УДК 004.832, 550.34

ПРИМЕНЕНИЕ МЕТОДОВ ОГРАНИЧЕННОЙ КЛАСТЕРИЗАЦИИ ДЛЯ ИССЛЕДОВАНИЯ ТЕХНОГЕННОЙ СЕЙСМИЧНОСТИ

А. А. Зуенко*, О. В. Фридман*, О. Г. Журавлева**

**Институт информатики и математического моделирования (обособленное подразделение Федерального исследовательского центра «Кольский научный центр Российской академии наук»)*
***Горный институт (обособленное подразделение Федерального исследовательского центра «Кольский научный центр Российской академии наук»)*

Поступила в редакцию 10.07.2019 г.

Аннотация. Целью работы является оценка применимости методов ограниченной кластеризации для исследования техногенной сейсмичности. В работе предлагается оригинальный метод ограниченной кластеризации, который в качестве базовой использует схему иерархической кластеризации (новые кластеры создаются путем объединения более мелких кластеров). В качестве меры различия используется расстояние между центроидами кластеров. Новизна метода заключается в возможности учитывать пользовательские ограничения на то, какие объекты обязательно должны попадать в один кластер, а какие не должны. Другими словами, в рамках предлагаемого подхода при отнесении объектов к одному или различным кластерам анализируются не только расстояния между объектами, но и значения их признаков. Предлагаемый метод позволяет останавливать процедуру кластеризации при нарушении пользовательских ограничений.

Проведенные исследования являются первоначальной попыткой оценить применимость методов ограниченной кластеризации для исследования техногенной сейсмичности и показать принципиальную возможность оценки с их помощью степени влияния горно-геологических факторов на сейсмоактивность участков массива горных пород. Исследованы различные комбинации стационарных и условно-стационарных влияющих факторов. Полученные результаты кластеризации можно использовать для предварительного выделения зон с различной сейсмоактивностью. Дальнейшее исследование должно производиться с учетом не только количества сейсмособытий, но и их энергии, расстояния между ними и др. Применение методов ограниченной кластеризации позволяет ускорить процесс вычислений за счет уменьшения количества альтернатив на каждом шаге кластеризации. Гарантированное выполнение пользовательских ограничений в рамках предлагаемого метода повышает доверие у экспертов предметной области к процедурам кластеризации. Полученные в ходе исследований результаты позволяют судить о перспективности дальнейшего развития методов ограниченной кластеризации для задач исследования техногенной сейсмичности.

Ключевые слова: ограниченная кластеризация, техногенная сейсмичность, кластерный анализ.

ВВЕДЕНИЕ

Постоянное техногенное воздействие на массив горных пород приводит к повышению сейсмической активности в пределах отрабатываемых месторождений. Для обеспечения безопасности и эффективности ведения горных работ необходимо проводить анализ сейсмической активности массива.

Сейсмические события на рудниках происходят неравномерно как в пространстве, так и во времени. Наблюдаются как отдельные события, так и группы (кластеры) сейсмических событий [1–3]. Причем кластеры сейсмособытий могут отличаться по таким характеристикам, как суммарная выделившаяся энергия, средняя выделившаяся энергия,

плотность распределения событий и др. Связано это с тем, что на сейсмичность массива влияет множество природных и техногенных факторов. И при различных сочетаниях влияющих факторов может наблюдаться различный уровень сейсмической активности.

Такими факторами являются геология и тектоника месторождения и прилегающего к нему района, геометрия и динамика горных работ [1], которые оказывают существенное влияние на напряженно-деформированное состояние массива горных пород и на проявление сейсмической активности. В связи с этим одним из подходов к анализу сейсмической активности и механизмов возникновения сейсмических событий при ведении горных работ является исследование влияющих факторов и определение возможных причин возникновения кластеров сейсмических событий.

Кластерный анализ зарегистрированных сейсмических событий позволяет разрабатывать классификации этих событий; проводить исследования схем группирования сейсмособытий; формулировать гипотезы возникновения сейсмических событий и осуществлять оценку выдвинутых гипотез.

В работе [4] предлагается для кластеризации применить метод *k-средних*, который позволяет выделить группы сейсмических событий и оценить их взаимосвязь с горно-геологическими факторами на примере высоконапряженного массива горных пород Кукисвумчоррского апатит-нефелинового месторождения. В общем случае с помощью данного метода строится ровно k различных кластеров, расположенных на возможно больших расстояниях друг от друга. Это является недостатком данного метода, поскольку, как правило количество кластеров, на которые могут разбиваться сейсмособытия изначально неизвестно.

Цель настоящей работы состоит в том, чтобы оценить применимость методов ограниченной кластеризации для исследования техногенной сейсмичности и показать принципиальную возможность оценки с их помощью степени влияния горно-геологических факторов на сейсмические события. Недостатком большинства существующих методов кластеризации является невозможность учитывать пользовательские ограничения на то, какие объекты обязательно должны попадать в один кластер, что подрывает доверие у экспертов предметной области к процедурам кластеризации, поскольку не всегда отнесение объекта в ближайший (по расстоянию) кластер является семантически корректной операцией. В рамках предлагаемого подхода при отнесении объектов к одному или различным кластерам анализируются не только расстояния между объектами, но и значения их признаков.

МАТЕРИАЛЫ И МЕТОДЫ

Прежде чем приступить к описанию разработанного метода, дадим начальное представление о задачах ограниченной кластеризации (Constrained Clustering) [5–8].

