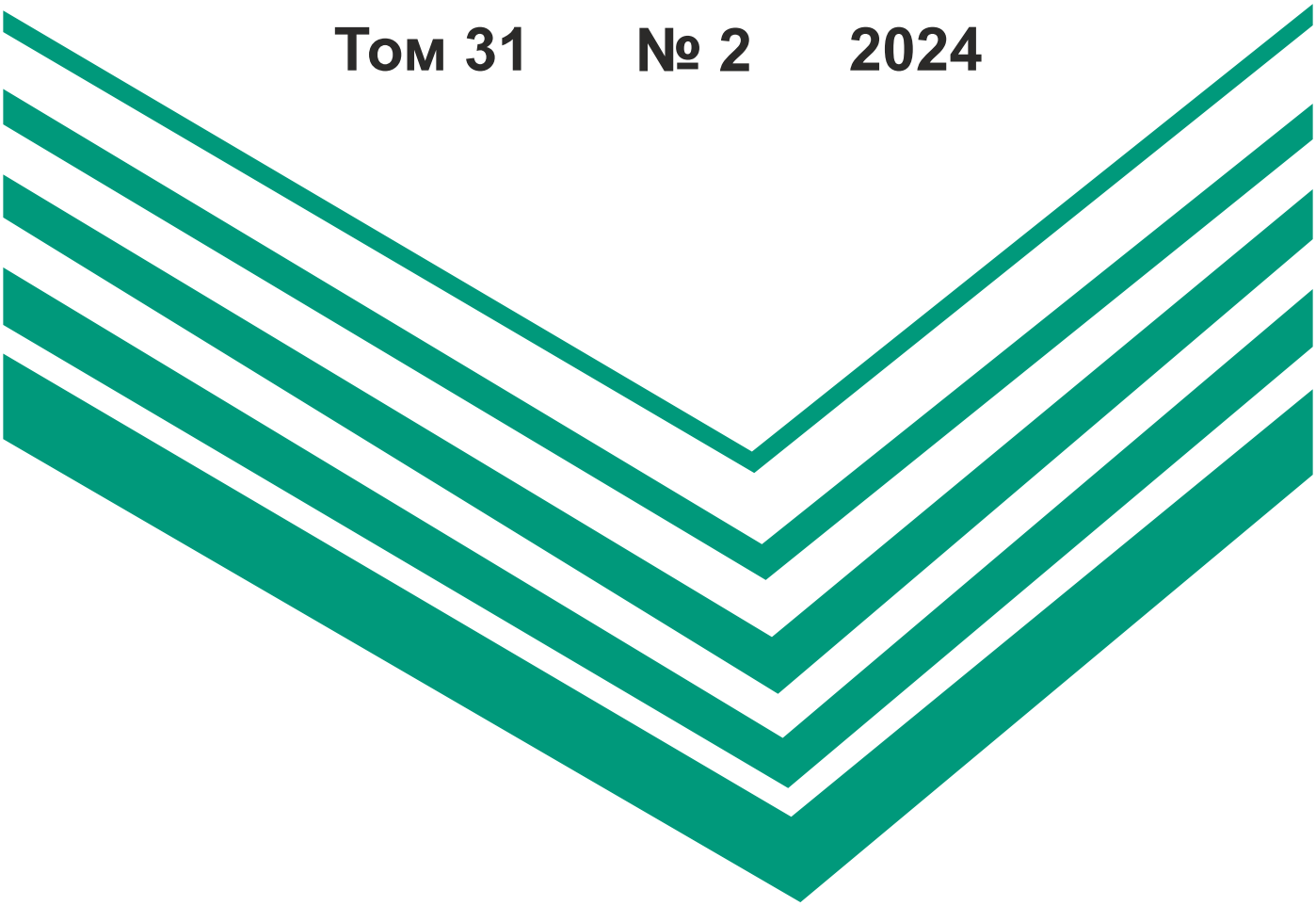
ISSN 2949-5598

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ

Том 31 № 2 2024



Новосибирск
Издательство Института математики

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Апрель–июнь 2024. Т. 31, № 2. С. 136–143

УДК 519.8+518.25 DOI: 10.33048/daio.2024.31.787

О СЛОЖНОСТИ ЗАДАЧИ ВЫБОРА КЛАСТЕРОВ БОЛЬШОГО РАЗМЕРА

А. В. Пяткин

Институт математики им. С. Л. Соболева,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия
E-mail: artempyatkin@gmail.com

Аннотация. Рассматривается задача выбора в данном множестве евклидовых векторов заданного числа кластеров с ограничением на максимальный разброс каждого кластера так, чтобы размер минимального из этих кластеров был максимальным. Под разбросом понимается сумма квадратов расстояний от элементов кластера до его центроида. Доказана NP-трудность этой задачи в случае, когда размерность пространства является частью входа. Библиогр. 15.

