ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 488, № 1, с. 16–20

МАТЕМАТИКА

УДК 519.16 + 519.72 + 519.85

NP-ПОЛНОТА НЕКОТОРЫХ ЗАДАЧ РАЗБИЕНИЯ КОНЕЧНОГО МНОЖЕСТВА ТОЧЕК ЕВКЛИДОВА ПРОСТРАНСТВА НА СБАЛАНСИРОВАННЫЕ КЛАСТЕРЫ

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

Представлено академиком РАН С.С. Гончаровым 13.05.2019 г.

Поступило 15.05.2019 г.

Рассматриваются три родственные между собой задачи разбиения N-элементного множества точек d-мерного евклидова пространства на два кластера так, чтобы сбалансировать значения: (1) внутрикластерного квадратичного разброса, нормированного на размер кластера, в первой задаче; (2) внутрикластерного квадратичного разброса, во второй задаче; (3) мощностно-взвешенного внутрикластерного разброса, в третьей задаче. Доказано, что все эти задачи NP-полны.

Ключевые слова: евклидово пространство, сбалансированное разбиение, квадратичный разброс, нормированный на размер кластера разброс, мощностно-взвешенный разброс, NP-полнота.

DOI: https://doi.org/10.31857/S0869-5652488116-20

Предметом исследования этой работы являются задачи дискретной оптимизации, а именно, мы анализируем три родственные квадратичные евклидовы задачи сбалансированного 2-разбиения конечного множества точек на кластеры. Цель исследования — выяснение вопроса о статусе сложности этих задач.

Исследование мотивировано открытостью указанного вопроса, а также важностью рассматриваемых задач как для математической теории оптимизации, так и для приложений, среди которых, в частности, анализ данных (Data analysis), интерпретация данных (Data mining) и статистика (Statistics). Рассматриваемые задачи разбиения на кластеры не эквивалентны ни одной из хорошо известных труднорешаемых кластеризационных геометрических задач — k-means [1], k-medians [2], k-center [3], k-Variance [4] и др. Насколько нам известно, рассматриваемые задачи также не эквивалентны ни одной из других труднорешаемых квадратичных евклидовых задач разбиения, выявленных в последние годы.

Рассматриваемые задачи имеют следующие формулировки.

Задача 1. Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\left| \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} \| y - \bar{y}(\mathcal{C}) \|^2 - \right.$$

$$\left. - \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \| y - \bar{y}(\mathcal{Y} \setminus \mathcal{C}) \|^2 \right| \le \varepsilon, \qquad (1)$$

где $\bar{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ и $\bar{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ — центроиды (геометрические центры) кластеров $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ соответственно?

Задача 2. Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\left| \sum_{y \in \mathcal{C}} \| y - \bar{y}(\mathcal{C}) \|^2 - \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \| y - \bar{y}(\mathcal{Y} \setminus \mathcal{C}) \|^2 \right| \le \varepsilon? \qquad (2)$$

Задача 3. Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\left| |\mathcal{C}| \sum_{y \in \mathcal{C}} \| y - \bar{y}(\mathcal{C}) \|^2 - \right.$$

$$\left. - |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \| y - \bar{y}(\mathcal{Y} \setminus \mathcal{C}) \|^2 \right| \le \varepsilon? \qquad (3)$$

[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*

В статистике хорошо известен F-критерий Фишера сравнения дисперсий по выборочным данным из двух распределений [5]. Если рассматривать кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ как выборки из двух нормальных распределений с неизвестными средними, то этот критерий позволяет сравнить (проверить на равенство) выборочные дисперсии

$$\frac{1}{|\mathcal{C}|-1}\sum_{y\in\mathcal{C}} \| y-\bar{y}(\mathcal{C})\|^2, \qquad (4)$$

$$\frac{1}{|\mathcal{Y}\setminus\mathcal{C}|-1}\sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \| y-\bar{y}(\mathcal{Y}\setminus\mathcal{C})\|^2 \qquad (5)$$

этих распределений по их отношению, которое близко к 1 в случае равенства.

