ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 25 № 4 2019

УДК 519.16+519.85

КВАДРАТИЧНАЯ ЕВКЛИДОВА ЗАДАЧА 2-КЛАСТЕРИЗАЦИИ 1-MEAN И 1-MEDIAN С ОГРАНИЧЕНИЕМ НА РАЗМЕРЫ КЛАСТЕРОВ: СЛОЖНОСТЬ И АППРОКСИМИРУЕМОСТЬ[1]

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

В работе рассматривается задача кластеризации 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. Установлено, что для нее не существует полностью полиномиальной аппроксимационной схемы, если P$\neq$ NP.

Ключевые слова: евклидово пространство, кластеризация, 2-разбиение, квадратичный разброс, центр, центроид, медиана, сильная NP-трудность, несуществование полностью полиномиальной приближенной схемы, эффективная сводимость с сохранением точности.

A. V. Kel'manov, A. V. Pyatkin, V. I. Khandeev. Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with constraints on the size of the clusters: Complexity and approximability.

We consider the problem of partitioning a set of N points in d-dimensional Euclidean space into two clusters minimizing the sum of the squared distances between each element and the center of the cluster to which it belongs. The center of the first cluster is its centroid (the geometric center). The center of the second cluster should be chosen among the points of the input set. We analyze the variant of the problem with given sizes (cardinalities) of the clusters; the sum of the sizes equals the cardinality of the input set. We prove that the problem is strongly NP-hard and there is no fully polynomial-time approximation scheme for its solution.

Keywords: Euclidean space, clustering, 2-partition, quadratic variation, center, centroid, median, strong NP-hardness, nonexistence of FPTAS, approximation-preserving reduction.

MSC: 68W25, 68Q25

DOI: 10.21538/0134-4889-2019-25-4-69-78

Введение

Предметом исследования в этой работе является экстремальная задача разбиения конечного множества точек евклидова пространства на два кластера. Цель — анализ вычислительной сложности задачи и выяснение некоторых вопросов ее аппроксимируемости.

Исследование мотивировано неизученностью задачи в математическом плане. А именно, до настоящего времени статус ее вычислительной сложности не был установлен. К тому же вопросы ее алгоритмической аппроксимируемости не были выяснены. Кроме того, задача актуальна не только для дискретной оптимизации, но также, в частности, для компьютерной геометрии и математической статистики. С практической точки зрения рассматриваемая задача важна, например, для решения известной междисциплинарной проблемы интерпретации данных (Data mining).

[1]Работа выполнена при финансовой поддержке РФФИ, проекты 19-01-00308 и 18-31-00398, программы ФНИ РАН, проекты 0314-2019-0014, 0314-2019-0015, а также программы Top-5-100 Министерства образования и науки РФ.

Работа имеет следующую структуру. В разд. 1 даны формулировки названной задачи и нескольких задач, которые в математическом плане близки к ней. В этом же разделе приведены некоторые трактовки задачи и замечания, поясняющие мотивацию настоящего исследования. В следующем разделе анализируется вычислительная сложность задачи. В разд. 3 рассматривается вопрос о существовании полностью полиномиальной аппроксимационной схемы (FPTAS), а также вопросы аппроксимируемости задачи алгоритмами для решения известной задачи.

1. Формулировка задачи и близкие по постановке задачи

Всюду далее $\|\cdot\|$ — евклидова норма.

Рассматриваемая задача имеет следующую формулировку.

З а д а ч а 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

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

где

$$\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y \text{ — центроид (геометрический центр) кластера } \mathcal{C}.$$

Ниже приведены 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. Отличие целевых функций в этих задачах определяется (см. далее) центрами квадратичного разброса точек искомых кластеров. В задаче 1 два оптимизируемых (неизвестных) центра разброса. Один из них — точка $\overline{y}(\mathcal{C})$ в $\mathbb{R}^d$, а другой — точка x в $\mathcal{Y}$.

Следующие три задачи 2-кластеризации относятся к числу известных и близких по постановке к задаче 1.

