данных с привязкой к объекту исследования;
- интерпретация образов.

Сбор данных представляет собой процесс регистрации сейсмодатчиками упругих волн, распространяемых в среде. Задача тривиальна, так как сейсморазведка, зародившаяся в 1920 г., на сегодняшний день может предложить сейсмодатчики и регистрирующую аппаратуру различных исполнений и широкого спектра характеристик (для применения в шахтах и рудниках опасных по газу и пыли, требуется взрывобезопасное исполнение как регистрирующей аппаратуры, так и средств электропитания и передачи информации).

Обработка сейсмограмм, в соответствии с принятой разработчиком системы методикой представляет наибольший интерес, так как, именно от правильности обработки зависит результат, который после интерпретации позволит решить главную задачу — осуществить прогноз состояния горного массива. Если в «большой» сейсморазведке, где с дневной поверхности производят изучение толщи земной коры, размещая датчики на несколько километров, применяются, в основном, стандартные методики, то в условиях подземного строительства, с применением микросейсмического подхода (то есть небольшого количества сейсмодатчиков, распределенных по площади в несколько десятков метров) рациональным и более эффективным являются авторские методики, разработанные и запатентованные для конкретных условий применения. Например, запатентованная и представленная в работе [3] методика позволяет определить не только структуру (зоны трещиноватости, геологические нарушения, зоны смены структуры пород), но и параметры исследуемого горного массива (напряженность, скорость волны в среде, прочность, пластичность) [4].

Визуализацию данных можно было бы отнести к пункту обработки данных, если бы процесс не был столь уникальным и не требовал разработки специальных средств для отображения обработанных данных. На этом этапе осуществляется соотнесение параметров исследованного участка горного массива и его ориентации в пространстве (привязка к координатным осям, пикетам и другим отметкам).

Финальным этапом процедуры является интерпретация результатов и, собственно, сам прогноз [19—21]. Если все предыдущие операции не требуют от пользователя глубоких знаний предмета и выполняются в автоматическом или полуавтоматическом режимах, то интерпретация требует наличия высококвалифицированных кадров (геологов, сейсмиков, геофизиков), одновременно владеющих

современными методами алгоритмизации и программирования, что само по себе является далеко непростой персонизированной специальной образовательной задачей.

Необходимо также иметь в виду, что в соответствии с правилами промышленной безопасности, геомониторинг должен осуществляться в непрерывном режиме. Процесс автоматического получения и обработки данных несомненно упрощает задачу специалиста, выполняющего прогноз, но не исключает необходимости постоянного наблюдения за результатами обработки, отображаемыми на мониторе.

Еще одной достаточно серьезной проблемой непрерывного геомониторинга с технической точки зрения, являются большие объемы данных, накапливаемые при работе системы, которые необходимо анализировать, архивировать и в ряде случаев, пересылать через глобальные сети заинтересованным людям.

С учетом вышесказанного становится актуальной задача предварительного экспресс анализа данных, заключающегося в подготовке набора характерных точек, на которые следует обратить внимание в первую очередь в рамках исследуемой части горного массива, что в итоге позволяет решить сразу несколько задач:

• наличие такого набора позволяет упростить обучение оператора /специалиста, работающего с системой визуализации и оценивающего состояние горного массива и уровень исходящей от него опасности;

• упрощает визуальный анализ обработанных данных и прогноз, снижая временные затраты ответственного за прогноз специалиста (геолога, геофизика);

• число таких точек, в сравнении с сейсмоданными, имеют значительно меньший объем, это позволяет без значительных затрат их хранить, анализировать и передавать в иные системы.

Одним из подходов к решению задачи экспресс анализа сейсмоданных может быть предложен кластерный анализ, результатом которого является нахождение центров кластеризации, которые могут быть использованы в качестве характерных точек.

Далее представим процедуру и результаты исследования применимости различных алгоритмов кластеризации для анализа сейсмоданных, необходимых и достаточных для решения конкретных задач функционала геоинформационной панели в системе оперативного контроля и прогноза возникновения и развития опасных динамических процессов в массиве.

В результате обработки сейсмоданных по методике [3], пользователю доступны несколько параметров, характеризующих исследуемый участок горного массива:

• *Stress* — напряженность массива (или горное давление). Относительный параметр, измеряется в условных единицах от 0 до 10, где 0 — означает декомпрессию, 2 — нормальное литостатическое давление для данных условий, а 10 — максимальную компрессию;

• *Water* — вероятность флюидопроявления. Это параметр позволяющий определить местоположение и вероятность нахождения флюида (газа / воды), измеряется в процентах от 0 до 100, где 100 — наибольшая вероятность нахождения флюида в отмеченной зоне, а 0 — минимальная вероятность;

• V_p — средняя скорость распространения продольных волн в среде, измеряется в м/с и принимает значения в зависимости от типа среды и ее состояния. Так для угольного массива средняя скорость продольных волн составляет, примерно, 2000 м/с;

• V_s — средняя скорость распространения поперечных волн в среде измеряется в м/с и принимает значения в

зависимости от типа среды и ее состояния;

• V_p/V_s — отношение скоростей, измеряется в условных единицах;

• *Poisson* — коэффициент Пуассона. Параметр принимает значение в диапазоне от 0 до 0,5, где 0 — это абсолютно хрупкие материалы, а 0,5 — абсолютно несжимаемые;

• *Young* — значения модуля Юнга (модуля упругости), измеряется в МПа;

• *Category* — категории устойчивости массива по шкале Бениявского. Параметр принимает значения от 1 до 5, где 1 — устойчивый массив, а 5 — неустойчивый, требующий максимального крепления и осторожности при разработке.

Методика предусматривает применение восьми сейсмодатчиков, рассредоточенных в пространстве горной выработки и позволяет исследовать массив размером (В×Ш×Г), м: 25×25×100. Дискретность сетки значений параметров составляет 1 м.

Предварительно выполненные работы

Нами был проведен сравнительный анализ различных алгоритмов кластеризации данных, получаемых от системы «Микон-ГЕО». Для сравнения алгоритмов были подготовлены выборки, содержащие координаты точек исследованной части горного массива и значения технологических параметров в них. Для оценки алгоритмов данные были предварительно размечены — в них, на основании экспертной оценки, были выбраны точки, на которые в первую очередь следует обратить внимание при их анализе специалистами. Результатом работы алгоритма кластеризации является набор центров кластеризации. Будем считать, что центр кластеризации совпадает с экспертной выборкой, если Евклидово расстояние между центром и точкой выборки составляет не более 2 м, данная величина определяется точностью измерительной аппаратуры.

Для оценки качества алгоритмов использовались следующие метрики:

• количество точек кластеризации, попавших в экспертную выборку — $c_к$;

• количество точек из экспертной выборки, обнаруженных алгоритмом кластеризации — c_m;

• индекс *Calinski-Harabasz* [5] — *VRC*;

• индекс *Davies-Bouldin* [6] — *DB*;

• силуэтный коэффициент [7] — *SWC*.