Формулы (4) и (5) в статистике известны как несмещённые оценки дисперсии по выборочным данным. В задаче 1 фигурируют смещённые оценки, которые отличаются от несмещённых лишь знаменателями. Однако это отличие, как известно, несущественно в асимптотическом смысле, так как обе оценки являются асимптотически несмещёнными.

Легко видеть, что в задаче 1 требуется разбить входное множество $\mathcal{Y}$ на два кластера по критерию сбалансированных выборочных дисперсий. Статистическая интерпретация задачи 1: можно ли разбить неоднородную выборку $\mathcal{Y}$ на две части (подвыборки) $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$ с выборочными дисперсиями, которые отличаются не более чем на некоторое заданное число $\varepsilon > 0$?

Задачи 2 и 3 аналогичны по смыслу задаче 1, но отличаются критериями разбиения. В задаче 2 критерием является суммарный квадратичный разброс относительно среднего (т.е. центроида). В задаче 3 таким критерием является мощностно-взвешенный разброс относительно центроида или сумма квадратов попарных расстояний, так как для любого конечного множества $\mathcal{Z} \subset \mathbb{R}^d$ справедливо легко проверяемое равенство

$$|\mathcal{Z}| \sum_{z\in\mathcal{Z}} \| z-\bar{z}(\mathcal{Z})\|^2 = \frac{1}{2}\sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{Z}} \| z-x\|^2, \qquad (6)$$

где $\bar{z}(\mathcal{Z})$ — центроид множества $\mathcal{Z}$.

Для выяснения статуса вычислительной сложности сформулированных задач мы рассматриваем следующую задачу, которая объединяет задачи 1–3.

Задача $\Pi(g(x))$. Дано: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и вещественное число $\varepsilon > 0$. Вопрос: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$, что

$$\left| \frac{g(|\mathcal{C}|)}{2}\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}} \| y-z\|^2 - \right.$$

$$\left. - \frac{g(|\mathcal{Y}\setminus\mathcal{C}|)}{2}\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\sum_{z\in\mathcal{Y}\setminus\mathcal{C}} \| y-z\|^2 \right| \le \varepsilon? \qquad (7)$$

С учётом (6), будем считать, что весовой коэффициент $g(x)$ у сумм в неравенстве (7) объединённой задачи равен $\frac{1}{x^2}$ для задачи 1, $\frac{1}{x}$ — для задачи 2 и 1 — для задачи 3, где x — мощность соответствующего кластера.

Далее мы доказываем NP-полноту задачи $\Pi(g(x))$ для каждой функции $g(x)$ из множества $\left\{\frac{1}{x^2}, \frac{1}{x}, 1\right\}$. Для этого используем известный NP-полный вариант классической задачи Разбиение, которая имеет следующую формулировку.

Задача РРД (Разбиение с равными долями). Дано: мультимножество из $2K$ целых неотрицательных чисел $a_1, \ldots, a_{2K}$, сумма которых равна $2W$. Вопрос: существует ли разбиение этого мультимножества на два мультиподмножества по K элементов каждое с суммой элементов в каждом из мультиподмножеств, равной W?

Теорема 1. *Для каждой функции*

$$g(x) \in \left\{\frac{1}{x^2}, \frac{1}{x}, 1\right\}$$

задача $\Pi(g(x))$ NP-*полна*.

Идея доказательства состоит в следующем. Мы рассматриваем произвольный пример задачи РРД, т.е. мультимножество $A = \{a_1, \ldots, a_{2K}\}$, сумма элементов которого равна $2W$. В этом примере можно считать, что $K > 3$, а также, что $W \ge 9$, так как в противном случае задача, очевидно, решается за линейное время с помощью полного перебора или с помощью динамического программирования соответственно.

