ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль–сентябрь 2023. Т. 30, № 3. С. 96–110

УДК 519.8+518.25 DOI 10.33048/daio.2023.30.763

# ПОЛИНОМИАЛЬНЫЕ АППРОКСИМАЦИОННЫЕ СХЕМЫ ДЛЯ ЗАДАЧ ВЫБОРА ВЕКТОРОВ И КЛАСТЕРИЗАЦИИ С РАЗНЫМИ ЦЕНТРАМИ

*А. В. Пяткин*

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

**Аннотация.** Рассматриваются две близкие в постановочном плане задачи. Во-первых, общая задача кластеризации, т. е. разбиения множества $d$-мерных евклидовых векторов на заданное число кластеров с разными типами центров, при котором суммарная дисперсия будет минимальной. Под дисперсией понимается сумма квадратов расстояний между элементами кластера и его центром. При этом для одной части кластеров центр может быть выбран произвольно (очевидно, что в этом случае следует выбрать центроид), для другой части в качестве центра должен быть выбран один из векторов исходного множества, а для остальных кластеров центрами являются заранее заданные точки. Также на входе задаются размеры каждого кластера. Вторая рассматриваемая задача — это задача выбора подмножества векторов заданной мощности с минимальной суммой квадратов расстояний от его элементов до центроида. В статье построены полиномиальные аппроксимационные схемы (PTAS) для обеих задач. Библиогр. 23.

**Ключевые слова:** кластеризация, центроид, медоид, аппроксимация, PTAS-схема, задача MSSC.

## Введение

В настоящей статье изучаются две задачи, связанные с кластеризацией множества векторов евклидова пространства на непустые подмножества похожих векторов и поиском в нём одного подмножества из наиболее похожих векторов. Такого рода задачи весьма актуальны для анализа и обработки данных, искусственного интеллекта, вычислительной геометрии, математической статистики и дискретной оптимизации [1–3].

Исследование выполнено в рамках государственного задания ИМ СО РАН (проект № FWNF–2022–0019).

© А. В. Пяткин, 2023

По сути, это целый класс задач, отличающихся способами выбора критериев похожести, а также ограничений на число и размеры кластеров.

Одним из наиболее часто встречающихся критериев является минимизация дисперсии, под которой понимается сумма квадратов расстояний между элементами кластера и некоторой точкой, которая называется *центром* кластера. Определим для произвольного кластера $\mathcal{C}$ и его центра $x$ функцию

$$f(\mathcal{C}, x) = \sum_{y \in \mathcal{C}} \|x - y\|^2.$$

Обычно рассматриваются следующие ограничения на выбор центра кластера:

а) произвольная точка (т. е. ограничений нет);

б) точка из исходного множества;

в) фиксированная (заданная) точка пространства.

Такой выбор центров можно обосновать следующим образом. Допустим, что в заданной географической области имеются несколько городков, в которых нужно мониторить ситуацию с помощью сенсоров. Некоторые сенсоры автономны, и их можно поместить куда угодно. Другие сенсоры требуют постоянного обслуживания людьми, и поэтому их следует помещать в поселениях (городках). Также в области могут иметься уже установленные ранее сенсоры (неважно каких типов), которые нельзя перемещать, но можно также использовать для мониторинга. Поскольку расход энергии пропорционален квадрату расстояния от сенсора до объекта, задачу кластеризации с разными типами сенсоров можно интерпретировать как проблему размещения сенсоров двух имеющихся типов с учётом ранее установленных сенсоров, минимизируя общий расход энергии.

Если все векторы из кластера $\mathcal{C}$ известны и центр кластера можно выбрать произвольно, то нетрудно показать с помощью частных производных, что оптимальным выбором будет так называемый *центроид*, определяемый как среднее значение всех векторов из кластера $\mathcal{C}$:

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

Заметим, что если это ограничение задано для всех центров кластеров, то получается классическая задача MSSC (minimum sum-squared clustering) [4, 5]. Если центр кластера должен быть выбран среди точек исходного множества, то такой центр назовём *медоидом*. Отметим, что термин медоид был введён в работе [6] как представитель кластера, чьё среднее отличие от остальных объектов кластера минимально. Хотя медоиды обычно ассоциируются с задачами, в которых используется сумма

расстояний, применение этого же термина для задач с квадратами расстояний не противоречит исходному определению. Наконец, третий тип центров кластеров (заданная точка пространства) будем называть *фиксированным*.

Основным объектом изучения является

**Задача 1** ($p$ центроидов, $q$ медоидов, $r$ фиксированных центров)**.** *Дано конечное множество векторов* $\mathcal{Y} = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$, *множество точек* $\{z_1, \ldots, z_r\} \subset \mathbb{R}^d$ *и положительные целые числа* $m_1, \ldots, m_k$, *удовлетворяющие соотношениям* $m_1 + \ldots + m_k = n$ *и* $p+q+r = k$. *Найти разбиение множества* $\mathcal{Y}$ *на* $k$ *кластеров* $\mathcal{C}_1, \ldots, \mathcal{C}_k$, *где* $|\mathcal{C}_i| = m_i$ *для всех* $i = 1, \ldots, k$ *такое, что значение*

$$F(\mathcal{C}_1, \ldots, \mathcal{C}_k; X) = \sum_{i=1}^{k} f(\mathcal{C}_i, x_i)$$

*было бы минимально, где* $X = \{x_1, \ldots, x_k\}$ *— множество центров кластеров, удовлетворяющих следующим ограничениям:*

- $x_i = \overline{y}(\mathcal{C}_i)$ *при* $1 \leqslant i \leqslant p$;
- $x_i \in \mathcal{Y}$ *при* $p + 1 \leqslant i \leqslant p + q$;
- $x_i = z_{i-p-q}$ *при* $p + q + 1 \leqslant i \leqslant k$.

