ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль—сентябрь 2016. Том 23, № 3. С. 21–34

УДК 519.16 + 519.85 DOI: 10.17377/daio.2016.23.520

ТОЧНЫЕ ПСЕВДОПОЛИНОМИАЛЬНЫЕ АЛГОРИТМЫ ДЛЯ ЗАДАЧИ СБАЛАНСИРОВАННОЙ 2-КЛАСТЕРИЗАЦИИ *)

А. В. Кельманов[1,2], *А. В. Моткова*[2]
[1]Институт математики им. С. Л. Соболева СО РАН,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия,
[2]Новосибирский гос. университет, ул. Пирогова, 2,
630090 Новосибирск, Россия

e-mail: kelm@math.nsc.ru, anitamo@mail.ru

Аннотация. Рассматривается NP-трудная в сильном смысле задача двухкластерного разбиения конечного множества точек евклидова пространства по критерию минимума суммы по обоим кластерам взвешенных сумм квадратов внутрикластерных расстояний от элементов кластеров до их центров. Весами сумм являются мощности искомых кластеров. Центр одного из кластеров задан на входе, а центр другого неизвестен и определяется как точка пространства, равная среднему значению элементов кластера (геометрический центр). Анализируются два варианта задачи, в которых мощности кластеров либо неизвестны, либо заданы на входе. Для случая задач, в которых входные данные целочисленны, а размерность пространства фиксирована, построены точные псевдополиномиальные алгоритмы. Библиогр. 24.

Ключевые слова: евклидово пространство, сбалансированная кластеризация, NP-трудность, целочисленный вход, фиксированная размерность пространства, точный псевдополиномиальный алгоритм.

Введение

Предметом исследования является одна из слабоизученных NP-трудных в сильном смысле квадратичных задач разбиения конечного множества точек евклидова пространства на два кластера. Цель исследования — обоснование точных псевдополиномиальных алгоритмов для специального случая двух вариантов этой задачи, в которых мощности кластеров либо неизвестны, либо заданы на входе.

*) Исследование выполнено при финансовой поддержке Российского Научного Фонда (проект 16–11–10041).

Исследование мотивировано отсутствием каких-либо эффективных алгоритмических результатов для рассматриваемой задачи и её актуальностью для многих приложений, среди которых, в частности, проблемы геометрии, статистические проблемы совместного оценивания и проверки гипотез по неоднородным выборкам, проблемы кластерного анализа данных, проблемы интерпретации данных (см., например, [7, 8] и цитированные там работы).

1. Формулировка задач, известные и полученные результаты

Всюду далее $\mathbb{R}$ — множество вещественных чиссел, $\|\cdot\|$ — евклидова норма в пространстве $\mathbb{R}^q$, а $\langle\cdot,\cdot\rangle$ — скалярное произведение.

Следуя [7, 8], сформулируем общую задачу, имеющую название Balanced Variance-based 2-Clustering with given center, в двух вариантах: без ограничений на мощности кластеров (далее — задача 1), с дополнительным ограничением на мощность кластеров (далее — задача 2).

Задача 1. Дано: *множество* $\mathcal{Y} = \{y_1, \ldots, y_N\}$ *точек из* $\mathbb{R}^q$.

Найти: *разбиение множества* $\mathcal{Y}$ *на два непустых кластера* $\mathcal{C}$ *и* $\mathcal{Y} \setminus \mathcal{C}$ *такое, что*

$$F(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min, \qquad (1)$$

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ *— геометрический центр (центроид) кластера* $\mathcal{C}$.

Задача 2. Дано: *множество* $\mathcal{Y} = \{y_1, \ldots, y_N\}$ *точек из* $\mathbb{R}^q$ *и натуральное число* M.

Найти: *разбиение множества* $\mathcal{Y}$ *на два непустых кластера* $\mathcal{C}$ *и* $\mathcal{Y} \setminus \mathcal{C}$ *такое, что целевая функция* (1) *минимальна, при ограничении* $|\mathcal{C}| = M$.

В сформулированных задачах центроид $\overline{y}(\mathcal{C})$ кластера $\mathcal{C}$ неизвестен (является оптимизируемой переменной), а центр кластера $\mathcal{Y} \setminus \mathcal{C}$ задан в начале координат. В формуле (1) мощности искомых кластеров являются множителями внутрикластерных сумм и интерпретируются как веса, обеспечивающие сбалансированную кластеризацию.

