*ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 484, № 4, с. 387–392*

**МАТЕМАТИКА**

*УДК 519.16 + 519.85*

# О СЛОЖНОСТИ НЕКОТОРЫХ ЗАДАЧ ПОИСКА СЕМЕЙСТВА НЕПЕРЕСЕКАЮЩИХСЯ КЛАСТЕРОВ

**А. В. Кельманов[1,2,*], А. В. Пяткин[1,2,**], В. И. Хандеев[1,2,***]**

Представлено академиком РАН С.С. Гончаровым 30.07.2018 г.

Поступило 01.08.2018 г.

Рассматриваются две родственные задачи поиска семейства непересекающихся подмножеств (кластеров) в конечном множестве точек евклидова пространства. В этих задачах требуется максимизировать размер минимального по мощности кластера так, чтобы в каждом кластере суммарный внутрикластерный квадратичный разброс точек не превышал заданной доли (константы) от суммарного квадратичного разброса точек во входном множестве. Доказано, что обе задачи NP-трудны даже на числовой прямой.

*Ключевые слова:* Евклидово пространство, кластеризация, максиминная задача, квадратичный разброс, NP-трудность.

**DOI:** https://doi.org/10.31857/S0869-56524844387-392

Предметом исследования работы являются две родственные экстремальные задачи поиска семейства непересекающихся кластеров в конечном множестве точек евклидова пространства. Цель исследования — анализ вычислительной сложности этих задач. Исследование мотивировано отсутствием опубликованных данных о сложностном статусе задач и их актуальностью как в теоретическом, так и в прикладном плане.

В плане теоретической мотивации обе рассматриваемые задачи не эквивалентны ни одной из близких по постановке хорошо известных труднорешаемых кластеризационных задач — *k*-means (или *k*-MSSC), *k*-median, *k*-medoid, *k*-center, *k*-center clustering with outliers и др. (см., например, [1–13]). Как и эти задачи дискретной оптимизации, рассматриваемые задачи имеют тесную связь с проблемами компьютерной геометрии, теории графов, статистики и аппроксимации.

Обе рассматриваемые задачи моделируют типичную для многих приложений проблему поиска в совокупности объектов семейства непересекающихся подмножеств, каждое из которых состоит из похожих по некоторому критерию объектов, и выделения вырожденных экземпляров — так называемых "выбросов" (outliers), которые не похожи между собой и не принадлежат ни одному из подмножеств семейства. К числу приложений, для которых эта проблема типична, относятся, в частности, анализ данных, машинное обучение, интерпретация данных, распознавание образов, статистика (Data analysis, Machine learning, Data mining, Pattern recognition, Statistics).

Рассматриваемые задачи имеют следующие формулировки.

З а д а ч а 1 (Maximum Min-Size Clustering with Outliers 1). Дано: $N$-элементное множество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, число $\alpha \in (0, 1)$ и точки $z_1, z_2 \in \mathbb{R}^d$. Найти: непустые непересекающиеся подмножества $\mathcal{C}_1, \mathcal{C}_2$ такие, что

$$\min\{|\mathcal{C}_1|, |\mathcal{C}_2|\} \to \max, \qquad (1)$$

при ограничениях

$$\sum_{y \in \mathcal{C}_i} \|y - z_i\|^2 \le \alpha \sum_{y \in \mathcal{Y}} \|y - \bar{y}(\mathcal{Y})\|^2, \ i = 1, 2, \qquad (2)$$

где

$$\bar{y}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} y$$

есть центроид (геометрический центр) множества $\mathcal{Y}$.