Другими словами, центры первых $p$ кластеров должны быть центроидами, центры следующих $q$ кластеров — медоидами, а центры последних $r$ кластеров — фиксированными точками $z_1, \ldots, z_r$.

Известно, что эта задача NP-трудна даже в случае $p \geqslant 1$ и $k = 2$. Действительно, при $p = 2$ и $q = r = 0$ получаем задачу MSSC с $k = 2$, NP-трудность которой была доказана в [7] для случая нефиксированных размеров кластеров (очевидно, что если вариант задачи с фиксированными размерами кластеров полиномиально разрешим, то и задача с нефиксированными размерами кластеров также полиномиально разрешима с помощью полного перебора всех размеров; отсюда NP-трудность задачи с нефиксированными мощностями кластеров влечёт NP-трудность задачи с фиксированными мощностями). Первая полиномиальная аппроксимационная схема (PTAS) для задачи MSSC с нефиксированными размерами кластеров при $k = 2$ была предложена в [8]; она находит $(1 + \varepsilon)$-приближённое решение за время $O(n(1/\varepsilon)^d)$. В работе [9] предложена PTAS-схема для той же задачи с произвольным числом кластеров $k$, находящая $(1+\varepsilon)$-приближённое решение за время $O(dn2^{(k/\varepsilon)^{O(1)}})$. Обе эти PTAS-схемы рандомизированные.

Случай $p = r = 1$, $q = 0$ был рассмотрен в [10–13]. NP-трудность задачи была сначала доказана для фиксированных [10, 11], а затем для нефиксированных [12, 13] размеров кластеров. Для обоих вариантов задачи известны 2-приближённые алгоритмы [14, 15] временной сложности

$O(n^2d)$. Кроме того, в [16] для задачи 1 с $p = r = 1$ и $q = 0$ построена PTAS-схема трудоёмкости $O(dn^{1+2/\varepsilon}(9/\varepsilon)^{3/\varepsilon})$.

Наконец, NP-трудность задачи 1 в случае $p = q = 1$, $r = 0$ была доказана в [17] для фиксированных размеров кластеров, а в [18] — для нефиксированных. В [18] также предложен 2-приближённый алгоритм трудоёмкости $O(n^2d + n^3)$ для последней задачи. Что касается случая $p = 0$, то он является полиномиально разрешимым для любого $k$, как показано в следующем разделе.

Основным результатом настоящей статьи является построение PTAS-схемы для общего случая задачи 1. Ранее для случая $p > 0$ и $q > 0$ таких схем известно не было.

Заметим, что в [19] была построена PTAS-схема для следующей близкой по смыслу задачи выбора подмножества похожих векторов.

**Задача 2** (выбор подмножества векторов)**.** *Дано множество векторов* $\mathcal{Y} = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ *и положительное целое* $m \leqslant n$. *Найти подмножество* $\mathcal{C} \subseteq \mathcal{Y}$ *размера* $m$, *минимизирующее функцию* $f(\mathcal{C}, \overline{y}(\mathcal{C}))$.

Предложенная в [19] полиномиальная схема находит $(1 + \varepsilon)$-приближённое решение задачи 2 за время $O(dn^{1+2/\varepsilon}(9/\varepsilon)^{3/\varepsilon})$. Отметим, что прямое применение указанной схемы для решения задачи 1 не позволяет построить для неё PTAS-схему, что может быть проиллюстрировано на следующем простом примере. Пусть $d = 1$ и $\mathcal{Y}$ содержит четыре числа $y_1 = -2$, $y_2 = y_3 = 0$ и $y_4 = 2$. Пусть также $p = q = 1$, $r = 0$ и $m_1 = m_2 = 2$. Очевидно, что при достаточно малом $\varepsilon > 0$ PTAS-схема найдёт решение $\mathcal{C} = \{y_2, y_3\}$ (которое является оптимальным решением задачи 2 с таким множеством $\mathcal{Y}$ и $m = 2$). Тогда $F(\mathcal{C}, \mathcal{Y} \setminus \mathcal{C}; \{0, 0\}) = 8$ (в качестве центра второго кластера берётся точка $y_2$), в то время как оптимальным решением является $\mathcal{C}_1 = \{y_1, y_2\}$ и $\mathcal{C}_2 = \{y_3, y_4\}$, поскольку $F(\mathcal{C}_1, \mathcal{C}_2; \{-1, 0\}) = 6$. Таким образом, при малом $\varepsilon > 0$ относительная погрешность остаётся равной $1/3$. Однако, используя одно из геометрических свойств, доказанных в [19], можно предложить иную технику, которая позволяет построить как PTAS-схему для задачи 1, так и лучшую PTAS-схему для задачи 2.

Статья имеет следующую структуру. В разд. 1 приводятся предварительные результаты, а также улучшенная PTAS-схема для задачи 2. Основной результат статьи (а именно, PTAS-схема для задачи 1) доказывается в разд. 2. Заключение завершает статью. Отметим, что укороченная версия данной статьи (без улучшенной PTAS-схемы для задачи 2) подавалась в виде тезисов на конференцию MOTOR-2023.