В [7, 8] установлено, что задачи 1 и 2 NP-трудны в сильном смысле. Поэтому согласно [19] для них не существует точных полиномиального и псевдополиномиального алгоритмов, если гипотеза P≠NP верна. Кроме того, в [7, 8] доказано, что для задач 1 и 2 не существует полностью полиномиальных приближённых схем, если P≠NP.

Подчеркнём, что каких-либо результатов алгоритмического плана для сформулированных задач ранее не было получено. Вместе с тем известен целый ряд результатов для близких в постановочном плане задач. Поскольку рассматриваемые задачи не эквивалентны известным задачам и не являются их частными случаями, сфокусируем внимание лишь на отличительных особенностях этих задач. Свойства алгоритмов, предложенных для этих задач, можно найти в цитируемых ниже работах.

К числу наиболее близких в постановочном плане относится труднорешаемая задача, в которой целевая функция отличается от (1) выражением для внутрикластерной суммы кластера $\mathcal{Y} \setminus \mathcal{C}$, а именно, под знаком суммы вместо $\|y\|^2$ фигурирует $\|y-\overline{y}(\mathcal{Y}\setminus\mathcal{C})\|^2$. В этой задаче, которую называют Balanced Variance-based 2-Clustering, центроиды обоих кластеров являются оптимизируемыми переменными. В силу известного равенства

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

справедливого для любого непустого множества $\mathcal{Z}$, задача минимизации суммы $\sum\limits_{x\in\mathcal{C}}\sum\limits_{z\in\mathcal{C}}\|x-z\|^2+\sum\limits_{x\in\mathcal{Y}\setminus\mathcal{C}}\sum\limits_{z\in\mathcal{Y}\setminus\mathcal{C}}\|x-z\|^2$ также близка по постановке к задаче 1. Её называют Min-Sum All-Pairs 2-Clustering, или Min-Sum 2-Clustering. Вопросы построения алгоритмов для этих задач исследовались, в частности, в [7, 8, 15–17, 21, 22, 24].

В последнее десятилетие активно исследовалась NP-трудная в сильном смысле задача Minimum Sum-of-Squares 2-Clustering with given center, выражение для целевой функции которой отличается от (1) отсутствием балансировочных множителей — мощностей искомых кластеров. Алгоритмические результаты для неё можно найти в [4–6, 11–13, 20] и цитированных там работах.

Близкой к рассматриваемым также является известная [18, 23] со времён Фишера труднорешаемая [14] задача Minimum Sum-of-Squares 2-Clustering, которую ещё называют 2-Means. В этой задаче, как и в задаче Balanced Variance-based 2-Clustering, центроиды обоих кластеров являются оптимизируемыми переменными, но балансировочные веса у внутрикластерных сумм отсутствуют. С исследованием этой задачи и её приложениями связаны тысячи публикаций.

В силу указанных отличий известные алгоритмические результаты для приведённых постановочно близких труднорешаемых задач не переносятся на задачи 1 и 2. Для этих задач требуются отдельные исследования.

В настоящей работе построены точные псевдополиномиальные алгоритмы для специального случая задач 1 и 2, в которых координаты точек входного множества целочисленны, а размерность пространства фиксирована. Алгоритмы находят оптимальное решение задачи 1 за время $\mathcal{O}(N^2(ND)^q)$, а задачи 2 — за время $\mathcal{O}(N(MD)^q)$, где D — максимальное абсолютное значение координат входных точек. Фактически, это первые алгоритмические результаты для рассматриваемых задач.

2. Псевдополиномиальные алгоритмы

Суть подхода к решению задачи 2 заключается в следующем. В области пространства, определяемой максимальным абсолютным значением координат входных точек, строится многомерная равномерная по каждой координате решётка (сетка) с рациональным шагом. Шаг решётки выбирается так, чтобы один из её узлов совпал с геометрическим центром одного из искомых кластеров. Для каждого узла построенной решётки решается задача минимизации вспомогательной целевой функции. В результате решения находится подмножество, доставляющее минимум этой функции. Найденное подмножество включается в семейство претендентов на решение исходной задачи. В качестве окончательного решения выбирается то подмножество из построенного семейства, для которого значение целевой функции исходной задачи минимально.

