

УДК 519.72

ПОЛИНОМИАЛЬНАЯ РАЗРЕШИМОСТЬ ОДНОМЕРНОГО СЛУЧАЯ ОДНОЙ NP-ТРУДНОЙ ЗАДАЧИ КЛАСТЕРИЗАЦИИ¹⁾

© 2019 г. А. В. Кельманов^{1,2,*}, В. И. Хандеев^{1,2,**}¹⁾ 630090 Новосибирск, пр-т акад. Коптюга, 4, Ин-т матем. им. С.Л. Соболева, Россия;²⁾ 630090 Новосибирск, ул. Пирогова, 2, Новосибирский гос. ун-т, Россия)

*e-mail: kelm@math.nsc.ru;

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

Поступила в редакцию 03.04.2019 г.
Переработанный вариант 03.04.2019 г.
Принята к публикации 15.05.2019 г.

Рассматривается задача разбиения конечного множества точек евклидова пространства на кластеры по критерию минимума суммы по всем кластерам внутрикластерных сумм квадратов расстояний между элементами кластеров и их центрами. Центры одной части кластеров заданы на входе задачи, а центры другой части кластеров неизвестны и определяются как центроиды (геометрические центры). Известно, что в общем случае эта задача NP-трудна в сильном смысле. В работе конструктивно доказано, что одномерный случай задачи разрешим за полиномиальное время. Библ. 20.

Ключевые слова: евклидово пространство, кластеризация, разбиение, минимум суммы квадратов расстояний, NP-трудная в сильном смысле задача, одномерный случай, полиномиальная разрешимость.

DOI: 10.1134/S0044466919090114

ВВЕДЕНИЕ

Предметом исследования настоящей работы является одна из труднорешаемых задач разбиения конечного множества точек евклидова пространства. Цель работы — анализ статуса вычислительной сложности одномерного случая задачи. Исследование мотивировано, с одной стороны, значимостью для дискретной оптимизации и вычислительной математики вопроса о разрешимости одномерного (специального) случая рассматриваемой труднорешаемой задачи, а с другой — важностью этой задачи для приложений, в частности, для статистики, анализа данных (Data analysis), интерпретации данных (Data mining), обработки больших данных (Big data processing).

1. ФОРМУЛИРОВКА ЗАДАЧИ, ЕЕ ИСТОКИ И СВЯЗАННЫЕ С НЕЙ ПРОБЛЕМЫ

Рассматриваемая задача имеет следующую формулировку.

Задача 1. Дано: N -элементное множество $\mathcal{U}$ точек в евклидовом пространстве размерности d , натуральное число K и набор $\{c_1, \dots, c_J\}$ точек. Найти: разбиение множества $\mathcal{U}$ на $K + J$ непустых кластеров $\mathcal{C}_1, \dots, \mathcal{C}_K, \mathcal{D}_1, \dots, \mathcal{D}_J$ такое, что

$$F(\mathcal{C}_1, \dots, \mathcal{C}_K; \mathcal{D}_1, \dots, \mathcal{D}_J) = \sum_{k=1}^K \sum_{y \in \mathcal{C}_k} \|y - \bar{y}(\mathcal{C}_k)\|^2 + \sum_{j=1}^J \sum_{y \in \mathcal{D}_j} \|y - c_j\|^2 \rightarrow \min,$$

где

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

есть центроид k -го кластера.

¹⁾ Разделы 2 и 3 выполнены при финансовой поддержке РФФИ, проекты 19-01-00308 и 18-31-00398, остальные разделы — при поддержке программы ФНИ РАН, проект 0314-2019-0015.

Задачу 1 можно рассматривать как модификацию известной задачи *K-Means* (*K-Средних*). В этой задаче дано N -элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и натуральное число K . Требуется найти разбиение входного множества $\mathcal{Y}$ на непустые непересекающиеся кластеры $\mathcal{C}_1, \dots, \mathcal{C}_K$, минимизирующее сумму

$$\sum_{k=1}^K \sum_{y \in \mathcal{C}_k} \|y - \bar{y}(\mathcal{C}_k)\|^2,$$

где $\bar{y}(\mathcal{C}_k)$ — центроид k -го кластера.