В качестве технологических параметров были выбраны *Stress* и *Water*. Количество точек, отмеченных экспертом для признака *Stress*, составляет двенадцать, для признака *Water* — три.

Общая схема исследования алгоритмов кластеризации представлена на рис. 1.

На первом этапе производится предобработка данных — масштабирование. Для приведения признаков к одному масштабу используем стандартизацию. Эксперименты с применением других подходов к масштабированию (нормализация и т.п.) показали, что они не оказывают существенного влияния на результат.

На втором этапе выполняется непосредственно кластеризация данных с использованием одного из выбранных алгоритмов. Проведен сравнительный анализ трех алгоритмов: *k-means* [8], *Mean shift* [9] и *DBSCAN* [10]. Для каждо-

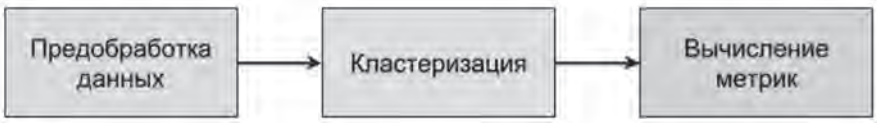

Рис. 1. Схема исследования алгоритмов кластеризации
Fig. 1. Research procedure of clustering algorithms

го алгоритма был предварительно выбран диапазон изменения гиперпараметров, оказывающих наибольшее влияние на результат работы алгоритма. Гиперпараметры, не попавшие в вариативную группу, установлены в константные значения. Количество запусков алгоритма определяется числом сочетаний различных значений модифицируемых гиперпараметров.

На третьем этапе вычисляются метрики. Список метрик приведен выше. Дополнительно для алгоритма *k-means* использовался «метод локтя» [11], основанный на анализе внутрикластерных искажений.

Далее произведен анализ применимости различных алгоритмов кластеризации при обработке сейсмических данных и определены параметры, характеризующие состояние горного массива.

Исследование применимости алгоритма *k-means* для анализа сейсмических данных

Алгоритм *k-means* является одним из наиболее простых и популярных алгоритмов кластеризации на сегодняшний день и относится к классу алгоритмов на основе прототипов. Каждый кластер в рамках алгоритма *k-means* определяется его центроидом — средним значением подобных точек. У данного алгоритма существует большое число вариаций [16, 17], которые могут быть использованы.

K-means имеет ряд недостатков, делающие его мало пригодным для кластеризации данных из системы «Микон-ГЕО», а именно: он плохо работает с кластерами сложной формы, размера и плотности, неустойчив к шумам, чувствителен к выбору первоначального распределения кластеров, но при этом работает быстро, поэтому был выбран для оценки нижней границы наших возможностей. Для снижения влияния конфигурации первоначального распределения кластеров используем модификацию этого алгоритма *k-means++* [12], которая дает лучший результат при прочих равных условиях.

Задание гиперпараметров для алгоритма *k-means*

Зададимся следующими значениями гиперпараметров алгоритма:

- количество кластеров: от 2-х до 20-ти, шаг изменения: 1;
- метод начальной расстановки центров кластеризации: *k-means++*;
- количество различных расстановок центров кластеризации: 10;
- максимальное количество итераций при прогоне алгоритма: 300.

В качестве верхней границы количества кластеров выбрано число 20, потому что на практике редко встречается количество областей, требующих большие значения.

Использование алгоритма *k-means* для кластеризации данных с признаком *Stress*

По полученным значениям внутрикластерных искажений невозможно определить оптимальное количество кластеров: нет явно выраженной точки, где искажения начинают увеличиваться быстрее по сравнению с предыдущими значениями (см. рис. 2).

На основе полученных данных вычислены заданные нами метрики. В табл. 1 приведены результаты — параметры алгоритма, при которых метрики принимают оптимальные значения. Там же указано количество точек из экспертной выборки, которые обнаруживает алгоритм при настройках, соответствующих данному значению метрики.

Алгоритм кластеризации *k-means* практически бесполезен для обработки данных от системы «Микон-ГЕО». При количестве кластеров 11, 12 или 13 он находит не более одной точки, отмеченной

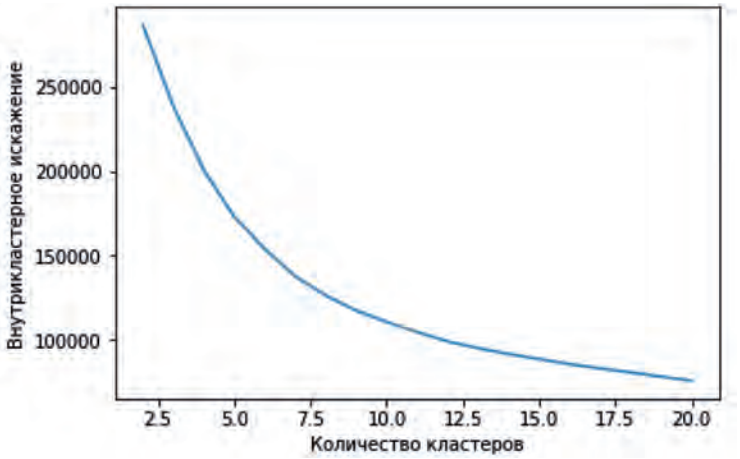

Рис. 2. Величина внутрикластерных искажений для различного количества кластеров (признак Stress, алгоритм k-means)

Fig. 2. Intracluster distortions for different number of clusters (Stress attribute, k-means algorithm)

Таблица 1

Численные значения метрик и соответствующее количество кластеров для признака Stress (алгоритм k-means)

Numerical values of metrics and the related number of clusters for attribute Stress (algorithm k-means)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	1	11	1
c_m	1	11	1
VRC	23 766,779	5	0
DB	1,124	7	0
SWC	0,237	12	1

экспертом. Увеличение количества кластеров, а также изменение других гиперпараметров алгоритма (количество различных расстановок центров кластеризации и максимальное количество итераций при прогоне алгоритма) не улучшили результат. На рис. 3 представлено расположение центров кластеризации найденных алгоритмом *k-means*, при количестве кластеров, равном 11.