## 1. Предварительные результаты

В данном разделе приводятся предварительные результаты, необходимые для обоснования PTAS-схемы. Отметим, что некоторые результаты известны (фольклорны), но их строгое доказательство приводится в тексте для полноты изложения.

Прежде всего покажем, что задача 1 с $k$ фиксированными центрами полиномиально разрешима для любого заданного $k$. Это известный фольклорный результат, доказательство которого можно найти, например, в [20]. Однако, приведённое ниже доказательство более короткое, хотя и основано на той же идее. Через $x(j)$ обозначим $j$-ю координату вектора $x$.

**Утверждение 1** [20]. *В случае $p = q = 0$ задача 1 разрешима за время $O(n^2d + n^3)$.*

Доказательство. Пусть $z_1, \ldots, z_r$ и $m_1, \ldots, m_r$ являются центрами и мощностями искомых кластеров соответственно. Заметим, что $m_1 + \ldots + m_r = n$. Для каждого $i = 1, \ldots, r$ зададим вектор $a_i \in \mathbb{R}^n$ по правилу $a_i(j) = \|z_i - y_j\|^2$ для всех $j = 1, \ldots, n$. Возьмём $m_i$ копий вектора $a_i$ для каждого $i = 1, \ldots, r$ и составим из этих $n$ векторов квадратную матрицу $A$ размера $n$. Тогда, очевидно, $F(\mathcal{Y})$ равно оптимальному решению задачи о назначениях для матрицы $A$. Поскольку задача о назначениях решается классическим Венгерским методом [21] за время $O(n^3)$, а построение матрицы $A$ требует $O(n^2d)$ операций, исходная задача разрешима за время $O(n^2d + n^3)$. Утверждение 1 доказано.

Заметим, что если все элементы матрицы $A$ являются целыми положительными числами, не превосходящими некоторой константы $C$, то задачу о назначениях можно решить быстрее с помощью алгоритма из работы [22], а именно, за время $O(M\sqrt{n}\log(nC))$, где через $M$ обозначено число ненулевых элементов матрицы $A$. Однако, в нашем случае этот алгоритм может работать хуже, поскольку $M = n^2 - n$, и даже в случае целочисленных координат векторов, константа $C$ может оказаться больше чем $2^n$.

Также заметим, что из утверждения 1 вытекает полиномиальная разрешимость задачи 1 в случае $p = 0$. Действительно, задача о поиске $q$ медоидов и $r$ фиксированных центров сводится к решению не более чем $n^q$ задач с $q + r$ фиксированными центрами (полным перебором всех вариантов выбора медоидов).

Ключевую роль играет

**Лемма 1.** *Пусть $\mathcal{C} \subset \mathcal{Y}$ — произвольный кластер мощности $m$. Тогда для любого положительного целого $t \leqslant m$ существует подмножество*

$\mathcal{T}^* \subset \mathcal{C}$ *мощности* $t$ *такое, что*

$$\|\overline{y}(\mathcal{C}) - \overline{y}(\mathcal{T}^*)\|^2 \leqslant \frac{f(\mathcal{C}, \overline{y}(\mathcal{C}))}{mt}.$$

Доказательство. Рассмотрим функцию $g(\mathcal{T}) = \|\overline{y}(\mathcal{C}) - \overline{y}(\mathcal{T})\|^2$. Выберем произвольное подмножество $\mathcal{T} = \{y_1, \ldots, y_t\}$. Тогда

$$g(\mathcal{T}) = \left\| \overline{y}(\mathcal{C}) - \frac{1}{t}\sum_{i=1}^{t} y_i \right\|^2 = \frac{1}{t^2}\left\| t\overline{y}(\mathcal{C}) - \sum_{i=1}^{t} y_i \right\|^2 = \frac{1}{t^2}\left\| \sum_{i=1}^{t} (\overline{y}(\mathcal{C}) - y_i) \right\|^2$$
$$= \frac{1}{t^2}\Bigg( \sum_{y \in \mathcal{T}} \|\overline{y}(\mathcal{C}) - y\|^2 + \sum_{y \in \mathcal{T}} \sum_{\substack{y' \in \mathcal{T}, \\ y' \neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle \Bigg).$$

Значит, среднее значение функции $g(\mathcal{T})$ по всем подмножествам $\mathcal{T} \subset \mathcal{C}$ мощности $t$ равно

$$\frac{1}{\binom{m}{t} t^2} \sum_{\substack{\mathcal{T} \subset \mathcal{C}, \\ |\mathcal{T}| = t}} \Bigg( \sum_{y \in \mathcal{T}} \|\overline{y}(\mathcal{C}) - y\|^2 + \sum_{y \in \mathcal{T}} \sum_{\substack{y' \in \mathcal{T}, \\ y' \neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle \Bigg).$$

Поскольку каждый $y \in \mathcal{C}$ лежит ровно в $\binom{m-1}{t-1}$ подмножествах $\mathcal{T} \subset \mathcal{C}$ размера $t$, имеем

$$\sum_{\substack{\mathcal{T} \subset \mathcal{C}, \\ |\mathcal{T}| = t}} \sum_{y \in \mathcal{T}} \|\overline{y}(\mathcal{C}) - y\|^2 = \binom{m-1}{t-1} \sum_{y \in \mathcal{C}} \|\overline{y}(\mathcal{C}) - y\|^2$$
$$= \binom{m-1}{t-1} f(\mathcal{C}, \overline{y}(\mathcal{C})). \quad (1)$$

