ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 24 № 4 2018

УДК 519.16+519.85

О СЛОЖНОСТИ НЕКОТОРЫХ МАКСИМИННЫХ ЗАДАЧ КЛАСТЕРИЗАЦИИ[1]

А. В. Кельманов, А. В. Пяткин, В. И. Хандеев

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 относительно центра кластера не превышал заданной доли (константы) от суммарного квадратичного разброса точек во входном множестве относительно его геометрического центра. В первой задаче центры внутрикластерных разбросов — произвольные, но заданные на входе точки пространства. Во второй задаче центры внутрикластерного разброса неизвестны (являются искомыми точками), но принадлежат входному множеству. Доказано, что обе задачи NP-трудны даже на числовой прямой как в общем случае, когда число кластеров является частью входа, так и в параметрическом случае, когда число кластеров фиксировано.

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

A. V. Kel'manov, A. V. Pyatkin, V. I. Khandeev. On the complexity of some max–min clustering problems.

Two similar problems of searching for a family of disjoint subsets (clusters) in a finite set of points in Euclidean space are considered. In these problems the size of the smallest cluster should be maximized so that in each cluster the intracluster quadratic variation of the points with respect to its center would not exceed a given (constant) fraction of the total quadratic variation of the points of the input set with respect to its centroid. In the first problem the centers of intracluster variations are arbitrary points given at the input. In the second problem the centers of the intracluster variation are unknown (to be found) but they must lie in the input set. It is proved that both problems are NP-hard even on the real line both in the general case when the number of the clusters is a part of the input and in the parametric case when the number of the clusters is fixed.

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

MSC: 68W25, 68Q25

DOI: 10.21538/0134-4889-2018-24-4-189-198

Введение

Предметом исследования являются две родственные экстремальные задачи поиска семейства непересекающихся кластеров в конечном множестве точек евклидова пространства. Цель исследования — анализ вычислительной сложности этих задач.

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 — "выбросов" (outliers), которые не похожи между собой и не принадлежат ни одному из подмножеств в семействе. К числу приложений, для которых эта проблема типична, относятся, например, анализ данных, машинное обучение, интерпретация данных, распознавание образов, статистика (data analysis, machine learning, data mining, pattern recognition, statistics).

Исследование мотивировано отсутствием опубликованных данных о сложностном статусе задач и их актуальностью как в теоретическом, так и в прикладном плане.

[1]Исследования поддержаны Российским научным фондом, грант 16-11-10041.

1. Формулировка задач и связанные с ними проблемы

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

$$\min\{|\mathcal{C}_1|, |\mathcal{C}_2|\} \to \max \tag{1.1}$$

при ограничениях

$$\sum_{y \in \mathcal{C}_i} \|y - z_i\|^2 \le \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2, \quad i = 1, 2, \tag{1.2}$$

где

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

— 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 непересекающиеся подмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки y_1, y_2 из $\mathcal{Y}$ такие, что имеет место соотношение (1.1), при ограничениях

$$\sum_{y \in \mathcal{C}_i} \|y - y_i\|^2 \le \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2, \quad i = 1, 2. \tag{1.3}$$

Задача 1 имеет простую трактовку в виде следующей прикладной проблемы. Имеется совокупность промышленных изделий (например, микросхем), содержащая две отличающиеся (по характеристикам) группы однородных изделий и группу бракованных изделий. Характеристики изделий в однородных группах имеют некоторый разброс, обусловленный технологией производства. Допустимое значение разброса задано пороговым значением. Изделия с характеристиками, превышающими этот порог, считаются бракованными. Соответствие изделия и группы неизвестно. Требуется найти две группы однородных изделий, содержащих наибольшее число допустимых элементов, и отделить группу бракованных изделий. Сходную трактовку имеет задача 2. В этих задачах координата точки соответствует характеристике изделий в изложенной прикладной проблеме.

Характеристики бракованных изделий соответствуют так называемым (в статистике, анализе данных и распознавании образов) "выбросам", т. е. случайным вырожденным элементам, которые "нарушают" внутрикластерную однородность выборочных данных.