Другое распространенное название задачи *K-Means* — *MSSS* (*Minimum Sum-of-Squares Clustering*), т.е. кластеризация по критерию минимума суммы квадратов. В статистике эта задача известна с прошлого века и связана с именем Фишера (см., например, [1], [2]). На практике (в самых разнообразных приложениях) она возникает в ситуациях, когда имеется гипотеза о том, что имеющаяся совокупность $\mathcal{Y}$ выборочных (входных) данных содержит K однородных кластеров (подмножеств) $\mathcal{C}_1, \dots, \mathcal{C}_K$, в которых точки разбросаны относительно соответствующих неизвестных средних значений $\bar{y}(\mathcal{C}_1), \dots, \bar{y}(\mathcal{C}_K)$. Однако соответствие данных и кластеров неизвестно. Очевидно, что в этой ситуации для корректного применения классических статистических методов (проверки гипотез и оценивания) к обработке выборочных данных требуется сначала разбить их на однородные группы (кластеры). Эта ситуация типична, в частности, для отмеченных выше (во введении) прикладных проблем, связанных с обработкой данных.

Сильная NP-трудность задачи *K-Means* доказана относительно недавно [3]. Однако полиномиальная разрешимость этой задачи на числовой прямой установлена еще в прошлом веке в [4]. В этой работе представлен полиномиальный алгоритм с временем работы $\mathcal{O}(KN^2)$, реализующий схему динамического программирования. Этот алгоритм опирается на хорошо известный точный полиномиальный алгоритм решения задачи о ближайшем соседе (*Nearest neighbor search problem*, NNSP) [5]. Заметим, что полиномиальная разрешимость одномерного случая задачи *K-Means* за время $\mathcal{O}(KN \log N)$ следует из более ранних (см. [4]) работ отечественных авторов [6]–[9], где представлены ускоренные эффективные алгоритмы для ряда специальных случаев задачи NNSP. Тем не менее в последующие годы для одномерного случая задачи *K-Means* были предложены новые алгоритмы с временем работы $\mathcal{O}(KN \log N)$. Обзор и свойства этих алгоритмов можно найти в [10], [11].

В отличие от задачи *K-Means*, Задача 1 моделирует прикладную проблему, в которой у части кластеров, а именно, у $\mathcal{D}_1, \dots, \mathcal{D}_J$, соответствующие центры $c_1, \dots, c_J$ квадратичного разброса данных известны заранее (заданы). Введенные обозначения позволяют назвать Задачу 1 как *K-Means and Given J-Centers*. Эта прикладная проблема также типична для анализа и интерпретации данных, распознавания образов и машинного обучения. В частности, двухкластерный вариант задачи — *1-Mean and Given 1-Center* — связан с решением прикладной проблемы обработки сигналов, а именно, с проблемой совместного обнаружения квазипериодически повторяющегося импульса неизвестной формы и оценивания этой формы в условиях гауссовского шума с нулевым средним [12]–[14]. В этом двухкластерном варианте задачи нулевое среднее соответствует кластеру с заданным в начале координат центром. Первое сообщение об этом двухкластерном варианте Задачи 1, по-видимому, было сделано в [12]. Заметим, что более простые экстремальные задачи, которые индуцируются прикладными проблемами помехоустойчивого обнаружения и различения импульсов заданных форм, характерны, в частности, для радиолокации, электронной разведки, гидроакустики, геофизики, технической и медицинской диагностики, космического мониторинга (см., например, [15]–[17]).

Сильная NP-трудность Задачи 1 установлена в [18]–[20]. Заметим, что задача *K-Means* не эквивалентна Задаче 1 и не является ее частным случаем. Поэтому вопрос о разрешимости Задачи 1 на числовой прямой требует самостоятельного изучения. Этот вопрос до последнего времени оставался открытым. В настоящей работе мы доказываем, что в одномерном случае Задача 1 разрешима за полиномиальное время.

2. СВОЙСТВА ОПТИМАЛЬНОГО РЕШЕНИЯ В ОДНОМЕРНОМ СЛУЧАЕ

Всюду далее будем считать, что $d = 1$. Одномерный случай Задачи 1 будем называть Задача 1D.

Наше доказательство опирается на несколько приведенных ниже вспомогательных утверждений, которые раскрывают структуру оптимального решения Задачи 1D.

Обозначим через $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ оптимальные кластеры в Задаче 1D.

Лемма 1. Если в Задаче 1D $c_m < c_\ell$, где $1 \leq m \leq J, 1 \leq \ell \leq J$, то для любых $x \in \mathcal{D}_m^*$ и $z \in \mathcal{D}_\ell^*$ выполнено неравенство $x \leq z$.