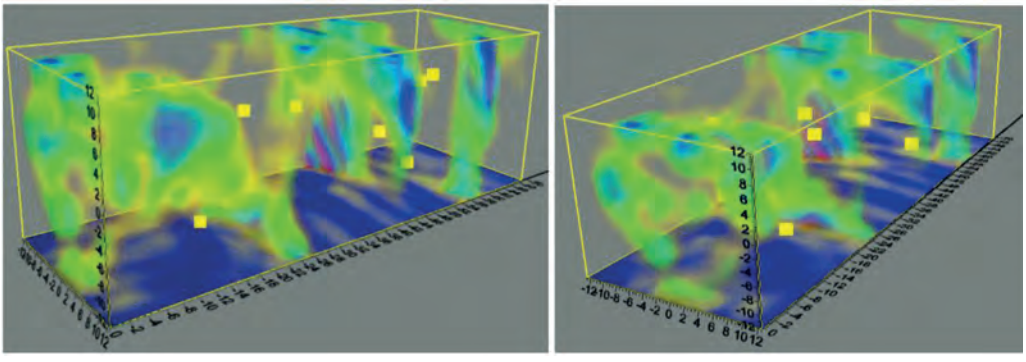

Рис. 3. Алгоритм k-means. Признак Stress. Количество кластеров 11

Fig. 3. Algorithm k-means. Attribute Stress. Number of clusters—11

Использование алгоритма *k-means* для кластеризации данных с признаком *Water*

Для признака *Water* «метод локтя» также не работает. По полученному графику зависимости внутрикластерных искажений от количества кластеров (рис. 4) невозможно определить оптимальное их (кластеров) количество.

Оптимальные значения метрик и соответствующие им параметры алгоритма *k-means* для признака *Water* представлены в табл. 2.

Для работы с признаком *Water* также, как и в случае со *Stress*, алгоритм *k-means* не может быть использован. Получен аналогичный результат, и улучшить его практически не удалось. Расположение центров кластеризации для признака *Water* (лучший результат по метрикам c_k и c_m), представлено на рис. 5.

Исследование применимости алгоритма *Mean shift* для анализа сейсмических данных

Алгоритм *Mean shift* является плотностным непараметрическим алгоритмом кластеризации. Количество кластеров при его настройке, в отличии от алгоритма *k-means*, не задается, оно определяется в процессе работы. Идея *Mean shift* заключается в поиске моды — максимума функции плотности вероятности. В процессе работы центры кластеризации смещаются по направлению к максимальной плотности. Основным гиперпараметром данного алгоритма являет-

Таблица 2

Численные значения метрик и соответствующее количество кластеров для признака Water (алгоритм k-means)

Numerical values of metrics and the related number of clusters for attribute Water (algorithm k-means)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	1	12	1
c_m	1	12	1
VRC	26 004,799	8	0
DB	1,057	8	0
SWC	0,255	9	0

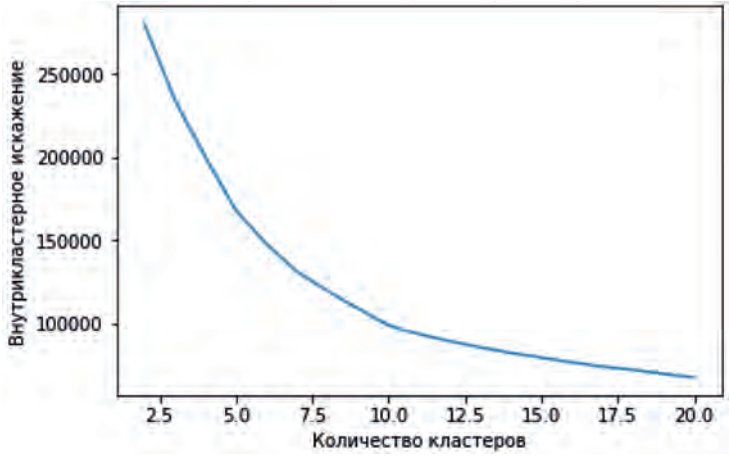

Рис. 4. Величина внутрикластерных искажений для различного количества кластеров (признак Water, алгоритм k-means)

Fig. 4. Intracluster distortions for different number of clusters (Water attribute, k-means algorithm)

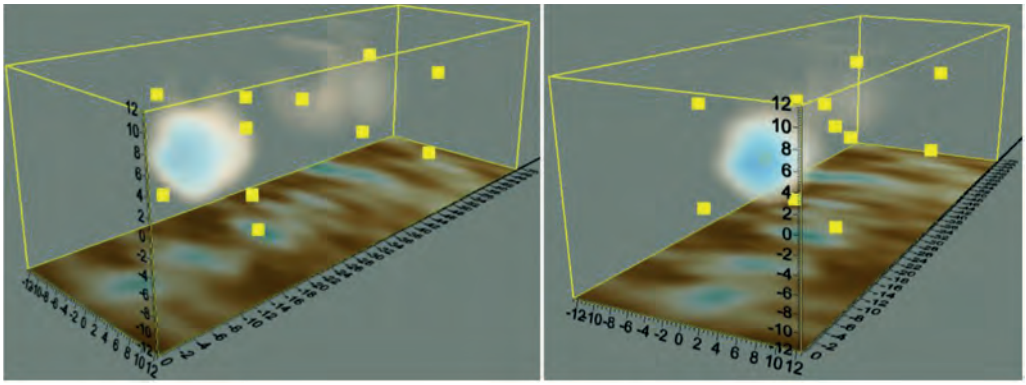

Рис. 5. Алгоритм k-means. Признак Water. Количество кластеров 12
Fig. 5. Algorithm k-means. Attribute Water. Number of clusters—12

ся величина ядра, которая, фактически, определяет размеры и количество кластеров.

Данный алгоритм не требует предварительного задания форм кластеров и может работать в произвольных признаковых пространствах. Его основной недостаток заключается в сложности выбора размера ядра, которое является определяющим гиперпараметром алгоритма.

Задание гиперпараметров для алгоритма *Mean shift*

Предварительные исследования показали, что при величине ядра меньше, чем 0,8 количество кластеров на рассматриваемых данных становится больше 20, что в нашем случае является верхней оценкой. Если величина ядра становится больше 1,3, то количество кластеров становится меньше 2-х, а это нижняя оценка. Поэтому указанный гиперпараметр подвергался изменению в диапазоне от 0,8 до 1,3 с шагом 0,5.

Использование алгоритма *Mean shift* для кластеризации данных с признаком *Stress*

В табл. 3 представлены оптимальные значения метрик для данного алгоритма кластеризации.

Качество алгоритма кластеризации *Mean shift*, лучше, по сравнению с *k-means*. Анализ полученных результатов кластеризации показал, что при величине ядра более 1,0 все центры кластеризации начинаю располагаться на расстоянии более 8 м от тех, что отмечены экспертом.