Обе задачи можно трактовать как поиск двух непересекающихся подмножеств (кластеров) 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 $\mathcal{Y} \subset \mathbb{R}^d$.

Из формулировки задач видно, что объединение искомых подмножеств может не покрывать входное множество. Правая часть неравенств (1.2) и (1.3) определяет α-долю от суммарного квадратичного разброса точек входного множества относительно его центроида. Левая часть неравенств (1.2) и (1.3) при каждом $i = 1, 2$ определяет суммарный внутрикластерный разброс точек из кластера $\mathcal{C}_i$ относительно заданной точки z_i из $\mathbb{R}^d$ в задаче 1 и относительно искомой точки y_i из $\mathcal{Y}$ в задаче 2. Уровень концентрации точек в кластерах определяется α-долей от суммарного квадратичного разброса точек во входном множестве. В обеих задачах требуется максимизировать минимальный размер кластера при ограничении сверху на уровень внутрикластерной концентрации точек.

Напомним, что кластеризация данных, т. е. поиск (выявление) подмножеств в данных и разбиение данных на подмножества по самым разнообразным критериям внутрикластерной

однородности (похожести элементов из одного кластера) и отделимости (непохожести) кластеров, является одной из ключевых проблем в машинном обучении (machine learning) [1], интерпретации данных (data mining) [2], распознавании образов (pattern recognition) [3] и обработке большеразмерных данных (big-scaling data processing) [4]. Поиск однородных подмножеств типичен также для редактирования совокупности данных (data editing) [5] и их очистки (data cleaning) [6] от непредвиденных случайных элементов в виде "выбросов". В упомянутых прикладных проблемах очистка имеющихся данных от "выбросов", как известно, является необходимой процедурой, так как эти, по сути, искаженные данные могут существенно ухудшить качество машинного обучения.

В плане теоретической мотивации подчеркнем, что рассматриваемые задачи не эквивалентны ни одной из хорошо известных кластеризационных геометрических задач — *k-means* (или k-MSSC) [7], *k-median* [8], *k-center* [9], *k-center clustering with outliers* [10]. В этих задачах, как известно, выпуклые оболочки оптимальных кластеров не пересекаются. В рассматриваемых задачах, в отличие от вышеупомянутых, при фиксированном разбросе точек входного множества в зависимости от заданного на входе значения α эти оболочки, очевидно, могут как пересекаться, так и не пересекаться. Именно этот факт привлекателен в плане практического анализа данных с невыясненной структурой. Изменяя значения α на входе и последовательно решая задачи при каждом α, можно найти набор кластеризаций данных и получить представление об их кластерной структуре. Задачи 1 и 2 также не эквивалентны недавно выявленным [11] геометрическим задачам разбиения на кластеры с центрами в начале координат, в которых выпуклые оболочки оптимальных кластеров могут пересекаться.

Наконец, заметим, что в области дискретной оптимизации ключевым является вопрос о сложностном статусе какой-либо вновь выявленной экстремальной задачи. Ниже мы показываем, что рассматриваемые экстремальные задачи и их многокластерные обобщения NP-трудны даже в одномерном случае, т. е. на прямой.

2. Анализ вычислительной сложности

Положим

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

$$f(\mathcal{C}_i, z_i) = \sum_{y \in \mathcal{C}_i} \|y - z_i\|^2, \quad \mathcal{C}_i \subseteq \mathcal{Y}, \quad z_i \in \mathbb{R}^d, \quad i = 1, 2. \tag{2.1}$$

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

З а д а ч а 1*A*.

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

$$\min\{|\mathcal{C}_1|, |\mathcal{C}_2|\} \geq M \tag{2.2}$$

при ограничениях

$$f(\mathcal{C}_i, z_i) \leq A, \quad i = 1, 2? \tag{2.3}$$

Теорема 1. *Задача* 1*A NP-полна даже в одномерном случае.*

Д о к а з а т е л ь с т в о. Напомним известную [12] NP-полную задачу.