Очевидно, что для любого $z \in \mathbb{R}^d$ выполняется тождество

$$\sum_{y \in \mathcal{C}} \langle \overline{y}(\mathcal{C}) - y, z \rangle = \left\langle m\overline{y}(\mathcal{C}) - \sum_{y \in \mathcal{C}} y,\, z \right\rangle = 0,$$

поэтому

$$0 = \sum_{y \in \mathcal{C}} \sum_{y' \in \mathcal{C}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle = \sum_{y \in \mathcal{C}} \|\overline{y}(\mathcal{C}) - y\|^2 + \sum_{y \in \mathcal{C}} \sum_{\substack{y' \in \mathcal{C}, \\ y' \neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle,$$

а значит,

$$\sum_{y \in \mathcal{C}} \sum_{\substack{y' \in \mathcal{C}, \\ y' \neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle = -\sum_{y \in \mathcal{C}} \|\overline{y}(\mathcal{C}) - y\|^2.$$

Отметим, что каждая упорядоченная пара $y, y' \in \mathcal{C}$, в которой $y \neq y'$, принадлежит ровно $\binom{m-2}{t-2}$ подмножествам $\mathcal{T} \subset \mathcal{C}$ размера $t$. Следовательно,

$$\sum_{\substack{\mathcal{T}\subset\mathcal{C},\\ |\mathcal{T}|=t}} \sum_{y\in\mathcal{T}} \sum_{\substack{y'\in\mathcal{T},\\ y'\neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle = \binom{m-2}{t-2} \sum_{y\in\mathcal{C}} \sum_{\substack{y'\in\mathcal{C},\\ y'\neq y}} \langle \overline{y}(\mathcal{C}) - y,\, \overline{y}(\mathcal{C}) - y' \rangle$$
$$= -\binom{m-2}{t-2} \sum_{y\in\mathcal{C}} \|\overline{y}(\mathcal{C}) - y\|^2 = -\binom{m-2}{t-2} f(\mathcal{C}, \overline{y}(\mathcal{C})). \quad (2)$$

Объединяя (1) и (2), получим формулу для среднего значения функции $g(\mathcal{T})$ по всем подмножествам мощности $t$:

$$\frac{\binom{m-1}{t-1} - \binom{m-2}{t-2}}{\binom{m}{t} t^2} f(\mathcal{C}, \overline{y}(\mathcal{C})) = \left(\frac{1}{mt} - \frac{t-1}{(m-1)mt}\right) f(\mathcal{C}, \overline{y}(\mathcal{C})$$
$$= \frac{(m-t)}{(m-1)mt} f(\mathcal{C}, \overline{y}(\mathcal{C})).$$

Поскольку минимальное значение функции $g(\mathcal{T})$ не превосходит её среднего значения, найдётся подмножество $\mathcal{T}^* \subset \mathcal{C}$ размера $t$, для которого выполняется неравенство

$$g(\mathcal{T}^*) \leqslant \frac{(m-t)}{(m-1)mt} f(\mathcal{C}, \overline{y}(\mathcal{C})) \leqslant \frac{f(\mathcal{C}, \overline{y}(\mathcal{C}))}{mt}.$$

Лемма 1 доказана.

Следующее утверждение было доказано в [19].

**Утверждение 2** [19]**.** *Для любого кластера $\mathcal{C} \subset \mathcal{Y}$ мощности $m$ и точки $z \in \mathbb{R}^d$ выполняется тождество*

$$f(\mathcal{C}, z) = f(\mathcal{C}, \overline{y}(\mathcal{C})) + m\|\overline{y}(\mathcal{C}) - z\|^2.$$

Из утверждения 2 и леммы 1 следует основной инструмент для построения PTAS-схемы для обеих рассматриваемых задач.

**Лемма 2.** *Обозначим через $\mathcal{Z}$ множество центроидов всех кластеров мощности не более $t$ в множестве $\mathcal{Y}$. Тогда для любого кластера $\mathcal{C} \subset \mathcal{Y}$ найдётся такая точка $z \in \mathcal{Z}$, что $f(\mathcal{C}, z) \leqslant (1 + 1/t) f(\mathcal{C}, \overline{y}(\mathcal{C}))$.*

Доказательство. Рассмотрим произвольный кластер $\mathcal{C} \subset \mathcal{Y}$ и положим $m = |\mathcal{C}|$. Ясно, что если $m \leqslant t$, то множество $\mathcal{Z}$ содержит центроид кластера $\mathcal{C}$ и можно положить $z = \overline{y}(\mathcal{C})$. В противном случае рассмотрим для кластера $\mathcal{C}$ подмножество $\mathcal{T}^*$ мощности $t$, существование которого

доказано в лемме 1. Тогда точка $z = \overline{y}(\mathcal{T}^*)$ лежит в $\mathcal{Z}$, и по утверждению 2 и лемме 1 получаем

$$f(\mathcal{C}, z) = f(\mathcal{C}, \overline{y}(\mathcal{C})) + m\|\overline{y}(\mathcal{C}) - z\|^2 \leqslant \left(1 + \frac{1}{t}\right) f(\mathcal{C}, \overline{y}(\mathcal{C})).$$

Лемма 2 доказана.

Таким образом, множество $\mathcal{Z}$, построенное в лемме 2, содержит не более чем $2n^t$ кандидатов на роль примерного центра каждого из первых $p$ кластеров в задаче 1. Более того, для каждого такого кластера $\mathcal{C}$ по крайней мере один из этих центров даёт $(1+\varepsilon)$-приближённое решение для $f(\mathcal{C}, \overline{y}(\mathcal{C}))$, где $\varepsilon = 1/t$.