Рассмотрим набор данных из n объектов $O = \{o_1; \ldots; o_n\}$. Каждый объект описывается значениями p атрибутов, также называемых переменными. Обозначим o_{ij} значение j-го атрибута объекта o_i. Большинство алгоритмов кластеризации опираются не на анализ свойств объектов, а используют только таблицы расстояний между объектами.

Предположим, что имеется мера различия между любыми двумя объектами $o_i; o_j \in O$, которую обозначим d_{ij}. Мера различия d_{ij} обычно рассчитывается с использованием метрики расстояния, определенной в пространстве атрибутов. Метрика расстояния должна удовлетворять следующим свойствам:

1. Неотрицательность:

$$d_{ij} \geq 0; \forall i; j \in [1; n].$$

2. Идентичность (совместимость):

$$d_{ij} = 0; \forall i \in [1; n].$$

3. Симметрия:

$$d_{ij} = d_{ji}; \forall i; j \in [1; n].$$

4. Неравенство треугольника:

$$d_{ij} \leq d_{it} + d_{tj}; \forall i; j; t \in [1; n].$$

Мера различия должна удовлетворять первым трем свойствам, но может не удов-

летворить неравенству треугольника. В кластерном анализе наиболее популярной метрикой расстояния является Евклидово расстояние и квадрат Евклидова расстояния [9]. Пусть дано два объекта $o_i; o_j$, тогда евклидово расстояние между ними определяется как:

$$d_{ij} = \sqrt{\begin{matrix}\left(o_{i1} - o_{j1}\right)^2 + \left(o_{i2} - o_{j2}\right)^2 + \\ \ldots + \left(o_{ip} - o_{jp}\right)^2\end{matrix}}.$$

Квадрат Евклидова расстояния определяется как:

$$d_{ij} = \left(o_{i1} - o_{j1}\right)^2 + \left(o_{i2} - o_{j2}\right)^2 + \ldots + \left(o_{ip} - o_{jp}\right)^2.$$

Другой популярной мерой расстояния является Манхэттенское расстояние [9], определяемое как:

$$d_{ij} = \left|o_{i1} - o_{j1}\right| + \left|o_{i2} - o_{j2}\right| + \ldots + \left|o_{ip} - o_{jp}\right|.$$

Аналогично, обозначим меру сходства между двумя объектами o_i и o_j как s_{ij}.

Мера подобия широко используется в спектральной кластеризации. Как правило, рассчитывается Гауссова функция различия [10]: $s_{ij} = \exp\left(-\frac{d_{ij}}{2\sigma_i^2}\right)$, где σ_i – параметр.

Существуют и другие меры сходства, например нормализованная корреляция Пирсона, мера Жакара, мера коэффициента игральной кости и др. [11].

Большинство методов кластеризации можно разделить на две категории: иерархические методы и методы разбиений (partitioning-based clustering methods). Методы разбиений нацелены на поиск некоторого разбиения P, в то время как иерархическая кластеризация направлена на то, чтобы найти множество вложенных разбиений $P_1; P_2; \ldots P_l$ исходного множества O.

Определение 1. Пусть $C_1; C_2; \ldots; C_k$ – подмножества множества O. $\{C_1; C_2; \ldots; C_k\}$ является разбиением O на k кластеров, если: для всех $c \in \{1; 2; \ldots; k\}$, $C_c \neq \varnothing$; $\bigcup_c C_c = 0$; для всех $c \neq c'$, $C_c \bigcap C_{c'} = \varnothing$.

Многие методы разбиений зависят от выбора репрезентативной точки (типичного представителя) для каждого кластера, это может быть центроид или медоид. Если объекты являются векторами p количественных атрибутов, центроид кластера вычисляется путем усреднения векторов, принадлежащих этому кластеру.

Определение 2. Центроид m_c кластера C_c является точкой, определенной как:

$$\forall j \in \{1 \ldots p\}\left(m_c\right)_j = \frac{1}{\left|C_c\right|}\sum_{o_i \in C_c} o_{ij}.$$

Когда атрибуты объектов качественные, центроид не может быть рассчитан и обычно для каждого кластера рассчитывают медоид.

Определение 3. Медоид x_c кластера C_c – это объект кластера, среднее значение меры различия которого для всех объектов в кластере минимально:

$$x_c = \arg\min_{o_i \in C_c} \frac{1}{\left|C_c\right|}\sum_{x_i \in C_c} d_{ij}.$$

Заметим, что центроид может не принадлежать множеству O, тогда как медоид принадлежит множеству O.

Рассмотрим методы ограниченной кластеризации.

Для того чтобы лучше смоделировать задачу и снизить ее сложность могут быть добавлены ограничения. В этом случае задача кластеризации становится задачей ограниченной кластеризации, целью которой является получение кластеров, которые удовлетворяют пользовательским ограничениям. Пользовательские ограничения могут быть классифицированы следующим образом: 1) ограничения на кластеры (cluster-level constraints), указывающие требования к кластерам; 2) ограничения на объекты кластеров (instance-level constraints), уточняющие требования к парам конкретных объектов.

Ограничения на объекты кластеров – это ограничения на пары объектов. Существует два типа ограничений на объекты кластеров, введенные впервые в [11]. Это ограничения *must-link* и *cannot-link*. Ограничение *must-link* между двумя объектами $o_i; o_j$, обозначаемое как $ML(o_i; o_j)$, предписывает, что оба объекта o_i и o_j должны быть в одном кластере. Напротив, ограничение *cannot-link* между объектами o_i и o_j, обозначаемое как $CL(o_i; o_j)$, предписывает, что эти два объекта не должны находиться в одном кластере.