Таблица 3

Численные значения метрик и соответствующее количество кластеров для признака Stress (алгоритм Mean shift)
Numerical values of metrics and the related number of clusters for attribute Stress (algorithm Mean shift)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	3	20 (размер ядра = 0,85)	3
c_m	3	20 (размер ядра = 0,85)	3
VRC	20 075,953	2 (размер ядра = 1,2)	0
DB	1,277	23 (размер ядра = 0,8)	3
SWC	0,193	15 (размер ядра = 0,9)	1

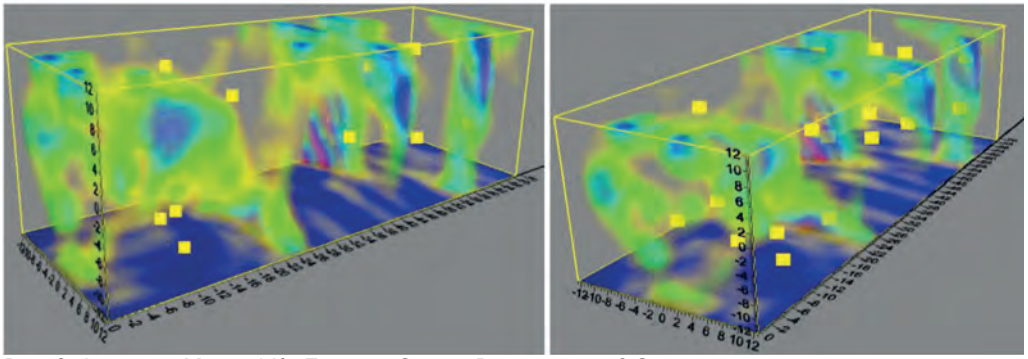

Рис. 6. Алгоритм Mean shift. Признак Stress. Размер ядра 0,8
Fig. 6. Algorithm Mean shift. Attribute Stress. Nucleus size—0.8

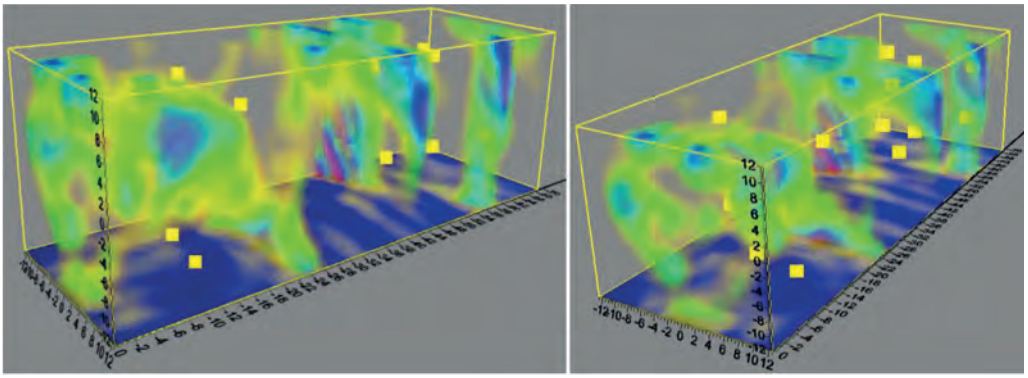

Рис. 7. Алгоритм Mean shift. Признак Stress. Размер ядра 0,85
Fig. 7. Algorithm Mean shift. Attribute Stress. Nucleus size—0.85

Несмотря на то, что максимальное количество найденных точек из экспертной выборке равно трем, среди центров кластеризации при размере ядра меньше 0,9 есть точки находящиеся на удалении не более 3 м, что близко к погрешности измерительной аппаратуры. Наложение полученных результатов на данные сейсмических наблюдений может помочь в их анализе.

На рис. 6 и 7 представлены расположения центров кластеризации, получен-

Таблица 4

Численные значения метрик и соответствующее количеству кластеров для признака Water (алгоритм Mean shift)
Numerical values of metrics and the related number of clusters for attribute Water (algorithm Mean shift)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	3	20 (размер ядра = 0,80)	2
c_m	2	20 (размер ядра = 0,80)	2
VRC	22 813,589	11 (размер ядра = 1,0)	1
DB	1,171	11 (размер ядра = 1,0)	1
SWC	0,232	11 (размер ядра = 1,0)	1

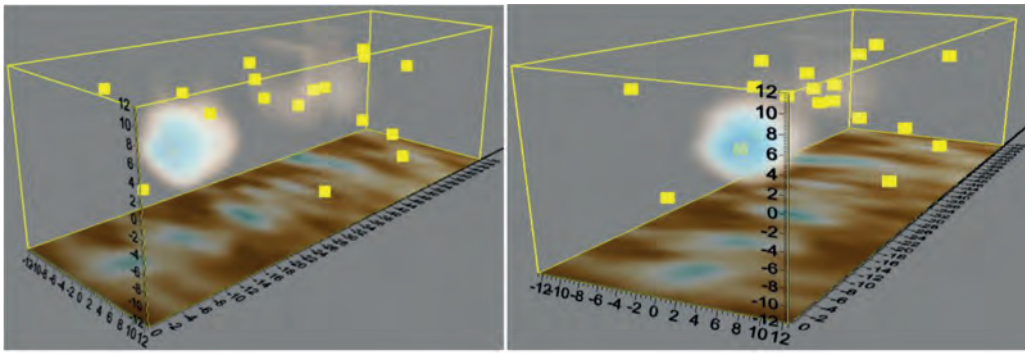

Рис. 8. Алгоритм Mean shift. Признак Water. Размер ядра 0,8
Fig. 8. Algorithm Mean shift. Attribute Water. Nucleus size—0.8

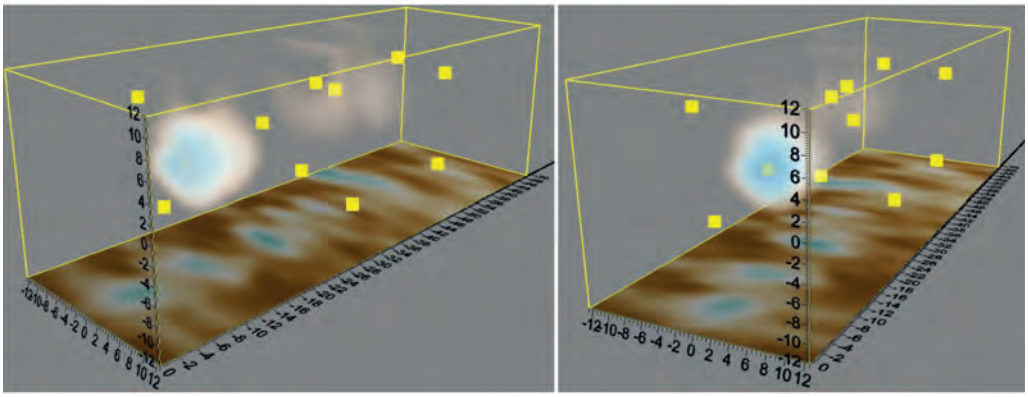

Рис. 9. Алгоритм Mean shift. Признак Water. Размер ядра 1,0
Fig. 9. Algorithm Mean shift. Attribute Water. Nucleus size—1.0

ные с помощью алгоритма *Mean shift*, для размеров ядра 0,8 (лучшее значение по метрике *DB*) и 0,85 (лучшее значение по метрикам c_k и c_m).

Использование алгоритма *Mean shift* для кластеризации данных с признаком *Water*

Численные значения метрик, вычисленные для результатов, полученных с помощью алгоритма *Mean shift* для признака *Water* представлены в табл. 4.

