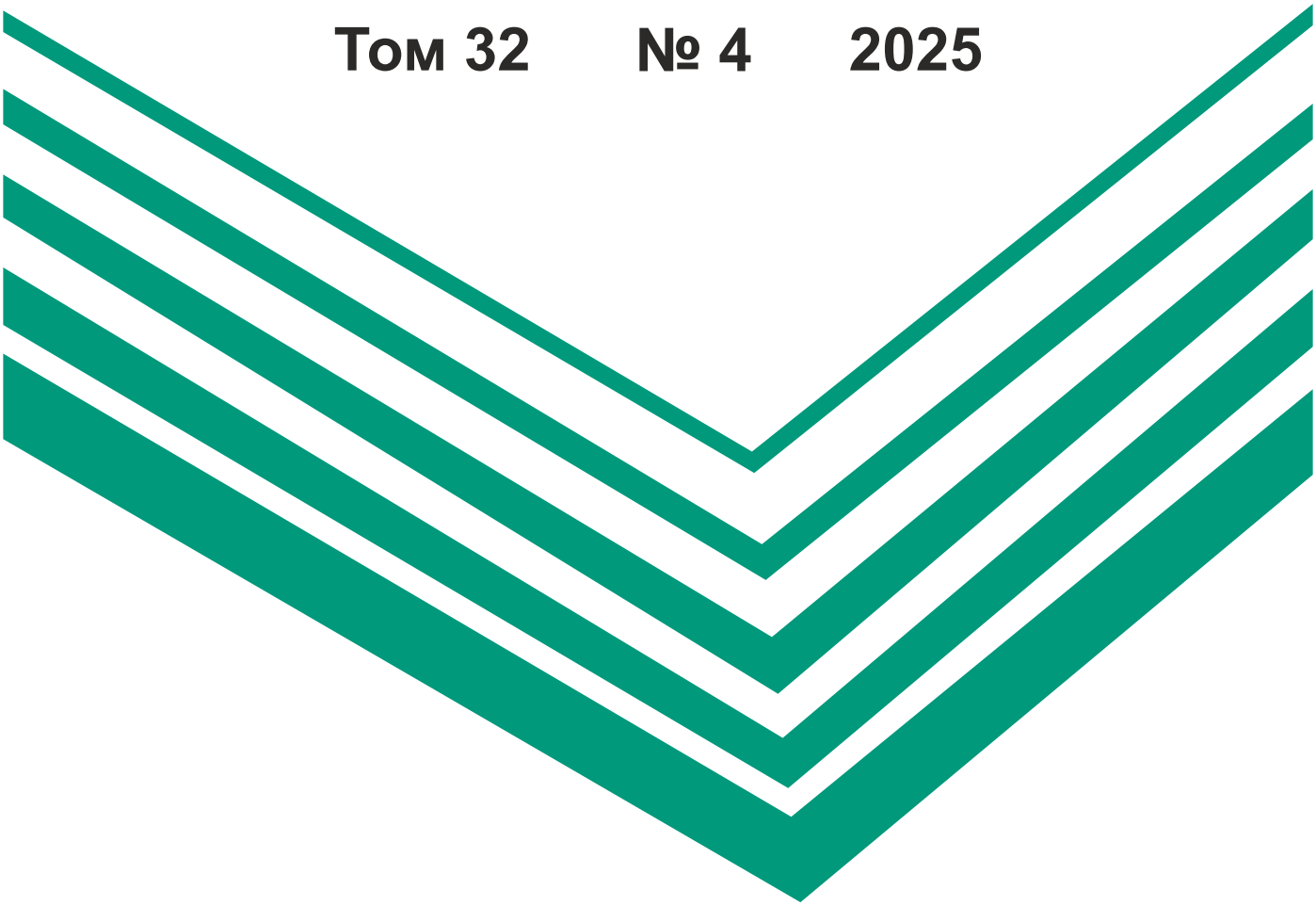
ISSN 2949-5598

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ

Том 32 № 4 2025



Новосибирск
Издательство Института математики

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Октябрь–декабрь 2025. Т. 32, № 4. С. 172–190

УДК 519.8 DOI: 10.33048/daio.2025.32.834

О СЛОЖНОСТИ ДВУХ ЗАДАЧ ПОИСКА КЛАСТЕРОВ С БОЛЬШОЙ МОЩНОСТЬЮ

С. М. Нещадим[1, a], *В. И. Хандеев*[2, b]

[1] Новосибирский гос. университет,
ул. Пирогова, 2, 630090 Новосибирск, Россия
[2] Институт математики им. С. Л. Соболева,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия
E-mail: [a] s.neshchadim@g.nsu.ru, [b] khandeev@math.nsc.ru

Аннотация. Для конечного множества точек евклидова пространства рассматриваются задачи поиска его непересекающихся подмножеств. В одной из задач мощность каждого из подмножеств должна быть не меньше заданного числа. В другой задаче мощность всех подмножеств одинакова, а их объединение должно совпадать с исходным множеством. В обеих задачах дополнительно требуется, чтобы для каждого подмножества сумма квадратов расстояний до центроида не превосходила заданной величины. Доказано, что задачи NP-полны в сильном смысле в случае, когда число кластеров равно двум, а размерность пространства является частью входа задачи. Кроме того, доказано, что задачи NP-полны в одномерном случае для любого фиксированного числа кластеров. Ил. 2, библиогр. 16.

Ключевые слова: кластеризация, наименьший размер кластера, ограниченный разброс, евклидово пространство, NP-полнота.

Введение

Предметом исследования настоящей статьи является задача поиска семейства непересекающихся подмножеств в конечном множестве точек евклидова пространства при ограничении снизу на мощность каждого подмножества и ограничении сверху на его разброс. Также рассматривается частный случай, в котором объединение искомого семейства совпадает со всем входным множеством. Целью исследования является установление сложностного статуса ранее не изученных подслучаев этих задач.

Одной из первых и наиболее известных задач кластеризации является задача k-means (k-средних), также именуемая MSSC (Minimum Sum of Squares Clustering) [1, 2]. Суть задачи состоит в том, чтобы разбить

множество точек в евклидовом пространстве на k кластеров таким образом, чтобы минимизировать сумму квадратов расстояний между каждой точкой и центром соответствующего ей кластера. Центром выступает центроид, т. е. среднее арифметическое точек внутри кластера. Для приближённого решения этой задачи часто применяют эвристический алгоритм k-means, который находит широкое применение в различных областях: от здравоохранения [3] до сегментации изображений [4] и анализа сетевого трафика [5]. При этом важно отметить, что для разбиения на два кластера задача NP-трудна [6], тогда как для одномерного случая существует алгоритм с полиномиальной сложностью [7]. Задача также разрешима за полиномиальное время при $k = 1$, поскольку в этом случае оптимальное решение представляет собой исходное множество точек целиком.