З а д а ч а "Разбиение 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

$$\sum_{x \in \mathcal{S}_1} x = \sum_{x \in \mathcal{S}_2} x? \tag{2.4}$$

Заметим, что задача $1A$, очевидно, лежит в классе NP. Сведем задачу "Разбиение 1" к задаче $1A$ и покажем, что это сведение полиномиально.

По произвольному входу задачи "Разбиение 1" построим следующий пример входа задачи $1A$. В задаче $1A$ положим

$$d = 1, \quad N = 2K, \quad M = K, \quad A = \frac{1}{2}\sum_{x \in \mathcal{X}} x, \quad z_1 = z_2 = 0, \quad \mathcal{Y} = \{y_1, \ldots, y_{2K}\},$$

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

Покажем, что в построенном примере задачи $1A$ мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ из $\mathcal{Y}$, удовлетворяющие неравенствам (2.2) и (2.3), 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 (2.4).

Пусть в задаче "Разбиение 1" существуют требуемые мультиподмножества

$$\mathcal{S}_i = \{x_k \mid k \in I_i\}, \quad i = 1, 2, \tag{2.5}$$

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

$$I_1 \cup I_2 = \{1, \ldots, 2K\}, \quad I_1 \cap I_2 = \varnothing. \tag{2.6}$$

В задаче $1A$ рассмотрим мультиподмножества

$$\mathcal{C}_i = \{y_k \mid k \in I_i\}, \quad i = 1, 2. \tag{2.7}$$

Поскольку в задаче "Разбиение 1" $|\mathcal{S}_1| = |\mathcal{S}_2| = K$, для мультиподмножеств (2.7) имеем $|\mathcal{C}_1| = |\mathcal{C}_2| = K$ и $\min\{|\mathcal{C}_1|, |\mathcal{C}_2|\} = K = M$. Кроме того, для внутрикластерных разбросов (2.1) имеем

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

Следовательно, условия (2.2) и (2.3) выполнены, и в задаче 1A требуемые мультиподмножества $\mathcal{C}_1$ и $\mathcal{C}_2$ существуют.

Пусть теперь в задаче $1A$ в мультимножестве $\mathcal{Y}$ существуют требуемые мультиподмножества (2.7), удовлетворяющие условиям (2.2) и (2.3). В задаче "Разбиение 1" рассмотрим мультиподмножества (2.5). Для этих мультиподмножеств имеем

$$|\mathcal{S}_1| = |\mathcal{S}_2|, \quad \sum_{k \in I_i} x_k = \sum_{x \in \mathcal{S}_i} x = \sum_{y \in \mathcal{C}_i} \|y\|^2 = \frac{1}{2}\sum_{x \in \mathcal{X}} x, \quad i = 1, 2.$$

Поэтому $\sum\limits_{x \in \mathcal{S}_1} x = \sum\limits_{x \in \mathcal{S}_2} x$. Следовательно, в задаче "Разбиение 1" требуемые мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ также существуют.

Теорема доказана.

З а м е ч а н и е 1. Для простоты доказательства мы использовали числа $c_k = \sqrt{x_k}$, $k = 1, \ldots, 2K$, которые могут быть иррациональными. Однако, легко видеть, что те же доказательные аргументы справедливы для рациональных чисел c_k таких, что

$$0 \leq \sqrt{x_k} - c_k < 1/(4K\lceil \max_k \sqrt{x_k} \rceil), \quad k = 1, \ldots, 2K,$$

так как x_k — целое число. Остается заметить, что построенное сведение, очевидно, полиномиально.

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

З а д а ч а 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 $\mathcal{C}_1, \ldots, \mathcal{C}_J$ во множестве $\mathcal{Y}$ такие, что

$$\min\{|\mathcal{C}_1|, \ldots, |\mathcal{C}_J|\} \to \max \tag{2.8}$$

при ограничениях

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

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 следующее утверждение.

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.*

Д о к а з а т е л ь с т в о. Сформулируем задачу в форме верификации свойств и покажем ее NP-полноту.