Подход к решению задачи 1 состоит в построении семейства решений задачи 2 для каждой допустимой мощности M кластера $\mathcal{C}$ и выборе в этом семействе наилучшего решения в смысле минимума целевой функции.

Этот по своей сути сеточный подход ранее успешно применялся в [1, 3, 9, 13] при построении точных псевдополиномиальных алгоритмов для задач кластеризации, сходных в постановочном плане с задачами 1 и 2. Настоящая работа демонстрирует результативность сеточного подхода к решению новой труднорешаемой задачи.

Для построения и обоснования алгоритмов сформулируем вспомогательные утверждения.

Лемма 1. *Для произвольной точки $x \in \mathbb{R}^q$ и непустого конечного множества $\mathcal{Z} \subset \mathbb{R}^q$ имеет место равенство*

$$\sum_{z \in \mathcal{Z}} \|z - x\|^2 = \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 + |\mathcal{Z}| \cdot \|x - \overline{z}\|^2,$$

где $\overline{z} = \frac{1}{|\mathcal{Z}|} \sum\limits_{z \in \mathcal{Z}} z$ — центроид множества $\mathcal{Z}$.

Утверждение леммы 1 относится к числу хорошо известных, а доказательство представлено во многих публикациях (см., например, [10]).

Лемма 2. *Пусть*

$$S(\mathcal{C}, x) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - x\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \quad \mathcal{C} \subseteq \mathcal{Y},\ x \in \mathbb{R}^q, \tag{2}$$

где $\mathcal{Y}$ *— входное множество точек задачи* 1. *Тогда*

$$S(\mathcal{C}, x) = F(\mathcal{C}) + |\mathcal{C}|^2 \|x - \overline{y}(\mathcal{C})\|^2. \tag{3}$$

Доказательство. Из леммы 1 для множества $\mathcal{Z} = \mathcal{C}$ и его центроида $\overline{y}(\mathcal{C})$ имеем

$$\sum_{y \in \mathcal{C}} \|y - x\|^2 = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{C}| \cdot \|x - \overline{y}(\mathcal{C})\|^2. \tag{4}$$

Подставляя (4) в (2), получим

$$\begin{aligned} S(\mathcal{C}, x) &= |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - x\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\ &= |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{C}|^2 \|x - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\ &= F(\mathcal{C}) + |\mathcal{C}|^2 \|x - \overline{y}(\mathcal{C})\|^2, \end{aligned}$$

что устанавливает справедливость (3). Лемма 2 доказана.

Всюду далее используем $f^x(y)$ для обозначения функции $f(x, y)$ при условии, что у этой функции аргумент x фиксирован; аналогичный смысл имеет обозначение $f^y(x)$.

Лемма 3. *Для условных минимумов функции* (2) *справедливы следующие утверждения*:

(i) *при любом непустом фиксированном подмножестве* $\mathcal{C} \subseteq \mathcal{Y}$ *минимум функции* $S^{\mathcal{C}}(x)$ *по* $x \in \mathbb{R}^q$ *достигается в точке* $x = \overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ *и равен* $F(\mathcal{C})$;

(ii) *если* $|\mathcal{C}| = M = \text{const}$, *то при любой фиксированной точке* $x \in \mathbb{R}^q$ *для минимума функции* $S^x(\mathcal{C})$ *по* $\mathcal{C} \subseteq \mathcal{Y}$ *имеет место равенство*

$$\arg\min_{\mathcal{C} \subseteq \mathcal{Y}} S^x(\mathcal{C}) = \arg\min_{\mathcal{C} \subseteq \mathcal{Y}} G^x(\mathcal{C}),$$

где $G^x(\mathcal{C}) = \sum\limits_{y \in \mathcal{C}} g^x(y)$,

$$g^x(y) = (2M - N)\|y\|^2 - 2M \langle y, x \rangle, \quad y \in \mathcal{Y}, \tag{5}$$

при этом

$$\min_{\mathcal{C} \subseteq \mathcal{Y}} G^x(\mathcal{C}) = \sum_{y \in \mathcal{B}^x} g^x(y), \tag{6}$$

а множество $\mathcal{B}^x$ *состоит из тех* M *точек множества* $\mathcal{Y}$*, в которых функция* $g^x(y)$ *имеет наименьшие значения.*

Доказательство. Справедливость утверждения (i) следует из леммы 2.