В работе рассматривается следующая похожая задача кластеризации.

Задача 1. *Даны множество* $\mathcal{Y} = \{y_1, \ldots, y_n\}$ *векторов из* $\mathbb{R}^d$ *и числа* $A \in \mathbb{R}_+$, $M \in \mathbb{N}$. *Существуют ли* k *попарно непересекающихся множеств* $\mathcal{C}_i \subset \mathcal{Y}$, $i = 1, \ldots, k$, *таких, что* $|\mathcal{C}_i| \geqslant M$ *и* $F(\mathcal{C}_i) \leqslant A$ *при* $i = 1, \ldots, k$, *где*

$$F(\mathcal{C}_i) = \sum_{y \in \mathcal{C}_i} \|y - \overline{y}(\mathcal{C}_i)\|^2,$$

$$\overline{y}(\mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|} \sum_{y \in \mathcal{C}_i} y, \quad i = 1, \ldots, k?$$

Заметим, что в отличие от задачи MSSC задача 1 в оптимизационной форме может быть записана как задача поиска семейства непересекающихся кластеров с максимальной мощностью минимального (по числу элементов) кластера и ограничением на разброс каждого кластера. При этом поиск кластеров большего размера, удовлетворяющих определённым ограничениям, применяется в таких областях, как, например, задачи анализа социальных сетей [8], биоинформатика [9] и т. д. Если же в задаче 1 дополнительно потребовать, чтобы все подмножества имели одинаковую мощность и в объединении совпадали со всем множеством $\mathcal{Y}$, то получим следующую задачу.

Задача 2. *Даны множество* $\mathcal{Y} = \{y_1, \ldots, y_n\}$ *векторов из* $\mathbb{R}^d$, *где* $n = kM$, *и число* $A \in \mathbb{R}_+$. *Существует ли разбиение множества* $\mathcal{Y}$ *на* k *множеств* $\mathcal{C}_i$, $i = 1, \ldots, k$, *такое, что* $|\mathcal{C}_i| = M$ *и* $F(\mathcal{C}_i) \leqslant A$?

Задача 2 является частным случаем задачи 1, а значит, задача 1 NP-полна во всех случаях, когда NP-полна задача 2.

В [10] рассмотрена задача VS-2 (в оптимизационной формулировке также известная как M-Variance [11]), которая представляет собой частный случай задачи 1 при $k = 1$. В [10] показано, что такая задача NP-

полна в сильном смысле, если размерность является частью входа задачи. При этом в одномерном случае ($d = 1$) при $k = 1$ задача M-Variance полиномиально разрешима [12].

Как и задача MSSC, задача 2 при $k = 1$ полиномиально разрешима: достаточно проверить разброс единственного возможного решения, а именно всего входного множества $\mathcal{Y}$.

В [13] построено полиномиальное сведение задачи о сбалансированной k-раскраске однородного графа к задаче 1. Тем самым из того, что задача о сбалансированной k-раскраске NP-полна при $k \geqslant 3$, следует NP-полнота задачи 1 в случае, когда размерность пространства является частью входа задачи, а число кластеров k фиксировано и больше либо равно трём. Более того, поскольку в сведении, представленном в [13], векторы в задаче 1 содержат только числа 0, 1 и -1, все кластеры имеют одинаковую мощность, а их объединение совпадает со всем входным множеством точек $\mathcal{Y}$, из этого сведения следует более строгое утверждение — NP-полнота в сильном смысле задачи 2 (как и задачи 1) при тех же условиях.

Таким образом, оставался открытым вопрос о сложностном статусе задач 1 и 2 в случае фиксированного $k = 2$. При этом похожие задачи, в которых в формуле разброса кластера используется не геометрический центр, а фиксированная точка евклидова пространства либо произвольная точка входного множества, NP-полны даже в одномерном случае [14].

В данной работе докажем, что задача 2 (а следовательно, и задача 1) NP-полна в сильном смысле в случае, когда $k = 2$ и размерность пространства является частью входа задачи. Также докажем, что обе эти задачи NP-полны в обычном смысле в случае одномерного пространства при произвольном фиксированном $k \geqslant 2$.

1. Вспомогательные задачи

В работе будет использоваться NP-полная [15] задача Максимальный разрез.

Задача Максимальный разрез. *Даны граф $G = (V, E)$ и целое $L_1 > 0$. Верно ли, что существует разбиение множества V на два таких непересекающихся множества V_1 и V_2, что число рёбер, соединяющих вершины из множеств V_1 и V_2, не меньше L_1?*

Кроме того, нам понадобится следующая вспомогательная задача.

Задача Два разреженных подграфа. *Даны граф $G = (V, E)$ и целое $L_2 > 0$. Верно ли, что существует разбиение множества V на два непересекающихся множества V_1 и V_2 таких, что $|V_1| = |V_2|$, а число*

рёбер для каждого $i = 1, 2$, *соединяющих вершины из* V_i *между собой, не превосходит* L_2?

Покажем, что данная задача NP-полна.

Теорема 1. *Задача* Два разреженных подграфа NP-*полна.*

Доказательство. Построим полиномиальное сведение задачи Максимальный разрез к задаче Два разреженных подграфа. Рассмотрим произвольный пример задачи Максимальный разрез — граф $G = (V, E)$ и число $L_1 \in \mathbb{Z}_+$. По этому примеру можно за полиномиальное время построить следующий пример задачи Два разреженных подграфа — граф $G^* = (V^*, E^*)$, состоящий из графа G и его копии, которую будем обозначать $\widetilde{G} = (\widetilde{V}, \widetilde{E})$, и число $L_2 = |E| - L_1$. Покажем, что исходный и построенный примеры одновременно либо имеют решения, либо нет.

Пусть для примера задачи Максимальный разрез существует решение — разбиение множества V вершин графа G на непересекающиеся подмножества V_1 и V_2 такие, что мощность множества

$$E_{12} = \{e \in E \mid \exists\, v_1 \in V_1,\, v_2 \in V_2 \colon e = (v_1, v_2)\}$$

рёбер, соединяющих вершины из V_1 и V_2, не меньше L_1. Рассмотрим следующую пару множеств вершин графа G^*:

$$V_1^* = V_1 \cup \widetilde{V}_2, \quad V_2^* = V_2 \cup \widetilde{V}_1,$$

где $\widetilde{V}_1$ и $\widetilde{V}_2$ — копии множеств V_1 и V_2, образующие вместе множество вершин графа $\widetilde{G}$. На рис. 1 показан пример с изображением графа G, его копии $\widetilde{G}$ и множеств V_1^*, V_2^*.