Доказательство. Пусть существуют $x \in \mathcal{D}_m^*$ и $z \in \mathcal{D}_\ell^*$ такие, что $x > z$. Тогда для точек x, z и c_m, c_ℓ имеем

$$\begin{aligned} & (x - c_m)^2 + (z - c_\ell)^2 - ((z - c_m)^2 + (x - c_\ell)^2) = \\ & = ((x - z) + (z - c_m))^2 + (z - c_\ell)^2 - ((z - c_m)^2 + ((x - z) + (z - c_\ell))^2) = \\ & = 2(x - z)(x - c_m) - 2(x - z)(x - c_\ell) = 2(x - z)(c_\ell - c_m) > 0. \end{aligned} \quad (2.1)$$

Положим $\mathcal{D}'_m = \mathcal{D}_m^* \setminus \{x\} \cup \{z\}$ и $\mathcal{D}'_\ell = \mathcal{D}_\ell^* \setminus \{z\} \cup \{x\}$. Используя (2.1), получаем оценку

$$\begin{aligned} & \sum_{y \in \mathcal{D}_m^*} (y - c_m)^2 + \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 = \sum_{y \in \mathcal{D}_m^* \setminus \{x\}} (y - c_m)^2 + \sum_{y \in \mathcal{D}_\ell^* \setminus \{z\}} (y - c_\ell)^2 + (x - c_m)^2 + (z - c_\ell)^2 > \\ & > \sum_{y \in \mathcal{D}_m^* \setminus \{x\}} (y - c_m)^2 + \sum_{y \in \mathcal{D}_\ell^* \setminus \{z\}} (y - c_\ell)^2 + (z - c_m)^2 + (x - c_\ell)^2 = \sum_{y \in \mathcal{D}'_m} (y - c_m)^2 + \sum_{y \in \mathcal{D}'_\ell} (y - c_\ell)^2, \end{aligned}$$

что противоречит предположению об оптимальности $\mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ и, как следствие, предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$. Лемма 1 доказана.

Лемма 2. Если в Задаче 1D $\bar{y}(\mathcal{C}_m^*) < \bar{y}(\mathcal{C}_\ell^*)$, где $1 \leq m \leq K, 1 \leq \ell \leq K$, то для любых $x \in \mathcal{C}_m^*$ и $z \in \mathcal{C}_\ell^*$ справедливо неравенство $x \leq z$.

Доказательство. Как и в доказательстве леммы 1, предположим, что существуют такие $x \in \mathcal{C}_m^*$ и $z \in \mathcal{C}_\ell^*$, для которых $x > z$. Тогда, аналогично доказательству (2.1), для точек x, z и $\bar{y}(\mathcal{C}_m^*), \bar{y}(\mathcal{C}_\ell^*)$ получим неравенство

$$(x - \bar{y}(\mathcal{C}_m^*))^2 + (z - \bar{y}(\mathcal{C}_\ell^*))^2 > (z - \bar{y}(\mathcal{C}_m^*))^2 + (x - \bar{y}(\mathcal{C}_\ell^*))^2. \quad (2.2)$$

Далее, для центроидов $\bar{y}(\mathcal{C}'_m)$ и $\bar{y}(\mathcal{C}'_\ell)$, где $\mathcal{C}'_m = \mathcal{C}_m^* \setminus \{x\} \cup \{z\}$, имеем

$$\begin{aligned} & \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 = \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m) + \bar{y}(\mathcal{C}'_m) - \bar{y}(\mathcal{C}_m^*))^2 = \\ & = \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 + 2 \sum_{y \in \mathcal{C}'_m} ((y - \bar{y}(\mathcal{C}'_m))(\bar{y}(\mathcal{C}'_m) - \bar{y}(\mathcal{C}_m^*)) + \left| \mathcal{C}'_m \right| (\bar{y}(\mathcal{C}'_m) - \bar{y}(\mathcal{C}_m^*))^2 = \\ & = \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 + \left| \mathcal{C}'_m \right| (\bar{y}(\mathcal{C}'_m) - \bar{y}(\mathcal{C}_m^*))^2, \end{aligned} \quad (2.3)$$

так как

$$\sum_{y \in \mathcal{C}'_m} ((y - \bar{y}(\mathcal{C}'_m))(\bar{y}(\mathcal{C}'_m) - \bar{y}(\mathcal{C}_m^*))) = 0,$$

в силу того, что

$$\sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m)) = 0.$$

Поэтому из (2.3) следует оценка

$$\sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}_m^*))^2 \geq \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2. \quad (2.4)$$

Аналогичным образом найдем

$$\sum_{y \in \mathcal{C}'_\ell} (y - \bar{y}(\mathcal{C}_\ell^*))^2 \geq \sum_{y \in \mathcal{C}_\ell'} (y - \bar{y}(\mathcal{C}_\ell'))^2, \quad (2.5)$$

где $\mathcal{C}'_\ell = \mathcal{C}_\ell^* \setminus \{z\} \cup \{x\}$.

