ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 489, № 4, с. 339–343

МАТЕМАТИКА

УДК 519.16 + 519.72 + 519.85

NP-ТРУДНОСТЬ КВАДРАТИЧНОЙ ЕВКЛИДОВОЙ ЗАДАЧИ 2-КЛАСТЕРИЗАЦИИ 1-MEAN И 1-MEDIAN С ОГРАНИЧЕНИЯМИ НА РАЗМЕРЫ КЛАСТЕРОВ

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

Представлено академиком РАН С.С. Гончаровым 06.09.2019 г.

Поступило 06.09.2019 г.

В работе рассматривается задача разбиения N-элементного множества точек в d-мерном евклидовом пространстве на два кластера. В этой задаче требуется найти 2-разбиение, минимизирующее сумму (по обоим кластерам) внутрикластерных квадратичных разбросов точек относительно искомых центров. Центр одного кластера определяется как центроид (геометрический центр), а центр другого кластера является искомой точкой во входном множестве. Анализируется вариант задачи, в котором размеры (т.е. мощности) кластеров заданы, а их суммарный размер совпадает с размером входного множества. Доказано, что задача NP-трудна в сильном смысле.

Ключевые слова: евклидово пространство, кластеризация, 2-разбиение, квадратичный разброс, центр, центроид, медиана, сильная NP-трудность.

DOI: https://doi.org/10.31857/S0869-56524894339-343

Предметом исследования этой работы является дискретная экстремальная задача разбиения конечного множества точек евклидова пространства на два кластера. Цель исследования — анализ вычислительной сложности задачи.

Исследование мотивировано неизученностью задачи в математическом плане. А именно до настоящего времени статус её вычислительной сложности не был установлен. Задача имеет приложения в компьютерной геометрии и математической статистике. С практической точки зрения рассматриваемая задача важна, например, для известной междисциплинарной проблемы интерпретации данных (Data mining).

1. ФОРМУЛИРОВКА ЗАДАЧИ И ПОХОЖИЕ ПРОБЛЕМЫ

Всюду далее $\|\cdot\|$ — евклидова норма. Рассматриваемая задача имеет следующую формулировку.

Задача 1 (Quadratic 1-Mean and 1-Median 2-Clustering with the Constraints on the Cluster Sizes). Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. Найти: такие точку $x \in \mathcal{Y}$ и разбиение $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ размеров M и $N - M$ соответственно, что

$$f(\mathcal{C}, x) = \sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - x\|^2 \to \min, \quad (1)$$

где

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

есть центроид (геометрический центр) подмножества $\mathcal{C}$.

Ниже приведены три известные задачи, математические формулировки которых наиболее близки к задаче 1. В этих задачах целевые функции отличны от целевой функции задачи 1, а входы идентичны входу этой задачи.

Задача 2 (2-Means Clustering with the Constraints on the Cluster Sizes). Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. Найти: такое разбиение $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ размеров M и $N - M$ соответственно, что

$$\sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \bar{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \to \min_{\mathcal{C} \subset \mathcal{Y}},$$

где $y(\mathcal{C})$ и $\bar{y}(\mathcal{Y} \setminus \mathcal{C})$ — центроиды кластеров $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ соответственно.

Сильная NP-трудность этой задачи следует из сильной NP-трудности [1] задачи 2-Means (без

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

[2] *Новосибирский национальный исследовательский государственный университет*

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

***E-mail: artempyatkin@gmail.com*

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

ограничений на размеры кластеров). Действительно, полиномиальная разрешимость задачи 2 влекла бы полиномиальную разрешимость задачи 2-Means. Достаточно было бы перебрать за полиномиальное время конечное число точных допустимых решений задачи для каждого M.