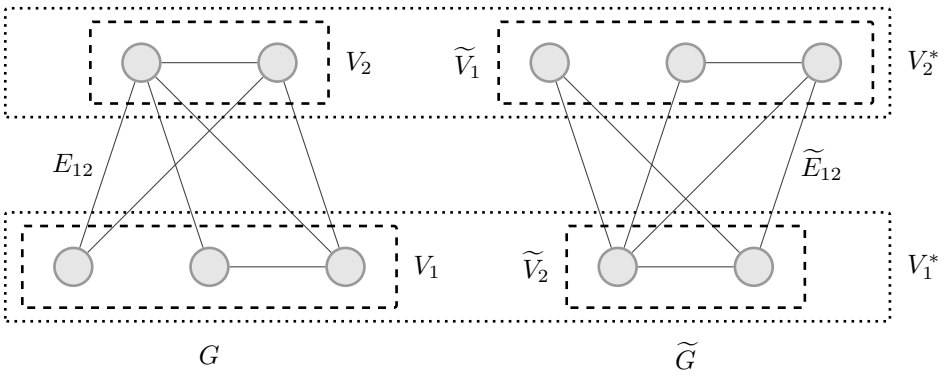


Рис. 1. Пример задачи Два разреженных подграфа

Покажем, что множества V_1^* и V_2^* являются решением построенного примера задачи Два разреженных подграфа. Множества V_i^* не пересекаются по построению; равенство мощностей $|V_1^*|$ и $|V_2^*|$ следует из того, что

$$|V_1| = |\widetilde{V}_1|, \quad |V_2| = |\widetilde{V}_2|.$$

Для каждого $i = 1, 2$ оценим число рёбер графа G^*, соединяющих вершины из V_i^* между собой. Обозначим множество всех таких рёбер через E_i^*. Множество E_1^* состоит из рёбер, соединяющих вершины из V_1, а также из рёбер, соединяющих вершины из $\widetilde{V}_2$. В свою очередь, множество E_2^* состоит из рёбер, соединяющих вершины из $\widetilde{V}_1$, а также из рёбер, соединяющих вершины из V_2. Следовательно,

$$|E_1^*| = |E_2^*| = |E_1| + |E_2|,$$

где E_i — множество рёбер графа G, соединяющих вершины из множества V_i между собой, $i = 1, 2$. Поскольку по предположению мощность множества E_{12} не меньше L_1, имеем

$$|E_1^*| = |E_2^*| = |E| - |E_{12}| \leqslant |E| - L_1 = L_2.$$

Таким образом, множества V_1^* и V_2^* являются решением построенного примера задачи Два разреженных подграфа.

Теперь предположим, что в задаче Два разреженных подграфа существует решение — требуемые множества V_1^* и V_2^* вершин. Пусть

$$V_1 = V_1^* \cap V, \quad V_2 = V_2^* \cap V, \quad \widetilde{V}_1 = V_2^* \cap \widetilde{V}, \quad \widetilde{V}_2 = V_1^* \cap \widetilde{V}.$$

Другими словами, разбиение вершин графа G^* на V_1^* и V_2^* порождает разбиение вершин графа G на V_1 и V_2, а вершин графа $\widetilde{G}$ — на $\widetilde{V}_2$ и $\widetilde{V}_1$. Множества E_1, E_2, E_{12} ($\widetilde{E}_1$, $\widetilde{E}_2$, $\widetilde{E}_{12}$) рёбер определим так же, как в первой части доказательства, используя множества вершин V_1, V_2 ($\widetilde{V}_1$, $\widetilde{V}_2$). Тогда по предположению $|E_1| + |\widetilde{E}_2| \leqslant L_2$, $|E_2| + |\widetilde{E}_1| \leqslant L_2$. Покажем, что как минимум одна из пар — V_1, V_2 или $\widetilde{V}_1, \widetilde{V}_2$ — является решением исходного примера задачи Максимальный разрез. Для этого оценим сумму мощностей множеств E_{12} и $\widetilde{E}_{12}$:

$$\begin{aligned} |E_{12}| + |\widetilde{E}_{12}| &= (|E| - |E_1| - |E_2|) + (|\widetilde{E}| - |\widetilde{E}_1| - |\widetilde{E}_2|) = \\ &= (|E| - |E_1| - |\widetilde{E}_2|) + (|E| - |E_2| - |\widetilde{E}_1|) \geqslant 2(|E| - L_2) = 2L_1. \end{aligned}$$

Из этой оценки следует, что хотя бы одна из мощностей $|E_{12}|$, $|\widetilde{E}_{12}|$ больше либо равна L_1, а значит, соответствующая пара множеств вершин (V_1, V_2 или $\widetilde{V}_1, \widetilde{V}_2$) будет решением исходной задачи Максимальный разрез.

Таким образом, произвольный пример задачи Максимальный разрез и построенный за полиномиальное время пример задачи Два разреженных подграфа одновременно либо имеют решения, либо не имеют, из чего следует, что задача Два разреженных подграфа NP-полна. Теорема 1 доказана.

Заметим, что задача Два разреженных подграфа эквивалентна следующей задаче на дополнении исходного графа.

Задача Два плотных подграфа. *Даны граф* $G = (V, E)$ *и целое* $L_3 > 0$. *Верно ли, что существует разбиение множества* V *на два непересекающихся множества* V_1 *и* V_2 *таких, что* $|V_1| = |V_2|$, *а число рёбер для каждого* $i = 1, 2$, *соединяющих вершины из* V_i *между собой, не меньше* L_3?

Действительно, поскольку в задаче Два разреженных подграфа $|V_1| = |V_2| = |V|/2$, существование разбиения на два равных по мощности подмножества вершин не более чем с L_2 рёбрами между элементами каждого из них эквивалентно существованию в дополнении графа разбиения на два равных по мощности подмножества вершин не менее чем с $\frac{1}{2} \cdot \frac{|V|}{2} \cdot \left(\frac{|V|}{2} - 1\right) - L_2$ рёбрами между элементами каждого из них. Таким образом, из теоремы 1 получаем

Следствие 1. *Задача* Два плотных подграфа NP-*полна.*

Наконец, в работе будет использоваться NP-полная [15] задача Разбиение. Заметим, что в этой задаче, добавив к входному множеству столько нулей, сколько элементов в этом множестве, можно дополнительно потребовать, чтобы искомое подмножество содержало ровно половину элементов входного множества. После этого можно считать, что все элементы ненулевые (в противном случае можно увеличить каждый элемент на 1). Таким образом, будет использоваться следующая NP-полная задача.

Задача Разбиение. *Дано множество* $\{n_1, \ldots, n_{2S}\} \subset \mathbb{N} \setminus \{0\}$ *натуральных чисел. Верно ли, что существует подмножество индексов* $\mathcal{I} \subset \{1, \ldots, 2S\}$, $|\mathcal{I}| = S$, *таких, что* $\sum\limits_{i \in \mathcal{I}} n_i = T$, *где* $2T = \sum\limits_{i=1}^{2S} n_i$?