Так как $|\mathcal{Y}| = N$ и $|\mathcal{C}| = M$, справедливость утверждения (ii) вытекает из следующей цепочки равенств:

$$\begin{aligned}
S^x(\mathcal{C}) &= M \sum_{y \in \mathcal{C}} \|y - x\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\
&= M \sum_{y \in \mathcal{C}} \|y\|^2 + M^2 \|x\|^2 - 2M \sum_{y \in \mathcal{C}} \langle y, x \rangle + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\
&= (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + (2M - N) \sum_{y \in \mathcal{C}} \|y\|^2 - 2M \sum_{y \in \mathcal{C}} \langle y, x \rangle \\
&= (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + \sum_{y \in \mathcal{C}} g^x(y) \\
&= (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + G^x(\mathcal{C}).
\end{aligned}$$

Остаётся заметить, что в последних двух равенствах первые два слагаемых не зависят от $\mathcal{C}$. Формула (6) очевидна. Лемма 3 доказана.

Допустим теперь, что координаты точек из множества $\mathcal{Y}$ целочисленны. Положим

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1, \ldots, q\}} |(y)^j|, \tag{7}$$

где $(y)^j$ — j-я координата точки y. Определим множество

$$\mathcal{D} = \left\{ x \in \mathbb{R}^q \mid \ (x)^j = \frac{1}{M}(v)^j, \ (v)^j \in \mathbb{Z}, \ |(v)^j| \leqslant MD, \ j = 1, \ldots, q \right\} \tag{8}$$

— многомерную решётку с равномерным рациональным шагом, равным $1/M$, по каждой координате. Заметим, что $|\mathcal{D}| = (2MD + 1)^q$.

Лемма 4. *Пусть элементы множества $\mathcal{Y}$ имеют целочисленные компоненты из интервала $[-D, D]$. Тогда центроид любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M лежит в множестве $\mathcal{D}$.*

Доказательство. Действительно, по определению (8) j-я координата центроида $\overline{y}(\mathcal{C})$ любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M, равная $\frac{1}{M}\sum\limits_{y\in\mathcal{C}}(y)^j$, является рациональным числом, так как j-я координата любой из входных точек целочисленна. Остаётся заметить, что в соответствии с (8) шаг решётки по каждой координате равен $1/M$ и, кроме того,

$$|(\overline{y}(\mathcal{C}))^j| = \frac{1}{|\mathcal{C}|}\Big|\sum_{y\in\mathcal{C}}(y)^j\Big| = \frac{1}{M}\Big|\sum_{y\in C}(y)^j\Big| \leqslant \frac{MD}{M} = D, \quad j = 1, \ldots, q,$$

для любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M. Лемма 4 доказана.

Запишем алгоритм решения задачи 2 в пошаговой форме.

Алгоритм $\mathcal{A}$

Вход: множество $\mathcal{Y}$, натуральное число M.

Шаг 1. Найдём значение D по формуле (7); построим решётку $\mathcal{D}$ по формуле (8).

Шаг 2. Для каждого узла $x \in \mathcal{D}$ вычислим $g^x(z)$, $z \in \mathcal{Y}$, по формуле (5); найдём подмножество $\mathcal{B}^x \subseteq \mathcal{Y}$ с M наименьшими значениями $g^x(z)$, вычислим значение $S(\mathcal{B}^x, x)$ по формуле (2).

Шаг 3. В семействе $\{S(\mathcal{B}^x, x), x \in \mathcal{D}\}$, построенном на шаге 2, найдём узел $x_A = \arg\min\limits_{x\in\mathcal{D}} S(\mathcal{B}^x, x)$ и подмножество $\mathcal{B}^{x_A} \subseteq \mathcal{Y}$. Положим $\mathcal{C}_A = \mathcal{B}^{x_A}$.

Выход: множество $\mathcal{C}_A$.

Теорема 1. *Пусть выполнены условия леммы* 4. *Тогда алгоритм $\mathcal{A}$ находит оптимальное решение задачи* 2 *за время $\mathcal{O}(qN(2MD+1)^q)$.*

Доказательство. Пусть подмножество $\mathcal{C}^*$ — оптимальное решение задачи 2, $y^* = \overline{y}(\mathcal{C}^*)$ — центроид подмножества $\mathcal{C}^*$, а $\mathcal{C}_A$ — подмножество, найденное алгоритмом.