З а д а ч а 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 мультиподмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ в мультимножестве $\mathcal{Y}$ такие, что

$$\min\{|\mathcal{C}_1|, \ldots, |\mathcal{C}_J|\} \ge M \tag{2.9}$$

при ограничениях

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

Построим полиномиальное сведение задачи $1A(J)$ к задаче $1A(J+1)$.

По входу задачи $1A(J)$ построим следующий пример входа задачи $1A(J+1)$. В задаче $1A(J+1)$ положим

$$\widetilde{\mathcal{Y}} = \mathcal{Y} \cup \mathcal{G}, \quad \widetilde{N} = N + M, \quad \widetilde{d} = d, \quad \widetilde{A} = A, \quad \widetilde{M} = M, \tag{2.10}$$

$$\mathcal{G} = \{g_1, \ldots, g_M\}, \quad g_i = L, \quad i = 1, \ldots, M, \tag{2.11}$$

$$\widetilde{z}_i = z_i, \quad i = 1, \ldots, J, \quad \widetilde{z}_{J+1} = L,$$

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

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

Пусть теперь в задаче $1A(J+1)$ существуют требуемые мультиподмножества $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_{J+1}$. Покажем, что ни одна из точек g_i, $i = 1, \ldots, M$, не входит в $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_J$. Предположим противное, например, $g_1 \in \widetilde{\mathcal{C}}_1$. Тогда

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

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

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

Теорема доказана.

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

З а д а ч а 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 (2.2) при ограничениях

$$f(\mathcal{C}_i, y_i) \le A, \quad i = 1, 2? \tag{2.12}$$

Справедлива следующая теорема.

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

Д о к а з а т е л ь с т в о. Напомним хорошо известную [12] NP-полную задачу.

З а д а ч а "Разбиение 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 (2.4)?

Задача 2A, очевидно, лежит в классе NP. Построим сведение задачи "Разбиение 2" к задаче 2A и покажем, что это сведение полиномиально.

По произвольному входу задачи "Разбиение 2" построим следующий пример входа задачи 2A. Положим

$$d = 1, \quad N = 2K + 2, \quad M = K + 1, \quad \mathcal{Y} = \{b_1, b_2, a_1, \ldots, a_K, c_1, \ldots, c_K\},$$

где

$$a_k = 0, \quad c_k = \sqrt{x_k}, \quad k = 1, \ldots, K, \quad b_1 = b_2 = -B,$$

причем

$$B > \max\left\{\frac{1}{4}\sum_{x\in\mathcal{X}} x, \sqrt{\frac{1}{2}\sum_{x\in\mathcal{X}} x}\right\}, \quad A = B^2 + \frac{1}{2}\sum_{x\in\mathcal{X}} x.$$

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

Покажем, что в построенном примере задачи 2A мультиподмножества $\mathcal{C}_1$, $\mathcal{C}_2$ и точки y_1, y_2 из $\mathcal{Y}$, удовлетворяющие неравенствам (2.2) и (2.12), существуют тогда и только тогда, когда в задаче "Разбиение 2" существуют мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ такие, что имеет место равенство (2.4).

Пусть в задаче "Разбиение 2" мультиподмножества $\mathcal{S}_1$ и $\mathcal{S}_2$ существуют. Пусть эти мультиподмножества, как и ранее, определяются формулами (2.5), в которых, в отличие от (2.6),

$$I_1, I_2 \subseteq \{1, \ldots, K\}, \quad I_1 \cup I_2 = \{1, \ldots, K\}, \quad I_1 \cap I_2 = \varnothing.$$

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

$$\mathcal{C}_i = \{b_i\} \cup \{c_k \mid k \in I_i\} \cup \{a_k \mid k \in I_{3-i}\}, \quad i = 1, 2. \tag{2.13}$$

Для этих мультиподмножеств из (2.1) имеем следующие внутрикластерные разбросы:

$$f(\mathcal{C}_i, y_i) = \|b_i - y_i\|^2 + \sum_{k\in I_i} \|c_k - y_i\|^2 = B^2 + \sum_{k\in I_i} x_k = B^2 + \frac{1}{2}\sum_{k\in\{1,\ldots,K\}} x_k = A, \quad i = 1, 2.$$

Это равенство означает, что условие (2.3) выполнено. Условие (2.2) выполнено по построению. Таким образом, в задаче 2A 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 2A. Покажем тогда, что точки b_1 и b_2 лежат в различных мультиподмножествах.