2. Случай двух кластеров при произвольной размерности пространства

Далее будем рассматривать задачу 2, в которой число кластеров фиксировано (не является частью входа задачи). В этом разделе рассмотрим случай двух кластеров и произвольной размерности пространства

и покажем, что в этом случае задача 2 (а значит, и задача 1) NP-полна в сильном смысле, используя задачу Два плотных подграфа.

Теорема 2. *Задача 2 при $k = 2$ NP-полна в сильном смысле.*

Доказательство. Построим полиномиальное сведение задачи Два плотных подграфа к задаче 2 при $k = 2$.

Рассмотрим произвольный пример задачи Два плотных подграфа — граф $G = (V, E)$ с чётным числом вершин $|V| = 2S$, $S \in \mathbb{N}$, и положительное целое число L_3.

Используя граф G, построим пример задачи 2. Для мощности кластера и ограничения на разброс положим

$$M = S, \quad A = (S-1)(2S-1) - \frac{2}{S}L_3.$$

Определим множество $\mathcal{Y}$. Пусть вершины V и рёбра E занумерованы так, что $V = \{v_1, \ldots, v_{2S}\}$, а $E = \{e_1, \ldots, e_{|E|}\}$. Тогда каждой вершине v_i поставим в соответствие элемент множества $\mathcal{Y}$ — точку $x_i \in \mathbb{R}^{|E|+D}$,

$$D = \sum_{v \in V} \overline{\deg}\, v,$$

где $\overline{\deg}\, v = |V| - 1 - \deg v$ — степень вершины $v \in V$ в дополнении графа G, а $\deg v$ — степень вершины v в самом графе G. Определим x_i следующим образом:

$$x_i = (x_i^{(1)}, \ldots, x_i^{(|E|)}, x_i^{(|E|+1)}, \ldots, x_i^{(|E|+D)}),$$

где при $j = 1, \ldots, |E|$ выполнено $x_i^{(j)} = 1$, если i-я вершина v_i инцидентна j-му ребру e_j, и $x_i^{(j)} = 0$ иначе; при $j = |E|+1, \ldots, |E|+D$ выполнено $x_i^{(j)} = 1$, если

$$\sum_{s=1}^{i-1} \overline{\deg}\, v_s < j - |E| \leqslant \sum_{s=1}^{i} \overline{\deg}\, v_s,$$

и $x_i^{(j)} = 0$ иначе.

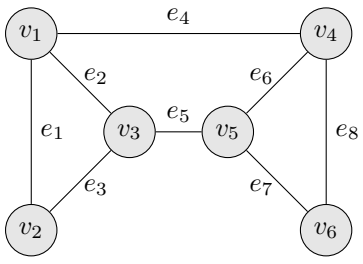


Рис. 2. Пример задачи Два плотных подграфа

Например, вершинам графа на 6 вершинах, изображённого на рис. 2, будут соответствовать точки

$$\begin{aligned}
x_1 &= (\overbrace{1101\,0000}^{|E|=8} \mid \overbrace{1100\,0000\,0000\,00}^{D=14}),\\
x_2 &= (1010\,0000 \mid 0011\,1000\,0000\,00),\\
x_3 &= (0110\,1000 \mid 0000\,0110\,0000\,00),\\
x_4 &= (0001\,0101 \mid 0000\,0001\,1000\,00),\\
x_5 &= (0000\,1110 \mid 0000\,0000\,0110\,00),\\
x_6 &= (0000\,0011 \mid 0000\,0000\,0001\,11).
\end{aligned}$$

Пусть для рассматриваемого примера задачи Два плотных подграфа существует решение, т. е. существует пара множеств $V_1, V_2 \subset V$ таких, что $V_1 \cap V_2 = \varnothing$, $|V_i| = S$, а каждое из сужений G_1, G_2 исходного графа G на эти множества содержит не менее L_3 рёбер.

Используя эти множества V_1, V_2, построим допустимое решение для примера задачи 2, а именно покажем, что множества

$$\mathcal{C}_i = \{x_j,\ j = 1, \ldots, 2S \mid v_j \in V_i\}, \quad i = 1, 2,$$

являются допустимым решением. Для этого достаточно показать, что выполняется ограничение на разбросы множеств $\mathcal{C}_i$, так как условие на мощности выполнено по построению. Рассмотрим произвольное S-элементное подмножество $\mathcal{X} \subset \mathcal{Y}$ и покажем, как его разброс выражается через вершины графа G, соответствующие его элементам, а затем применим получившуюся формулу к множествам $\mathcal{C}_i$. Для вычисления разброса $F(\mathcal{X})$ будем использовать следующую формулу (см., например, равенство (9) в работе [10]):

$$\sum_{y \in \mathcal{X}} \|y - \overline{y}(\mathcal{X})\|^2 = \frac{1}{2|\mathcal{X}|} \sum_{y,z \in \mathcal{X}} \|y - z\|^2. \tag{1}$$

Найдём, чему равна величина $\|y - z\|^2$ для пары различных точек $y, z \in \mathcal{X}$. Прежде всего, для этой величины справедливо равенство

$$\|y - z\|^2 = \sum_{j=1}^{|E|} (y^{(j)} - z^{(j)})^2 + \sum_{j=|E|+1}^{|E|+D} (y^{(j)} - z^{(j)})^2, \tag{2}$$

где $y^{(j)}$ и $z^{(j)}$ — j-е координаты точек y и z соответственно. При этом первые $|E|$ координат каждой точки имеют столько единиц, какова степень соответствующей вершины в графе. Кроме того, у двух точек среди этих координат может быть общая единица, только если между соответствующими вершинами есть ребро. Таким образом, если точкам y, z

соответствуют вершины v_y, v_z, то первая сумма в правой части (2) равна

$$\sum_{j=1}^{|E|}(y^{(j)} - z^{(j)})^2 = \deg v_y + \deg v_z - 2I(v_y v_z \in E), \tag{3}$$

где $I(v_y v_z \in E) = 1$, если вершины v_y и v_z смежны, и $I(v_y v_z \in E) = 0$ иначе. При этом вторая сумма в правой части (2) равна

$$\sum_{j=|E|+1}^{|E|+D}(y^{(j)} - z^{(j)})^2 = \overline{\deg}\, v_y + \overline{\deg}\, v_z, \tag{4}$$

так как в последних D координатах точек y и z единицы стоят на разных местах. Объединяя (2)–(4), получаем

$$\|y - z\|^2 = \deg v_y + \deg v_z - 2I(v_y v_z \in E) + \overline{\deg}\, v_y + \overline{\deg}\, v_z = \\ = 2(2S - 1 - I(v_y v_z \in E)).$$