Задача 3 (Quadratic 2-Medians Clustering with the Constraints on the Cluster Sizes). Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. Найти: такие точки x, $z \in \mathcal{Y}$ и разбиение $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ размеров M и $N - M$ соответственно, что

$$\sum_{y \in \mathcal{C}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - z\|^2 \to \min.$$

Эта задача сходна с известной (см., например, [2]) задачей 2-Medians минимизации суммы

$$\sum_{y \in \mathcal{C}} \|y - x\| + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - z\|.$$

Легко видеть, что задача 3 разрешима за время $\mathcal{O}(dN^3)$. Достаточно перебрать N^2 пар x, z точек, для каждой пары найти допустимое решение и в полученном семействе найти наилучшее. Допустимое решение строится за $\mathcal{O}(dN)$ операций:

1) проектируем все точки множества $\mathcal{Y}$ на прямую, соединяющую точки x и z;

2) разбиваем индуцированный этим проектированием отрезок на два примыкающих отрезка по M и $N - M$ точек.

Задача 4 (Quadratic 1-Mean and Given 1-Center 2-Clustering with the Constraints on the Cluster Sizes). Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. Найти: такое разбиение $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ размеров M и $N - M$ соответственно, что

$$g(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min,$$

где $y(\mathcal{C})$ — центроид кластера $\mathcal{C}$.

В этой задаче центр квадратичного разброса точек кластера $\mathcal{Y} \setminus \mathcal{C}$ совпадает с началом координат. Очевидно, что эта задача полиномиально эквивалентна задаче, в которой этим центром является произвольная точка из $\mathbb{R}^d$, заданная на входе. Сильная NP-трудность задачи 4 следует из равенства

$$g(\mathcal{C}) = \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} \left\| \sum_{y \in \mathcal{C}} y \right\|^2,$$

в котором первый член в правой части не зависит от $\mathcal{C}$, и сильной NP-трудности [3–5] известной задачи Longest M-Vector Sum. Напомним, что в этой задаче дано N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и натуральное число M. Требуется найти подмножество $\mathcal{C} \subseteq \mathcal{Y}$ размера M, доставляющее максимум норме $\left\| \sum_{y \in \mathcal{C}} y \right\|$ суммы элементов из этого подмножества.

Все приведённые экстремальные задачи, включая задачу 1, имеют геометрический характер, который ясен непосредственно из формулировок задач. В каждой из задач 1–4 оптимальному разбиению на кластеры соответствует разделяющая поверхность. Этими поверхностями являются оптимальные гиперплоскости, которые перпендикулярны отрезку, соединяющему центры квадратичного разброса. Этот факт легко устанавливается при анализе структурных свойств оптимальных решений задач. Поэтому все задачи можно трактовать как поиск оптимальной гиперплоскости, разделяющей на две части входное множество точек.

Напомним, что построение оптимальных разделяющих поверхностей по имеющимся данным — типичная прикладная проблема машинного обучения (Machine learning) [6], кластеризации данных (Data clustering) [7] и распознавания образов (Pattern recognition) [8]. В отмеченных приложениях эта проблема возникает всякий раз, когда в руках у исследователя-прикладника оказываются данные с неясной структурой.

Выяснение структуры данных с помощью так называемого разведочного поиска подходящего (адекватного) описания (т.е. интерпретации) данных в виде модели порождения данных типичен для прикладных проблем Data mining [9] и математической статистики. В классической статистике, в отличие от Data mining, предполагается, что данные однородны, т.е. являются выборкой из одного распределения. Напротив, в Data mining предполагается, что данные неоднородны, т.е. являются выборкой из нескольких распределений, причём априорное соответствие данных распределениям неизвестно. Отсутствие этого соответствия обусловливает создание математических инструментов в виде эффективных алгоритмов решения необозримого множества задач разбиения данных с самой разнообразной структурой на однородные по какому-либо фиксированному критерию кластеры, а также инструментов в виде критериев проверки адекватности аппроксимационных моделей разбиения имеющимся данным. Например, чтобы выяснить, какая из сформулированных выше задач (моделей аппроксимации) разбиения адекватна данным (входному множеству

точек) или ни одна из них не адекватна данным, в первую очередь необходимы эффективные алгоритмы решения этих кластеризационных задач. Очевидно, что создание эффективных в вычислительном плане алгоритмов является одной из ключевых проблем для Data mining. В свою очередь, создание таких алгоритмов обусловливает исследование сложностного статуса задач разбиения. Приведённые замечания поясняют мотивацию настоящего исследования. Фактически наша работа отвечает на вопрос — можно ли эффективно (за полиномиальное время) разбить имеющиеся данные в соответствии с (1).