В самом деле, если, например, $b_1, b_2 \in \mathcal{C}_1$, то поскольку $|\mathcal{C}_1| = K+1 \geq 4$, для любого выбора y_1 имеем $f(\mathcal{C}_1, y_1) \geq 2B^2 > A$, что противоречит условию $f(\mathcal{C}_1, y_1) \leq A$.

Кроме того, покажем, что в паре $\{\mathcal{C}_1, \mathcal{C}_2\}$ нет мультиподмножества, содержащего все точки c_k, $k = 1, \ldots, K$. Действительно, в противном случае положим, например, $c_k \in \mathcal{C}_1$, $k = 1, \ldots, K$. Тогда, если $y_1 = -B$ или $y_1 = c_r$ для некоторого $r \in \{1, \ldots, K\}$, то

$$f(\mathcal{C}_1, y_1) \geq (B + \min_{k=1,\ldots,K} c_k)^2 > B^2 + 2B > A,$$

что противоречит условию $f(\mathcal{C}_1, y_1) \leq A$. Поэтому $y_1 = 0$. Но тогда

$$f(\mathcal{C}_1, y_1) = B^2 + \sum_{k=1,\ldots,K} c_k^2 = B^2 + \sum_{x \in \mathcal{X}} x > A.$$

Получили противоречие.

Далее, так как оба мультиподмножества $\mathcal{C}_1$ и $\mathcal{C}_2$ содержат не менее одной точки c_k, $k \in \{1, \ldots, K\}$, и одну из точек b_1, b_2, то те же самые доказательные аргументы, как и выше, влекут равенства $y_1 = y_2 = 0$.

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 (2.13). Тогда, используя эту формулу и равенства $y_i = 0$, $i = 1, 2$, получим следующую оценку для внутрикластерных разбросов:

$$f(\mathcal{C}_i, y_i) = \|b_i - y_i\|^2 + \sum_{k \in I_i} \|c_k - y_i\|^2 = B^2 + \sum_{k \in I_i} x_k \leq A = B^2 + \frac{1}{2} \sum_{k \in \{1,\ldots,K\}} x_k, \quad i = 1, 2.$$

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 $\mathcal{S}_1$ and $\mathcal{S}_2$ также существуют.

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

Теорема доказана.

Из теоремы 3 следует, что задача 2 NP-трудна, как и сформулированное ниже ее 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 (2.8) при ограничениях

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

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 следующее утверждение.

Теорема 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.*

Д о к а з а т е л ь с т в о. Сформулируем задачу в форме верификации свойств и покажем ее NP-полноту.

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 $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и точки $y_1, \ldots, y_J$ во множестве $\mathcal{Y}$ такие, что имеет место соотношение (2.9) при ограничениях

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

Построим полиномиальное сведение задачи $2A(J)$ к задаче $2A(J+1)$.

По входу задачи $2A(J)$ построим следующий пример входа задачи $2A(J+1)$. Определим вход задачи $2A(J+1)$ равенствами (2.10) и (2.11), в которых

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

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

Пусть теперь в задаче $2A(J+1)$ существуют требуемые мультиподмножества $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_{J+1}$ и точки $\widetilde{y}_1, \ldots, \widetilde{y}_{J+1}$. Покажем, что из мультимножеств $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_{J+1}$ не более чем одно содержит точки из $\mathcal{G}$.

Предположим противное, например, $g_1 \in \widetilde{\mathcal{C}}_1$, $g_2 \in \widetilde{\mathcal{C}}_2$. Тогда из $|\mathcal{G}| = M$ и условия (2.9) следует, что $\widetilde{\mathcal{C}}_1$ (как и $\widetilde{\mathcal{C}}_2$) содержит хотя бы одну точку $y \in \mathcal{Y}$. Поэтому, если $y_1 \in \mathcal{G}$, то

$$f(\mathcal{C}_1, y_1) \ge \|y - y_1\|^2 > A,$$

что противоречит условию $f(\mathcal{C}_1, y_1) \le A$. Если же $y_1 \in \mathcal{Y}$, то

$$f(\mathcal{C}_1, y_1) \ge \|g_1 - y_1\|^2 > A,$$

что также противоречит условию $f(\mathcal{C}_1, y_1) \le A$.