Согласно лемме 4 центроид любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M лежит в множестве $\mathcal{D}$. Поэтому центроид y^* лежит в этом же множестве и, следовательно, на шаге 2 при переборе элементов решётки он будет проанализирован алгоритмом.

Для центроида y^* и узла x_A решётки по определению шага 3 имеем неравенство

$$S(\mathcal{C}_A, x_A) \leqslant S(\mathcal{B}^{y^*}, y^*). \tag{9}$$

Кроме того, из первого утверждения леммы 3 следует оценка

$$F(\mathcal{C}_A) \leqslant S(\mathcal{C}_A, x_A), \tag{10}$$

а из второго — неравенство

$$S(\mathcal{B}^{y^*}, y^*) \leqslant F(\mathcal{C}^*). \tag{11}$$

Объединяя (9)–(11), получаем

$$F(\mathcal{C}_A) \leqslant F(\mathcal{C}^*).$$

С другой стороны, так как $\mathcal{C}_A$ — допустимое решение задачи, а $\mathcal{C}^*$ — оптимальное, справедливо обратное неравенство

$$F(\mathcal{C}^*) \leqslant F(\mathcal{C}_A).$$

Таким образом, $F(\mathcal{C}_A) = F(\mathcal{C}^*)$, т. е. $\mathcal{C}_A$ — оптимальное решение задачи 2.

Оценим временну́ю сложность алгоритма. На шаге 1 для отыскания значения D потребуется $\mathcal{O}(qN)$ операций, а для построения решётки $\mathcal{D}$ — $\mathcal{O}(q|\mathcal{D}|)$ операций.

На шаге 2 вычисление значений $g^x(z)$ выполняется за время $\mathcal{O}(qN)$. Для поиска M наименьших элементов в множестве из N элементов потребуется $\mathcal{O}(N)$ операций (например, с помощью алгоритма отыскания n-го наименьшего значения в неупорядоченном массиве [2]). Вычисление значения $S(\mathcal{B}^x, x)$ выполняется за время $\mathcal{O}(qN)$. Следовательно, для каждого узла x решётки $\mathcal{D}$ требуется $\mathcal{O}(qN)$ операций. Поскольку решётка содержит узлов $|\mathcal{D}|$, выполнение второго шага осуществляется за время $\mathcal{O}(qN|\mathcal{D}|)$.

Шаг 3 выполняется за $|\mathcal{D}| = (2MD+1)^q$ операций.

Суммируя затраты на всех шагах, находим, что временна́я сложность алгоритма есть величина $\mathcal{O}(qN(2MD+1)^q)$. Теорема 1 доказана.

Заметим, что $(2MD+1)^q \leqslant (3MD)^q = 3^q(MD)^q$, поэтому при фиксированной размерности q пространства время работы алгоритма равно $\mathcal{O}(N(MD)^q)$. Поскольку D — числовое значение, алгоритм $\mathcal{A}$ решения задачи 2 псевдополиномиален.

Ясно, что построенный алгоритм $\mathcal{A}$ можно применить для отыскания точного решения задачи 1. Для этого достаточно найти N решений задачи 2 для каждого $M = 1, \ldots, N$ и среди найденных решений выбрать наилучшее в смысле минимума целевой функции. Трудоёмкость такого алгоритма, очевидно, равна $\mathcal{O}(N^2(ND)^q)$.

Заключение

В работе построены точные псевдополиномиальные алгоритмы для одной из актуальных NP-трудных в сильном смысле квадратичных евклидовых задач сбалансированной 2-кластеризации в случае, когда входные данные целочисленны, а размерность пространства фиксирована.

Поскольку рассмотренная задача относится к числу слабо изученных в алгоритмическом плане, делом ближайшей перспективы является построение эффективных приближённых алгоритмов с гарантированными оценками точности.

ЛИТЕРАТУРА

1. **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.
2. **Вирт Н.** Алгоритмы + структуры данных = программы. М.: Мир, 1985. 360 р.
3. **Гимади Э. Х., Глазков Ю. В., Рыков И. А.** О двух задачах выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы размерности // Дискрет. анализ и исслед. опер. 2008. Т. 15, № 4. С. 30–43.
4. **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.
5. **Долгушев А. В., Кельманов А. В.** Приближённый алгоритм решения одной задачи кластерного анализа // Дискрет. анализ и исслед. опер. 2011. Т. 18, № 2. С. 29–40.
6. **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 УрО РАН. 2015. Т. 21, № 3. С. 100–109.
7. **Кельманов А. В., Пяткин А. В.** NP-трудность некоторых квадратичных евклидовых задач 2-кластеризации // Докл. АН. 2015. Т. 464, № 5. С. 535–538.
8. **Кельманов А. В., Пяткин А. В.** О сложности некоторых квадратичных евклидовых задач 2-кластеризации // Журн. вычисл. математики и мат. физики. 2016. Т. 56, № 3. С. 150–156.
9. **Кельманов А. В., Романченко С. М.** Псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. 2012. № 2. С. 156–162.