Из неравенств (2.2), (2.4) и (2.5) получим следующую оценку:

$$\begin{aligned} \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}_\ell^*} (y - \bar{y}(\mathcal{C}_\ell^*))^2 &> \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}_\ell'} (y - \bar{y}(\mathcal{C}_\ell'))^2 \geq \\ &\geq \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}_\ell'} (y - \bar{y}(\mathcal{C}_\ell'))^2, \end{aligned}$$

которая противоречит предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*$ и, как следствие, — предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$. Лемма 2 доказана.

Лемма 3. Для оптимального решения Задачи 1D справедливы следующие утверждения:

а) если $\bar{y}(\mathcal{C}_m^*) < c_\ell$, где $1 \leq m \leq K, 1 \leq \ell \leq J$, то для любых $x \in \mathcal{C}_m^*$ и $z \in \mathcal{D}_\ell^*$ справедливо неравенство $x \leq z$;

б) если $\bar{y}(\mathcal{C}_m^*) > c_\ell$, где $1 \leq m \leq K, 1 \leq \ell \leq J$, то для любых $x \in \mathcal{C}_m^*$ и $z \in \mathcal{D}_\ell^*$ выполнено неравенство $x \geq z$.

Доказательство. Докажем лемму для случая а). Как и в доказательстве леммы 1, предположим, что существуют такие $x \in \mathcal{C}_m^*$ и $z \in \mathcal{D}_\ell^*$, для которых $x > z$. Тогда, аналогично доказательству (2.1), для точек x, z и $\bar{y}(\mathcal{C}_m^*), c_\ell$ получим

$$(x - \bar{y}(\mathcal{C}_m^*))^2 + (z - c_\ell)^2 > (z - \bar{y}(\mathcal{C}_m^*))^2 + (x - c_\ell)^2.$$

Из этого неравенства и (2.4) следует оценка

$$\begin{aligned} \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 &> \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{D}_\ell'} (y - c_\ell)^2 \geq \\ &\geq \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{D}_\ell'} (y - c_\ell)^2, \end{aligned}$$

где $\mathcal{C}'_m = \mathcal{C}_m^* \setminus \{x\} \cup \{z\}$ и $\mathcal{D}'_\ell = \mathcal{D}_\ell^* \setminus \{z\} \cup \{x\}$. Эта оценка противоречит предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$.

Случай б) доказывается аналогичным образом. Лемма 3 доказана.

Лемма 4. Для каждого $k \in \{1, \dots, K\}$ и каждого $j \in \{1, \dots, J\}$ в Задаче 1D справедливо: $\bar{y}(\mathcal{C}_k^*) \neq c_j$.

Доказательство. Пусть существуют такие $m \in \{1, \dots, K\}$ и $\ell \in \{1, \dots, J\}$, для которых $\bar{y}(\mathcal{C}_m^*) = c_\ell$.

Рассмотрим произвольные точки $x \in \mathcal{C}_m^*$ и $z \in \mathcal{D}_\ell^*$. Положим $\mathcal{C}'_m = \mathcal{C}_m^* \setminus \{x\} \cup \{z\}$ и $\mathcal{D}'_\ell = \mathcal{D}_\ell^* \setminus \{z\} \cup \{x\}$. Тогда справедливо равенство

$$\sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 = \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 + \sum_{y \in \mathcal{D}'_\ell} (y - c_\ell)^2. \quad (2.6)$$

Далее, поскольку $\mathcal{U}$ — множество, $x \neq z$ и $\bar{y}(\mathcal{C}_m^*) \neq \bar{y}(\mathcal{C}'_m)$. Поэтому из (2.3) следует оценка

$$\sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 > \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2. \quad (2.7)$$

Наконец, применяя (2.7) к правой части (2.6), получаем

$$\sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 > \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 + \sum_{y \in \mathcal{D}'_\ell} (y - c_\ell)^2,$$

что противоречит предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$. Лемма 4 доказана.

Лемма 5. Для каждой $k, j \in \{1, \dots, K\}$ таких, что $k \neq j$, в Задаче 1D справедливо $\bar{y}(\mathcal{C}_k^*) \neq \bar{y}(\mathcal{C}_j^*)$.

Доказательство. Как и в доказательстве леммы 4, предположим, что существуют такие $m, \ell \in \{1, \dots, K\}$, для которых $\bar{y}(\mathcal{C}_m^*) = \bar{y}(\mathcal{C}_\ell^*)$.