Несмотря на то, что для признака *Water* алгоритм *Mean shift* находит две из трех точек, отмеченных экспертом, среди 20-ти полученных кластеров их будет трудно найти без проведения дополнительного анализа. Центры кластеризации для размера ядра равному 0,8 показаны на рис. 8, для размера ядра 1,0 — на рис. 9.

Исследование применимости алгоритма *DBSCAN* для анализа сейсмических данных

DBSCAN является плотностным алгоритмом кластеризации. Он также как и *Mean shift* не требует предварительно задания кластеров, их количество определяется в процессе работы и зависит от гиперпараметров алгоритма и обрабатываемых данных. *DBSCAN* может находить кластера произвольной формы, устойчив к выбросам и хорошо работает на больших объемах данных. Из недостатков можно отметить, что алгоритм не очень хорошо работает на данных, плотность элементов в которых сильно отли-

чается, требует предварительного масштабирования признаков [18].

Задание гиперпараметров для алгоритма *DBSCAN*

Два основных гиперпараметра алгоритма — это минимальное количество точек, которые образуют плотную область (*minPts*) и радиус окрестности (*esp*). Минимальное значение *minPts* = 3, в качестве верхней оценки можно ориентироваться на величину *minPts* = 2**dim* [13], где *dim* — это размерность признакового пространства. В нашем случае размерность равна четырем поэтому максимальное значение *minPts* = 8. Таким образом, для величины *minPts* был взят ряд значений 3—8. Для выбора минимального значения *esp* использовался график зависимости расстояния до *k*-го соседа от количества точек, расстояние до *k*-го соседа у которых меньшее. Максимальное значение *esp* определяется по итоговому количество кластеров: их должно быть не меньше трех.

Использование алгоритма *DBSCAN* для кластеризации данных с признаком *Stress*

Построим график для оценки минимального значения *esp* (рис. 10). Из графика видно, что начиная с определенного количества точек, у которых расстояние до *k*-го соседа меньше заданного радиуса окрестности, происходит резкий рост расстояния (0,22). Эта вели-

Таблица 5

Численные значения метрик и соответствующее количество кластеров для признака Stress (алгоритм DBSCAN)

Numerical values of metrics and the related number of clusters for attribute Stress (algorithm DBSCAN)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	16	73 (*eps* = 0,22, *minPts* = 8)	9
c_m	9	73 (*eps* = 0,22, *minPts* = 8)	9
VRC	245,797	16 (*eps* = 0,26, *minPts* = 8)	5
DB	0,455	4 (*eps* = 0,35, *minPts* = 3)	2
SWC	0,417	3 (*eps* = 0,35, *minPts* = 7)	1

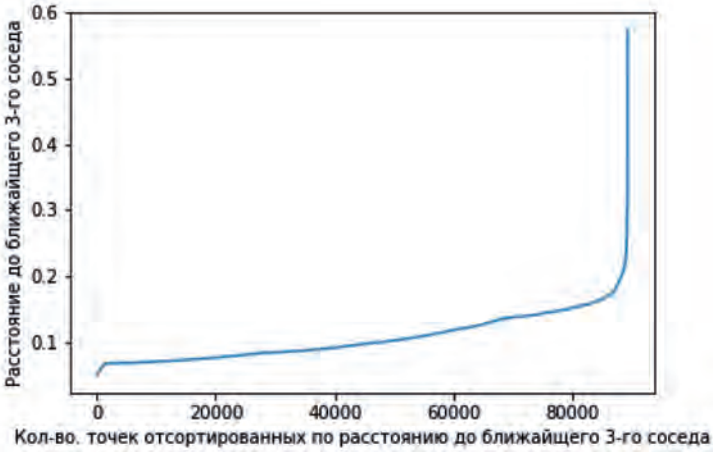

Рис. 10. График для оценки величины радиуса окрестности для признака Stress
Fig. 10. Neighborhood radius evaluation chart for attribute Stress

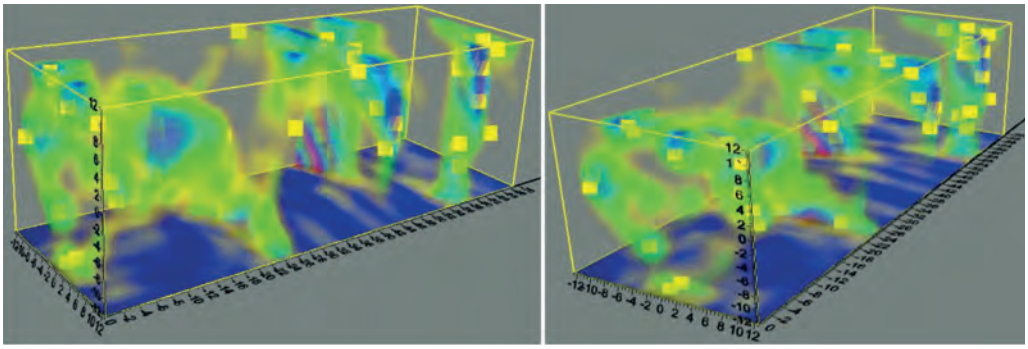

Рис. 11. Алгоритм DBSCAN. Признак Stress. Значения параметров: eps = 0,22, minPts = 8
Рис. 11. Algorithm DBSCAN. Attribute Stress. Values of parameters: eps = 0,22, minPts = 8

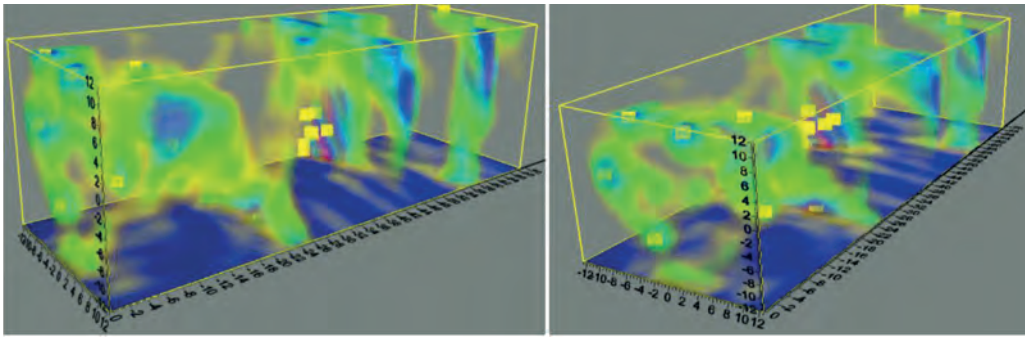

Рис. 12. Алгоритм DBSCAN. Признак Stress. Значения параметров: eps = 0,26, minPts = 8
Рис. 12. Algorithm DBSCAN. Attribute Stress. Values of parameters: eps = 0,26, minPts = 8

чина используется в качестве опорной. Максимальное значение для параметра *Stress* по предварительно проведенным экспериментам составляет 0,35.