З а д а ч а 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

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

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

В этой задаче центрами квадратичного разброса точек являются центроиды (в $\mathbb{R}^d$) искомых кластеров. Сильная NP-трудность задачи следует из сильной NP-трудности (см. [1]) задачи 2-*Means* (без ограничений на мощности кластеров). Действительно, полиномиальная разрешимость 2-*Means with the Constraints on the Cluster Sizes* влекла бы полиномиальную разрешимость задачи 2-*Means*. Достаточно было бы перебрать за полиномиальное время конечное число точных допустимых решений задачи для каждого M.

З а д а ч а 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 $\sum\limits_{y\in\mathcal{C}}\|y-x\|+\sum\limits_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-z\|$. Легко видеть, что задача 3 разрешима за время $\mathcal{O}(dN^3)$. Достаточно перебрать N^2 пар точек x, z, 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.

З а д а ч а 4 (1-*Mean and Given* 1-*Center with the Constraints on the Cluster Sizes*).

Дано: N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. Найти разбиение $\mathcal{Y}$ на непустые подмножества $\mathcal{C}$ и $\mathcal{Y}\setminus\mathcal{C}$ размеров M и $N-M$ соответственно такие, что

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

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

В этой задаче неизвестен только один центр квадратичного разброса точек. Этот центр $\overline{y}(\mathcal{C})$ является центроидом искомого кластера $\mathcal{C}$. Центр квадратичного разброса точек второго кластера $\mathcal{Y}\setminus\mathcal{C}$ совпадает с началом координат. Заметим, что задача с центром разброса, который задан в некоторой точке $x\in\mathbb{R}^d$, не равной 0, очевидно, полиномиально сводится к этой задаче. Достаточно перенести в точку x начало координат и пересчитать координаты точек входного множества. Сильная NP-трудность этой задачи следует из равенства

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

в котором первый член в правой части не зависит от $\mathcal{C}$, и сильной NP-трудности (см. [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 $\|\sum\limits_{y\in\mathcal{C}}y\|$ суммы элементов из этого подмножества.

Все приведенные экстремальные задачи, включая задачу 1, имеют геометрический характер, который ясен непосредственно из их формулировок. 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.

Заметим, что построение оптимальных разделяющих поверхностей по имеющимся данным — типичная прикладная проблема машинного обучения (Machine learning), распознавания образов (Pattern recognition) и кластеризации данных (Data clustering) (см. [6–8] соответственно). В отмеченных приложених 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 (см. [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, причем априорное соответствие данных распределениям неизвестно. Из-за отсутствия этого соответствия создаются математические инструменты в виде эффективных алгоритмов решения необозримого множества задач разбиения данных с самой разнообразной структурой на однород-

ные по какому-либо фиксированному критерию кластеры, а также инструменты 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. Приведенные замечания поясняют мотивацию настоящей статьи. Фактически наша работа отвечает на вопрос, можно ли эффективно (за полиномиальное) время разбить имеющиеся данные в соответствии с (1).

В заключение этого раздела подчеркнем, что рассматриваемая задача 1 не эквивалентна ни одной из приведенных 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

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

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.

З а д а ч а 1A. 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

$$f(\mathcal{C}, x) = \frac{1}{2|\mathcal{C}|} \sum_{y\in\mathcal{C}} \sum_{z\in\mathcal{C}} \|y - z\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y - x\|^2 \leq B\,? \tag{3}$$

Напомним следующую NP-полную задачу (см. [11]).

З а д а ч а 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 (см. [12]).

Справедлива следующая теорема.

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

Д о к а з а т е л ь с т в о. Для доказательства теоремы построим полиномиальное сведение задачи *Clique* в однородном графе к задаче 1A.

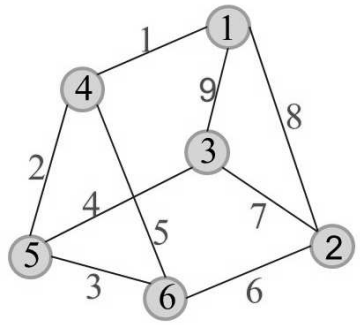
Рассмотрим однородный граф $G = (V, E)$ степени Δ на n вершинах. Будем считать, что $\Delta > 2$ и $n - K > 1$, так как в противном случае задача *Clique*, очевидно, решается за полиномиальное время.