Ключевые слова: кластер, центроид, разброс, NP-трудность.

Введение

Объектом изучения настоящей статьи является задача, в которой для заданного множества векторов $\mathcal{Y} \subseteq \mathbb{R}^d$ необходимо построить семейство попарно не пересекающихся подмножеств (кластеров) «похожих» векторов из $\mathcal{Y}$ при ограничении на разброс внутри каждого кластера, при этом размер минимального из этих кластеров максимизируется. Целью работы является определение алгоритмической сложности данной задачи в случае, когда разброс определяется как сумма квадратов расстояний от элементов кластера до его центроида, а размерность пространства является частью входа.

Задачи кластеризации весьма актуальны для анализа и обработки данных, искусственного интеллекта, вычислительной геометрии, математической статистики и дискретной оптимизации [1–3]. Одним из наиболее известных примеров является классическая задача MSSC (minimum sum-square clustering), также иногда называемая k-means [4–6]. В ней требуется разбить данное множество векторов евклидова пространства на k непересекающихся кластеров, минимизируя суммарный разброс (сумму квадратов расстояний от элемента каждого кластера

до его центроида) всех кластеров. Известно, что эта задача NP-трудна [7] даже при $k = 2$. Задача также NP-трудна в случае, когда заданы размеры кластеров [8], даже если размеры всех кластеров равны 3. В [9] доказано, что задача разбиения на два кластера равного размера также NP-трудна.

Однако в случае, когда исходные данные содержат случайные выбросы, более актуальна не задача разбиения множества на кластеры, а задача выбора k непересекающихся подмножеств с заданными ограничениями на их размеры и разброс, причём число кластеров k в этом случае является не частью входа, а параметром задачи. В случае $k = 1$ задача NP-трудна [10]; при $k \geqslant 2$ её сложностной статус неизвестен.

В работе исследуется

Задача 1 (Поиск k кластеров)**.** *Даны множество $\mathcal{Y} = \{y_1, \ldots, y_n\}$ векторов в пространстве $\mathbb{R}^d$ и число $A > 0$. Выделить k попарно не пересекающихся кластеров $\mathcal{C}_i \subseteq \mathcal{Y}$, $i = 1, \ldots, k$, максимизируя* $\min\{|\mathcal{C}_1|, \ldots, |\mathcal{C}_k|\}$ *при условии*

$$f(\mathcal{C}_i) = \sum_{y \in \mathcal{C}_i} \|y - \overline{y}(\mathcal{C}_i)\|^2 \leqslant A, \quad i = 1, \ldots, k,$$

где $\overline{y}(\mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|} \sum\limits_{y \in \mathcal{C}_i} y$ — центроид кластера $\mathcal{C}_i$.

Близкие в постановочном плане задачи, когда вместо центроида используются заданные точки пространства или точки исходного множества, рассматривались в работе [11], где была доказана их NP-трудность даже в одномерном случае (т. е. при $d = 1$). В [12] для задачи 1 был предложен приближённый полиномиальный алгоритм с оценкой точности $1/k$ для произвольного $k \geqslant 2$ и размерности пространства $d = 1$. Там же было отмечено, что сложностной статус задачи 1 неизвестен.

В настоящей статье доказывается, что задача 1 NP-трудна в случае, когда размерность пространства d является частью входа. В разд. 1 приводятся предварительные сведения, необходимые для доказательства основного результата в разд. 2. В заключении даются финальные комментарии.

1. Предварительные результаты

Пусть задан граф $G = (V, E)$ и целое число $k \geqslant 2$. Тогда *k-раскраской* графа G называется разбиение множества вершин V на k независимых подмножеств $V_1, \ldots, V_k$, называемых *цветами*. Известно, что задача существования в графе k-раскраски NP-полна [13] при $k \geqslant 3$. Более того, эта задача остаётся NP-полной для r-однородных графов при $r > k$, поскольку в [14] было доказано, что проблема 3-раскраски 4-однородного плоского графа NP-полна. В *сбалансированной* k-раскраске требуется

найти разбиение на k независимых подмножеств с дополнительным условием, что $|V_i| - |V_j| \in \{-1, 0, 1\}$ для всех i, j, т. е. размеры всех независимых множеств должны быть примерно одинаковы. Очевидно, что эта задача тоже NP-трудна, так как задача k-раскраски графа легко сводится к сбалансированной версии добавлением нужного числа независимых вершин. Покажем, что сбалансированная k-раскраска NP-трудна в следующем частном случае.