В табл. 5 представлены оптимальные значения метрик для данного алгоритма кластеризации.

Анализ результатов кластеризации алгоритмом *DBSCAN* показал, что помимо центров кластеризации, совпадающих (с заданной погрешностью) с точками из экспертной выборки, есть довольно много центров, которые находятся на небольшом от них отдалении (в радиусе 4—6 м). Эти центры образуют скопления, позволяющие выявить области, на которые следует обратить внимание при анализе. Также необходимо отметить, что количество ложных срабатываний, когда центр кластеризации находится на удалении более 6—7 м от точки, отмеченной экспертом, составляет, как правило, не более 20% процентов от общего количества центров. В итоге по метрике *VRC* можно получить оптимальный результат с подтверждением его пригодности в процессе дальнейших практических исследований.

На рис. 11 и 12 представлены расположения центров кластеризации, найденные алгоритмом *DBSCAN*, при (*eps* = = 0,22, *minPts* = 8) и (*eps* = 0,26, *minPts* = = 8).

Первая группа соответствует лучшему варианту согласно метрикам $c_к$ и c_m, вторая выбрана по метрике *VRC*.

Использование алгоритма *DBSCAN* для кластеризации данных с признаком *Water*

Для оценки *esp* построен график (рис. 13), аналогичный тому, что был соз-

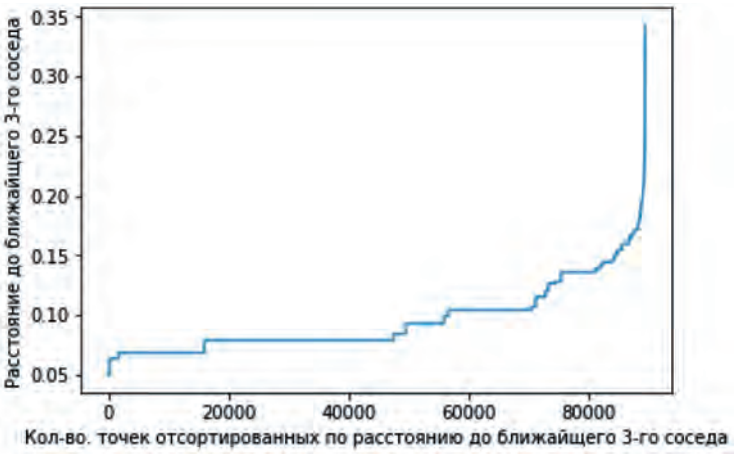

Рис. 13. График для оценки величины радиуса окрестности для признака Water
Fig. 13. Neighborhood radius evaluation chart for attribute Water

дан для признака *Stress*. Из графика определяется минимальное значение радиуса окрестности по методике, изложенной для признака *Stress*, (0,18). Максимальное значение *esp* = 0,29. Полученные численные значения метрик представлены в табл. 6. Результаты кластеризации данных с признаком *Water*

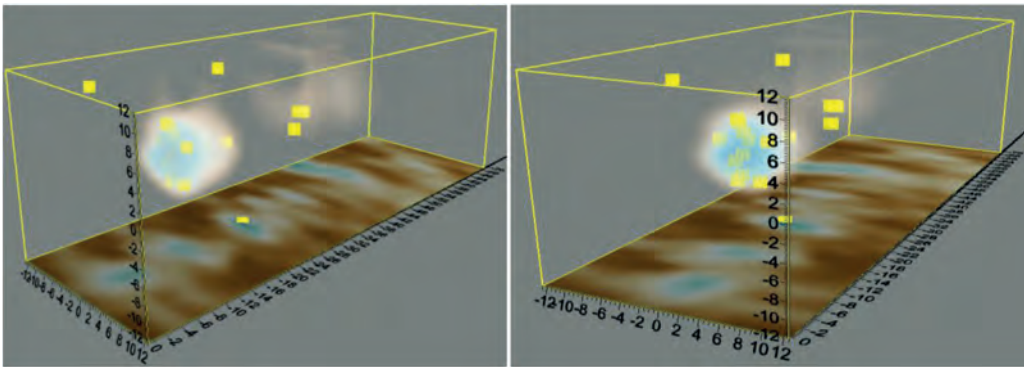

Рис. 14. Алгоритм DBSCAN. Признак Water. Значения параметров: eps = 0,24, minPts = 8
Рис. 14. Algorithm DBSCAN. Attribute Water. Values of parameters: eps = 0,24, minPts = 8

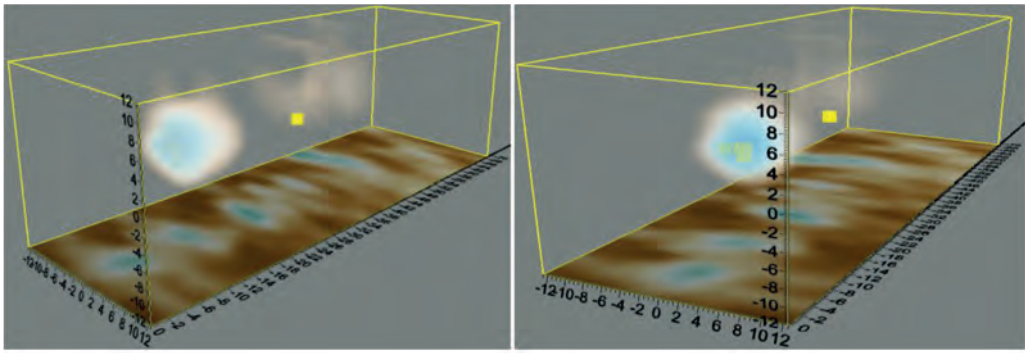

Рис. 15. Алгоритм DBSCAN. Признак Water. Значения параметров: eps = 0,29, minPts = 7
Рис. 15. Algorithm DBSCAN. Attribute Water. Values of parameters: eps = 0,29, minPts = 7

Таблица 6

Численные значения метрик и соответствующее количество кластеров для признака Water (алгоритм DBSCAN)

Numerical values of metrics and the related number of clusters for attribute Water (algorithm DBSCAN)

Метрика	Значение	Количество кластеров	Количество точек из экспертной выборки
c_k	14	77 (*eps* = 0,2, *minPts* = 4)	2
c_m	2	24 (*eps* = 0,24, *minPts* = 8)	2
VRC	551,665	6 (*eps* = 0,29, *minPts* = 7)	1
DB	0,368	6 (*eps* = 0,29, *minPts* = 7)	1
SWC	0,733	6 (*eps* = 0,29, *minPts* = 7)	1

алгоритмом *DBSCAN* аналогичны тем, что получены для признака *Stress*: вместе с центрами, совпадающими с экспертной выборкой. Есть центры, формирующие скопления возле них.