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

$$d = |E|, \quad N = n + 1, \quad M = K, \quad B = (n-1)\Delta - K + 1, \quad y_N = 0; \quad y_i = a_i, \quad i = 1, \ldots, n, \tag{4}$$

где a_i — i-я строка матрицы $A = \{a_{i,j}\}$ $(i = 1, \ldots, n,\ j = 1, \ldots, d)$ — 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.

Для пояснения ниже приведен пример 3-регулярного графа G на шести вершинах с девятью ребрами и матрица A для этого графа:



3-регулярный граф на шести вершинах.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Матрица A для графа слева.

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

$$\|y_i - y_j\|^2 = \begin{cases} 2\Delta - 2, \text{ если ребро } v_i v_j \in E(G), \\ 2\Delta \text{ в противном случае,} \end{cases} \qquad 1 \le i < j \le n. \tag{5}$$

$$\|y_i - y_N\|^2 = \Delta, \qquad i = 1, \ldots, n. \tag{6}$$

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 1A в качестве центра кластера $\mathcal{Y} \setminus \mathcal{C}$ выберем точку $x = y_N$. При этом выборе из (3)–(6) для целевой функции задачи 1A 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 (3) выполнено в виде равенства, т. е. в примере задачи 1A 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 $\mathcal{C} \subset \mathcal{Y}$ размера $M = K$ и точка $x \in \mathcal{Y}$ такие, что $f(\mathcal{C}, x) \le B$.

Сначала покажем от противного, что в этом случае $y_N \notin \mathcal{C}$. Опираясь на свойства (5), (6) элементов в построенном примере множества $\mathcal{Y}$, найдем следующую оценку для первого слагаемого целевой функции задачи 1A

$$\frac{1}{2|\mathcal{C}|} \sum_{y_i \in \mathcal{C}} \sum_{y_j \in \mathcal{C}} \|y_i - y_j\|^2 \ge \frac{1}{2M}\big((M-1)(M-2)(2\Delta - 2) + 2(M-1)\Delta\big)$$
$$= \frac{\Delta - 2}{M} + M\Delta - M - 2\Delta + 3 = \frac{\Delta - 2}{M} + B + M\Delta - n\Delta - \Delta + 2$$
$$= \frac{\Delta - 2}{M} + B - (n - M + 1)\Delta + 2 > B - (n - M + 1)\Delta + 2. \tag{7}$$

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

$$\sum_{y_i \in \mathcal{Y} \setminus \mathcal{C}} \|y_i - x\|^2 \geq (n - M + 1)\Delta. \tag{8}$$

Объединяя (7) и (8), получим, что если $y_N \in \mathcal{C}$, $x = y_N$, то для целевой функции задачи 1A выполнено

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

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

В случае $y_N \in \mathcal{C}$, $x \neq y_N$ для второго слагаемого целевой функции задачи 1A имеем оценку

$$\sum_{y_i \in \mathcal{Y} \setminus \mathcal{C}} \|y_i - x\|^2 \geq (n - M)(2\Delta - 2). \tag{9}$$

Объединяя (7) и (9), получим оценку

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

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

Наконец, допустим, что в задаче *Clique* подмножество $V_C \subseteq V$ содержит k пар несмежных вершин. Тогда для кластера $\mathcal{C} \subset \mathcal{Y}$ из (5) и (6) для целевой функции задачи 1A имеем

$$f(\mathcal{C}, x) \geq \frac{1}{2M}\big(M(M-1)(2\Delta - 2) + 4k\big) + (n - M)\Delta = B + \frac{2k}{M},$$

что при $k > 0$ противоречит сделанному предположению $f(\mathcal{C}, x) \leq B$. 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 $f(\mathcal{C}, x) \leq B$, 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 $K = M$.

Остается заметить, что поскольку координаты точек y_i, а также числа B и M в построенном сведении ограничены полиномом от размера графа, в соответствии с [11] задача 1A NP-полна в сильном смысле.

Теорема 1 доказана.