Согласно [12], ограничения на объекты кластеров имеют следующие свойства:

1. Ограничения *must-link* транзитивны: Пусть CC_a и CC_b являются связанными компонентами (подграфами, полностью связанными посредством ограничений *must-link*), пусть o_i и o_j будут объектами в CC_a и CC_b соответственно, тогда:

$$ML(o_i;o_j);o_i \in CC_a;o_j \in CC_b \Rightarrow ML(o_x;o_y);$$

$$\forall o_x;o_y : o_x \in CC_a;o_y \in CC_b.$$

2. Для ограничений *cannot-link* справедливо следующее: пусть CC_a и CC_b являются связанными компонентами (подграфами, полностью связанными посредством ограничений *must-link*), пусть o_i и o_j будут объектами в CC_a и CC_b соответственно, тогда:

$$CL(o_i;o_j);o_i \in CC_a;o_j \in CC_b \Rightarrow CL(o_x;o_y);$$

$$\forall o_x;o_y : o_x \in CC_a;o_y \in CC_b.$$

Идея о введении ограничений *must-link* и *cannot-link* может показаться простой, но на самом деле эти ограничения являются мощным инструментом для многих приложений. Увеличение количества ограничений на пары объектов может существенно улучшить точность результата кластеризации.

Ограничения *must-link* и *cannot-link* также могут использоваться для выражения других пользовательских ограничений. В работах [13, 14] введены ограничения δ-*constraint* и ε-*constraint*, которые могут быть обработаны совместно с ограничениями на пары объектов:

• ограничение δ-constraint (также называемое minimum split constraint) выражает, что расстояние между любой парой точек, которые находятся в двух разных кластерах, должно быть не меньше, чем значение δ. Это ограничение может быть представлено в виде конъюнкции ограничений *must-link* для всех пар объектов с расстоянием меньше, чем δ.

• ограничение ε-*constraint*: для любого кластера C_c, содержащего более одного объекта, для каждого объекта $o_i \in C_c$ должен существовать другой объект $o_j \in C_c$, такой, что расстояние между o_i и o_j не превышает $d_{ij} : d_{ij} \le \varepsilon$. В работе [14] показано, что это ограничение эквивалентно дизъюнкции ограничений *must-link*. Для каждой точки o_i вычисляется множество χ объектов o_j таких, что: $d_{ij} \le \varepsilon$. Ограничение ε-*constraint* может быть представлено как дизъюнкция ограничений *must-link* между объектом o_i и точками множества χ.

Другим ограничением, которое относится к ограничениям на объекты кластеров, является ограничение на максимальный диаметр (максимально возможное расстояние между объектами одного кластера). Это ограничение задает верхнюю границу γ на диаметр кластеров, таким образом, расстояние между любой парой точек, принадлежащих одному и тому же кластеру, должно быть не более значения γ. Это ограничение может рассматриваться как конъюнкция ограничений *cannot-link* между всеми парами объектов с расстоянием, превышающим γ [15].

В данной работе предлагается оригинальный метод ограниченной кластеризации, который в качестве базовой использует схему иерархической кластеризации (новые кластеры создаются путем объединения более мелких кластеров). В качестве меры различия используется расстояние между центроидами кластеров. Новизна метода заключается в возможности учитывать пользовательские ограничения на то, какие объекты обязательно должны попадать в один кластер, а какие не должны. Другими словами, в рамках предлагаемого подхода при отнесении объектов к одному или различным кластерам анализируются не только расстояния между объектами, но и значения их признаков. Предлагаемый метод позволяет останавливать процедуру кластеризации при нарушении пользовательских ограничений.

В ходе кластеризации применяются следующие правила.

Пусть A, B, C – кластеры, с областями определения D_A, D_B, D_C.

$D_A = \{1,2,\ldots,N\}$, $D_B = \{1,2,\ldots,N\}$, $D_C = \{1,2,\ldots,N\}$, где N – номер кластера. Тогда:

1. $A = B, B = C \vDash A = C$,

2. $A = B, B \ne C \vDash A \ne C$.

3. Объединение кластеров невозможно, если они не имеют общих свойств.

Для наглядного представления хода кластеризации, на каждом шаге формируется матрица, в которой указываются рассчитанные расстояния между кластерами и с помощью раскраски ячеек наглядно демонстрируется возможность или невозможность объединения кластеров на следующем шаге кластеризации с учетом введенных ограничений *must-link* и *cannot-link*.

С помощью ограничения *cannot-link* определяются недопустимые комбинации кластеров (по отсутствию общих свойств).

При объединении кластеров с ненулевым расстоянием используется ограничение *must-link*, т. е. на объекты объединенного кластера распространяются запреты, наложенные ограничением *cannot-link* на исходные кластеры.

В ходе кластеризации на каждом шаге проверяется выполнение правила, согласно которому при объединении кластеров внутрикластерное расстояние не должно превосходить межкластерное.

При возникновении альтернативы объединения кластеров рассчитывается расстояние между объединяемыми кластерами для каждой из них, и по принципу компактности кластеров и их наибольшей удаленности друг от друга по свойствам, выбирается та альтернатива, где минимальное расстояние между кластерами будет больше.

На каждом шаге для объединения выбирается пара кластеров с наименьшим межкластерным расстоянием и не противоречащая накладываемым пользовательским ограничениям. При объединении кластеров в матрице расстояний сливаются соответствующие столбцы и строки, причем пользовательские ограничения исходных кластеров распространяются на полученный объединенный кластер. Таким образом, на следующем шаге кластеризации существенно сокращается объем вычислений, как это будет видно из представленного ниже примера. Далее рассчитывается центроид полученного кластера с учетом населенности исходных кластеров, а затем расстояния до других кластеров.