Рассмотрим произвольные точки $x \in \mathcal{C}_m$ и $z \in \mathcal{C}_\ell$. Пусть $\mathcal{C}'_m = \mathcal{C}_m^* \setminus \{x\} \cup \{z\}$, $\mathcal{C}'_\ell = \mathcal{C}_\ell^* \setminus \{z\} \cup \{x\}$. Тогда справедливо равенство

$$\sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}_\ell^*} (y - c_\ell)^2 = \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}'_\ell} (y - c_\ell)^2. \quad (2.8)$$

По аналогии с доказательством (2.7), найдем оценки

$$\sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}_m^*))^2 > \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 \quad (2.9)$$

и

$$\sum_{y \in \mathcal{C}'_\ell} (y - \bar{y}(\mathcal{C}_\ell^*))^2 > \sum_{y \in \mathcal{C}'_\ell} (y - \bar{y}(\mathcal{C}'_\ell))^2. \quad (2.10)$$

Наконец, применяя оценки (2.9) и (2.10) к правой части (2.8), получаем

$$\sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{y \in \mathcal{C}_\ell^*} (y - \bar{y}(\mathcal{C}_\ell^*))^2 > \sum_{y \in \mathcal{C}'_m} (y - \bar{y}(\mathcal{C}'_m))^2 + \sum_{y \in \mathcal{C}'_\ell} (y - \bar{y}(\mathcal{C}'_\ell))^2,$$

что противоречит предположению об оптимальности $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$. Лемма 5 доказана.

Заметим, что леммы 4 и 5 справедливы также и для многомерного случая.

Леммы 1–5 устанавливают взаимное расположение на числовой прямой оптимальных кластеров $\mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ с заданными центрами и кластеров $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*$ с неизвестными центрами. Эти леммы служат базой для доказательства следующего утверждения.

Теорема 1. Пусть в Задаче 1D точки $y_1, \dots, y_N$ входного множества $\mathcal{U}$, а также точки $c_1, \dots, c_J$ упорядочены так, что

$$y_1 < \dots < y_N; \quad c_1 < \dots < c_J.$$

Тогда оптимальному разбиению $\mathcal{U}$ на кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*, \mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ соответствует разбиение последовательности $1, \dots, N$ натуральных чисел на непересекающиеся целочисленные отрезки.

Доказательство. Справедливость теоремы 1 следует из лемм 1–5. Теорема 1 доказана.

3. ПОЛИНОМИАЛЬНАЯ РАЗРЕШИМОСТЬ ЗАДАЧИ В ОДНОМЕРНОМ СЛУЧАЕ

Следующая теорема является основным результатом работы.

Теорема 2. Существует алгоритм, который находит оптимальное решение Задачи 1D за полиномиальное время.

Доказательство теоремы проводится конструктивно, а именно: мы обосновываем алгоритм, который реализует схему динамического программирования и позволяет находить точное решение задачи за время $\mathcal{O}(KJN^2)$.

Без ограничения общности допустим, что точки $y_1, \dots, y_N$ входного множества $\mathcal{U}$, а также точки $c_1, \dots, c_J$ упорядочены как в теореме 1.

Пусть $\mathcal{U}_{s,t} = \{y_s, \dots, y_t\}$, где $1 \leq s \leq t \leq N$, — подмножество из $t - s + 1$ точек входного множества $\mathcal{U}$ с номерами от s до t .

Положим

$$f_{s,t}^j = \sum_{i=s}^t (y_i - c_j)^2, \quad j = 1, \dots, J,$$

$$f_{s,t} = \sum_{i=s}^t (y_i - \bar{y}(\mathcal{Y}_{s,t}))^2 \equiv \sum_{i=s}^t y_i^2 - \frac{1}{t-s+1} \left(\sum_{i=s}^t y_i \right)^2,$$

где $\bar{y}(\mathcal{Y}_{s,t})$ — центроид подмножества $\mathcal{Y}_{s,t}$.

В соответствии с теоремой 1, пусть $\{i^*, i^* + 1, \dots, N\}$ — номера элементов оптимального кластера, содержащего точку y_N .

Заметим, что для оптимального значения F^* целевой функции Задачи 1D возможны два случая:

$$F^* = \begin{cases} \sum_{k=1}^{K-1} \sum_{y \in \mathcal{C}_k^*} (y - \bar{y}(\mathcal{C}_k^*))^2 + \sum_{j=1}^J \sum_{y \in \mathcal{D}_j^*} (y - c_j)^2 + f_{i^*,N}, & \text{если } \mathcal{C}_K^* = \{y_{i^*}, y_{i^*+1}, \dots, y_N\}, \\ \sum_{k=1}^K \sum_{y \in \mathcal{C}_k^*} (y - \bar{y}(\mathcal{C}_k^*))^2 + \sum_{j=1}^{J-1} \sum_{y \in \mathcal{D}_j^*} (y - c_j)^2 + f_{i^*,N}^J, & \text{если } \mathcal{D}_J^* = \{y_{i^*}, y_{i^*+1}, \dots, y_N\}. \end{cases} \quad (3.11)$$