Задача 2 (Сбалансированная k-раскраска однородного графа)**.** *Даны r-однородный граф $G = (V, E)$ на n вершинах и целое число $k \geqslant 3$, причём $n = kt$ для некоторого целого $t > 1$. Существует ли разбиение V на k независимых подмножеств $V_1, \ldots, V_k$, мощность каждого из которых равна t?*

Лемма 1. *Задача* 2 NP*-полна.*

Доказательство. Сведём к задаче 2 проблему существования k-раскраски в r-однородном графе. Рассмотрим произвольный r-однородный t-вершинный граф $H = (V_0, E_0)$, который требуется раскрасить в k цветов. Возьмём в качестве G граф, полученный объединением k копий графа H. Очевидно, что G является r-однородным графом на kt вершинах. Покажем, что в нём есть сбалансированная k-раскраска тогда и только тогда, когда граф H k-раскрашиваем.

Очевидно, что ограничение сбалансированной k-раскраски графа G на любую из копий графа H задаёт его k-раскраску.

Предположим, что граф H можно раскрасить в k цветов. Тогда раскрасим первую копию в графе G в соответствии с этой раскраской, а в каждой последующей копии сделаем циклический сдвиг цветов на 1 (т. е. в i-й копии цвет j имеют те и только те вершины, которые в 1-й копии имеют цвет $j - i + 1$, где разность берётся по модулю k). Нетрудно видеть, что каждым цветом раскрашено ровно t вершин (например, цветом 1 в i-й копии раскрашены вершины, которые в графе H раскрашены цветом i) и каждый цвет образует независимое множество. Лемма 1 доказана.

Нам также потребуется следующее известное тождество (доказательство можно найти, например, в [15]).

Утверждение 1 [15]**.** *Имеет место формула*

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

2. Основной результат

Сформулируем задачу 1 в виде задачи верификации свойств.

Задача 3. *Даны множество* $\mathcal{Y} = \{y_1, \ldots, y_n\}$ *векторов в пространстве* $\mathbb{R}^d$ *и числа* $A > 0$, $M > 1$. *Существует ли* k *попарно не пересекающихся кластеров* $\mathcal{C}_i \subseteq \mathcal{Y}$, $i = 1, \ldots, k$, *удовлетворяющих условиям* $|\mathcal{C}_i| \geqslant M$ *и* $f(\mathcal{C}_i) \leqslant A$ *для всех* $i = 1, \ldots, k$?

Основным результатом работы является

Теорема 1. *Задача* 3 NP-*полна.*

Доказательство. Построим сведение задачи 2 к задаче 3. Рассмотрим произвольный r-однородный граф $G = (V, E)$ на n вершинах и целое число $k \geqslant 3$, причём $n = kt$. Зафиксируем произвольную ориентацию рёбер графа G. Положим $d = |E|$ и для каждого $j = 1, \ldots, d$ обозначим j-ю дугу графа G через e_j. Далее для всех $i = 1, \ldots, n$ поставим в соответствие вершине v_i этого графа следующий d-мерный вектор y_i. Пусть j-я координата вектора y_i равна -1, если дуга e_j исходит из вершины v_i; равна 1, если дуга e_j заходит в вершину v_i; и равна 0, если дуга e_j не инцидентна вершине v_i. Положим $M = t$ и $A = r(t-1)$. Поскольку граф G r-однороден, то

$$\|y_i - y_j\|^2 = \begin{cases} 0, & \text{если } i = j, \\ 2r, & \text{если } v_i v_j \notin E, \\ 2r + 2, & \text{если } v_i v_j \in E. \end{cases}$$

Покажем, что k попарно не пересекающихся кластеров с требуемыми условиями существуют тогда и только тогда, когда G допускает сбалансированную k-раскраску.

Если сбалансированная k-раскраска существует, то каждым цветом раскрашено ровно t вершин. Положим $\mathcal{C}_i = \{y_j \mid v_j \text{ имеет цвет } i\}$. Тогда $|\mathcal{C}_i| = t = M$ и по утверждению 1 $f(\mathcal{C}_i) = 2r(t-1)t/2t = r(t-1) = A$ для всех $i = 1, \ldots, k$, что и требуется.