Положим $\varepsilon = \frac{1}{3K}$ и выберем рациональные числа b_i, $i = 1, 2, \ldots, 2K$, удовлетворяющие соотношению

$$\sqrt{a_i} \le b_i \le \sqrt{a_i} + \delta, \qquad (8)$$

где

$$\delta = \frac{\varepsilon}{K(K-1)Wg(K)}. \qquad (9)$$

По произвольному входу задачи РРД построим пример входа задачи $\Pi(g(x))$. Положим $N = 2K$, $d = 4K$, $\mathcal{Y} = \{y_1, \ldots, y_{2K}\}$, где для всех $i = 1, 2, \ldots, 2K$

точка y_i содержит рациональное число (параметр) $M > 0$ в компоненте i и число b_i в компоненте $2K + i$, а все остальные компоненты этой точки равны 0. Значения M уточним в зависимости от функции $g(x)$.

Далее мы устанавливаем следующие свойства элементов семейства $\mathcal{Y}$ точек в построенном примере входа задачи $\Pi(g(x))$.

Свойство 1. Пусть подмножество $I \subset \{1, 2, ..., 2K\}$ содержит номера точек из кластера $\mathcal{C} \subset \{\mathcal{Y}\}$. Тогда для точек из $\mathcal{C}$ справедливо равенство

$$\sum_{y\in\mathcal{C}}\sum_{x\in\mathcal{C}} \| z - x \|^2 =$$

$$= 2 | \mathcal{C} | (| \mathcal{C} | - 1) M^2 + 2(| \mathcal{C} | - 1)\sum_{i\in I} b_i^2. \qquad (10)$$

Свойство 2. Для суммы квадратов координатных элементов (8) справедлива оценка

$$\sum_{i=1}^{2K} b_i^2 \le 2W(1 + K\delta). \qquad (11)$$

Свойство 3. Для любого разбиения множества $\{1, 2, ..., 2K\}$ индексов на подмножества I_1, I_2 мощности K справедливо

$$\left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| - KW\delta \le \left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \le$$
$$\le \left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| + KW\delta. \qquad (12)$$

Допустим теперь, что подмножество I_1 индексов содержит номера точек из $\mathcal{C}$, а подмножество I_2 — номера точек из $\mathcal{Y} \setminus \mathcal{C}$, причём $| I_1 | = c$. Тогда в задаче $\Pi(g(x))$ неравенство (7) при $x = c$ в соответствии с (10) принимает вид

$$\Big|((c^2 - c)g(c) - ((2K - c)^2 - (2K - c))g(2K - c))M^2 +$$
$$+ (c - 1)g(c)\sum_{i\in I_1} b_i^2 - (2K - c - 1)g(2K - c)\sum_{i\in I_2} b_i^2\Big| \le \varepsilon. \qquad (13)$$

Если в задаче РРД существует разбиение мультимножества A на два требуемых подмножества, то для соответствующих подмножеств индексов выполнено $c = | I_1 | = | I_2 | = K$ и

$$\sum_{i\in I_1} a_i = \sum_{i\in I_2} a_i. \qquad (14)$$

Тогда, подставив $c = K$ в (13) и учитывая (12) и (14), легко заметить, что для выполнения неравенства (7) в задаче $\Pi(g(x))$ достаточно, чтобы

$$(K - 1)g(K)\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \le K(K - 1)g(K)W\delta \le \varepsilon,$$

что выполняется, если имеет место равенство (9).

Таким образом, если в задаче РРД существует требуемое разбиение, то и в задаче $\Pi(g(x))$ требуемое разбиение также существует при выборе (9) параметра δ.

Для доказательства обратной импликации мы указываем значение параметра M для каждой из трёх функций $g(x) \in \left\{\frac{1}{x^2}, \frac{1}{x}, 1\right\}$ в объединённой задаче $\Pi(g(x))$.