На рис. 14 представлены распределения центров кластеризации при (*eps* = = 0,24, *minPts* = 8), что соответствует лучшему результату по метрике c_m. На рис. 15 — оптимальный результат по метрикам *VRC*, *DB* и *SWC*.

Ввиду того, что из анализируемых данных нельзя выборочно исключать объекты, а их количество в наборе составляет 89 250, при этом у каждого объекта имеется четыре признака, в анализ не попал ряд алгоритмов, таких как Affinity Propagation [14], Spectral Clustering [15], так как их сложность по памяти составляет $O(N^2)$ (объем памяти, который требуется для алгоритма, с ростом объема входных данных, в худшем случае растет по квадратичному закону (с точностью до константы).

По полученным результатам можно сделать вывод, что наиболее результативным алгоритмом кластеризации для экспресс анализа сейсмических данных в системе «Микон-ГЕО» является *DBSCAN*. Центры кластеризации, получаемые с помощью него, в значительной своей массе ложатся либо в точках, отмеченных экспертом, либо рядом с ними на удалении 5—6 м.

Выводы

Из исследованных алгоритмов кластеризации наиболее подходящим для решения задачи экспресс анализа сейсмических данных является *DBSCAN*.

Использование рассмотренной в статье методологии позволяет применить алгоритмы кластеризации не только для задач получения и обработки сейсмических данных, но и для решения задач их визуализации и интерпретации образов, что в итоге позволит получить оперативные значения параметров горного массива, необходимые для решения задачи прогноза опасного состояния системы «градиент горного давления — газовый поток».

СПИСОК ЛИТЕРАТУРЫ

1. *Приказ* Ростехнадзора от 19.11.2013 № 550 «Об утверждении Федеральных норм и правил в области промышленной безопасности «Правила безопасности в угольных шахтах».

2. *Официальный* сайт компании ООО «ИНГОРТЕХ» [электронный ресурс], http://ingortech.ru/produktsiya/statsionarnye-sistemy/paragraf-41-pb/kontrol-gornogo-massiva-p-41-pb.

3. *Патент*, https://patents.google.com/patent/US6498989.

4. *Лапин Э. С.*, *Писецкий В. Б.*, *Бабенко А. Г.*, *Патрушев Ю. В.* «Микон-ГЕО» — система оперативного обнаружения и контроля состояния зон развития опасных геогазодинамических яв-

лений при разработке месторождений полезных ископаемых подземным способом // Безопасность труда в промышленности. — 2012. — № 4. — С. 18—22.

5. Calinski T., Harabasz J. A dendrite method for cluster analysis // Communications in Statistics. 1974, vol. 3, pp. 1—27.

6. *Davies D. L.*, *Bouldin D. W.* A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1. 1979, pp. 224—227.

7. *Rousseeuw P. J.* Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis // Computational and Applied Mathematics 20. 1987, pp. 53—65.

8. *Lloyd S.* Least square quantization in PCM's. IEEE Transactions on Information Theory. vol. 28, pp. 129—137.

9. *Comaniciu D., Meer P.* Mean shift. A robust approach toward feature space analysis // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.

10. *Ester M., Kriegel H. P., Sander J., Xu X.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press., 1996, pp. 226—231.

11. *Sebastian Raschka*. Python Machine Learning, 1st Edition. Packt Publishing Ltd. 2015, 454 p.

12. *Arthur D., Vassilvitskii S.* k-means++: The advantages of careful seeding / Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics. 2007, pp. 1027—1035.

13. *Sander J., Ester M., Kriegel H. P., Xu X.* Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications // Data Mining and Knowledge Discovery. Berlin: Springer-Verlag. 1998 (2), pp. 169—194.

14. *Brendan J. F.*, *Delbert D.* Clustering by passing messages between data points // Science. 2007. No 15, pp. 972—979.

15. *Ng A., Jordan M., Weiss Y.* On spectral clustering: analysis and an algorithm / NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001, pp. 849—856.

16. *Amorim R. C.*, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors // Information Sciences. 324. — 2015. — 126—145.

17. *Hamerly G., Drake J.* Accelerating Lloyd's algorithm for k-means clustering / Partitional Clustering Algorithms. 2015, pp. 41—78.

18. *Campello R. J. G. B., Moulavi D, Zimek A., Sander J.* Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection // ACM Transactions on Knowledge Discovery from Data. 2015. vol 10, pp. 5:1—5:51.

19. *Писецкий В. Б.*, *Лапин С. Э.*, *Зудилин А. Э.*, *Патрушев Ю. В.*, *Шнайдер И. В.* Методика и результаты промышленного применения системы сейсмического контроля состояния горного массива «Микон-ГЕО» в процессе подземной разработки рудных и угольных месторождений // Проблемы недропользования. — 2016. — С. 58—64.

20. *Писецкий В. Б.*, *Robert Huang*, *Патрушев Ю. В.*, *Зудилин А. Э.*, *Шнайдер И. В.*, *Широбоков М. П.* Результаты испытаний сейсмических систем контроля состояния устойчивости горного массива в процессах строительства автодорожных тоннелей в Китае // Добывающая промышленность. — 2017. — № 2 (06). — С. 108.

21. *Писецкий В. Б.*, *Власов В. В.*, *Черепанов В. П.*, *Абатурова И. В.*, *Зудилин А. Э.*, *Патрушев Ю. В.*, *Александрова А. В.* Прогноз устойчивости горного массива на основе метода сейсмической локации в процессах строительства подземных сооружений // Инженерные изыскания. — 2014. — № 9—10. — С. 46—51.

22. *Яковлев Д. В.*, *Лазаревич Т. И.*, *Поляков А. Н.* Принципы построения систем контроля состояния горного массива на основе анализа актуальных рисков осуществления подземной добычи // Горный информационно-аналитический бюллетень. — 2015. — СВ 7. — С. 471—481. ГИАБ

REFERENCES

1. *Prikaz Rostekhnadzora ot 19.11.2013 № 550 «Ob utverzhdenii Federal'nykh norm i pravil v oblasti promyshlennoy bezopasnosti «Pravila bezopasnosti v ugol'nykh shakhtakh»* [Approval of

Federal Code on Industrial Safety: Safety Regulations for Coal Mines. Rostekhnadzor Order No. 550 dated November 19, 2013]. [In Russ].

2. Official site OOO «INGORTEH», http://ingortech.ru/produktsiya/statsionarnye-sistemy/paragraf-41-pb/kontrol-gornogo-massiva-p-41-pb.

3. Patent US6498989, https://patents.google.com/patent/US6498989.

4. Lapin E.S., Pisetskiy V.B., Babenko A.G., Patrushev Yu.V. Mikon-GEO—on-line detection and monitoring of hazardous geo-gas-dynamic event initiation and growth in underground mineral mining. *Bezopasnost' truda v promyshlennosti*. 2012, no 4, pp. 18—22. [In Russ].