Лемма 2 также позволяет построить новую PTAS-схему для задачи 2 с лучшей оценкой трудоёмкости, чем в [19]. Действительно, рассмотрим следующий простой алгоритм.

**Алгоритм 1**

Вход: множество векторов $\mathcal{Y} = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$, размер кластера $m \leqslant n$ и положительный целый параметр $t \leqslant m$.

Шаг 0. Положим $\mathcal{Z} = \varnothing$ и рекорд $R = \infty$.

Шаг 1. Рассмотрим все подмножества $\mathcal{T} \subset \mathcal{Y}$ размера $t$ и добавим в $\mathcal{Z}$ центроид $\overline{y}(\mathcal{T})$ каждого из рассмотренных подмножеств.

Шаг 2. Для каждого $z \in \mathcal{Z}$ рассмотрим подмножество $\mathcal{C}_z$, состоящее из $m$ ближайших к $z$ векторов из $\mathcal{Y}$. Если $f(\mathcal{C}_z, z) < R$, то обновляем рекорд $R := f(\mathcal{C}_z, z)$.

Шаг 3. Обозначим через $z^*$ и $\mathcal{C}_{z^*}$ точку и подмножество, на которых достигается рекорд $R$.

Выход: точка $z^*$ и множество $\mathcal{C}_{z^*}$.

**Теорема 1.** *Для любого $\varepsilon > 0$ алгоритм 1 находит $(1 + \varepsilon)$-приближённое решение для задачи 2 за время $O(n^{1+1/\varepsilon}d)$.*

Доказательство. Обозначим через $\mathcal{C}^*$ оптимальный кластер в задаче 2 и положим $t = \lceil 1/\varepsilon \rceil$. Тогда по лемме 2 существует такая точка $z' \in \mathcal{Z}$, что $f(\mathcal{C}^*, z') \leqslant (1 + 1/t) f(\mathcal{C}^*, \overline{y}(\mathcal{C}^*))$. Обозначим через $\mathcal{C}'$ подмножество из $m$ ближайших к $z'$ векторов из $\mathcal{Y}$ (напомним, что это подмножество рассматривалось на шаге 2 алгоритма 1). Тогда

$$f(\mathcal{C}_{z^*}, \overline{y}(\mathcal{C}_{z^*})) \leqslant f(\mathcal{C}_{z^*}, z^*) \leqslant f(\mathcal{C}', z') \leqslant f(\mathcal{C}^*, z') \leqslant (1 + \varepsilon) f(\mathcal{C}^*, \overline{y}(\mathcal{C}^*)),$$

так как $\varepsilon \geqslant 1/t$.

Множество $\mathcal{Z}$ содержит не более чем $n^t$ элементов. Поиск $m$ ближайших к точке $z \in \mathcal{Z}$ векторов из $\mathcal{Y}$ занимает $O(nd)$ операций (детали можно найти, например, в [23]). Значит, временная сложность алгоритма 2 для решения задачи 2 не превосходит $O(dn^{1+1/\varepsilon})$. Теорема 1 доказана.

## 2. Основной результат

Предлагаемая PTAS-схема для решения задачи 1 действует следующим образом. После выбора параметра $t$ строится множество $\mathcal{Z}$, содержащее центроиды всех подмножеств мощности не более $t$ в $\mathcal{Y}$. Далее множество $\mathcal{Z}$ рассматривается как множество кандидатов на роль приближённого центра каждого из первых $p$ кластеров (для которых центрами являются центроиды) в задаче 1, множество $\mathcal{Y}$ — как множество возможных центров для следующих $q$ кластеров (медоиды) и одноэлементные множества $\{z_j\}$ — как множества центров оставшихся $r$ кластеров (с фиксированными центрами). Далее рассматриваются все возможные способы выбора по одному центру для каждого кластера из соответствующего ему множества кандидатов, и для каждого такого выбранного набора центров решается задача с $k$ фиксированными центрами с помощью утверждения 1. Тогда наилучшее из найденных решений является $(1+\varepsilon)$-приближённым решением задачи 1.

Приведём теперь формальное описание аппроксимационной схемы.

**Алгоритм 2**

Вход: множество векторов $\mathcal{Y} = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$, множество точек (фиксированных центров) $\{z_1, \ldots, z_r\} \subset \mathbb{R}^d$, положительные целые числа $p, q, r, m_1, \ldots, m_k$, удовлетворяющие соотношениям $m_1 + \ldots + m_k = n$ и $p + q + r = k$, а также положительный целый параметр $t$.

Шаг 0. Положим $\mathcal{Z} = \varnothing$ и рекорд $R = \infty$.

Шаг 1. Рассмотрим каждое подмножество $\mathcal{T} \subset \mathcal{Y}$ мощности не более $t$ и добавим к $\mathcal{Z}$ его центроид $\overline{y}(\mathcal{T})$.

Шаг 2. Положим $\mathcal{Z}_i = \mathcal{Z}$ для всех $i = 1, \ldots, p$, а также $\mathcal{Z}_i = \mathcal{Y}$ для всех $i = p+1 \ldots, p+q$ и $\mathcal{Z}_i = \{z_{i-p-q}\}$ для всех $i = p+q+1, \ldots, k$.