Предположим, что существует требуемое разбиение в задаче 3. Поскольку $|\mathcal{Y}| = Mk$, имеем $|\mathcal{C}_i| = M$ для всех $i = 1, \ldots, k$. Пусть подграф V_i, порождённый всеми вершинами v_j, для которых $y_j \in \mathcal{C}_i$, содержит k_i рёбер. Тогда по утверждению 1 имеем

$$f(\mathcal{C}_i) = \frac{2rM(M-1) + 2k_i}{2M} = A + \frac{k_i}{M}.$$

Поскольку $f(\mathcal{C}_i) \leqslant A$, получаем $k_i = 0$, откуда вытекает, что V_i независимо, т. е. имеет место сбалансированная раскраска графа G в k цветов. Теорема 1 доказана.

Заключение

В работе доказана NP-полнота задачи выбора в некотором множестве векторов d-мерного евклидова пространства k кластеров большого размера с заданным ограничением на максимальный разброс в случае, когда размерность пространства является частью входа. Ранее сложностной статус этой задачи был неизвестен при $k \geqslant 2$. Интересной открытой проблемой остаётся выяснение сложности задачи в случае фиксированной размерности пространства.

Финансирование работы

Работа выполнена в рамках государственного задания Институт математики им. С. Л. Соболева СО РАН (проект № FWNF–2022–0019).

Конфликт интересов

Автор заявляет, что у него нет конфликта интересов.

Литература

1. **Berkhin P.** A survey of clustering data mining techniques // Grouping multidimensional data: Recent advances in clustering. Heidelberg: Springer, 2006. P. 25–71. DOI: 10.1007/3-540-28349-8_2.
2. **Jain A. K., Dubes R. C.** Algorithms for clustering data. Englewood Cliffs, NJ: Prentice Hall, 1988. 320 p.
3. **Ghoreyshi S., Hosseinkhani J.** Developing a clustering model based on K-means algorithm in order to creating different policies for policyholders in insurance industry // Int. J. Adv. Comput. Sci. Inf. Technol. 2015. V. 4, No. 2. P. 46–53.
4. **Fisher W. D.** On grouping for maximum homogeneity // J. Am. Stat. Assoc. 1958. V. 53, No. 284. P. 789–798.
5. **MacQueen J.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability (Berkeley, USA, June 21 – July 18, 1965; Dec. 27, 1965 – Jan. 7, 1966). V. 1. Berkeley: Univ. Calif. Press, 1967. P. 281–297.
6. **Blömer J., Lammersen C., Schmidt M., Sohler C.** Theoretical analysis of the k-means algorithm — A survey // Algorithm engineering: Selected results and surveys. Cham: Springer, 2016. P. 81–116. (Lect. Notes Comput. Sci.; V. 9220). DOI: 10.1007/978-3-319-49487-6.
7. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. 2009. V. 75, No. 2. P. 245–248.
8. **Pyatkin A. V., Aloise D., Mladenovic N.** NP-hardness of balanced minimum sum-of-squares clustering // Pattern Recognit. Lett. 2017. V. 97. P. 44–45.
9. **Bertoni A., Goldwurm M., Lin J., Saccà F.** Size constrained distance clustering: Separation properties and some complexity results // Fund. Inform. 2012. V. 115, No. 1. P. 125–139. DOI: 10.3233/FI-2012-620.

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

12. Khandeev V. I., Neshchadim S. M. Constant-factor approximation algorithms for some maximin multi-clustering problems // Mathematical optimization theory and operations research. Proc. 22nd Int. Conf. (Yekaterinburg, Russia, July 2–8, 2023). Cham: Springer, 2023. P. 85–100. (Lect. Notes Comput. Sci.; V. 13930). DOI: 10.1007/978-3-031-35305-5_6.

13. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.

14. Dailey D. P. Uniqueness of colorability and colorability of planar 4-regular graphs are NP-complete // Discrete Math. 1980. V. 30, No. 3. P. 289–293. DOI: 10.1016/0012-365X(80)90236-8.

15. Кельманов А. В., Пяткин А. В. Об одном варианте задачи выбора подмножества векторов // Дискрет. анализ и исслед. операций. 2008. Т. 15, № 5. С. 20–34.

Пяткин Артём Валерьевич

Статья поступила
8 ноября 2023 г.

После доработки —
20 ноября 2023 г.