Следовательно, хотя бы J из мультимножеств $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_{J+1}$ не содержат точек из $\mathcal{G}$. Будем считать, что это мультимножества $\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_J$. Тогда требуемыми мультиподмножествами и точками в задаче 2A(J) являются $\mathcal{C}_1 = \widetilde{\mathcal{C}}_1, \ldots, \mathcal{C}_J = \widetilde{\mathcal{C}}_J$, $y_1 = \widetilde{y}_1, \ldots, y_J = \widetilde{y}_J$.

Теорема доказана.

Таким образом, все рассмотренные задачи кластеризации NP-трудны даже на числовой прямой.

Легко заметить, что NP-трудность задачи 1 следует из NP-трудности задачи 2. Действительно, для решения задачи 2 достаточно для каждой из N^2 пар $z_1, z_2 \in \mathcal{Y}$ решить пример задачи 1. Тем не менее, мы привели независимые доказательства труднорешаемости этих задач с использованием различных классических NP-трудных задач. На наш взгляд, независимые доказательства будут полезны в плане выяснения вопросов алгоритмической аппроксимируемости рассматриваемых задач.

Заключение

В работе показана NP-трудность некоторых ранее не изученных максиминных задач поиска непересекающихся множеств в конечном множестве точек евклидова пространства. Показано, что эти задачи NP-трудны даже в одномерном случае. Построение эффективных алгоритмов с теоретическими гарантиями качества (точности, временной сложности, надежности) для этих задач представляется делом ближайшей перспективы.

СПИСОК ЛИТЕРАТУРЫ

1. An introduction to statistical learning / G. James, D. Witten, T. Hastie, R. Tibshirani. N Y: Springer Science+Business Media, LLC, 2013. 426 p.
2. **Aggarwal C. C.** Data mining: The textbook. N Y: Springer International Publishing, 2015. 734 p. doi: 10.1007/978-3-319-14142-8 .
3. **Bishop C. M.** Pattern recognition and machine learning. N Y: Springer Science+Business Media, LLC, 2006. 738 p.
4. Big data clustering: A review / A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, T. Herawan // Lecture Notes in Computer Science. 2014. Vol. 8583. P. 707–720. doi: 10.1007/978-3-319-09156-3_49 .
5. **Alessio Farcomeni, Greco Luca.** Robust methods for data reduction. Boca Raton: CRC Press, 2015. 297 p.
6. **Osborne, Jason W.** Best practices in data cleaning: A complete quide to everything you need to do before and after collecting your data. Los Angeles: SAGE Publication, Inc., 2013. 296 p.
7. NP-hardness of Euclidean sum-of-squares clustering / D. Aloise, A. Deshpande, P. Hansen, P. Popat // Machine Learning, 2009. Vol. 75, no. 2. P. 245–248. doi: 10.1007/s10994-009-5103-0 .
8. **Papadimitriou C. H.** Worst-case and probabilistic analysis of a geometric location problem // SIAM J. Comput. 1981. Vol. 10, no. 3. P. 542–557. doi: 10.1137/0210040 .
9. **Masuyama S., Ibaraki T., Hasegawa T.** The computational complexity of the m-center problems in the plane // Trans. IECE Jpn. 1981. Vol. 64, no. 2. P. 57–64.
10. **Hatami B, Zarrabi-Zade H.** A streaming algorithm for 2-center with outliers in high dimensions // Computational Geometry. 2017. Vol. 60. P. 26–36. doi: 10.1016/j.comgeo.2016.07.002 .
11. **Кельманов А. В., Пяткин А. В.** NP-трудность некоторых евклидовых задач разбиения конечного множества точек // Журн. вычисл. математики и мат. физики. 2018. Vol. 58, no. 5. С. 852–856. doi: 10.7868/S0044466918050149 .
12. **Garey M. R., Johnson D. S.** Computers and intractability: A quide to the theory of NP-completeness. San Francisco: Freeman, 1979. 338 p.