З а д а ч а 2 (Maximum Min-Size Clustering with Outliers 2). Дано: $N$-элементное множество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве и число $\alpha \in (0, 1)$. Найти: непустые

[1] *Институт математики им. С.Л. Соболева Сибирского отделения Российской Академии наук, Новосибирск*

[2] *Новосибирский государственный университет*

* *E-mail: kelm@math.nsc.ru*

** *E-mail: artem@math.nsc.ru*

*** *E-mail: khandeev@math.nsc.ru*

непересекающиеся подмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки $y_1, y_2$ из $\mathcal{Y}$ такие, что имеет место соотношение (1), при ограничениях

$$\sum_{y\in\mathcal{C}_i}\|y-y_i\|^2\le\alpha\sum_{y\in\mathcal{Y}}\|y-\bar{y}(\mathcal{Y})\|^2,\ i=1,2. \qquad (3)$$

Обе задачи можно трактовать как поиск двух непересекающихся подмножеств точек, сконцентрированных относительно двух точек; в задаче 1 это заданные точки $z_1$, $z_2$ из $\mathbb{R}^d$, а в задаче 2 — неизвестные точки $y_1$, $y_2$ из входного множества $\mathcal{Y}\subset\mathbb{R}^d$. В обеих задачах требуется максимизировать минимальный размер кластера при ограничении сверху на суммарный внутрикластерный квадратичный разброс (левая часть неравенств (2) и (3)). Это ограничение определяется через α-долю от суммарного квадратичного разброса точек во входном множестве (правая часть неравенств (2) и (3)). Наконец, обе задачи можно интерпретировать как поиск совокупности сгущений (кластеров) точек в евклидовом пространстве.

В области дискретной оптимизации ключевым вопросом является вопрос о сложностном статусе какой-либо вновь выявленной экстремальной задачи. Ниже мы показываем, что задачи 1 и 2, а также их многокластерные обобщения NP-трудны даже в одномерном случае (т.е. на прямой).

Для доказательства мы показываем NP-полноту задач, из которой, как известно [14], следует NP-трудность. Положим

$$A=\alpha\sum_{y\in\mathcal{Y}}\|y-\bar{y}(\mathcal{Y})\|^2,$$

$$f(\mathcal{C}_i,z_i)=$$

$$=\sum_{y\in\mathcal{C}_i}\|y-z_i\|^2,\ \mathcal{C}_i\subseteq\mathcal{Y},\ z_i\in\mathbb{R}^d,\ i=1,2. \qquad (4)$$

Далее будем считать, что $\mathcal{Y}$ — мультимножество. Сформулируем задачу 1 в форме верификации свойств.

З а д а ч а 1А. Дано: $N$-элементное мультимножество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, число $A > 0$, точки $z_1, z_2\in\mathbb{R}^d$ и натуральное число $M$. Вопрос: существуют ли в $\mathcal{Y}$ непустые непересекающиеся мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ такие, что

$$\min\{|\mathcal{C}_1|,|\mathcal{C}_2|\}\ge M, \qquad (5)$$

при ограничениях

$$f(\mathcal{C}_i,z_i)\le A,\ \ i=1,2\,? \qquad (6)$$

Сложностной статус этой задачи устанавливает следующая

Т е о р е м а 1. *Задача* 1А NP-*полна даже в одномерном случае.*

Для доказательства труднорешаемости задачи 1А мы строим полиномиальное сведение к ней следующей NP-полной задачи.

Р а з б и е н и е 1 (P a r t i t i o n 1). Дано: мультимножество $\mathcal{X}=\{x_1,\ldots,x_{2K}\}$ натуральных чисел. Вопрос: существует ли разбиение $\mathcal{X}$ на мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ такое, что $|\mathcal{S}_1|=|\mathcal{S}_2|$ и

$$\sum_{x\in\mathcal{S}_1}x=\sum_{x\in\mathcal{S}_2}x\,? \qquad (7)$$

По произвольному входу задачи Разбиение 1 строится следующий пример входа задачи 1А. В задаче 1А полагаем

$$d=1,\ \ N=2K,\ \ M=K,\ \ A=\frac{1}{2}\sum_{x\in\mathcal{X}}x,\ \ z_1=z_2=0,$$

$$\mathcal{Y}=\{c_1,\ldots,c_{2K}\},$$

где

$$c_k=\sqrt{x_k},\ \ k=1,2,\ldots,2K.$$

Далее мы показываем, что в построенном примере задачи 1А мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ из $\mathcal{Y}$, удовлетворяющие неравенствам (5) и (6), существуют тогда и только тогда, когда в задаче Разбиение 1 существует разбиение $\mathcal{X}$ на мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ такое, что $|\mathcal{S}_1|=|\mathcal{S}_2|$ и имеет место равенство (7).