Обозначим Задачу 1D через $\langle K, J, N \rangle$. Погрузим эту задачу в семейство $\{\langle k, j, n \rangle, k = 0, \dots, K; j = 0, \dots, J; n = 1, \dots, N\}$ подзадач, где $\langle k, j, n \rangle$ — подзадача нахождения разбиения множества $\{y_1, \dots, y_n\}$ на непустые кластеры $\mathcal{C}_1, \dots, \mathcal{C}_k, \mathcal{D}_1, \dots, \mathcal{D}_j$ такие, что

$$\sum_{m=1}^k \sum_{y \in \mathcal{C}_m} (y - \bar{y}(\mathcal{C}_m))^2 + \sum_{\ell=1}^j \sum_{y \in \mathcal{D}_\ell} (y - c_\ell)^2 \rightarrow \min.$$

Пусть $F_{k,j}(n)$ — оптимальное значение целевой функции подзадачи $\langle k, j, n \rangle$. Тогда, в соответствии с (3.11), возможны два случая:

$$F_{k,j}(n) = \begin{cases} \sum_{m=1}^{k-1} \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{\ell=1}^j \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 + f_{i^*,n}, & \text{если } \mathcal{C}_k^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\}, \\ \sum_{m=1}^k \sum_{y \in \mathcal{C}_m^*} (y - \bar{y}(\mathcal{C}_m^*))^2 + \sum_{\ell=1}^{j-1} \sum_{y \in \mathcal{D}_\ell^*} (y - c_\ell)^2 + f_{i^*,n}^j, & \text{если } \mathcal{D}_j^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\}, \end{cases} \quad (3.12)$$

где $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$ — оптимальные кластеры в подзадаче $\langle k, j, n \rangle$, а $\{i^*, i^* + 1, \dots, n\}$ — номера элементов оптимального кластера, содержащего точку y_n .

Для проведения вычислений определим начальные и граничные условия, которые следуют из свойств оптимального решения:

$$F_{k,j}(n) = \begin{cases} 0, & \text{если } n = k = j = 0, \\ +\infty, & \text{если } n = 0, \quad k = 0, \dots, K, \quad j = 0, \dots, J, \quad k + j \neq 0, \\ +\infty, & \text{если } k = -1, \quad j = -1, \dots, J, \quad n = 0, \dots, N, \\ +\infty, & \text{если } j = -1, \quad k = -1, \dots, K, \quad n = 0, \dots, N. \end{cases} \quad (3.13)$$

Тогда, в соответствии с принципом оптимальности Беллмана, для двух случаев (3.12) имеем

$$F_{k,j}(n) = \min_{i=1}^n \{F_{k-1,j}(i-1) + f_{i,n}\}, \quad k = 0, \dots, K, \quad j = 0, \dots, J, \quad n = 1, \dots, N, \quad (3.14)$$

в первом случае, когда $\mathcal{C}_k^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\}$. Во втором случае, когда $\mathcal{D}_j^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\}$, имеем

$$F_{k,j}(n) = \min_{i=1}^n \{F_{k,j-1}(i-1) + f_{i,n}^j\}, \quad k = 0, \dots, K, \quad j = 0, \dots, J, \quad n = 1, \dots, N. \quad (3.15)$$

Объединяя (3.14) и (3.15), получаем, что для $F_{k,j}(n)$ выполнено равенство

$$F_{k,j}(n) = \min \left\{ \min_{i=1}^n \{F_{k-1,j}(i-1) + f_{i,n}\}, \min_{i=1}^n \{F_{k,j-1}(i-1) + f_{i,n}^j\} \right\}, \quad (3.16)$$

$$k = 0, \dots, K, \quad j = 0, \dots, J, \quad n = 1, \dots, N.$$

Таким образом, для оптимального значения целевой функции Задачи 1D справедлива формула

$$F^* = F_{K,J}(N).$$

Таблица значений функции

$$F_{k,j}(n), \quad k = -1, 0, 1, \dots, K, \quad j = -1, 0, 1, \dots, J, \quad n = 0, \dots, N,$$

вычисляется по формуле (3.13) и рекуррентной формуле (3.16). В целом формулы (3.13), (3.16) реализуют прямой ход алгоритма.

Далее, из принципа оптимальности Беллмана следует, что кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_K^*$ и $\mathcal{D}_1^*, \dots, \mathcal{D}_J^*$ находятся по следующему рекуррентному правилу, которое реализует обратный ход алгоритма, начиная с $k = K, j = J, n = N$.

Пошаговая запись этого правила выглядит следующим образом.