Кластеризация проводится до тех пор, пока имеются разрешенные комбинации кластеров.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Проведена ограниченная кластеризация для 14 условных ячеек, на которые разбит один из участков высоконапряженного массива горных пород Кукисвумчоррского апатит-нефелинового месторождения с целью выявления влияния стационарных и условно-стационарных факторов на происходящие сейсмособытия. Такими факторами являются геология и тектоника месторождения и прилегающего к нему района, геометрия и динамика горных работ. В качестве объектов кластеризации выступали сейсмические события. Каждое сейсмическое событие, отнесенное к некоторой пространственной ячейке, описывалось определенным набором признаков, каждый из которых был сопоставлен некоторому фактору, оказывающему, по мнению экспертов, влияние на возникновение сейсмических событий. В первом приближении учитывалось только наличие/отсутствие факторов в ячейке: соответствующей переменной присваивалось значение «1», если признак наблюдался, «0» – если не наблюдался. Кластеризация проводилась только с учетом количества произошедших сейсмических событий, не учитывалась энергия событий, расстояние до объектов. В этом случае естественным пользовательским ограничением является отнесение к одному кластеру только тех объектов (событий), которые имеют хотя бы один общий признак со всеми другими объектами данного кластера.

В качестве стационарных критериев используются влияющие факторы, которые действуют постоянно и не изменяются во времени:

1) наличие структурных нарушений (разломы, окисленные зоны);

2) типы пород в ячейке: рудное тело, вмещающие породы, рудное тело и вмещающие породы.

В качестве условно-стационарных критериев рассматриваются факторы, которые могут изменяться в процессе ведения горных работ:

1) границы очистного (выработанного) пространства текущего горизонта;

2) границы очистного пространства вышележащего горизонта;

3) выработки.

Далее приняты следующие условные обозначения: Р1 – разлом 1 (ячейки 2–14); Р2 – разлом 2 (ячейки 13-14); ОП – границы очистного пространства; ОПв – границы очистного пространства вышележащего горизонта; В – выработки; РТ – рудное тело; ВП – вмещающие породы; РТ/ВП – рудное тело/вмещающие породы;

На рис. 1 представлено схематичное изображение рассматриваемого участка Кукисвумчоррского месторождения с разбиением на ячейки, а также произошедшие сейсмические события.

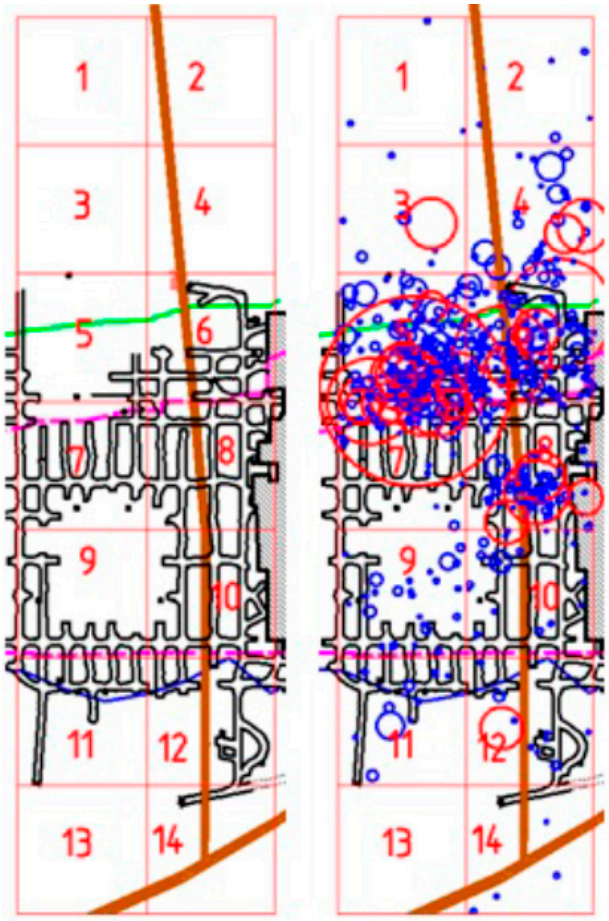
В табл. 1 представлено описание каждой ячейки с точки зрения наличия в ней стационарных и условно-стационарных критериев, а так же количество сейсмособытий, произошедших в ячейках (далее – населенность кластера или *N*).

На первом этапе кластеризации все сейсмособытия (объекты), произошедшие в ячейке считаются принадлежащими одному кластеру, а каждая ячейка – отдельным кластером.

Анализируя данные табл. 1, определяем пользовательское ограничение *cannot-link*, которое накладывается на недопустимые комбинации кластеров. Каждая строка табл. 1 содержит булевый вектор, который описывает набор свойств соответствующей ячейки. Сравнение этих векторов позволяет выявить их недопустимые сочетания, такие, что компоненты сравниваемых векторов полностью различны, а значит, общие свойства у соответствующих кластеров отсутствуют.

Кроме того, до начала расчетов вводится пользовательское ограничение *must-link*, согласно которому объединяются кластера, имеющие нулевое расстояние между центроидами, а значит полностью совпадающий набор свойств.

Используя ограничение *cannot-link* недопустимые сочетания кластеров помечаем цветом в табл. 2 (и далее), в которой представлены результаты расчета расстояний между кластерами 1–14. Расстояние между



а) *б)*

Рис. 1. Участок Кукисвумчоррского месторождения (а – влияющие факторы, б) – влияющие факторы и произошедшие сейсмические события),

– сейсмические события энергетического класса 3–5,

– сейсмические события энергетического класса 6–8,

– разлом,

– контакты рудного тела,

– границы очистного пространства текущего горизонта,

– границы очистного пространства вышележащего горизонта,

– выработки

кластерами определяется как квадрат Евклидова расстояния между их центроидами.