В заключение этого раздела подчеркнём, что рассматриваемая задача 1 не эквивалентна ни одной из приведённых выше близких по постановке кластеризационных задач. Насколько нам известно, она не входит в список других изучавшихся ранее задач дискретной оптимизации. К тому же она не является ни частным случаем, ни обобщением какой-либо из этих задач. Поэтому вопрос о статусе сложности задачи 1 требует отдельного исследования.

2. АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ

Как известно (см., например, [10]), для любого конечного множества точек $\mathcal{Z} \subset \mathbb{R}^d$ справедливо равенство

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

где $\overline{z}(\mathcal{Z})$ — центроид множества $\mathcal{Z}$. Используя это равенство, запишем целевую функцию (1) в эквивалентном виде и сформулируем задачу 1 в форме верификации свойств.

Задача 1А. Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве, натуральное $M < N$ и число $B > 0$. Вопрос: существуют ли в $\mathcal{Y}$ такие кластер $\mathcal{C}$ размера M и точка $x \in \mathcal{Y}$, что имеет место неравенство

$$f(\mathcal{C}, x) = \frac{1}{2|\mathcal{C}|}\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}} \|y - z\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y - x\|^2 \le B? \quad (2)$$

Напомним следующую NP-полную [11] задачу.

Задача Клика (Clique). Дано: n-вершинный граф $G = (V, E)$ и положительное число K. Вопрос: существует ли в графе G клика размера не меньше K?

Для выяснения вопроса о сложности рассматриваемой задачи нам потребуется специальный случай задачи Clique для однородного графа, степень Δ которого не фиксирована. Эта задача также относится к числу NP-полных задач [12].

Справедлива следующая

Теорема 1. *Задача* 1А NP-*полна в сильном смысле*.

Для доказательства теоремы мы строим полиномиальное сведение задачи Clique в однородном графе к задаче 1А.

Рассмотрим однородный граф $G = (V, E)$ степени Δ на n вершинах. Будем считать, что $\Delta > 2$ и $n - K > 1$, так как в противном случае задача Clique решается за полиномиальное время.

По произвольному входу задачи Clique построим следующий пример входа задачи 1А. В задаче 1А положим

$$\begin{gathered} d = |E|, \quad N = n + 1, \\ M = K, \quad B = (n-1)\Delta - K + 1, \\ y_N = 0; \quad y_i = a_i, \quad i = 1, 2, \ldots, n, \end{gathered} \quad (3)$$

где a_i — i-я строка матрицы $A = \{a_{i,j}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d$, инцидентности графа G, в которой $a_{i,j} = 1$, если j-е ребро графа G инцидентно вершине v_i, и $a_{i,j} = 0$ в противном случае.

Ввиду однородности графа G имеем следующие свойства элементов в построенном примере множества $\mathcal{Y}$:

$$\|y_i - y_j\|^2 = \begin{cases} 2\Delta - 2, & \text{если ребро } v_iv_j \in E(G), \\ 2\Delta & \text{в противном случае,} \end{cases} \quad (4)$$
$$1 \le i < j \le n,$$

$$\|y_i - y_N\|^2 = \Delta, \quad i = 1, 2, \ldots, n. \quad (5)$$

Кроме того, заметим, что (по построению) любому кластеру $\mathcal{C} \subseteq \mathcal{Y} \setminus \{y_N\}$ однозначно соответствует подмножество $V_C \subseteq V$ вершин графа G.