Поступила 30.07.2018
После доработки 08.10.2018
Принята к публикации 12.11.2018

Кельманов Александр Васильевич
д-р физ.-мат. наук, зав. лабораторией
Институт математики им. С. Л. Соболева СО РАН;
Новосибирский государственный университет,
г. Новосибирск
e-mail: kelm@math.nsc.ru

Пяткин Артем Валерьевич
д-р физ.-мат. наук, зав. лабораторией
Институт математики им. С. Л. Соболева СО РАН;
Новосибирский государственный университет,
г. Новосибирск
e-mail: artem@math.nsc.ru

Хандеев Владимир Ильич
канд. физ.-мат. наук, науч. сотрудник
Институт математики им. С. Л. Соболева СО РАН;
Новосибирский государственный университет,
г. Новосибирск
e-mail: khandeev@math.nsc.ru

REFERENCES

1. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning*. N Y: Springer Science+Business Media, LLC, 2013, 426 p. ISBN: 978-1461471370 .

2. Aggarwal C.C. *Data Mining: The Textbook*. N Y: Springer International Publishing, 2015, 734 p. doi: 10.1007/978-3-319-14142-8 .
3. Bishop C.M. *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, LLC, 2006, 738 p. ISBN: 978-0-387-31073-2 .
4. Shirkhorshidi A.S., Aghabozorgi S, Wah T,Y., and Herawan T. Big data clustering: A review. In: Murgante B. et al. (eds) *Computational Science and Its Applications - ICCSA 2014. Lecture Notes in Computer Science*, 2014, vol. 8583, pp. 707–720. doi: 10.1007/978-3-319-09156-3_49 .
5. Alessio Farcomeni, Greco Luca. *Robust methods for data reduction*. Boca Raton: CRC Press, 2015, 297 p. ISBN: 9781466590625 .
6. Osborne, Jason W. *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*, Los Angeles: SAGE Publication, Inc., 2013, 296 p. ISBN: 9781412988018 .
7. Aloise D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 2009, vol. 75, no. 2, pp. 245–248. doi: 10.1007/s10994-009-5103-0 .
8. Papadimitriou C. H. Worst-case and probabilistic analysis of a geometric location problem. *SIAM J. Comput.*, 1981, vol. 10, no. 3, pp. 542–557. doi: 10.1137/0210040 .
9. Masuyama S., Ibaraki T., Hasegawa T. The computational complexity of the m-center problems in the plane. *Trans. IECE Jpn.*, 1981, vol. 64, no. 2, pp. 57–64.
10. Hatami B., Zarrabi-Zade H. A streaming algorithm for 2-center with outliers in high dimensions. *Computational Geometry*, 2017, vol. 60, pp. 26–36. doi: 10.1016/j.comgeo.2016.07.002 .
11. Kel'manov A. V., Pyatkin A. V. NP-hardness of some Euclidean problems of partitioning a finite set of points. *Comput. Math. Math. Phys.*, 2018, vol. 58, no. 5, pp. 822–826. doi: 10.1134/S0965542518050123 .
12. Garey M.R., Johnson D.S. Computers and intractability: A guide to the theory of NP-completeness. San Francisco: Freeman, 1979, 340 p. ISBN: 978-0716710455 .

Received July 30, 2018
Revised October 08, 2018
Accepted November 12, 2018

Funding Agency: This work was supported by the Russian Science Foundation (project no. 16-11-10041).

Alexander Vasil'evich Kel'manov, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630090 Russia, e-mail: kelm@math.nsc.ru .

Artem Valer'evich Pyatkin, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630090 Russia, e-mail: artem@math.nsc.ru .

Vladimir Il'ich Khandeev, Cand. Sci. (Phys.-Math.), Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630090 Russia, e-mail: khandeev@math.nsc.ru .