5. Calinski T., Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*. 1974, vol. 3, pp. 1—27.

6. Davies D.L., Bouldin D.W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1*. 1979, pp. 224—227.

7. Rousseeuw P.J. *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis.* Computational and Applied Mathematics 20. 1987, pp. 53—65.

8. Lloyd S. Least square quantization in PCM's. *IEEE Transactions on Information Theory*. vol. 28, pp. 129—137.

9. Comaniciu D., Meer P. Mean shift. A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

10. Ester M., Kriegel H.P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland*, OR, AAAI Press., 1996, pp. 226—231.

11. Sebastian Raschka. *Python Machine Learning*, 1st Edition. Packt Publishing Ltd. 2015, 454 p.

12. Arthur D., Vassilvitskii S. k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics. 2007, pp. 1027—1035.

13. Sander J., Ester M., Kriegel H.P., Xu X. *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery*. Berlin: Springer-Verlag. 1998 (2), pp. 169—194.

14. Brendan J.F., Delbert D. Clustering by passing messages between data points. *Science*. 2007. No 15, pp. 972—979.

15. Ng A., Jordan M., Weiss Y. On spectral clustering: analysis and an algorithm. *NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems*: Natural and Synthetic. 2001, pp. 849—856.

16. Amorim R.C., Hennig C. *Recovering the number of clusters in data sets with noise features using feature rescaling factors*. Information Sciences. 324. 2015. 126—145.

17. Hamerly G., Drake J. Accelerating Lloyd's algorithm for k-means clustering. *Partitional Clustering Algorithms*. 2015, pp. 41—78.

18. Campello R.J. G. B., Moulavi D, Zimek A., Sander J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*. 2015. vol 10, pp. 5:1—5:51.

19. Pisetskiy V.B., Lapin S.E., Zudilin A.E., Patrushev Yu.V., Shnayder I.V. Procedure and commercial application results of Mikon-GEO seismic monitoring system in underground mining of ore and coal deposits. *Problemy nedropol'zovaniya*. 2016, pp. 58—64. [In Russ].

20. Pisetskiy V.B., Robert Huang, Patrushev Yu.V., Zudilin A.E., Shnayder I.V., Shirobokov M.P. Test data of seismic monitoring systems for rock mass stability in construction of highway tunnels in China. *Dobyvayushchaya promyshlennost'*. 2017, no 2 (06), pp. 108. [In Russ].

21. Pisetskiy V.B., Vlasov V.V., CHerepanov V.P., Abaturova I.V., Zudilin A.E., Patrushev Yu.V., Aleksandrova A.V. Rock mass stability prediction based on seismic location method in underground construction. *Inzhenernye izyskaniya*. 2014, no 9—10, pp. 46—51. [In Russ].

22. Yakovlev D.V., Lazarevich T.I., Polyakov A.N. Principles of constructing rock mass monitoring systems based on analysis of actual risks in underground mineral mining. *Gornyy informatsionno-analiticheskiy byulleten'*. 2015. Special edition 7, pp. 471—481. [In Russ].

ИНФОРМАЦИЯ ОБ АВТОРАХ

Абдрахманов Марат Ильдусович — канд. техн. наук,
главный специалист, e-mail: marat-ab@mail.ru,
ООО «Информационные горные технологии»,
Лапин Сергей Эдуардович[1] — канд. техн. наук,
старший научный сотрудник, e-mail: sergei.l@bk.ru,
Шнайдер Иван Владимирович[1] — аспирант,
[1] Уральский государственный горный университет.
Для контактов: Абдрахманов М.И., e-mail: marat-ab@mail.ru.

INFORMATION ABOUT THE AUTHORS

M.I. Abdrakhmanov, Cand. Sci. (Eng.), Chief Specialist,
e-mail: marat-ab@mail.ru,
LLC «Information mining technologies», Ekaterinburg, Russia,
S.E. Lapin[1], Cand. Sci. (Eng.), Senior Researcher, e-mail: sergei.l@bk.ru,
I.V. Shnayder[1], Graduate Student,
[1] Ural State Mining University, 620144, Ekaterinburg, Russia.
Corresponding author: M.I. Abdrakhmanov, e-mail: marat-ab@mail.ru.

РУКОПИСИ, ДЕПОНИРОВАННЫЕ В ИЗДАТЕЛЬСТВЕ «ГОРНАЯ КНИГА»

ОПЫТ ПРОВЕДЕНИЯ ВСКРЫВАЮЩИХ ВЫРАБОТОК В УСЛОВИЯХ МНОГОЛЕТНЕМЕРЗЛЫХ РОССЫПНЫХ МЕСТОРОЖДЕНИЙ СЕВЕРО-ВОСТОКА РФ

(№ 1185/06–19 от 17.05.2019; 11 с.)

Марков Валерий Степанович[1] — канд. техн. наук, доцент, e-mail: marko-valeri@mail.ru,
Петрова Любовь Владимировна[1] — старший преподаватель, e-mail: eL_Pi@mail.ru,
[1] Горный институт, Северо-Восточный федеральный университет им. М.К. Аммосова.

Приведен анализ опыта и особенности проходки вскрывающих выработок в условиях многолетнемерзлых россыпных месторождений Северо-Востока РФ. Рассмотрены способы проведения россыпных месторождений, приведено их описание. Установлено, что основной вскрывающей выработкой в условиях подземной разработки россыпей является наклонный ствол. Освещены вопросы техники и технологии проходки наклонных стволов в зависимости от горно-геологических и горнотехнических условий залегания, а также приведены достигнутые технико-экономические показатели.

Ключевые слова: многолетнемерзлые россыпные месторождения, вскрывающая выработка, наклонный ствол, скреперные лебедки, проходческий комбайн, ленточный конвейер, погрузочно-доставочные машины, буровзрывные работы, производительность труда.

THE EXPERIENCE OF MINING OF OPENING WORKING IN PERMAFROST PLACER DEPOSITS IN THE NORTH-EAST OF THE RUSSIAN FEDERATION

V.S. Markov[1], Cand. Sci. (Eng.), Assistant Professor, e-mail: marko-valeri@mail.ru,
L.V. Petrovala[1], Senior Lecturer, e-mail: eL_Pi@mail.ru,
[1] Mining Institute, North-Eastern Federal University named after M.K. Ammosov, 678000, Yakutsk, Russia.

The article presents an analysis of the experience and peculiarities of mining of opening workings in the conditions of permafrost placer deposits of the North-East of the Russian Federation. The methods of sinking of permafrost placers are considered, and their description is given. It is established that incline is the main opening working in the conditions of underground development of placers. The issues of technology of incline sinking are covered depending on the geological and mining conditions of occurrence, and there are the achieved technical and economic indicators.

Key words: permafrost placer deposits, opening working, incline, scraper winch, roadheading machine, extensible belt conveyer, load-haul-dumper, drilling and blasting, duty of labour.