Пусть в задаче Разбиение 1 существуют требуемые мультиподмножества

$$\mathcal{S}_i=\{x_k\mid k\in I_i\},\ i=1,2, \qquad (8)$$

где $I_1, I_2\subseteq\{1,2,\ldots,2K\}$,

$$I_1\cup I_2=\{1,2,\ldots,2K\},\ \ I_1\cap I_2=\phi. \qquad (9)$$

Легко видеть, что в задаче 1А в качестве $\mathcal{C}_1$ и $\mathcal{C}_2$ можно взять

$$\mathcal{C}_i=\{c_k\mid k\in I_i\},\ i=1,2. \qquad (10)$$

Для этих мультиподмножеств имеем

$$\min\{|\mathcal{C}_1|,|\mathcal{C}_2|\}=\min\{K,K\}=K=M,$$

$$f(\mathcal{C}_i,z_i)=\sum_{y\in\mathcal{C}_i}\|y\|^2=\frac{1}{2}\sum_{x\in\mathcal{X}}x=A,\ i=1,2.$$

Следовательно, условия (5) и (6) выполнены и в задаче 1А требуемые мультиподмножества $\mathcal{C}_1$ и $\mathcal{C}_2$ существуют.

Пусть теперь в задаче 1А в мультимножестве $\mathcal{Y}$ существуют требуемые мультиподмножества (10), удовлетворяющие условиям (5) и (6).

В задаче Разбиение 1 рассмотрим мультимножества (8). Для этих мультимножеств имеем $|\mathcal{S}_1|=|\mathcal{S}_2|$ и

$$\sum_{k\in I_i} x_k = \sum_{x\in\mathcal{S}_i} x = \sum_{y\in\mathcal{C}_i} \| y \|^2 = \frac{1}{2}\sum_{x\in\mathcal{X}} x,\ i=1,2.$$

Поэтому

$$\sum_{x\in\mathcal{S}_1} x = \sum_{x\in\mathcal{S}_2} x.$$

Следовательно, в задаче Разбиение 1 требуемые мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ также существуют.

Замечание 1. Для простоты доказательства мы использовали числа $c_k = \sqrt{x_k}$, $k$ = 1, 2, ..., $K$, которые могут быть иррациональны. Однако легко видеть, что те же доказательные аргументы справедливы для рациональных чисел $c_k$ таких, что $0 \le \sqrt{x_k} - c_k < \dfrac{1}{4K\lceil \max_k \sqrt{x_k} \rceil}$, $k$ = 1, 2, ..., $K$, так как $x_k$ — целое число. Остаётся заметить, что построенное сведение, очевидно, полиномиально.

Из теоремы 1 следует, что задача 1 NP-трудна, как и сформулированное ниже ее многокластерное обобщение.

Задача 1′. Дано: $N$-элементное множество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, натуральное число $J$, число $\alpha \in (0, 1)$ и точки $z_1, \ldots z_J \in \mathbb{R}^d$. Найти: непустые непересекающиеся подмножества $\mathcal{C}_1$, ..., $\mathcal{C}_J$ во множестве $\mathcal{Y}$ такие, что

$$\min\{|\mathcal{C}_1|, \ldots, |\mathcal{C}_J|\} \to \max, \tag{11}$$

при ограничениях

$$\sum_{y\in\mathcal{C}_i} \| y - z_i \|^2 \le \alpha \sum_{y\in\mathcal{Y}} \| y - \bar{y}(\mathcal{Y}) \|^2,\ i=1,2,\ldots,J.$$

В задаче 1′ число $J$ кластеров является частью входа. Для варианта задачи 1, в котором число кластеров не является частью входа, т.е. для параметрической задачи, которую далее обозначим через 1($J$), имеет место следующая

Теорема 2. *Если число J кластеров не является частью входа, то для любого фиксированного параметра* $J \ge 2$ *задача* 1(J) NP-*трудна даже в одномерном случае.*

Идея доказательства базируется на индукции. Мы показываем NP-полноту задачи 1($J$) в форме верификации свойств. При этом из NP-полноты задачи следует NP-трудность её оптимизационного варианта. В форме верификации свойств задача 1($J$) имеет следующую формулировку.