Используя предыдущую формулу, а также (1), получаем, что разброс множества $\mathcal{X}$ может быть записан в следующем виде:

$$F(\mathcal{X}) = \frac{1}{2S}\sum_{\substack{y,z\in\mathcal{X},\\ y\neq z}}\|y - z\|^2 = (S-1)(2S-1) - \frac{1}{S}\sum_{\substack{y,z\in\mathcal{X},\\ y\neq z}} I(v_y v_z \in E). \tag{5}$$

Используя (5), можем переписать ограничение на разброс $F(\mathcal{X}) \leqslant A$ для произвольного S-элементного подмножества $\mathcal{X}$ в следующем эквивалентном виде:

$$\sum_{\substack{y,z\in\mathcal{X},\\ y\neq z}} I(v_y v_z \in E) \geqslant S((S-1)(2S-1) - A) = 2L_3. \tag{6}$$

Так как в правой части (6) каждая пара вершин из подмножества $\mathcal{X}$ учитывается дважды, а соответствующие множествам $\mathcal{C}_1$, $\mathcal{C}_2$ графы G_1, G_2 содержат не менее L_3 рёбер, множества $\mathcal{C}_1, \mathcal{C}_2$ удовлетворяют требуемым ограничениям на разброс, а значит, образуют решение построенного примера задачи 2.

Предположим, что у построенного примера задачи 2 существует решение — множества $\mathcal{C}_1$ и $\mathcal{C}_2$. Поскольку $|\mathcal{Y}| = 2S$ и каждое из множеств должно содержать не менее S элементов, то $|\mathcal{C}_1| = |\mathcal{C}_2| = S$. Рассмотрим множества вершин

$$V_i = \{v_j,\, j = 1, \ldots, 2S \mid x_j \in \mathcal{C}_i\}, \quad i = 1, 2.$$

Так как множества $\mathcal{C}_i$, $i = 1, 2$, удовлетворяют ограничению на разброс, из (6) следует, что графы G_1, G_2, индуцированные множествами V_1, V_2, содержат как минимум L_3 рёбер и оба состоят из S вершин.

Другими словами, V_1, V_2 являются решением задачи Два плотных подграфа. Тем самым задача 2 при $k = 2$ NP-полна.

Сильная NP-полнота следует из того факта, что частный случай задачи 2, индуцированный построенными примерами, не является задачей с числовыми параметрами, поскольку все входные значения либо равны единице или нулю (координаты точек входного множества $\mathcal{Y}$), либо ограничены полиномиальной функцией по длине входа (ограничение на разброс A кластеров). Теорема 2 доказана.

3. Одномерный случай

В этом разделе, в отличие от предыдущего, дополнительно предположим, что размерность d пространства равна единице — в предыдущем разделе величина d была частью входа задачи. Таким образом, будем рассматривать задачу 2 в случае, когда требуется найти два кластера, а все элементы входного множества $\mathcal{Y}$ имеют только одну вещественную координату.

Теорема 3. *Задача* 2 *при* $k = 2$ NP-*полна в случае фиксированной размерности пространства* $d = 1$.

Доказательство. Рассмотрим произвольный пример задачи Разбиение — $\{n_1, \ldots, n_{2S}\} \subset \mathbb{N} \setminus \{0\}$, $\sum\limits_{i=1}^{2S} n_i = 2T$. Без ограничения общности будем считать, что $S \geqslant 6$.

Построим следующий пример задачи 2 при $k = 2$ и $d = 1$: множество

$$\mathcal{Y} = \{B, B, n_1, \ldots, n_{2S}\},$$

где

$$B = -\max\left\{4T,\, N\left\lceil\sqrt{2S(S+1)}\,\right\rceil,\, \frac{1}{2}(2S+1)SN^2\right\} - 1, \quad N = \max_{i=1,\ldots,2S} n_i,$$

мощность кластеров $M = S + 1$, ограничение на разброс

$$A = \frac{S}{S+1}B^2 - \frac{2B}{S+1}T + SN^2.$$

Пусть у рассматриваемого примера задачи Разбиение имеется решение, т. е. существует множество индексов $\mathcal{I}$, $|\mathcal{I}| = S$, такое, что $\sum\limits_{i\in\mathcal{I}} n_i = T$. Рассмотрим пару множеств

$$\mathcal{C}_i = \{B\} \cup \{n_j \mid j \in \mathcal{I}_i\}, \quad i = 1, 2, \tag{7}$$

где $\mathcal{I}_1 = \mathcal{I}$, а $\mathcal{I}_2 = \{1, \ldots, 2S\} \setminus \mathcal{I}$.

Оценим разбросы $F(C_i)$, $i = 1, 2$, этих множеств. Для произвольного $\mathcal{C} \subset \mathbb{R}^d$ справедлива формула (см., например, цепочку (7) равенств в работе [10])

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(C)\|^2 = \sum_{y \in \mathcal{C}} \|y\|^2 - |\mathcal{C}| \cdot \|\overline{y}(\mathcal{C})\|^2. \tag{8}$$

Подставляя в (8) $\mathcal{C} = \mathcal{C}_i$, получаем, что для произвольного множества $\mathcal{C}_i$, имеющего структуру (7), выполнено

$$F(\mathcal{C}_i) = \sum_{y \in \mathcal{C}_i} |y - \overline{y}(\mathcal{C}_i)|^2 = \sum_{y \in \mathcal{C}_i} y^2 - |\mathcal{C}_i| \cdot (\overline{y}(\mathcal{C}_i))^2 =$$

$$= B^2 + \sum_{j \in \mathcal{I}_i} n_j^2 - \frac{1}{S+1}\Big(\sum_{j \in \mathcal{I}_i} n_j + B\Big)^2.$$

После раскрытия скобок и приведения подобных слагаемых получаем

$$F(\mathcal{C}_i) = \frac{S}{S+1}B^2 - \frac{2B}{S+1}\sum_{j \in \mathcal{I}_i} n_j + \sum_{j \in \mathcal{I}_i} n_j^2 - \frac{1}{S+1}\Big(\sum_{j \in \mathcal{I}_i} n_j\Big)^2. \tag{9}$$

Чтобы оценить величину $F(\mathcal{C}_i)$ сверху, применим равенство $\sum\limits_{j \in \mathcal{I}} n_j = T$, а также неравенства $n_j^2 \leqslant N^2$, $j = 1, \ldots, 2S$. Тогда из (9) следует оценка

$$F(\mathcal{C}_i) \leqslant \frac{S}{S+1}B^2 - \frac{2B}{S+1}T + SN^2 = A,$$

что означает, что множества $\mathcal{C}_i$, $i = 1, 2$, являются решением построенного примера задачи 2.

Пусть построенный пример задачи 2 имеет решение — множества $\mathcal{C}_i$, $i = 1, 2$.