Принята к публикации
22 декабря 2023 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
April–June 2024. Vol. 31, No. 2. P. 136–143

UDC 519.8+518.25 DOI: 10.33048/daio.2024.31.787

ON THE COMPLEXITY OF THE PROBLEM OF CHOICE OF LARGE CLUSTERS

A. V. Pyatkin

Sobolev Institute of Mathematics,
4 Acad. Koptyug Avenue, 630090 Novosibirsk, Russia
E-mail: artempyatkin@gmail.com

Abstract. The paper considers the following problem. Given a set of Euclidean vectors, find several clusters with a restriction on the maximum scatter of each cluster so that the size of the minimum cluster would be maximum. Here the scatter is the sum of squared distances from the cluster elements to its centroid. The NP-hardness of this problem is proved in the case where the dimension of the space is a part of the input. Bibliogr. 15.

Keywords: cluster, centroid, scatter, NP-hardness.

References

1. **P. Berkhin,** A survey of clustering data mining techniques, in *Grouping Multidimensional Data: Recent Advances in Clustering* (Springer, Heidelberg, 2006), pp. 25–71, DOI: 10.1007/3-540-28349-8_2.
2. **A. K. Jain** and **R. C. Dubes,** *Algorithms for Clustering Data* (Prentice Hall, Englewood Cliffs, NJ, 1988).
3. **S. Ghoreyshi** and **J. Hosseinkhani,** Developing a clustering model based on K-means algorithm in order to creating different policies for policyholders in insurance industry, *Int. J. Adv. Comput. Sci. Inf. Technol.* **4** (2), 46–53 (2015).
4. **W. D. Fisher,** On grouping for maximum homogeneity, *J. Am. Stat. Assoc.* **53** (284), 789–798 (1958).
5. **J. MacQueen,** Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability, Berkeley, USA, June 21 – July 18, 1965; Dec. 27, 1965 – Jan. 7, 1966,* Vol. 1 (Univ. Calif. Press, Berkeley, 1967), pp. 281–297.

English transl.: *Journal of Applied and Industrial Mathematics* **18** (2), 312–315 (2024), DOI: 10.1134/S1990478924020121.

6. **J. Blömer, C. Lammersen, M. Schmidt,** and **C. Sohler,** Theoretical analysis of the k-means algorithm — A survey, in *Algorithm Engineering: Selected Results and Surveys* (Springer, Cham, 2016), pp. 81–116 (Lect. Notes Comput. Sci., Vol. 9220), DOI: 10.1007/978-3-319-49487-6_3.
7. **D. Aloise, A. Deshpande, P. Hansen,** and **P. Popat,** NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.* **75** (2), 245–248 (2009).
8. **A. V. Pyatkin, D. Aloise,** and **N. Mladenovic,** NP-hardness of balanced minimum sum-of-squares clustering, *Pattern Recognit. Lett.* **97**, 44–45 (2017).
9. **A. Bertoni, M. Goldwurm, J. Lin,** and **F. Saccà,** Size constrained distance clustering: Separation properties and some complexity results, *Fund. Inform.* **115** (1), 125–139 (2012), DOI: 10.3233/FI-2012-620.
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)].
11. **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), DOI: 10.21538/0134-4889-2018-24-4-189-198 [Russian] [*Proc. Steklov Inst. Math.* **309** (Suppl. 1), S65–S73 (2020), DOI: 10.1134/S0081543820040082].
12. **V. I. Khandeev** and **S. M. Neshchadim,** Constant-factor approximation algorithms for some maximin multi-clustering problems, in *Mathematical Optimization Theory and Operations Research* (Proc. 22nd Int. Conf., Yekaterinburg, Russia, July 2–8, 2023) (Springer, Cham, 2023), pp. 85–100 (Lect. Notes Comput. Sci., Vol. 13930), DOI: 10.1007/978-3-031-35305-5_6.
13. **M. R. Garey** and **D. S. Johnson,** *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979; Mir, Moscow, 1982 [Russian]).
14. **D. P. Dailey,** Uniqueness of colorability and colorability of planar 4-regular graphs are NP-complete, *Discrete Math.* **30** (3), 289–293 (1980), DOI: 10.1016/0012-365X(80)90236-8.
15. **A. V. Kel'manov** and **A. V. Pyatkin,** On a version of the problem of choosing a vector subset, *Diskretn. Anal. Issled. Oper.* **15** (5), 20–34 (2008) [Russian] [*J. Appl. Ind. Math.* **3** (4), 447–455 (2009)].

Artem V. Pyatkin

Received November 8, 2023
Revised November 20, 2023
Accepted December 22, 2023