Шаг 3. Рассмотрим все возможные способы выбора точек $x_i \in \mathcal{Z}_i$ для каждого $i = 1, \ldots, k$ и для каждого такого набора точек решим задачу 1 с $k$ фиксированными центрами для множества $\mathcal{Y}$ с помощью алгоритма из утверждения 1. Обозначим через $\mathcal{C}_1, \ldots, \mathcal{C}_k$ найденные кластеры и положим $X = \{x_1, \ldots, x_k\}$. Если $F(\mathcal{C}_1, \ldots, \mathcal{C}_k; X) < R$, то обновляем рекорд $R := F(\mathcal{C}_1, \ldots, \mathcal{C}_k; X)$.

Шаг 4. Обозначим через $\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A$ и $X^A = \{x_1, \ldots, x_k\}$ кластеры множество центров, на которых достигается рекорд $R$.

Выход: кластеры $\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A$ и $X^A = \{x_1, \ldots, x_k\}$.

Осталось доказать корректность и оценить трудоёмкость алгоритма 2.

**Теорема 2.** *Алгоритм* 2 *находит* $(1+\varepsilon)$*-приближённое решение задачи* 1 *за время* $O((d+n)n^{q+2+p/\varepsilon})$, *где* $\varepsilon = 1/t$.

Доказательство. Докажем корректность алгоритма. Пусть кластеры $\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A$ и множество их центров $X^A = \{x_1, \ldots, x_k\}$ получены на выходе алгоритма 2. Обозначим через $\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*$ и $X^* = \{x_1^*, \ldots, x_k^*\}$ набор кластеров и их центров в оптимальном решении задачи 1. По лемме 2 для каждого $i = 1, \ldots, p$ существует такая точка $x_i' \in \mathcal{Z}_i$, что $f(\mathcal{C}_i^*, x_i') \leqslant (1+\varepsilon) f(\mathcal{C}_i^*, x_i^*)$. Положим $x_i' = x_i^*$ для всех $i = p+1, \ldots, k$, и пусть $X' = \{x_1', \ldots, x_k'\}$. Тогда, очевидно, что

$$F(\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*; X') \leqslant (1+\varepsilon) F(\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*; X^*). \tag{3}$$

Заметим, что набор центров $X'$ рассматривался на шаге 3 алгоритма 2. Обозначим через $\mathcal{C}_1', \ldots, \mathcal{C}_k'$ множество кластеров, найденных для набора $X'$ на шаге 3. Поскольку этот набор является оптимальным решением задачи с $k$ фиксированными центрами (заданными множеством $X'$) для множества $\mathcal{Y}$, а разбиение $\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A$ и множество центров $X^A$ — наилучшее решение, найденное алгоритмом 2, имеют место следующие неравенства:

$$F(\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A; X^A) \leqslant F(\mathcal{C}_1', \ldots, \mathcal{C}_k'; X') \leqslant F(\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*; X'). \tag{4}$$

Объединяя (3) и (4), получаем требуемое соотношение

$$F(\mathcal{C}_1^A, \ldots, \mathcal{C}_k^A; X^A) \leqslant (1+\varepsilon) F(\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*; X^*).$$

Оценим трудоёмкость алгоритма. Очевидно, что шаги 0, 2 и 4 выполняются за константное время, а шаг 1 алгоритма 2 требует $O(dn^t)$ операций. Наиболее трудоёмким является шаг 3, на котором для каждого из возможных наборов центров решается задача с фиксированными центрами. Поскольку всего имеется не более $(2n^t)^p n^q$ возможных способов выборов центров, а решение задачи с $k$ фиксированными центрами требует $O(n^2 d + n^3)$ операций по утверждению 1, трудоёмкость алгоритма 2 равна $O((d+n)n^{pt+q+2})$. Теорема 2 доказана.

## Заключение

Основным результатом статьи является полиномиальная аппроксимационная схема (PTAS) трудоёмкости $O((d+n)n^{q+2+p/\varepsilon})$ для достаточно общей задачи кластеризации, в которой центры первых $p$ кластеров являются центроидами, следующих $q$ кластеров — медоидами, а центры оставшихся $r$ кластеров фиксированы (числа $p$, $q$ и $r$ известны заранее, т. е. не являются частью входа). Мощности искомых кластеров составляют часть входа. Ключевая идея схемы заключается в построении не слишком большого множества кандидатов на роли центроидов, содержащего хорошее приближение для центров всех таких кластеров. Схема рассматривает все возможные комбинации выбора центров и решает

для каждого набора задачу с $k$ фиксированными центрами сведением последней к задаче о назначениях.

В качестве дополнительного результата получена улучшенная PTAS-схема для задачи выбора подмножества векторов.

Автор благодарит рецензента за тщательное изучение работы и сделанные замечания, позволившие улучшить её представление.