Сначала покажем, что оба элемента B из входного $\mathcal{Y}$ не могут лежать в одном подмножестве. Для этого предположим обратное — что одно из подмножеств, например $\mathcal{C}_1$, содержит оба элемента B, т. е. имеет вид

$$\mathcal{C}_1 = \{B, B\} \cup \{n_j \mid j \in \mathcal{I}_1\}, \quad |\mathcal{I}_1| = S - 1.$$

Чтобы прийти к противоречию, покажем, что множество $\mathcal{C}_1$ не удовлетворяет ограничению на разброс. Для этого понадобится вспомогательное неравенство: предварительно покажем, что разброс A удовлетворяет неравенству $A < B^2$. Действительно, используя определение A, это неравенство можно переформулировать следующим образом:

$$\frac{S}{S+1}B^2 - \frac{2B}{S+1}T + SN^2 < B^2,$$

или эквивалентно

$$-\frac{2T}{S+1}B + SN^2 < \frac{1}{S+1}B^2. \tag{10}$$

Из определения B следует, что выполнены следующие два неравенства:

$$-\frac{2B}{S+1}T < \frac{1}{2(S+1)}B^2,$$

$$SN^2 < \frac{1}{2(S+1)}B^2,$$

после сложения которых получим, что неравенство (10) выполняется, а значит, $A < B^2$.

Оценим разброс $\mathcal{C}_1$ снизу:

$$F(\mathcal{C}_1) = \sum_{y\in\mathcal{C}_1} |y - \overline{y}(\mathcal{C}_1)|^2 \geqslant 2|B - \overline{y}(\mathcal{C}_1)|^2. \tag{11}$$

Заметим, что

$$\overline{y}(\mathcal{C}_1) = \frac{1}{S+1}\Big(2B + \sum_{j\in\mathcal{I}_1} n_j\Big) > \frac{2B}{S+1}.$$

Так как $B < \frac{2B}{S+1}$ и $S \geqslant 6$, оценку (11) можно продолжить:

$$F(\mathcal{C}_1) \geqslant 2|B - \overline{y}(\mathcal{C}_1)|^2 > 2\Big(B - \frac{2B}{S+1}\Big)^2 = 2B^2\Big(\frac{S-1}{S+1}\Big)^2 \geqslant B^2,$$

что противоречит неравенству $A < B^2$. Значит, оба элемента B не могут лежать в одном множестве, т. е. оба множества имеют вид (7).

Предположим, что сумма элементов d_i, $i \in \mathcal{I}_1$, не равна T. Без ограничения общности можно считать, что $\sum\limits_{i\in\mathcal{I}_1} d_i \geqslant T+1$, так как в противном случае можно поменять $\mathcal{I}_1$ и $\mathcal{I}_2$ местами.

Оценим разброс множества $\mathcal{C}_1$ снизу, используя равенство (9):

$$F(\mathcal{C}_1) = \frac{S}{S+1}B^2 - \frac{2B}{S+1}\sum_{j\in\mathcal{I}_1} n_j + \sum_{j\in\mathcal{I}_1} n_j^2 - \frac{1}{S+1}\Big(\sum_{j\in\mathcal{I}_1} n_j\Big)^2 \geqslant$$

$$\geqslant \frac{S}{S+1}B^2 - \frac{2B}{S+1}(T+1) + 0 - \frac{1}{S+1}(SN)^2 =$$

$$= A - \frac{2B}{S+1} - SN^2 - \frac{1}{S+1}(SN)^2 = A - \frac{1}{S+1}(2B + (2S+1)SN^2).$$

Так как для выбранного B выполнено $-2B > (2S+1)SN^2$, имеем $F(\mathcal{C}_1) > A$. Таким образом, наше предположение приводит к противоречию с допустимостью множеств $\mathcal{C}_1$, $\mathcal{C}_2$. Значит, сумма элементов n_j, $j \in \mathcal{I}_1$, равна T.

Следовательно, исходный пример задачи Разбиение и построенный пример задачи 2 одновременно либо имеют решения, либо не имеют. Значит, задача 2 при $k = 2$ NP-трудна в одномерном случае. Теорема 3 доказана.

Обобщим полученный в теореме 3 результат на случай произвольного фиксированного числа кластеров $k \geqslant 2$, используя индукцию по k.

Теорема 4. *Для произвольного $k \geqslant 2$ задача 2 с фиксированным числом кластеров k* NP*-полна даже в одномерном случае.*

Доказательство. Докажем утверждение индукцией по числу кластеров k. NP-полнота в случае $k = 2$ доказана в теореме 3.

Допустим, что задача 2 для k кластеров NP-полна в одномерном случае, и покажем, что задача 2 для $k+1$ кластеров также NP-полна в этом случае. Для этого построим полиномиальное сведение случая k кластеров к $k + 1$.

Рассмотрим произвольный одномерный пример задачи 2 для k кластеров — множество $\mathcal{Y}$, $|\mathcal{Y}| = kM$, ограничение на разброс A. Построим следующий пример задачи 2 для $k + 1$ кластеров: ограничение на разброс A остаётся таким же, входное множество $\widetilde{\mathcal{Y}}$ равно

$$\widetilde{\mathcal{Y}} = \mathcal{Y} \cup \{B_1, \ldots, B_M\}, \quad B_i = B,\ i = 1, \ldots, M,$$

где $B = \left\lceil \sqrt{2M(A+1)} \right\rceil + N$, а $N = \max\limits_{y \in \mathcal{Y}} |y|$ — максимальное абсолютное значение элементов из $\mathcal{Y}$.

Покажем, что оба примера имеют решения либо не имеют одновременно. Пусть существует решение задачи 2 для k кластеров — $\mathcal{C}_i$, $i = 1, \ldots, k$. Рассмотрим $k + 1$ множеств

$$\widetilde{\mathcal{C}}_i = \mathcal{C}_i,\ i = 1, \ldots, k, \quad \widetilde{\mathcal{C}}_{k+1} = \{B_1, \ldots, B_M\}.$$

Очевидно, что мощности всех этих множеств равны M и они попарно не пересекаются. Разбросы множеств $\widetilde{\mathcal{C}}_i$ не превосходят порога A по предположению, а разброс $\widetilde{\mathcal{C}}_{k+1}$ равен 0, так как все его элементы одинаковы. Следовательно, множества $\widetilde{\mathcal{C}}_i$, $i = 1, \ldots, k + 1$, являются решением построенного примера задачи 2 для $k + 1$ кластеров.