Задача 1A($J$). Дано: $N$-элементное мультимножество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, число $A > 0$, точки $z_1, \ldots z_J \in \mathbb{R}^d$ и натуральное число $M$. Вопрос: существуют ли непустые непересекающиеся подмножества $\mathcal{C}_1$, ..., $\mathcal{C}_J$ во множестве $\mathcal{Y}$ такие, что

$$\min\{|\mathcal{C}_1|, \ldots, |\mathcal{C}_J|\} \ge M, \tag{12}$$

при ограничениях

$$f(\mathcal{C}_i, z_i) \le A,\ i=1,2,\ldots,J\ ?$$

Строим сведение задачи 1A($J$) к задаче 1A($J$ + *1*). Для этого по входу задачи 1A($J$) построим следующий пример входа задачи 1A($J$ + *1*). В задаче 1A($J$ + *1*) положим

$$\widetilde{\mathcal{Y}} = \mathcal{Y} \cup \mathcal{G},\ \widetilde{N} = N + M,\ \widetilde{d} = d,\ \widetilde{A} = A,\ \widetilde{M} = M, \tag{13}$$

$$\mathcal{G} = \{g_1, \ldots, g_M\},\quad g_i = L,\quad i=1,2,\ldots,M, \tag{14}$$
$$\tilde{z}_i = z_i,\ i=1,2,\ldots,J;\ \tilde{z}_{J+1} = L,$$

где

$$L > \max_{i=1,\ldots,J} z_i + \sqrt{A}.$$

Пусть в задаче 1A($J$) существуют требуемые мультиподмножества $\mathcal{C}_1$, ..., $\mathcal{C}_J$. Легко проверить, что в задаче 1A($J$ +*1*) в качестве требуемых мультиподмножеств $\tilde{\mathcal{C}}_1$, ..., $\tilde{\mathcal{C}}_{J+1}$ можно взять $\tilde{\mathcal{C}}_1 = \mathcal{C}_1$, ... ..., $\tilde{\mathcal{C}}_J = \mathcal{C}_J$, $\tilde{\mathcal{C}}_{J+1} = \mathcal{G}$.

Пусть теперь в задаче 1A($J$ + *1*) существуют требуемые мультиподмножества $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_{J+1}$.

Покажем, что ни одна из точек $g_1$, $i$ = 1, 2, ..., $M$, не входит в $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_J$. Предположим противное, например, $g_1 \in \tilde{\mathcal{C}}_1$. Тогда

$$f(\tilde{\mathcal{C}}_1, \tilde{z}_1) \ge \| g_1 - \tilde{z}_1 \|^2 \ge L^2 > \tilde{A},$$

что противоречит условию $f(\tilde{\mathcal{C}}_1, \tilde{z}_1) \le \tilde{A}$.

Поэтому мультимножества $\mathcal{C}_1 = \tilde{\mathcal{C}}_1$, ..., $\mathcal{C}_J = \tilde{\mathcal{C}}_J$ являются требуемыми мультиподмножествами в задаче 1A($J$).

Проанализируем теперь сложность задачи 2. Сформулируем эту задачу в форме верификации свойств.

Задача 2A. Дано: $N$-элементное мультимножество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, число $A > 0$ и натуральное число $M$. Вопрос: существуют ли непустые непересекающиеся мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки $y_1$, $y_2$ в $\mathcal{Y}$ такие, что имеет место неравенство (5), при ограничениях

$$f(\mathcal{C}_i, y_i) \le A,\ i=1,2\ ? \tag{15}$$

Справедлива следующая

Теорема 3. *Задача* 2A NP-*полна даже в одномерном случае.*

Факт труднорешаемости задачи 2A устанавливается путём полиномиального сведения к ней следующей NP-полной задачи [14].

Разбиение 2 (Partition 2). Дано: мультимножество $\mathcal{X} = \{x_1, \ldots, x_K\}$ натуральных чисел. Вопрос: существует ли разбиение мультимножества

$\mathcal{X}$ на мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ такое, что имеет место равенство (7)?