I. При $g(x) = 1$ из (9) имеем $\delta = \dfrac{\varepsilon}{K(K - 1)W}$.

При этом неравенство (13) в задаче $\Pi(g(x))$ принимает вид

$$\varepsilon \ge \Big|(4Kc - 4K^2 + 2K - 2c)M^2 +$$
$$+ (c - 1)\sum_{i\in I_1} b_i^2 - (2K - c - 1)\sum_{i\in I_2} b_i^2\Big|.$$

Из этого неравенства с учётом (11) следует

$$\varepsilon \ge | K - c | M^2 - 4KW(K\delta + 1). \qquad (15)$$

Выберем M так, что $M^2 > \varepsilon + 4KW(K\delta + 1)$. Тогда коэффициент при M^2 в правой части (15) должен быть равен нулю. Действительно, если $K \ne c$, то $|K - c| \ge 1$, и из неравенства (15) имеем $\varepsilon \ge M^2 - 4KW(K\delta + 1) > \varepsilon$, противоречие. Таким образом, $|K - c| = 0$, а значит, $K = c$.

Далее, из целочисленности элементов мультимножества A следует: если в задаче РРД

$$\sum_{i\in I_1} a_i \ne \sum_{i\in I_2} a_i, \qquad (16)$$

то

$$\left|\sum_{i\in I_1} a_i - \sum_{i\in I_2} a_i\right| \ge 1. \qquad (17)$$

Поэтому из неравенства (13) с учётом (12) имеем

$$\varepsilon \ge (K - 1)\left|\sum_{i\in I_1} b_i^2 - \sum_{i\in I_2} b_i^2\right| \ge (K - 1)(1 - KW\delta) =$$
$$= K - 1 - \varepsilon,$$

что противоречит условию $\varepsilon = \frac{1}{3K}$. Следовательно, в задаче РРД выполнено (14).

II. Если $g(x) = \frac{1}{x}$, то из (9) имеем $\delta = \frac{\varepsilon}{(K-1)W}$. При этом неравенство (13) принимает вид

$$\varepsilon \geq \left| (c - 1 - (2K - c - 1))M^2 + \right.$$

$$\left. + \frac{c-1}{c}\sum_{i \in I_1} b_i^2 - \frac{2K-c-1}{2K-c}\sum_{i \in I_2} b_i^2 \right|,$$

откуда с учётом (11) имеем

$$\varepsilon \geq 2 \mid K - c \mid M^2 - 2W(K\delta + 1). \qquad (18)$$

Выберем M так, что $M^2 > \varepsilon + 2W(K\delta + 1)$. Тогда при этом выборе параметра M из (18) имеем $K = c$, как и в случае I.

Далее, как и в рассмотренном случае I, если в задаче РРД имеет место неравенство (16), то из этого неравенства следует (17). Поэтому из (13) с учётом (12) имеем

$$\varepsilon \geq \frac{K-1}{K}\left|\sum_{i \in I_1} b_i^2 - \sum_{i \in I_2} b_i^2\right| \geq \frac{K-1}{K}(1 - KW\delta) =$$

$$= \frac{K-1}{K} - \varepsilon,$$

что противоречит условиям $\varepsilon = \frac{1}{3K}$ и $K > 3$. Таким образом, как и в случае I, в задаче РРД выполнено (14).

III. Наконец, при $g(x) = \frac{1}{x^2}$ из (9) имеем $\delta = \frac{K\varepsilon}{(K-1)W}$. При этом неравенство (13) имеет вид

$$\varepsilon \geq \left| \left(\frac{c-1}{c} - \frac{2K-c-1}{2K-c}\right)M^2 + \right.$$

$$\left. + \frac{c-1}{c^2}\sum_{i \in I_1} b_i^2 - \frac{2K-c-1}{(2K-c)^2}\sum_{i \in I_2} b_i^2 \right|.$$

Из этого неравенства с учётом (11) получаем

$$\varepsilon \geq \frac{\mid K - c \mid}{2K^2} M^2 - 2W(K\delta + 1). \qquad (19)$$

Выберем теперь M так, что $M^2 > 2K^2(\varepsilon + 2W(K\delta + 1))$. Тогда из (19) имеем $K = c$, как и в случаях I и II.