Из табл. 2 очевидно, что на первом шаге объединяются кластеры 1,3; 2,4; 7,9; 8,10, так как они имеют одинаковый набор свойств, а, следовательно, – нулевое расстояние между центроидами (используется ограничение

Таблица 1

Влияющие факторы (критерии) и количество сейсмособытий (N)

Номер ячейки Cell number	P1	P2	ОП	ОПв	РТ	ВП	РТ/ВП	В	N
1	0	0	0	0	0	1	0	0	2
2	1	0	0	0	0	1	0	0	2
3	0	0	0	0	0	1	0	0	6
4	1	0	0	0	0	1	0	0	10
5	0	0	0	0	0	0	1	1	107
6	1	0	1	1	0	0	1	1	126
7	0	0	0	1	1	0	0	1	45
8	1	0	1	1	1	0	0	1	108
9	0	0	0	1	1	0	0	1	25
10	1	0	1	1	1	0	0	1	43
11	0	0	0	1	0	0	1	1	10
12	1	0	0	1	0	0	1	1	32
13	0	1	0	0	0	1	0	0	0
14	1	1	0	0	0	1	0	0	7

Таблица 2

Исходные данные для иерархической кластеризации

	1	3	2	4	5	6	7	9	8	10	11	12	13	14
1	0	0	1	1	3	6	4	4	6	6	4	5	1	2
3	0	0	1	1	3	5	4	4	6	6	4	4	1	2
2	1	1	0	0	4	4	5	5	5	5	5	5	2	1
4	1	1	0	0	4	4	4	4	5	5	5	4	2	1
5	3	3	4	4	0	3	3	3	5	5	1	2	4	5
6	6	5	4	4	3	0	4	4	2	2	2	1	7	6
7	4	4	5	4	3	4	0	0	2	2	1	3	5	6
9	4	4	5	4	3	4	0	0	2	2	1	2	5	6
8	6	6	5	5	5	2	2	2	0	0	4	2	7	6
10	6	6	5	5	5	2	2	2	0	0	4	2	7	6
11	4	4	5	5	1	2	1	1	4	4	0	1	5	6
12	5	4	5	4	2	1	3	2	2	2	1	0	6	5
13	1	1	2	2	4	7	5	5	7	7	5	6	0	1
14	2	2	1	1	5	6	6	6	6	6	6	5	1	0

must-link). В табл. 3 представлены результаты первого шага кластеризации.

Очевидно, что количество строк и столбцов в табл. 3 сократилось, по сравнению с табл. 2 исходных данных, при этом полученные после объединения строки и столбцы сохраняют раскраску исходных, что говорит о применении пользовательского ограничения *cannot-link*.

Далее происходит объединение кластеров 1,3 и 2,4; 11,12; 13,14.

Рассчитаем центроиды вновь полученных кластеров, с учетом населенности кластеров, и расстояния между кластерами. Центроиды

Таблица 3

Результаты первого шага кластеризации

	1,3	2,4	5	6	7,9	8,10	11	12	13	14	*N*
1,3	0	1	3	5	4	6	4	5	1	2	8
2,4	1	0	4	4	5	5	5	4	2	1	12
5	3	4	0	3	3	5	1	2	4	5	107
6	5	4	3	0	4	2	2	1	7	6	126
7,9	4	5	3	4	0	2	1	3	5	6	70
8,10	6	5	5	2	2	0	4	2	7	6	151
11	4	5	1	2	1	4	0	1	5	6	10
12	5	4	2	1	3	2	1	0	6	5	32
13	1	2	4	7	5	7	5	6	0	1	0
14	2	1	5	6	6	6	6	5	1	0	7

Центроид 1, 2, 3, 4

Номер ячейки	Р1	Р2	ОП	ОПв	РТ	ВП	РТ/ВП	В	*N*
1,3	0	0	0	0	0	1	0	0	8
2,4	1	0	0	0	0	1	0	0	12

Расстояние между кластерами 5 и 1, 2, 3, 4

Номер ячейки	Р1	Р2	ОП	ОПв	РТ	ВП	РТ/ВП	В	*N*
5	0	0	0	0	0	0	1	1	107
1,2,3,4	0.6	0	0	0	0	1	0	0	20

кластеров рассчитываются как показано в *Определении 2*, а расстояния между кластерами, как уже упоминалось, определяется как квадрат Евклидова расстояния между центроидами.

Приведем пример расчета (Центроид 1, 2, 3, 4).

$$C_{1,2,3,4} = \{(08/20+12/20), 0, 0, 0, 0, 1, 0, 0\},$$

где 8 –населенность кластера $C_{1,3}$, 12 – населенность кластера $C_{2,4}$, а 20 – населенность объединенного кластера $C_{1,2,3,4}$.

Получим центроид объединенного кластера 1,2.3,4 – $C_{1,2,3,4} = \{0.6, 0, 0, 0, 0, 1, 0, 0\}$.

Рассчитаем расстояние между кластерами 5 и 1, 2, 3, 4.

$$R_{5-1,2,3,4} = (0-0.6)^2 + 0 + 0 + 0 + \\ + 0 + 1 + 1 + 1 = 3.36.$$

Аналогично рассчитываются расстояния между другими кластерами и центроиды вновь получаемых кластеров.

Результаты расчетов на втором и последующих шагах кластеризации представлены в табл. 4–9. На каждом шаге для объединения выбирается разрешенная комбинация кластеров с наименьшим межкластерным расстоянием.