По произвольному входу задачи Разбиение 2 строится следующий пример входа задачи 2А. Положим

$$d=1,\ N=2K+2,\ M=K+1,$$

$$\mathcal{Y}=\{b_1,\ b_2,\ a_1,\ \ldots,\ a_K,\ c_1,\ \ldots,\ c_K\},$$

где

$$a_k=0,\ c_k=\sqrt{x_k},\ k=1,2,\ \ldots,\ K,$$

$$b_1=b_2=-B,$$

причём

$$B>\max\left\{\frac{1}{4}\sum_{x\in\mathcal{X}}x,\ \sqrt{\frac{1}{2}\sum_{x\in\mathcal{X}}x}\right\},$$

и

$$A=B^2+\frac{1}{2}\sum_{x\in\mathcal{X}}x.$$

Без ограничения общности будем считать, что $K\geq 3$.

Далее мы показываем, что в построенном примере задачи 2А мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки $y_1$, $y_2$ из $\mathcal{Y}$, удовлетворяющие неравенствам (5) и (15), существуют тогда и только тогда, когда в задаче Разбиение 2 существуют мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ такие, что имеет место равенство (7).

Пусть в задаче Разбиение 2 мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ существуют. Пусть эти мультиподмножества, как и ранее, определяются формулами (8), в которых, в отличие от (9),

$$I_1,\ I_2\subseteq\{1,2,\ldots,K\},\ I_1\cup I_2=\{1,2,\ldots,K\},$$

$$I_1\cap I_2=\phi.$$

В задаче 2А рассмотрим точки $y_1 = 0$, $y_2 = 0$ и следующие мультиподмножества мощности $M = K + 1$ каждое:

$$\mathcal{C}_i=\{b_i\}\cup\{c_k\mid k\in I_i\}\cup\{a_k\mid k\in I_{3-i}\},\ i=1,2. \quad (16)$$

Для этих мультиподмножеств из (4) имеем следующие внутрикластерные разбросы:

$$f(\mathcal{C}_i,y_i)=\|b_i-y_i\|^2+\sum_{k\in I_i}\|c_k-y_i\|^2=B^2+\sum_{k\in I_i}x_k=$$

$$=B^2+\frac{1}{2}\sum_{k\in\{1,\ \ldots,\ K\}}x_k=A,\ i=1,2.$$

Это равенство означает, что условие (15) выполнено. Условие (5) выполнено по построению. Таким образом, в задаче 2А требуемые мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки $y_1$, $y_2$ существуют.

Допустим теперь, что требуемые мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки $y_1$, $y_2$ существуют в задаче 2А. Тогда мы сначала показываем, что точки $b_1$ и $b_2$ лежат в различных мультиподмножествах. Кроме того, устанавливаем, что в паре $\{\mathcal{C}_1, \mathcal{C}_2\}$ нет мультиподмножества, содержащего все точки $c_k$, $k = 1, 2, \ldots, K$.

Легко установить, что $y_1 = y_2 = 0$, так как оба мультиподмножества $\mathcal{C}_1$ и $\mathcal{C}_2$ содержат не менее одной точки $c_k$, $k\in\{1,2,\ldots,K\}$, и одну из точек $b_1$, $b_2$.

Наконец, так как точки $b_1$ и $b_2$ лежат в разных мультиподмножествах и оба мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ содержат точки из $\{c_1,\ldots,c_K\}$ и $\{a_1,\ldots,a_K\}$, можно считать, что структура этих мультиподмножеств определяется формулой (16). Тогда, используя эту формулу и равенства $y_i = 0$, $i = 1, 2$, получим следующую оценку для внутрикластерных разбросов:

$$f(\mathcal{C}_i,y_i)=\|b_i-y_i\|^2+\sum_{k\in I_i}\|c_k-y_i\|^2=B^2+\sum_{k\in I_i}x_k\leq$$

$$\leq A=B^2+\frac{1}{2}\sum_{k\in\{1,\ldots,K\}}x_k,\ i=1,2.$$

Следовательно,

$$\sum_{k\in I_1}x_k=\sum_{k\in I_2}x_k=\frac{1}{2}\sum_{k\in\{1,\ldots,K\}}x_k.$$

Поэтому в задаче Разбиение 2 требуемые мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ также существуют.