I. Допустим сначала, что в задаче Clique подмножество V_C вершин образует клику размера K. В задаче 1А в качестве центра кластера $\mathcal{Y} \setminus \mathcal{C}$ выберем точку $x = y_N$. При этом выборе из (2)–(5) для целевой функции задачи 1А в построенном примере имеем

$$f(\mathcal{C}, x) = \frac{1}{2|\mathcal{C}|}\sum_{y_i\in\mathcal{C}}\sum_{y_j\in\mathcal{C}} \|y_i - y_j\|^2 + \sum_{y_i\in\mathcal{Y}\setminus\mathcal{C}} \|y_i - y_N\|^2 =$$
$$= \frac{1}{2M} M(M-1)(2\Delta - 2) + (n - M)\Delta =$$
$$= n\Delta - \Delta - M + 1 = B.$$

Это значит, что условие (2) выполнено в виде равенства, т.е. в примере задачи 1А найдутся соответствующие этому условию кластер $\mathcal{C}$ размера $M = K$ и точка $x = y_N$, если в задаче Clique существует клика размера K.

II. Допустим теперь, что в построенном примере задачи 1A существуют некоторый кластер $\mathcal{C} \subset \mathcal{Y}$ мощности M и точка $x \in \mathcal{Y}$ такие, что $f(\mathcal{C}, x) \le B$.

Сначала мы показываем от противного, что $y_N \notin \mathcal{C}$. Опираясь на свойства (4), (5) элементов в построенном примере множества $\mathcal{Y}$, находим следующую оценку для первого слагаемого целевой функции задачи 1A:

$$\frac{1}{2|\mathcal{C}|}\sum_{y_i\in\mathcal{C}}\sum_{y_j\in\mathcal{C}}\|y_i - y_j\|^2 \ge$$

$$\ge \frac{1}{2M}((M-1)(M-2)(2\Delta-2) + 2(M-1)\Delta) >$$

$$> B - (n - M + 1)\Delta + 2. \qquad (6)$$

Далее мы доказываем, что в случае $y_N \in \mathcal{C}$, $x = y_N$ для второго слагаемого целевой функции задачи 1A справедлива оценка

$$\sum_{y_i\in\mathcal{Y}\setminus\mathcal{C}}\|y_i - x\|^2 \ge (n - M + 1)\Delta. \qquad (7)$$

Объединяя (6) и (7), получим, что если $y_N \in \mathcal{C}$, $x = y_N$, то для целевой функции задачи 1A выполнено

$$f(\mathcal{C}, x) \ge B + 2 > B,$$

что противоречит условию $f(\mathcal{C}, x) \le B$.

Затем мы доказываем, что в случае $y_N \in \mathcal{C}$, $x \ne y_N$ для второго слагаемого целевой функции задачи 1A справедлива оценка

$$\sum_{y_i\in\mathcal{Y}\setminus\mathcal{C}}\|y_i - x\|^2 \ge (n - M)(2\Delta - 2). \qquad (8)$$

Объединяя (6) и (8), получим оценку

$$f(\mathcal{C}, x) \ge B + (n - M - 1)(\Delta - 2) > B,$$

которая также противоречит условию $f(\mathcal{C}, x) \le B$. Таким образом, $y_N \in \mathcal{Y} \setminus \mathcal{C}$.

Пусть в задаче Clique подмножество $V_C \subseteq V$ содержит k пар несмежных вершин. Тогда для кластера $\mathcal{C} \subset \mathcal{Y}$ из (4) и (5) для целевой функции задачи 1A имеем

$$f(\mathcal{C}, x) \ge \frac{1}{2M}(M(M-1)(2\Delta-2) + 4k) +$$

$$+ (n - M)\Delta = B + \frac{2k}{M},$$

что при $k > 0$ противоречит сделанному предположению $f(\mathcal{C}, x) \le B$. Следовательно, $k = 0$, а это значит, что множество V_C образует клику. Иными словами, если в построенном примере задачи 1A существуют некоторые кластер $\mathcal{C} \subset \mathcal{Y}$ размера M и точка $x \in \mathcal{Y}$ такие, что $f(\mathcal{C}, x) \le B$, то и в задаче Clique существует клика размера $K = M$.