Шаг 1. Если

$$\min_{i=1}^n (F_{k-1,j}(i-1) + f_{i,n}) \leq \min_{i=1}^n (F_{k,j-1}(i-1) + f_{i,n}^j), \quad (3.17)$$

то

$$\mathcal{C}_k^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\}, \quad (3.18)$$

где

$$i^* = \arg \min_{i=1}^n (F_{k-1,j}(i-1) + f_{i,n}), \quad k := k-1, \quad n := i^* - 1. \quad (3.19)$$

Если, однако,

$$\min_{i=1}^n (F_{k-1,j}(i-1) + f_{i,n}) > \min_{i=1}^n (F_{k,j-1}(i-1) + f_{i,n}^j), \quad (3.20)$$

то

$$\mathcal{D}_j^* = \{y_{i^*}, y_{i^*+1}, \dots, y_n\},$$

где

$$i^* = \arg \min_{i=1}^n (F_{k,j-1}(i-1) + f_{i,n}^j); \quad j := j-1, \quad n := i^* - 1.$$

Шаг 2. Если $k > 0$ или $j > 0$, то идти на Шаг 1, иначе — конец вычислений.

Докажем индукцией по k и j , что для произвольных k, j, n это правило находит кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$, которые являются оптимальным решением подзадачи $\langle k, j, n \rangle$.

Рассмотрим случай (3.17), когда наибольшая точка y_n из поднабора $y_1 < \dots < y_n$ входит в кластер $\mathcal{C}_k^*$. Случай (3.20), когда точка y_n входит в кластер $\mathcal{D}_j^*$, анализируется аналогично.

Заметим, что в силу формул (3.16) и (3.14) существует оптимальное решение $\mathcal{C}'_1, \dots, \mathcal{C}'_{k-1}, \mathcal{C}_k^*, \mathcal{D}'_1, \dots, \mathcal{D}'_j$ подзадачи $\langle k, j, n \rangle$, в котором для последнего кластера $\mathcal{C}_k^*$ справедливы формулы (3.18) и (3.19). Поэтому для кластеров $\mathcal{C}'_1, \dots, \mathcal{C}'_{k-1}, \mathcal{C}_k^*, \mathcal{D}'_1, \dots, \mathcal{D}'_j$ имеем

$$F_{k,j}(n) = F(\mathcal{C}'_1, \dots, \mathcal{C}'_{k-1}, \mathcal{C}_k^*; \mathcal{D}'_1, \dots, \mathcal{D}'_j) = F(\mathcal{C}'_1, \dots, \mathcal{C}'_{k-1}; \mathcal{D}'_1, \dots, \mathcal{D}'_j) + f_{i^*,n} = F_{k-1,j}(i^*-1) + f_{i^*,n}. \quad (3.21)$$

По индукционному предположению, для подзадачи $\langle k-1, j, i^*-1 \rangle$ правило обратного хода находит оптимальные кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$, то есть кластеры, для которых

$$F(\mathcal{C}_1^*, \dots, \mathcal{C}_{k-1}^*; \mathcal{D}_1^*, \dots, \mathcal{D}_j^*) = F_{k-1,j}(i^*-1). \quad (3.22)$$

Тогда из (3.21) и (3.22) получим

$$F(\mathcal{C}_1^*, \dots, \mathcal{C}_k^*; \mathcal{D}_1^*, \dots, \mathcal{D}_j^*) = F(\mathcal{C}_1^*, \dots, \mathcal{C}_{k-1}^*; \mathcal{D}_1^*, \dots, \mathcal{D}_j^*) + f_{i^*,n} = F_{k-1,j}(i^*-1) + f_{i^*,n} = F_{k,j}(n),$$

из чего следует оптимальность кластеров $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$.

Сформулируем следующий алгоритм решения Задачи 1D.

Алгоритм А

Вход: N -элементное множество $\mathcal{U}$ точек, натуральное число K и набор $\{c_1, \dots, c_J\}$ точек.

Шаг 1. Отсортируем по возрастанию точки $y_1, \dots, y_N$ и точки $c_1, \dots, c_J$.

Шаг 2. Вычислим значения $f_{s,t}^j$ и $f_{s,t}$.

Шаг 3. Найдем оптимальные значения $F_{k,j}(n)$, используя формулы (3.13) и (3.16).

Шаг 4. Найдем оптимальные кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$, используя правила обратного хода.

Выход: кластеры $\mathcal{C}_1^*, \dots, \mathcal{C}_k^*, \mathcal{D}_1^*, \dots, \mathcal{D}_j^*$.