З а м е ч а н и е 2. В доказательстве этой теоремы, как и при доказательстве теоремы 1, фигурируют числа $c_k=\sqrt{x_k}$, $k = 1, 2, \ldots, K$, которые могут быть иррациональны. Однако все выкладки остаются справедливыми для рациональных чисел $c_k$ таких, что $0\leq\sqrt{x_k}-c_k<\dfrac{1}{2K\lceil max_k\sqrt{x_k}\rceil}$, $k = 1, 2, \ldots, K$, так как $x_k$ — целое число. Остаётся заметить, что построенное сведение, очевидно, полиномиально.

Из теоремы 3 следует, что задача 2 NP-трудна, как и сформулированное ниже её многокластерное обобщение.

З а д а ч а 2′. Дано: $N$-элементное множество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, натуральное число $J$ и число $\alpha\in(0,1)$. Найти: непустые непересекающиеся подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и точки $y_1, \ldots, y_J$ во множестве $\mathcal{Y}$ такие, что имеет место соотношение (11), при ограничениях

$$\sum_{y\in\mathcal{C}_i}\|y-y_i\|^2\leq\alpha\sum_{y\in\mathcal{Y}}\|y-\bar{y}(\mathcal{Y})\|^2,\ i=1,2,\ldots,J.$$

В этой задаче число $J$ кластеров является частью входа. Для варианта задачи 2, в котором число кластеров не является частью входа, т.е. для параметрической задачи, которую далее обозначим через 2($J$), имеет место следующая

Т е о р е м а 4. *Если число J кластеров не является частью входа, то для любого фиксированного параметра $J \geq 2$ задача* 2($J$) NP-*трудна даже в одномерном случае.*

Как и при доказательстве теоремы 2, справедливость утверждения теоремы 4 устанавливается индукцией по числу кластеров. Мы показываем NP-полноту следующей задачи.

З а д а ч а 2A($J$). Дано: $N$-элементное мультимножество $\mathcal{Y}$ точек в $d$-мерном евклидовом пространстве, число $A > 0$ и натуральное число $M$. Вопрос: существуют ли непустые непересекающиеся подмножества $\mathcal{C}_1$, ..., $\mathcal{C}_J$ и точки $y_1, \ldots, y_J$ во множестве $\mathcal{Y}$ такие, что имеет место соотношение (12), при ограничениях

$$f(\mathcal{C}_i, y_i) \leq A,\ i = 1, 2, \ldots, J\ ?$$

Строим полиномиальное сведение задачи 2$A(J)$ к задаче 2$A(J+1)$. Определим вход задачи 2$A(J+1)$ равенствами (13) и (14), где

$$L > \max_{y \in \mathcal{Y}} y + \sqrt{A}.$$

Если в задаче 2A($J$) существуют требуемые мультиподмножества $\mathcal{C}_1$, ..., $\mathcal{C}_J$, то легко проверить, что в задаче 2$A(J+1)$ требуемыми мультиподмножествами $\tilde{\mathcal{C}}_1$, ..., $\tilde{\mathcal{C}}_{J+1}$ и точками $\tilde{y}_1, \ldots, \tilde{y}_{J+1}$ являются $\tilde{\mathcal{C}}_1 = \mathcal{C}_1$, ..., $\tilde{\mathcal{C}}_J = \mathcal{C}_J$, $\tilde{\mathcal{C}}_{J+1} = \mathcal{G}$, $\tilde{y}_1 = y_1$, ... ..., $\tilde{y}_J = y_J$, $\tilde{y}_{J+1} = g_1$.

Пусть теперь в задаче 2$A(J+1)$ существуют требуемые мультиподмножества $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_{J+1}$ и точки $\tilde{y}_1, \ldots, \tilde{y}_{J+1}$. В этом случае, действуя от противного, мы показываем, что хотя бы $J$ из мультимножеств $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_{J+1}$ не содержат точек из $\mathcal{G}$. При этом можно считать, что этими мультиподмножествами являются $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_J$. Тогда легко установить, что требуемыми мультиподмножествами и точками в задаче 2$A(J)$ являются $\mathcal{C}_1 = \tilde{\mathcal{C}}_1$, ..., $\mathcal{C}_J = \tilde{\mathcal{C}}_J$, $y_1 = \tilde{y}_1$, ..., $y_J = \tilde{y}_J$.