Далее, как и в рассмотренных случаях I и II, если в задаче РРД имеет место неравенство (16), то из этого неравенства следует (17). Поэтому из (13) с учётом (12) имеем

$$\varepsilon \geq \frac{K-1}{K^2}\left|\sum_{i \in I_1} b_i^2 - \sum_{i \in I_2} b_i^2\right| \geq \frac{K-1}{K^2}(1 - KW\delta) =$$

$$= \frac{K-1}{K^2} - \varepsilon,$$

т.е. $K - 1 \leq 2K^2\varepsilon = \frac{2K}{3}$, что неверно при $K > 3$. Поэтому в задаче РРД выполнено (14), как и в случаях I и II.

Таким образом, в задаче $\Pi(g(x))$ для каждой функции $g(x)$ выполнение неравенства (7) влечёт в задаче РРД существование разбиения мультимножества A на две равные по мощности доли с одинаковыми суммами элементов. Из теоремы 1 и тождества (6) непосредственно вытекает основной результат настоящей работы.

Следствие 1. *Задачи* 1–3 NP-*полны*.

Очевидно, что рассмотренные задачи 1–3 можно обобщить на случай, когда число кластеров больше, чем 2. В этих обобщениях задач вопросом является — можно ли разбить входное множество так, чтобы соответствующие неравенства (1), (2) и (3) выполнялись для всех пар кластеров. Ясно, что когда число кластеров является частью входа, эти обобщения тоже являются NP-полными задачами. Вопрос о статусе сложности параметрического случая этих задач, когда число кластеров не является частью входа, остаётся открытым. Построение эффективных приближённых алгоритмов с гарантированными оценками точности является делом ближайшей перспективы.

Источники финансирования. Работа выполнена при финансовой поддержке РФФИ, проекты 19–01–00308 и 18–31–00398, программы ФНИ РАН, проекты 0314–2019–0015 и 0314–2019–0014, а также программы Проекта 5–100 Министерства образования и науки РФ.

СПИСОК ЛИТЕРАТУРЫ

1. *Aloise D.*, *Deshpande A.*, *Hansen P.*, *Popat P.* // NP-hardness of Euclidean sum-of-squares clustering. Machine Learning. 2009. V. 75. № 2. P. 245–248.

2. *Papadimitriou C.H.* // Worst-Case and Probabilistic Analysis of a Geometric Location Problem. SIAM J. Comput. 1981. V. 10. № 3. P. 542–557.
3. *Masuyama S., Ibaraki T., Hasegawa T.* // The computational complexity of the m-center problems in the plane. IEEE Trans. IECE Jpn. 1981. V. 64. № 2. P. 57–64.
4. *Aggarwal H., Imai N., Katoh N., Suri S.* // Finding k points with minimum diameter and related problems. J. Algorithms. 1991. V. 12. № 1. P. 38–56.
5. *Snedecor G.W., Cochran W.G.* // Statistical Methods, Eighth Edition, Iowa State University Press, 1989.
6. *Garey M.R., Johnson D.S.* // Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman. San Francisco, 1979.

ON THE COMPLEXITY OF SOME PARTITION PROBLEMS OF A FINITE SET OF POINTS IN EUCLIDEAN SPACE INTO BALANCED CLUSTERS

A. V. Kel'manov[1,2], A. V. Pyatkin[1,2], V. I. Khandeev[1,2]

[1] *Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russian Federation*

[2] *Novosibirsk State University, Novosibirsk, Russian Federation*

Presented by Academician of the RAS S.S. Goncharov May 13, 2019

Received May 5, 2019

We consider some problems of partitioning a finite set of N points in d-dimension Euclidean space into two clusters balancing the value of (1) the quadratic variance normalized by a cluster size, (2) the quadratic variance, and (3) the size-weighted quadratic variance. We have proved the NP-completeness of all these problems.

Keywords: euclidean space, balanced partition, quadratic variance, normalized by the cluster size, sized-weighted, NP-completeness.