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

3. Алгоритмическая аппроксимируемость

Мы показали, что задача 1 NP-трудна в сильном смысле. Из этого результата, как известно, следует, что для этой задачи не существует точного псевдополиномиального алгоритма, если только классы P и NP не совпадают. Однако эта задача имеет числовые входы. Поэтому важный вопрос о существовании для нее полностью полиномиальной аппроксимационной схемы (FPTAS) требует дополнительного анализа (в соответствии, например, с [11; 13]). На этот вопрос отвечает следующая теорема.

Теорема 2. *Если P≠NP, то для задачи* 1 *не существует схемы* FPTAS.

Доказательство. Согласно [11; 13] для доказательства нам достаточно показать справедливость двух условий: (a) для целочисленных входных данных значение целевой функции задачи 1 целочисленно, (b) это значение ограничено полиномом от входных целочисленных значений.

Умножив обе части (1) на $2|\mathcal{C}|$ и применив тождество (2), получим

$$2|\mathcal{C}|f(\mathcal{C},x)=\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2+2|\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-x\|^2. \tag{10}$$

Рассмотрим задачу поиска подмножества $\mathcal{C}\subset\mathcal{Y}$ размера M и точки $x\in\mathcal{Y}$, для которых значение правой части (10) минимально. Поскольку $|\mathcal{C}|=M$, эта задача эквивалентна задаче 1, а значит в силу теоремы 1 является NP-трудной в сильном смысле.

Далее, легко видеть, что значение целевой функции этой задачи целочисленно при целочисленных входных данных, т. е. условие (a) выполнено. Кроме того, очевидно, что в этой задаче значение целевой функции ограничено полиномом от максимального (по всем точкам входного множества и их координатам) абсолютного значения целочисленной координаты, т. е. условие (b) также выполнено. Поэтому для этой задачи не существует схемы FPTAS, если P≠NP. Следовательно, если P≠NP, то и для эквивалентной задачи 1 также не существует схемы FPTAS.

Теорема 2 доказана.

Вопрос об аппроксимируемости задачи другими видами алгоритмов раскрывает следующее утверждение.

Утверждение. *Задача* 1 *эффективно аппроксимируема алгоритмами решения задачи* 4 *с той же точностью и вероятностью несрабатывания.*

Д о к а з а т е л ь с т в о. Действительно, очевидно, что эффективные алгоритмы решения задачи 4 легко переносятся на решение задачи 1. А именно, для этого достаточно:
1) перебрать N кандидатов, т. е. точек множества $\mathcal{Y}$, на роль центра разброса точек в кластере $\mathcal{Y}\setminus\mathcal{C}$; 2) перенести начало координат в перебираемые точки-кандидаты и получить таким образом входы задачи 4; 3) найти приближенное решение задачи 4 с помощью какого-либо из существующих эффективных алгоритмов; 4) в семействе допустимых решений-кандидатов, полученных путем последовательного выполнения отмеченных пп. 1)–3), найти наилучшее в смысле наименьшего значения целевой функции задачи 1.

Утверждение доказано.

Это простое утверждение позволяет применять существующие алгоритмические результаты для известной задачи 4 к рассмотренной в настоящей работе задаче 1. Среди этих результатов отметим, в частности, 2-приближенный полиномиальный алгоритм [14], полиномиальную аппроксимационную схему (PTAS) [15], рандомизированный алгоритм [16] и схему PTAS для случая, когда размерность пространства является медленно растущей функцией от размера входного множества [17].

Заключение

В работе доказана сильная NP-трудность ранее не исследованной квадратичной задачи 2-кластеризации конечного множества точек евклидова пространства. Установлено, что для этой задачи не существует схемы FPTAS, если P≠NP. Получены ответы на некоторые актуальные вопросы об алгоритмической аппроксимируемости задачи. Продолжение исследований этих вопросов — дело ближайшей перспективы.

Ясно, что по точному (экспоненциальному) или эффективному приближенному решению задачи 1 можно найти соответствующее решение варианта задачи с оптимизируемыми размерами кластеров. Для этого достаточно перебрать N соответствующих (точных или приближенных) решений в семействе задач с заданными размерами кластеров и выбрать наилучшее в этом семействе.

Тем не менее в математическом плане значительный интерес представляет вопрос о статусе сложности и аппроксимируемости варианта задачи 1, в котором мощности кластеров оптимизируются вместе с искомыми кластерами. Выяснение этого открытого вопроса — предмет будущих исследований.