## ЛИТЕРАТУРА

1. **Berkhin P.** A survey of clustering data mining techniques // Grouping multidimensional data: Recent advances in clustering. Heidelberg: Springer, 2006. P. 25–71.
2. **Jain A. K., Dubes R. C.** Algorithms for clustering data. Englewood Cliffs, NJ: Prentice Hall, 1988.
3. **Ghoreyshi S., Hosseinkhani J.** Developing a clustering model based on $K$-means algorithm in order to creating different policies for policyholders // Int. J. Adv. Comput. Sci. Inf. Tech. 2015. V. 4, No. 2. P. 46–53.
4. **Fisher W. D.** On grouping for maximum homogeneity // J. Am. Stat. Assoc. 1958. V. 53, No. 284. P. 789–798.
5. **MacQueen J.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability (Berkeley, USA, June 21 — July 18, 1965; Dec. 27, 1965 — Jan. 7, 1966). V. 1. Berkeley: Univ. California Press, 1967. P. 281–297.
6. **Kaufman L., Rousseeuw P. J.** Clustering by means of medoids // Statistical data analysis based on the $L_1$-norm. Amsterdam: North-Holland, 1987. P. 405–416.
7. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. 2009. V. 75, No. 2. P. 245–248.
8. **Inaba M., Katoh N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based $k$-clustering // Proc. 10th Annu. Symp. Computational Geometry (Stony Brook, NY, USA, June 6–8, 1994). New York: ACM Press, 1994. P. 332–339.
9. **Kumar A., Sabharwal Y., Sen S.** A simple linear time $(1+\varepsilon)$-approximation algorithm for geometric $k$-means clustering in any dimensions // Proc. 45th Annu. IEEE Symp. Foundations of Computer Science (Rome, Italy, Oct. 17–19, 2004). Los Alamitos, CA: IEEE Comp. Soc., 2004. P. 454–462.
10. **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.
11. **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. С. 55–74.

**12. Кельманов А. В., Пяткин А. В.** О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Докл. Акад. наук. 2008. Т. 421, № 5. С. 590–592.

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

**14. Долгушев А. В., Кельманов А. В.** Приближённый алгоритм решения одной задачи кластерного анализа // Дискрет. анализ и исслед. операций. 2011. Т. 18, № 2. С. 29–40.

**15. Кельманов А. В., Хандеев В. И.** Полиномиальный алгоритм с оценкой точности 2 для решения одной задачи кластерного анализа // Дискрет. анализ и исслед. операций. 2013. Т. 20, № 4. С. 36–45.

**16. Долгушев А. В., Кельманов А. В., Шенмайер В. В.** Полиномиальная аппроксимационная схема для одной задачи разбиения конечного множества на два кластера // Тр. Ин-та математики и механики. 2015. Т. 21, № 3. С. 100–109.

**17. Кельманов А. В., Пяткин А. В., Хандеев В. И.** NP-трудность квадратичной евклидовой задачи 2-кластеризации 1-mean и 1-median с ограничениями на размеры кластеров // Докл. Акад. наук. 2019. Т. 489, No. 4. С. 339–343.

**18. Pyatkin A. V.** 1-Mean and 1-medoid 2-clustering problem with arbitrary cluster sizes: Complexity and approximation // Yugoslav J. Oper. Res. 2023. V. 33, No. 1. P. 59–69.

**19. Шенмайер В. В.** Аппроксимационная схема для одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. 2012. Т. 19, № 2. С. 93–101.

**20. Галашов А. Е., Кельманов А. В.** 2-Приближённый алгоритм для одной задачи поиска семейства непересекающихся подмножеств векторов // Автоматика и телемеханика. 2014. № 4. С. 5–19.

**21. Edmonds J., Karp R. M.** Theoretical improvements in algorithmic efficiency for network flow problems // J. ACM. 1972. V. 19, No. 2. P. 248–264.

**22. Gabow H. N., Tarjan R. E.** Faster scaling algorithms for network problems // SIAM J. Comput. 1989. V. 18, No. 5. P. 1013–1036.

**23. Wirth H.** Algorithms + data structures = programs. Englewood Cliffs, NJ: Prentice Hall, 1976.

*Пяткин Артём Валерьевич*

Статья поступила
7 февраля 2023 г.

После доработки —
24 апреля 2023 г.

Принята к публикации
25 апреля 2023 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
July–September 2023. Vol. 30, No. 3. P. 96–110

UDC 519.8+518.25 DOI 10.33048/daio.2023.30.763

# PTAS FOR PROBLEMS OF VECTOR CHOICE AND CLUSTERING WITH DIFFERENT CENTERS

*A. V. Pyatkin*

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

**Abstract.** Two close in statements problems are considered. The first one is clustering, i. e. partitioning the set of $d$-dimensional Euclidean vectors into the given number of clusters with different types of centers so that the total dispersion would be minimum. By dispersion here we mean the sum of squared distances between the elements of the clusters and their centers. There are three types of centers: an arbitrary point (clearly, the centroid is the best choice), a point of the initial set (so-called medoid) or a fixed point of the space given in advance. The sizes of the clusters are also given as a part of the input. The second problem is the vector subset choice problem, which is finding a subset of vectors of fixed cardinality having the minimum sum of squared distances between its elements and the centroid. For each of these problems a PTAS is constructed. Bibliogr. 23.

**Keywords:** clustering, centroid, medoid, approximation, PTAS, MSSC.

## REFERENCES

1. **P. Berkhin,** A survey of clustering data mining techniques, in *Grouping Multidimensional Data: Recent Advances in Clustering* (Springer, Heidelberg, 2006), pp. 25–71.
2. **A. K. Jain** and **R. C. Dubes,** *Algorithms for Clustering Data* (Prentice Hall, Englewood Cliffs, NJ, 1988).
3. **S. Ghoreyshi** and **J. Hosseinkhani,** Developing a clustering model based on $K$-means algorithm in order to creating different policies for policyholders, *Int. J. Adv. Comput. Sci. Inf. Tech.* **4** (2), 46–53 (2015).

This research is carried out within the framework of the state contract of the Sobolev Institute of Mathematics (Project FWNF–2022–0019).

English version: Journal of Applied and Industrial Mathematics **17** (3) (2023).