Таким образом, из п. I и п. II следует, что в построенном примере задачи 1A кластер $\mathcal{C}$ размера M и точка x, удовлетворяющие условию (1), существуют тогда и только тогда, когда в задаче Clique существует клика размера $K = M$.

Остаётся заметить, что поскольку координаты точек y_i, а также числа B и M в построенном сведении ограничены полиномом от размера графа, задача 1A NP-полна в сильном смысле.

Из теоремы 1 следует, что задача 1 NP-трудна в сильном смысле.

ЗАКЛЮЧЕНИЕ

В работе доказана сильная NP-трудность ранее неисследованной квадратичной задачи 2-кластеризации конечного множества точек евклидова пространства.

В математическом плане значительный интерес представляет вопрос о статусе сложности варианта задачи 1, в котором мощности кластеров оптимизируются вместе с искомыми кластерами. Выяснение этого открытого вопроса представляется предметом будущих исследований.

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 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. *Kariv O.*, *Hakimi S.L.* An Algorithmic Approach to Network Location Problems. Pt. II: The *p*-Medians // SIAM J. Appl. Math. 1979. V. 37. P. 513–538.
3. *Гимади Э.Х.*, *Кельманов А.В.*, *Кельманова М.А.*, *Хамидуллин С.А.* Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. журн. индустр. математики. 2006. Т. 9. № 1 (25). С. 55–74.
4. *Gimadi E.Kh.*, *Kel'manov A.V.*, *Kel'manova M.A.*, *Khamidullin S.A.* A Posteriori Detecting a Quasiperiodic Fragment in a Numerical Sequence // Pattern Recognition and Image Analysis. 2008. V. 18. № 1. P. 30–42.
5. *Бабурин А.Е.*, *Гимади Э.Х.*, *Глебов Н.И.*, *Пяткин А.В.* Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. операций. Сер. 2. 2007. Т. 14. № 1. С. 32–42.

6. *James G.*, *Witten D.*, *Hastie T.*, *Tibshirani R.* An Introduction to Statistical Learning. N.Y.: Springer Science+Business Media, LLC, 2013. 426 p.
7. *Shirkhorshidi A.S.*, *Aghabozorgi S.*, *Wah T.Y.*, *Herawan T.* Big Data Clustering: A Review // LNCS. 2014. V. 8583. P. 707–720.
8. *Bishop C.M.* Pattern Recognition and Machine Learning. N.Y.: Springer Science+Business Media, LLC, 2006. 738 p.
9. *Aggarwal C.C.* Data Mining: The Textbook. Springer International Publishing, 2015. 734 p.
10. *Edwards A.W.F.*, *Cavalli-Sforza L.L.* A Method for Cluster Analysis // Biometrics. 1965. V. 21. P. 362–375.
11. *Garey M.R.*, *Johnson D.S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman, 1979. 338 p.
12. *Papadimitriou C.H.* Computational complexity. N.Y.: Addison-Wesley, 1994. 523 p.

NP-HARDNESS OF QUADRATIC EUCLIDEAN 1-MEAN AND 1-MEDIAN 2-CLUSTERING PROBLEM WITH THE CONSTRAINTS ON THE CLUSTER SIZES

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 September 6, 2019

Received September 6, 2019

In the paper, we consider a problem of clustering a finite set of N points in d-dimensional Euclidean space into two clusters minimizing the sum over all clusters of the intracluster sums of the distances between clusters elements and their centers. The center of one cluster is defined as centroid (geometric center). The center of the other one is a sought point in the input set. We analyze the variant of the problem with the given clusters sizes. We have proved the strong NP-hardness of this problem.

Keywords: Euclidean space, clustering, 2-partition, quadratic variation, center, centroid, median, strong NP-hardness.