СПИСОК ЛИТЕРАТУРЫ

1. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Machine Learning. 2009. Vol. 75, no. 2. P. 245–248. doi: 10.1007/s10994-009-5103-0 .
2. **Kariv O., Hakimi S.L.** An algorithmic approach to network location problems. Pt. II: The p-Medians // SIAM J. Appl. Math. 1979. Vol. 37, no. 3. P. 513–538. doi: 10.1137/0137041 .
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 Anal. 2008. Vol. 18, no. 1. P. 30–42. doi: 10.1134/S1054661808010057 .
5. **Бабурин А.Е., Гимади Э.Х., Глебов Н.И., Пяткин А.В.** Задача отыскания подмножества векторов с максимальным суммарным весом // Дискретный анализ и исследование 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. **Bishop C.M.** Pattern recognition and machine learning. N Y: Springer, Science+Business Media, LLC, 2006. 738 p.
8. **Shirkhorshidi A.S., Aghabozorgi S, Wah T.Y., Herawan T.** Big data clustering: A review // Computational Science and Its Applications (ICCSA 2014): Proc. / eds. B. Murgante et al. 2014. P. 707–720. (Lecture Notes in Computer Science; vol. 8583). doi: 10.1007/978-3-319-09156-3_49 .
9. **Aggarwal C.C.** Data mining: The Textbook. N Y etc.: Springer, International Publishing, 2015. 734 p.
10. **Edwards A.W.F., Cavalli-Sforza L.L.** A method for cluster analysis // Biometrics. 1965. Vol. 21. P. 362–375. doi: 10.2307/2528096 .
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.
13. **Vazirani V.V.** Approximation algorithms. Berlin; Heidelberg; N Y: Springer-Verlag, 2001. 380 p.
14. **Долгушев А.В., Кельманов А.В.** Приближенный алгоритм решения одной задачи кластерного анализа // Дискретный анализ и исследование операций. 2011. Т. 18, №2. С. 29–40.
15. **Долгушев А.В., Кельманов А.В., Шенмайер В.В.** Приближенная полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: 9-я междунар. конф. (Респ. Черногория, г. Будва, 16–22 сентября 2012 г.): сб. докл. М.: Торус Пресс, 2012. С. 242–244.
16. **Кельманов А.В., Хандеев В.И.** Рандомизированный алгоритм для одной задачи двухкластерного разбиения множества векторов // Журн. вычисл. математики и мат. физики. 2015. Т. 55, №2. С. 335–344.
17. **Kel'manov A.V., Motkova A.V., Shenmaier V.V.** An approximation scheme for a weighted two-cluster partition problem // Analysis of Images, Social Networks and Texts - 6th Internat. Conf. (AIST 2017): Revised Selected Papers. 2018. P. 323–333. (Lecture Notes in Computer Science; vol. 10716.) doi: 10.1007/978-3-319-73013-4_30 .

Поступила 12.08.2019
После доработки 10.09.2019
Принята к публикации 16.09.2019

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

Пяткин Артем Валерьевич
д-р физ.-мат. наук,

зав. лабораторией
Институт математики им. С. Л. Соболева СО РАН;
Новосибирский государственный университет
г. Новосибирск
e-mail: artem@math.nsc.ru