4. **W. D. Fisher,** On grouping for maximum homogeneity, *J. Am. Stat. Assoc.* **53** (284), 789–798 (1958).
5. **J. MacQueen,** Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability, Berkeley, USA, June 21 — July 18, 1965; Dec. 27, 1965 — Jan. 7, 1966,* Vol. 1 (Univ. California Press, Berkeley, 1967), pp. 281–297.
6. **L. Kaufman** and **P. J. Rousseeuw,** Clustering by means of medoids, in *Statistical Data Analysis Based on the $L_1$-Norm* (North-Holland, Amsterdam, 1987), pp. 405–416.
7. **D. Aloise, A. Deshpande, P. Hansen,** and **P. Popat,** NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.* **75** (2), 245–248 (2009).
8. **M. Inaba, N. Katoh,** and **H. Imai,** Applications of weighted Voronoi diagrams and randomization to variance-based $k$-clustering, in *Proc. 10th Annu. Symp. Computational Geometry, Stony Brook, NY, USA, June 6–8, 1994* (ACM Press, New York, 1994), pp. 332–339.
9. **A. Kumar, Y. Sabharwal,** and **S. Sen,** A simple linear time $(1 + \varepsilon)$-approximation algorithm for geometric $k$-means clustering in any dimensions, in *Proc. 45th Annu. IEEE Symp. Foundations of Computer Science, Rome, Italy, Oct. 17–19, 2004* (IEEE Comp. Soc., Los Alamitos, CA, 2004), pp. 454–462.
10. **A. E. Baburin, É. Kh. Gimadi, N. I. Glebov,** and **A. V. Pyatkin,** The problem of finding a subset of vectors with the maximum total weight, *Diskretn. Anal. Issled. Oper., Ser. 2,* **14** (1), 32–42 (2007) [Russian] [*J. Appl. Ind. Math.* **2** (1), 32–38 (2008)].
11. **É. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova,** and **S. A. Khamidullin,** A posteriori detection of a quasiperiodic fragment with a given number of repetitions in a numerical sequence, *Sib. Zh. Ind. Mat.* **9** (1), 55–74 (2006) [Russian].
12. **A. V. Kel'manov** and **A. V. Pyatkin,** On the complexity of a search for a subset of «similar» vectors, *Dokl. Akad. Nauk* **421** (5), 590–592 (2008) [Russian] [*Dokl. Math.* **78** (1), 574–575 (2008)].
13. **A. V. Kel'manov** and **A. V. Pyatkin,** On a version of the problem of choosing a vector subset, *Diskretn. Anal. Issled. Oper.* **15** (5), 20–34 (2008) [Russian] [*J. Appl. Ind. Math.* **3** (4), 447–455 (2009)].
14. **A. V. Dolgushev** and **A. V. Kel'manov,** An approximation algorithm for solving a problem of cluster analysis, *Diskretn. Anal. Issled. Oper.* **18** (2), 29–40 (2011) [Russian] [*J. Appl. Ind. Math.* **5** (4), 551–558 (2011)].
15. **A. V. Kel'manov** and **V. I. Khandeev,** A 2-approximation polynomial algorithm for a clustering problem, *Diskretn. Anal. Issled. Oper.* **20** (4), 36–45 (2013) [Russian] [*J. Appl. Ind. Math.* **7** (4), 515–521 (2013)].
16. **A. V. Dolgushev, A. V. Kel'manov,** and **V. V. Shenmaier,** A polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters, *Tr. Inst. Mat. Mekh.* **21** (3), 100–109 (2015) [Russian] [*Proc. Steklov Inst. Math.* **295** (Suppl. 1), 47–56 (2016)].

**17.** **A. V. Kel'manov, A. V. Pyatkin,** and **V. I. Khandeev,** NP-hardness of quadratic Euclidean 1-mean and 1-median 2-clustering problem with constraints on the cluster sizes, *Dokl. Akad. Nauk* **489** (4), 339–343 (2019) [Russian] [*Dokl. Math.* **100** (3), 545–548 (2019)].

**18.** **A. V. Pyatkin,** 1-Mean and 1-medoid 2-clustering problem with arbitrary cluster sizes: Complexity and approximation, *Yugoslav J. Oper. Res.* **33** (1), 59–69 (2023).

**19.** **V. V. Shenmaier,** An approximation scheme for a problem of search for a vector subset, *Diskretn. Anal. Issled. Oper.* **19** (2), 93–101 (2012) [Russian] [*J. Appl. Ind. Math.* **6** (3), 381–386 (2012)].

**20.** **A. E. Galashov** and **A. V. Kel'manov,** A 2-approximate algorithm to solve one problem of a family of disjoint vector subsets, *Avtom. Telemekh.,* No. 4, 5–19 (2014) [Russian] [*Autom. Remote Control* **75** (4), 595–606 (2014)].

**21.** **J. Edmonds** and **R. M. Karp,** Theoretical improvements in algorithmic efficiency for network flow problems, *J. ACM* **19** (2), 248–264 (1972).

**22.** **H. N. Gabow** and **R. E. Tarjan,** Faster scaling algorithms for network problems, *SIAM J. Comput.* **18** (5), 1013–1036 (1989).

**23.** **H. Wirth,** *Algorithms + Data Structures = Programs* (Prentice Hall, Englewood Cliffs, NJ, 1976).

*Artem V. Pyatkin*

Received February 7, 2023
Revised April 24, 2023
Accepted April 25, 2023