Допустим, что построенный пример задачи 2 с $k+1$ кластерами имеет решение — множества $\widetilde{\mathcal{C}}_i$, $|\widetilde{\mathcal{C}}_i| = M$, $i = 1, \ldots, k + 1$. Покажем, что все элементы B_i лежат в одном множестве. Для этого рассмотрим множество, содержащее хотя бы один элемент B_i (такое множество существует, так как по предположению $\widetilde{\mathcal{C}}_i$, $i = 1, \ldots, k + 1$, — это разбиение $\widetilde{\mathcal{Y}}$). Без ограничения общности можно считать, что это множество $\widetilde{\mathcal{C}}_{k+1}$. Допустим, что в $\widetilde{\mathcal{C}}_{k+1}$ есть элемент не из B_i. Тогда разброс множества $\widetilde{\mathcal{C}}_{k+1}$ можно

оценить снизу, используя формулу (1):

$$F(\widetilde{\mathcal{C}}_{k+1}) = \frac{1}{2M} \sum_{x,y \in \widetilde{\mathcal{C}}_{k+1}} |x-y|^2 \geqslant \frac{1}{2M} |x^* - B|^2,$$

где x^* — некоторый элемент из $\mathcal{Y}$. Учитывая определение B и то, что $|x^*| \leqslant N$, последнее неравенство можно продолжить:

$$F(\widetilde{\mathcal{C}}_{k+1}) \geqslant \frac{1}{2M}(B-N)^2 \geqslant A+1 > A.$$

Таким образом, разброс $F(\widetilde{\mathcal{C}}_{k+1})$ строго больше A, что противоречит допустимости решения $\widetilde{\mathcal{C}}_i$, $i = 1, \ldots, k+1$. Стало быть, все элементы B_i содержатся в множестве $\widetilde{\mathcal{C}}_{k+1}$. Тогда множества $\mathcal{C}_i = \widetilde{\mathcal{C}}_i$, $i = 1, \ldots, k$, являются решением задачи 2 для k кластеров.

Таким образом, задачи 2 для k и $k+1$ кластеров одновременно либо имеют решения, либо не имеют. Следовательно, доказаны база и шаг индукции, что означает, что задача 2 с фиксированным числом кластеров k NP-полна в одномерном случае для произвольного $k \geqslant 2$. Теорема 4 доказана.

Таким образом, для произвольного числа кластеров $k \geqslant 2$ задачи 1 и 2 для k кластеров NP-полны даже в одномерном случае. Заметим, что для этих задач существует [16] псевдополиномиальный алгоритм, поэтому они не будут NP-полными в сильном смысле.

Заключение

В работе рассмотрены задачи поиска непересекающихся подмножеств большой мощности при ограничении на их разброс. Доказано, что задачи NP-полны в сильном смысле в случае поиска двух кластеров и размерности пространства, являющейся частью входа задачи, а также что задачи NP-полны в одномерном случае.

Интересным направлением дальнейших исследований является выяснение сложности задач в случае фиксированной мощности искомых кластеров.

Финансирование работы

Исследование выполнено в рамках государственного задания Института математики им. С. Л. Соболева (проект № FWNF–2022–0015). Дополнительных грантов на проведение или руководство этим исследованием получено не было.

Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

Литература

1. **Pérez-Ortega J., Almanza-Ortega N. N., Vega-Villalobos A.** [et al.]. The K-means algorithm evolution // Introduction to data science and machine learning. Rijeka: IntechOpen, 2019. 22 p. DOI: 10.5772/intechopen.85447.
2. **Ikotun A. M., Ezugwu A. E., Abualigah L.** [et al.]. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data // Inf. Sci. 2023. V. 622. P. 178–210. DOI: 10.1016/j.ins.2022.11.139.
3. **Grant R. W., McCloskey J., Hatfield M.** [et al.]. Use of latent class analysis and k-means clustering to identify complex patient profiles // JAMA Netw. Open. 2020. V. 3, No. 12. Article ID e2029068. 13 p. DOI: 10.1001/jamanetworkopen.2020.29068.
4. **Dhanachandra N., Manglem K., Chanu Y. J.** Image segmentation using k-means clustering algorithm and subtractive clustering algorithm // Proc. Comput. Sci. 2015. V. 54. P. 764–771. DOI: 10.1016/j.procs.2015.06.090.
5. **Kumari R., Sheetanshu, Singh M. K., Jha R.** Anomaly detection in network traffic using K-mean clustering // 2016 3rd Int. Conf. Recent Advances in Information Technology (Dhanbad, India, Mar. 3–5, 2016). Piscataway: IEEE, 2016. P. 387–393. DOI: 10.1109/RAIT.2016.7507933.
6. **Aloise D., Deshpande A., Hansen P.** [et al.]. NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. 2009. V. 75, No. 2. P. 245–248. DOI: 10.1007/s10994-009-5103-0.
7. **Jørgensen A. G., Larsen K. G., Mathiasen A.** [et al.]. Fast exact k-means, k-medians and Bregman divergence clustering in 1D. Ithaca, NY, 2017. 16 p. (e-Print Archive / Cornell Univ.; arXiv:1701.07204). DOI: 10.48550/arXiv.1701.07204.
8. **Khodadadi A., Saeidi S.** Discovering the maximum k-clique on social networks using bat optimization algorithm // Comput. Soc. Netw. 2021. V. 8, No. 1. Article ID 6. 15 p. DOI: 10.1186/s40649-021-00087-y.
9. **Tomita E., Akutsu T., Matsunaga T.** Efficient algorithms for finding maximum and maximal cliques: Effective tools for bioinformatics // Biomedical engineering, trends in electronics, communications and software. Rijeka: Intech Open, 2011. P. 625–640. DOI: 10.5772/13245.
10. **Кельманов А. В., Пяткин А. В.** NP-полнота некоторых задач выбора подмножества векторов // Дискрет. анализ и исслед. операций. 2010. Т. 17, № 5. С. 37–45.
11. **Aggarwal A., Imai H., Katoh N., Suri S.** Finding k points with minimum diameter and related problems // J. Algorithms. 1991. V. 12, No. 1. P. 38–56. DOI: 10.1016/0196-6774(91)90022-Q.
12. **Kel'manov A. V., Ruzankin P. S.** An accelerated exact algorithm for the one-dimensional M-variance problem // Pattern Recognit. Image Anal. 2019. V. 29, No. 4. P. 573–576. DOI: 10.1134/S1054661819040072.
13. **Пяткин А. В.** О сложности задачи выбора кластеров большого размера // Дискрет. анализ и исслед. операций. 2024. Т. 31, № 2. С. 136–143.

14. Кельманов А. В., Пяткин А. В., Хандеев В. И. О сложности некоторых максиминных задач кластеризации // Тр. Ин-та математики и механики. 2018. Т. 24, № 4. С. 189–198.

15. Garey M. R., Johnson D. S. Computers and intractability: A guide to the theory of NP-completeness. San Francisco: Freeman, 1979. 338 p.