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

REFERENCES

1. Aloise D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 2009, vol. 75, no. 2, pp. 245–248. doi: 10.1007/s10994-009-5103-0 .
2. Kariv O., Hakimi S. An algorithmic approach to network location problems. Part II: The p-Medians. *SIAM J. Appl. Math.*, 1979, vol. 37, no. 3, pp. 539–560. doi: 10.1137/0137041 .
3. Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A. A posteriori detection of a quasiperiodic fragment with a given number of repetitions in a numerical sequence. *Sib. Zh. Ind. Mat.*, 2006, vol. 9, no. 1, pp. 55–74 (in Russian).
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, vol. 18, no. 1, pp. 30–42. doi: 10.1134/S1054661808010057 .
5. Baburin A.E., Gimadi E.Kh., Glebov, N.I., Pyatkin, A.V. The problem of finding a subset of vectors with the maximum total weight. *J. Appl. Industr. Math.*, 2008, vol. 2, no. 1, pp. 32–38. doi: 10.1007/s11754-008-1004-3 .
6. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning*. N Y: Springer Science+Business Media, LLC, 2013, 426 p. ISBN: 978-1461471370 .
7. Bishop C.M. *Pattern Recognition and Machine Learning*. N Y: Springer Science+Business Media, LLC, 2006, 738 p. ISBN: 978-0-387-31073-2 .
8. Shirkhorshidi A.S., Aghabozorgi S, Wah T,Y., and Herawan T. Big data clustering: A review. In: Murgante B. et al. (eds), Computational Science and Its Applications (ICCSA 2014), *Lecture Notes in Computer Science*, 2014, vol. 8583, pp. 707–720. doi: 10.1007/978-3-319-09156-3_49 .
9. Aggarwal C.C. *Data mining: The textbook*. Cham: Springer, 2015, 734 p. doi: 10.1007/978-3-319-14142-8 .
10. Edwards A.W.F., Cavalli-Sforza L.L. A method for cluster analysis. *Biometrics*, 1965, vol. 21, pp. 362–375. doi: 10.2307/2528096 .
11. Garey M.R., Johnson D.S. *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco: Freeman, 1979, 338 p. ISBN: 0716710447 .
12. Papadimitriou C.H. *Computational complexity*. N Y: Addison-Wesley, 1994, 523 p. ISBN: 0-201-53082-1 .
13. Vazirani V.V. *Approximation Algorithms*. Berlin; Heidelberg; N Y: Springer-Verlag, 2003, 380 p. doi: 10.1007/978-3-662-04565-7 .
14. Dolgushev A.V., Kel'manov A.V. An approximation algorithm for solving a problem of cluster analysis. *J. Appl. Indust. Math.*, 2011, vol. 5, no. 4, pp. 551–558. doi: 10.1134/S1990478911040107 .
15. Dolgushev A.V., Kel'manov A.V., Shenmaier V.V. A polynomial-time approximation scheme for one problem of cluster analysis In: K. V. Vorontsov (ed.) Intelligent Data Processing: Proc. of the 9th Internat. Conf. (Republic of Montenegro, Budva, September 16–22, 2012), Moscow: Torus Press, 2012, pp. 242–244 (in Russian).
16. Kel'manov A.V., Khandeev V.I. A Randomized algorithm for two-cluster partition of a set of vectors. *Comput. Math. Math. Phys.*, 2015, vol. 55, no. 2, pp. 330–339. doi: 10.1134/S096554251502013X .

17. Kel'manov A.V., Motkova A.V., Shenmaier V.V. An approximation scheme for a weighted two-cluster partition problem. Analysis of Images, Social Networks and Texts - 6th Internat. Conf. (AIST 2017), Revised Selected Papers, *Lecture Notes in Computer Science*, 2018. Vol. 10716. P. 323–333. doi: 10.1007/978-3-319-73013-4_30 .

Received August 12, 2019
Revised September 10, 2019
Accepted September 16, 2019

Funding Agency: This work was supported by the Russian Foundation for Basic Research (project nos. 19-01-00308 and 18-31-00398), by Program I.5.1 for Fundamental Research of the Siberian Branch of the Russian Academy of Sciences (project nos. 0314-2019-0014 and 0314-2019-0015), and by the Ministry of Education and Science of the Russian Federation within the Russian Academic Excellence Project.

Alexander Vasil'evich Kel'manov, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: kelm@math.nsc.ru .

Artem Valer'evich Pyatkin, Dr. Phys.-Math. Sci., Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: artem@math.nsc.ru .

Vladimir Il'ich Khandeev, Cand. Sci. (Phys.-Math.), Sobolev Institute of Mathematics; Novosibirsk State University, Novosibirsk, 630990 Russia, e-mail: khandeev@math.nsc.ru .

Cite this article as: A. V. Kel'manov, A. V. Pyatkin, V. I. Khandeev. Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with constraints on the size of the clusters: Complexity and approximability, *Trudy Instituta Matematiki i Mekhaniki URO RAN*, 2019, vol. 25, no. 4, pp. 69–78 .