Таким образом, несмотря на простоту формулировок, все рассмотренные задачи максиминной кластеризации NP-трудны даже в одномерном случае (на числовой прямой). Построение эффективных алгоритмов с теоретическими гарантиями качества (точности, временной сложности, надёжности) для этих задач представляется делом ближайшей перспективы.

**Источник финансирования.** Работа выполнена при финансовой поддержке РНФ (проект 16–11–10041).

## СПИСОК ЛИТЕРАТУРЫ

1. *Aloise D., Deshpande A., Hansen P., Popat P.* NP-hardness of Euclidean Sum-of-Squares clustering // Machine Learning. 2009. № 75 (2). P. 245–248.
2. *Drineas P., Frieze A., Kannan R., Vempala S., Vinay V.* Clustering Large Graphs via the Singular Value Decomposition // Machine Learning. 2004. № 56. P. 9–33.
3. *Arora S., Raghavan P., Rao S.* Approximation Schemes for Euclidean $k$-Medians and Related Problems. Proc. 30th Annual ACM Symposium on Theory of Computing. Dallas (TX), 1998. P. 106–113.
4. *Kariv O., Hakimi S.* An Algorithmic Approach to Network Location Problems. Pt 1. The $p$-Centers // SIAM J. Appl. Math. 1979. V. 37. P. 513–538.
5. *Feder T., Greene D.* Optimal Algorithms for Approximate Clustering. In: Proc. of the 20th ACM Symposium on Theory of Computing. N.Y., 1988. P. 434–444.
6. *Hochbaum D.S., Shmoys D.B.* A Best Possible Heuristic for the $k$-Center Problem // Math. Operations Res. 1985. V. 10(2). P. 180–184.
7. *Kaufman L., Rousseeuw P.J.* Clustering by Means of Medoids. In: Statistical Data Analysis Based on the $L_1$-Norm and Related Methods. Amsterdam: North-Holland, 1987. P. 405–416.
8. *Krarup J., Pruzan P.* The Simple Plant Location Problem: Survey and Synthesis // Europ. J. Oper. Res. 1983. № 12. V. (1). P. 36–81.
9. Discrete Location Theory/Mirchandani P., Francis. R. Eds. L.: Wiley-Intersience, 1990.
10. *Charikar M., Khuller S., Mount D.M., Narasimhan G.* Algorithms for Facility Location Problems with Outliers. In: Proc. 12th ACM-SIAM Symps. Discrete Algorithms. Wash. (D.C.), 2001. P. 642–651.
11. *Agarwal P.K., Phillips J.M.* An Efficient Algorithm for 2D Euclidean 2-Center with Outliers. In: Proc. 16th Annu. European Sympos. Algorithms. Karlsruhe, 2008. P. 64–75.
12. *McCutchen R.M., Khuller S.* Streaming Algorithms for $k$-Center Clustering with Outliers and with Anonymity. In: Proc. 11th Intern. Workshop Approx. Algorithms. Karlsruhe, 2008. P. 165–178.
13. *Hatami B., Zarrabi-Zade H.* A Streaming Algorithm for 2-center with Outliers in High Dimensions // Comput. Geom. 2017. V. 60. P. 26–36.
14. *Garey M.R., Johnson D.S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman, 1979.

# ON THE COMPLEXITY OF SOME PROBLEMS OF SEARCHING FOR A FAMILY OF DISJOINT CLUSTERS

**A. V. Kel'manov, A. V. Pyatkin, V. I. Khandeev**

Presented by Academician of the RAS S.S. Goncharov July 30, 2018

Received August 1, 2018

We consider some consimilar problems of searching for disjoint subsets (clusters) in the nite set of points in Euclidean space. In these problems, it is required to maximize the minimum subset size such that the value of each intracluster quadratic variation would not exceed a given fraction (constant) of the total quadratic variation of the points of the input set with respect to its centroid. In the paper, we have proved that all the problems are NP-hard even on a line.

*Keywords:* Euclidean space, clustering, max-min problem, quadratic variation, NP-hardness.