16. Khandeev V. I., Neshchadim S. M. Pseudo-polynomial algorithms for some problems of searching for the largest subsets // Mathematical optimization theory and operations research: Recent trends. Rev. Sel. Pap. 23th Int. Conf. (Omsk, Russia, June 30 – July 6, 2024). Cham: Springer, 2024. P. 319–333. (Commun. Comput. Inf. Sci.; V. 2239). DOI: 10.1007/978-3-031-73365-9_22.

Нещадим Сергей Михайлович
Хандеев Владимир Ильич

Статья поступила
5 мая 2025 г.
После доработки —
25 августа 2025 г.
Принята к публикации
22 сентября 2025 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
October–December 2025. Vol. 32, No. 4. P. 172–190

UDC 519.8 DOI: 10.33048/daio.2025.32.834

ON THE COMPLEXITY OF TWO PROBLEMS OF FINDING CLUSTERS OF LARGE CARDINALITY

S. M. Neshchadim[1, a] and *V. I. Khandeev*[2, b]

[1] Novosibirsk State University,
2 Pirogov Street, 630090 Novosibirsk, Russia
[2] Sobolev Institute of Mathematics,
4 Acad. Koptyug Avenue, 630090 Novosibirsk, Russia
E-mail: [a] s.neshchadim@g.nsu.ru, [b] khandeev@math.nsc.ru

Abstract. Clustering problems for a finite point set in Euclidean space are considered. The first problem requires each subset to have cardinality no smaller than a given threshold (not necessarily covering the entire set), while the second one requires all subsets to have the same cardinality and form a partition of the given set. In both problems for each subset, it is additionally required that the sum of squared distances to the centroid does not exceed a given value. Both problems are proven to be strongly NP-complete when the number of clusters is two and the space dimension is part of the input. Furthermore, NP-completeness is established for the one-dimensional case with an arbitrary fixed number of clusters. Illustr. 2, bibliogr. 16.

Keywords: clustering, bounded scatter, minimum cluster size, Euclidean space, NP-completeness.

References

1. **J. Pérez-Ortega, N. N. Almanza-Ortega, A. Vega-Villalobos,** [et al.], The K-means algorithm evolution, in *Introduction to Data Science and Machine Learning* (IntechOpen, Rijeka, 2019), DOI: 10.5772/intechopen.85447.
2. **A. M. Ikotun, A. E. Ezugwu, L. Abualigah,** [et al.], K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Inf. Sci.* **622**, 178–210 (2023), DOI: 10.1016/j.ins.2022.11.139.

English transl.: *Journal of Applied and Industrial Mathematics* **19** (4) (2025).

3. **R. W. Grant, J. McCloskey, M. Hatfield,** [et al.], Use of latent class analysis and k-means clustering to identify complex patient profiles, *JAMA Netw. Open.* **3** (12), ID e2029068 (2020), DOI: 10.1001/jamanetworkopen.2020.29068.
4. **N. Dhanachandra, K. Manglem,** and **Y. J. Chanu,** Image segmentation using k-means clustering algorithm and subtractive clustering algorithm, *Proc. Comput. Sci.* **54**, 764–771 (2015), DOI: 10.1016/j.procs.2015.06.090.
5. **R. Kumari, M. K. Sheetanshu, Singh,** and **R. Jha,** Anomaly detection in network traffic using K-mean clustering, in *2016 3rd Int. Conf. Recent Advances in Information Technology* (Dhanbad, India, Mar. 3–5, 2016) (IEEE, Piscataway, 2016), pp. 387–393, DOI: 10.1109/RAIT.2016.7507933.
6. **D. Aloise, A. Deshpande, P. Hansen,** [et al.], NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.* **75** (2), 245–248 (2009), DOI: 10.1007/s10994-009-5103-0.
7. **A. G. Jørgensen, K. G. Larsen, A. Mathiasen,** [et al.], Fast exact k-means, k-medians and Bregman divergence clustering in 1D (Ithaca, NY, 2017) (e-Print Archive / Cornell Univ., arXiv:1701.07204), DOI: 10.48550/arXiv.1701.07204.
8. **A. Khodadadi** and **S. Saeidi,** Discovering the maximum k-clique on social networks using bat optimization algorithm, *Comput. Soc. Netw.* **8** (1), ID 6 (2021), DOI: 10.1186/s40649-021-00087-y.
9. **E. Tomita, T. Akutsu,** and **T. Matsunaga,** Efficient algorithms for finding maximum and maximal cliques: Effective tools for bioinformatics, in *Biomedical Engineering, Trends in Electronics, Communications and Software* (Intech Open, Rijeka, 2011), pp. 625–640, DOI: 10.5772/13245.
10. **A. V. Kel'manov** and **A. V. Pyatkin,** NP-completeness of some problems of choosing a vector subset, *Diskretn. Anal. Issled. Oper.* **17** (5), 37–45 (2010) [Russian] [*J. Appl. Ind. Math.* **5** (3), 352–357 (2011) DOI: 10.1134/S1990478911030069].
11. **A. Aggarwal, H. Imai, N. Katoh,** and **S. Suri,** Finding k points with minimum diameter and related problems, *J. Algorithms* **12** (1), 38–56 (1991), DOI: 10.1016/0196-6774(91)90022-Q.
12. **A. V. Kel'manov** and **P. S. Ruzankin,** An accelerated exact algorithm for the one-dimensional M-variance problem, *Pattern Recognit. Image Anal.* **29** (4), 573–576 (2019), DOI: 10.1134/S1054661819040072.
13. **A. V. Pyatkin,** On the complexity of the problem of choice of large clusters, *Diskretn. Anal. Issled. Oper.* **31** (2), 136–143 (2024) [Russian] [*J. Appl. Ind. Math.* **18** (2), 312–315 (2024), DOI: 10.1134/S1990478924020121].
14. **A. V. Kel'manov, A. V. Pyatkin,** and **V. I. Khandeev,** On the complexity of some max–min clustering problems, *Tr. Inst. Mat. Mekh.* **24** (4), 189–198 (2018) [Russian] [*Proc. Steklov Inst. Math.* **309** (Suppl. 1), S65–S73 (2020) DOI: 10.1134/S0081543820040082].
15. **M. R. Garey** and **D. S. Johnson,** *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).

16. V. I. Khandeev and **S. M. Neshchadim,** Pseudo-polynomial algorithms for some problems of searching for the largest subsets, in *Mathematical Optimization Theory and Operations Research: Recent Trends*, Rev. Sel. Pap. 23th Int. Conf. (Omsk, Russia, June 30 – July 6, 2024) (Springer, Cham, 2024), pp. 319–333 (Commun. Comput. Inf. Sci., V. 2239), DOI: 10.1007/978-3-031-73365-9_22.

Sergey M. Neshchadim
Vladimir I. Khandeev

Received May 5, 2025
Revised August 25, 2025
Accepted September 22, 2025