Оценим трудоемкость алгоритма. Шаг 1 требует $\mathcal{O}(N \log N)$ операций. Шаг 2 может быть выполнен за время $\mathcal{O}(JN^2)$ с использованием префиксного суммирования. Время работы шага 3 определяется трудоемкостью вычислений по формуле (3.16). Эта формула вычисляется $\mathcal{O}(KJN)$ раз, и каждое вычисление значения $F_{k,j}(n)$ требует $\mathcal{O}(N)$ операций. Наконец, шаг 4 требует $\mathcal{O}((K+J)N)$ операций. Суммируя затраты на всех шагах, получаем оценку $\mathcal{O}(KJN^2)$ времени работы алгоритма, то есть алгоритм полиномиален. Теорема 2 доказана.

ЗАКЛЮЧЕНИЕ

В настоящей работе мы доказали полиномиальную разрешимость одномерного случая одной NP-трудной в сильном смысле задачи разбиения конечного множества точек евклидова пространства. Построение эффективных приближенных алгоритмов с гарантированными оценками качества для общего случая задачи, а также ускоренных полиномиальных алгоритмов для ее одномерного случая, представляется делом ближайшей перспективы.

СПИСОК ЛИТЕРАТУРЫ

1. Fisher R.A. Statistical Methods and Scientific Inference. New York. Hafner, 1956. 350 p.
2. MacQueen J.B. Some Methods for Classification and Analysis of Multivariate Observations, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967. V. 1. Univ. of California Press, Berkeley, 1967. P. 281–297.
3. Alois D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. 2009. V. 75. № 2. P. 245–248.
4. Rao M. Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. 1971. V. 66. P. 622–626.
5. Bellman R. Dynamic Programming. Princeton University Press, Princeton, New Jersey, 1957. 342 p.
6. Глебов Н.И. О выпуклых последовательностях // Дискретный анализ. 1965. Вып. 4. Изд-во ИМ СО РАН, Новосибирск. С. 10–22.
7. Гимадутинов Э.Х. О свойствах решений одной задачи оптимального размещения точек на отрезке // Управляемые системы. 1969. Вып. 2. Изд-во ИМ СО РАН, Новосибирск. С. 77–91.
8. Гимадутинов Э.Х. Об одном классе задач нелинейного программирования // Управляемые системы. 1969. Вып. 3. Изд-во ИМ СО РАН, Новосибирск. С. 101–113.
9. Гимади (Гимадутинов) Э.Х. Выбор оптимальных шкал в одном классе задач типа размещения, унификации и стандартизации // Управляемые системы. 1970. Вып. 6. Изд-во ИМ СО РАН, Новосибирск. С. 57–70.
10. Xiaolin Wu. Optimal quantization by matrix searching // J. Algorithms. 1991. V. 12. № 4. P. 663–673.
11. Grönlund A., Larsen K.G., Mathiasen A., Nielsen J.S., Schneider S., Song M. Fast exact k -means, k -medians and Bregman divergence clustering in 1D // CoRR arXiv:1701.07204. 2017.
12. Кельманов А.В., Кельманова М.А., Хамидуллин С.А. Совместное обнаружение и оценивание повторяющегося фрагмента в зашумленной числовой последовательности при заданном числе квазипериодических повторов // Тез. докл. Российской конф. “Дискретный анализ и исследование операций” (DAOR-4). Новосибирск, 2004. Изд-во ИМ СО РАН, Новосибирск. С. 185.
13. Гимади Э.Х., Кельманов А.В., Кельманова М.А., Хамидуллин С.А. Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. ж. индустр. математики. 2006. Т. 9. № 1(25). С. 55–74.

14. *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.
15. *Кельманов А.В., Хамидуллин С.А.* Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // *Ж. вычисл. матем. и матем. физ.* 2001. Т. 41. № 5. С. 807–820.
16. *Kel'manov A.V., Jeon B.* A Posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // *IEEE Trans. on Signal Proc.* 2004. V. 52. № 3. P. 645–656.
17. *Carter J.A., Agol E., et al.* Kepler-36: A pair of planets with neighboring orbits and dissimilar densities // *Science*. 2012. V. 337. № 6094. P. 556–559.
18. *Кельманов А.В., Пяткин А.В.* О сложности одного из вариантов задачи выбора подмножества “похожих” векторов // *Докл. АН*. 2008. Т. 421. № 5. С. 590–592.
19. *Кельманов А.В., Пяткин А.В.* Об одном варианте задачи выбора подмножества векторов // *Дискретн. анализ и исслед. опер.* 2008. Т. 15. № 5. С. 20–34.
20. *Кельманов А.В., Пяткин А.В.* О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // *Ж. вычисл. матем. и матем. физ.* 2009. Т. 49. № 11. С. 2059–2065.