Очевидно, что имеется возможность объединения кластеров 6 и 11, 12.

Теперь объединяем кластеры 1, 2, 3, 4, 13, 14 (табл. 6).

Объединим кластеры 7, 9 и 8, 10 (табл. 7).

Объединим кластеры 6, 11, 12 и 7, 9, 8, 10 (табл. 8).

Объединим кластеры 5 и 6–12 (табл. 9).

Очевидно, что дальнейшее объединение невозможно, т.к. нет разрешенных комбинаций кластеров. На рис. 2 представлена дендрограмма, полученная в результате кластеризации.

Объединение кластеров 1–4 и 13, 14 произошло на третьем шаге, что представляется не

Таблица 4

Результаты второго шага кластеризации

	1,2,3,4	5	6	11,12	7,9	8,10	13,14	*N*
1,2,3,4	0	3.36	5.16	1.03	4.36	5.16	1.16	20
5	3.36	0	3	1.6	3	5	4	107
6	5.16	3	0	1.06	4	2	6	126
11,12	4.03	1.6	1.06	0	2.6	3.06	5.06	42
7,9	4.36	3	4	2.6	0	2	6	70
8,10	5.16	5	2	3.06	2	0	6	151
13,14	1.16	4	6	5.06	6	6	0	7

Таблица 5

Результаты третьего шага кластеризации

	1,2,3,4	13,14	5	6,11,12	7,9	8,10	*N*
1,2,3,4	0	1.16	3.56	4.56	4.33	5.36	20
13,14	1.16	0	4	5.56	5	6	7
5	3.56	4	0	3.56	3	5	107
6,11,12	4.56	4.56	3.56	0	3.45	2.56	168
7,9	4.33	5	3	3.45	0	2	70
8,10	5.36	6	5	2.56	2	0	151

Таблица 6

Результаты четвертого шага кластеризации

	1,2,3,4,13,14	5	6,11,12	7,9	8,10	*N*
1,2,3,4,13,14	0	3.38	4.96	5.11	5.41	27
5	3.38	0	3.56	3	5	107
6,11,12	4.96	3.56	0	3.45	2.56	168
7,9	5.11	3	3.45	0	2	70
8,10	5.41	5	2.56	2	0	151

Таблица 7

Результаты пятого шага кластеризации

	1,2,3,4,13,14	5	6,11,12	7,9,8,10	*N*
1,2,3,4,13,14	0	3.33	3.86	4.88	27
5	3.33	0	3.56	3.46	107
6,11,12	3.86	3.56	0	2.075	168
7,9,8,10	4.88	3.46	2.075	0	221

Таблица 8

Результаты шестого шага кластеризации

	1,2,3,4,13,14	5	6–12	*N*
1,2,3,4,13,14	0	3.33	3.86	27
5	3.33	0	2.94	107
6-12	3.86	2.94	0	389

Таблица 9

Результаты седьмого шага кластеризации

	1,2,3,4,13,14	5–12	*N*
1,2,3,4,13,14	0	5.94	27
5-12	5.94	0	496

очень логичным, поскольку в условиях Кукисвумчоррского месторождения, как правило, уровень сейсмоактивности участков висячего бока месторождения значительно превышает уровень сейсмоактивности участков лежачего бока месторождения. В связи с этим было сделано предположение о необходимости добавления дополнительных признаков: висячий бок – ВБ (ячейки 1–6) и лежачий бок – ЛБ (ячейки 11–14). В табл. 10 представлены влияющие факторы и количество сейсмособытий для анализа по увеличенному числу признаков.

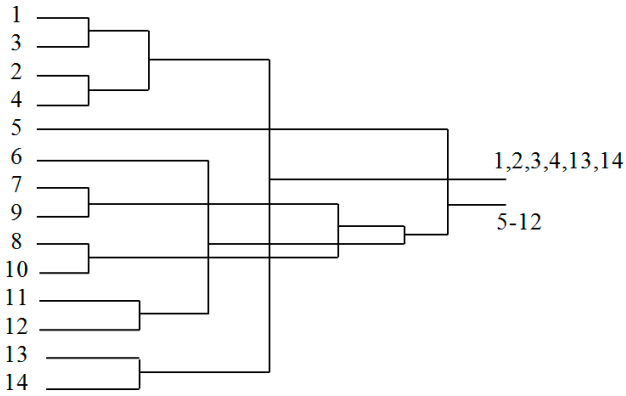


Рис. 2. Дендрограмма, полученная в результате кластеризации

На рис. 3 представлена дендрограмма, полученная в результате кластеризации по десяти признакам для двух выборок по высоте.

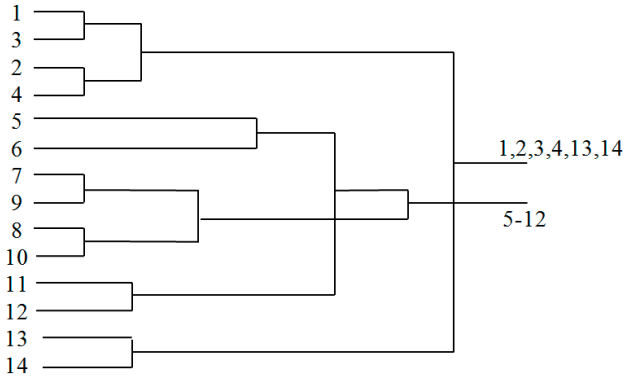


Рис. 3. Дендрограмма, полученная в результате кластеризации по десяти признакам для двух выборок по высоте

Сравнительный анализ полученных дендрограмм позволяет сделать следующие выводы:

1. Добавление двух признаков изменяет ход кластеризации. Так в первом случае (для восьми признаков) кластеры 1–4 и 13, 14 объединяются уже на третьем шаге, а во втором (для десяти признаков) – на последнем. В данном примере в этих кластерах происходит наименьшее число сейсмособытий. Итогом кластеризации является разбиение участка массива горных пород на два кластера: в один попадают ячейки, где проводятся добычные работы, в другой – вмещающие породы, т.е. добычные работы не проводятся.

2. Объединение кластеров 5 и 6 (в которых происходит наибольшее количество сейсмособытий) в первом случае происходит после промежуточного объединения с другими кластерами: на втором шаге для кластера 6 и на последнем шаге для кластера 5. Во втором случае (для десяти признаков) эти кластеры объединяются уже на третьем шаге кластеризации. Общими факторами для данных ячеек являются типы пород, приуроченность к висячему боку месторождения, а также пройденные выработки. В ячейке 6 есть также факторы: разлом-1, границы очистного пространства текущего и вышележащего горизонта. Существенное влияние на результат кластеризации также оказывает признак – населенность кластера, т.е. количество сейсмических событий. При этом в ячейке 6 число сейсмособытий выше, однако, по данным рис. 1 видно, что наиболее сильные события произошли в ячейке 5, т.е. в дальнейшем необходимо также учитывать и энергию событий. Отметим, что наиболее сильные события произошли близко к границе очистного пространства вышележащего горизонта, которая относится к ячейке 7 (вблизи границы ячеек 5 и 7), т.е. результат кластеризации чувствителен к первоначальному разбиению на пространственные ячейки.

Таблица 10

Влияющие факторы и количество сейсмособытий

Номер ячейки Cell number	Р1	Р2	ОП	ОПв	РТ	ВП	РТ_ВП	В	ВБ	ЛБ	*N*
1	0	0	0	0	0	1	0	0	1	0	2
2	1	0	0	0	0	1	0	0	1	0	2
3	0	0	0	0	0	1	0	0	1	0	6
4	1	0	0	0	0	1	0	0	1	0	10
5	0	0	0	0	0	0	1	1	1	0	107
6	1	0	1	1	0	0	1	1	1	0	126
7	0	0	0	1	1	0	0	1	0	0	45
8	1	0	1	1	1	0	0	1	0	0	108
9	0	0	0	1	1	0	0	1	0	0	25
10	1	0	1	1	1	0	0	1	0	0	43
11	0	0	0	1	0	0	1	1	0	1	10
12	1	0	0	1	0	0	1	1	0	1	32
13	0	1	0	0	0	1	0	0	0	1	0
14	1	1	0	0	0	1	0	0	0	1	7

ЗАКЛЮЧЕНИЕ

Проведенные исследования являются первоначальной попыткой оценить применимость методов ограниченной кластеризации для исследования техногенной сейсмичности и показать принципиальную возможность оценки с их помощью степени влияния горно-геологических факторов на сейсмические события. Исследованы различные комбинации стационарных и условно-стационарных факторов, влияющих на сейсмичность массива горных пород. Полученные результаты кластеризации можно использовать для предварительного выделения зон с различной сейсмоактивностью. Гарантированное выполнение пользовательских ограничений в рамках предлагаемого метода позволяет ускорить процесс вычислений счет уменьшения количества альтернатив на каждом шаге кластеризации и повышает доверие у экспертов предметной области к процедурам кластеризации.

Анализ полученных результатов кластеризации для 14 условных ячеек, на которые разбит один из участков высоконапряженного массива горных пород Кукисвумчоррского апатит-нефелинового месторождения позволяет сделать несколько выводов:

1. Введение дополнительных признаков изменяет ход кластеризации, но не результат.

2. Конечные кластеры отличаются только по свойству «выработки», причем в кластере, где выработки отсутствуют, сейсмических событий мало и наоборот. Следовательно, данное свойство оказывает наибольшее влияние на сейсмоактивность.

3. Двигаясь по дереву по направлению к листьям можно проследить, какие свойства являлись наиболее значимыми на каждом шаге кластеризации.

4. На основе полученных результатов можно оценить степень сейсмической активности в каждом кластере, а значит и в каждой ячейке.

5. Вероятно, кластеризацию следует остановить при получении числа кластеров большего, чем два, т. е. в перспективе одним из подходов может быть определение некоторого порогового значения, при котором необходимо останавливать кластеризацию.

6. Для получения более надежного результата необходимо применение более сложного критерия, учитывающего не только число событий и их местоположение по ячейкам, но и энергию событий, их пространственное расположение и другие характеристики.

7. Другим перспективным направлением представляется более гибкий учет влияющих факторов, а также определение степени их влияния (в том числе, возможно, выявление и последующее исключение из рассмотрения незначительно влияющих факторов). Так, например, необходимо учитывать не только наличие разлома в ячейке, но и удаленность сейсмических событий от него. Аналогично необходимо учитывать удаленность сейсмических событий от границ очистного пространства и др.

С точки зрения развития методов ограниченной кластеризации в дальнейшем предполагается разработать точные методы (методы систематического поиска), опирающиеся на парадигму программирования в ограничениях (Constraint Programming). Полученные в ходе исследований результаты позволяют судить о перспективности дальнейшего развития методов ограниченной кластеризации для задач исследования техногенной сейсмичности.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-07-00615а.

СПИСОК ЛИТЕРАТУРЫ

1. Сейсмичность при горных работах / А. А. Козырев [и др.]; – Апатиты : Изд-во Кольского научного центра РАН, 2002. – 325 с.

2. *Hudyma, M. R.* Seismic Hazard Mapping at Mount Charlotte Mine / M. R. Hudyma, P. A. Mikula, M. Owen.// Proceedings of the 5th North American Rock Mechanics Symposium. Hammah, R., Bawden, W. F., Curran, J., and Telesnicki, M. (eds.), University of Toronto Press. – 2002. – P. 1087–1094.

3. *Gibowicz, SJ* Seismicity induced by mining: ten years later / SJ Gibowicz, S. Lasocki // Adv Geophys 44. – 2001. – P. 39–181.

4. *Журавлева, О. Г.* Кластеризация сейсмических событий в условиях удароопасных месторождений Хибинского массива / О. Г. Журавлева // Проблемы недропользования. – 2017. – № 1. С. 14–20.

5. *Babaki, Behrouz.* Constrained Clustering using Column Generation / Babaki, Behrouz, Tias Guns and Siegfried Nijssen // In Proceedings of the 11th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, 2014. – P. 438–454.

6. *Banerjee, Arindam.* Scalable Clustering Algorithms with Balancing Constraints / Banerjee, Arindam and Joydeep Ghosh // Data Mining and Knowledge Discovery. – 2006. – V. 13, no. 3. – P. 365–395.

7. *Bilenko M.* Integrating constraints and metric learning in semi-supervised clustering / M. Bilenko, S. Basu and R. J. Mooney // In Proceedings of the 21st International Conference on Machine Learning, 2004. pp. 11–18.

8. *Ward, J. H.* Hierarchical grouping to optimize an objective function / J. H. Ward // J. of the American Statistical Association, 1963. – 236 p.

9. *Han, Jiawei.* Data mining: Concepts and techniques / Han, Jiawei, Micheline Kamber and Jian Pei. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition. – 2011. – 703 p.

10. *Luxburg, Ulrike.* A Tutorial on Spectral Clustering / Ulrike Luxburg // Statistics and Computing, 2007. – V. 17, no. 4. – P. 395–416.

11. *Wagsta, K.* Clustering with instance-level constraints / K. Wagsta, C. Cardie // In Proceedings of the 17th International Conference on Machine Learning. – 2000. – P. 1103–1110.

12. *Davidson, Ian and Sugato Basu.* A survey of clustering with instance level constraints // ACM Transactions on Knowledge Discovery from Data. – 2007. – P. 1–41.

13. *Davidson, I. and S. S. Ravi.* Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. // In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. – 2005. – P. 59–70.

14. *Ester, Martin, Hans P. Kriegel, Jorg Sander and Xiaowei Xu.* 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. – 1996. – P. 226–231.

15. *Dao, Thi-Bich-Hanh, Khanh-Chuong Duong, Christel Vrain.* Constrained clustering by constraint programming // Artificial Intelligence. – 2017. – V. 244. – P. 70–94.

Зуенко Александр Анатольевич – канд. техн. наук, ведущий научный сотрудник ИИММ КНЦ РАН, e-mail: zuenko@iimm.ru

Фридман Ольга Владимировна – канд. техн. наук, старший научный сотрудник ИИММ КНЦ РАН, e-mail: ofridman@iimm.ru

Журавлева Ольга Геннадьевна – канд. техн. наук, старший научный сотрудник, ГоИ КНЦ РАН, e-mail: zhuravlevaog@goi.kolasc.net.ru

APPLICATION OF CONSTRAINED CLUSTERING METHODS FOR RESEARCHING MINING-INDUCED SEISMICITY

A. A. Zuenko*, O. V. Fridman*, O. G. Zhuravleva**

**Institute of Informatics and Mathematical Modeling – Subdivision of the Federal Research Centre «Kola Science Centre of the Russian Academy of Sciences»*
***Mining Institute – Subdivision of the Federal Research Centre «Kola Science Centre of the Russian Academy of Sciences»*

Annotation. The purpose of this study is to assess the applicability of constrained clustering methods for researching mining-induced seismicity. The paper proposes an original method of constrained clustering, which uses the hierarchical clustering scheme as a base (new clusters are created by combining smaller clusters). The distance between the centroids of the clusters is used as a measure of the difference. The novelty of the method lies in the ability to take into account user constraints on which objects must fall into one cluster, and which should not. In other words, within the framework of the proposed approach, when assigning objects to one or different clusters, not only the distances between the objects, but also the values of their attributes are analyzed. The proposed method allows stopping the clustering procedure in case of violation of user constraints. The studies are an initial attempt to assess the applicability of constrained clustering methods for researching mining-induced seismicity and to show a possibility in principle to assess the degree of mining and geological factors impact to seismicity. Various combinations of stationary and semi-stationary factors impacting to rock mass seismicity are investigated. The results of clustering can be used to pre-allocate areas with different seismic activity. Further research should be carried out taking into account not only the number of seismic events, but also their energy, the distance between them, etc. The use of constrained clustering methods allows us to speed up the computation process by reducing the number of alternatives at each clustering step. Guaranteed implementation of user restrictions within the framework of the proposed method increases the confidence of subject matter experts in clustering procedures. The obtained results make it possible to judge the prospects for the further development of limited clustering methods for the study of mining-induced seismicity.
Keywords: constrained clustering, mining-induced seismicity, cluster analysis.

Zuenko Alexander Anatolievich – PhD, leading researcher, IIMM KSC RAS, e-mail: zuenko@iimm.ru

Fridman Olga Vladimirovna – PhD, senior researcher, IIMM KSC RAS, e-mail: ofridman@iimm.ru

Zhuravleva Olga Genadievna – PhD, senior researcher, MI KSC RAS, e-mail: zhuravlevaog@goi.kolasc.net.ru