10. **Кельманов А. В., Романченко С. М.** FPTAS для одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. 2014. Т. 21, № 3. С. 41–52.
11. **Кельманов А. В., Хандеев В. И.** Полиномиальный алгоритм с оценкой точности 2 для решения одной задачи кластерного анализа // Дискрет. анализ и исслед. опер. 2013. Т. 20, № 4. С. 36–45.
12. **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.
13. **Кельманов А. В., Хандеев В. И.** Точный псевдополиномиальный алгоритм для одной задачи двухкластерного разбиения множества векторов // Дискрет. анализ и исслед. опер. 2015. Т. 22, № 3. С. 36–48.
14. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. 2009. Vol. 75, No. 2. P. 245–248.
15. **Brucker P.** On the complexity of clustering problems // Optimization and Operations Research. Proc. Workshop Held Univ. Bonn (Bonn, Germany, Oct. 2–8, 1977). Heidelberg: Springer-Verl., 1978. P. 45–54. (Lect. Notes Econom. Math. Systems; Vol. 157).
16. **De la Vega W. F., Karpinski M., Kenyon C., Rabani Y.** Polynomial time approximation schemes for metric min-sum clustering // Electron. Colloq. Comput. Complexity (ECCC), Report No. 25. Potsdam: Hasso-Plattner-Inst. Softwaresystemtechnik, 2002.
17. **De la Vega W. F., Kenyon C.** A randomized approximation scheme for metric Max-Cut // J. Comput. Syst. Sci. 2001. Vol. 63. P. 531–541.
18. **Fisher R. A.** Statistical methods and scientific inference. New York: Hafner Press, 1959. 350 p.
19. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. San Francisco: Freeman, 1979. 314 p.
20. **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 Recognit. Image Anal. 2008. Vol. 18, No. 1. P. 30–42.
21. **Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J.** Efficient algorithms for variance-based k-clustering // Proc. 1st Pac. Conf. Comput. Graphics Appl. (Seoul, Korea, Aug. 30 – Sept. 2, 1993). River Edge, NJ: World Scientific, 1993. Vol. 1. P. 75–89.
22. **Inaba M., Katoh N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering // Proc. 10th Symp. Comput. Geom. (Stony Brook, NY, USA, June 6—8, 1994). New York: ACM,1994. P. 332–339.
23. **Rao M.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. 1971. Vol. 66. P. 622–626.

24. Sahni S., Gonzalez T. P-complete approximation problems // J. ACM. 1976. Vol. 23. P. 555–566.

Кельманов Александр Васильевич,
Моткова Анна Владимировна

Статья поступила
25 мая 2016 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
July–August 2016. Volume 23, No. 3. P. 21–34

UDC 519.16 + 519.85 DOI: 10.17377/daio.2016.23.520

EXACT PSEUDOPOLINOMIAL ALGORITHMS FOR A BALANCED 2-CLUSTERING PROBLEM

A. V. Kel'manov[1,2], *A. V. Motkova*[2]
[1]Sobolev Institute of Mathematics,
4 Acad. Koptyug Ave., 630090 Novosibirsk, Russia
[2]Novosibirsk State University,
2 Pirogov St., 630090 Novosibirsk, Russia

e-mail: kelm@math.nsc.ru, anitamo@mail.ru

Abstract. We consider the strongly NP-hard problem of partitioning a set of Euclidean points into two clusters so as to minimize the sum (over both clusters) of the weighted sum of the squared intracluster distances from the elements of the clusters to their centers. The weights of sums are the sizes of the clusters. The center of one cluster is given as input, while the center of the other cluster is unknown and determined as the average value over all points in the cluster (the geometric center). The two versions of the problems are analyzed in which the cluster sizes are either parts of the input or optimization variables. We present and prove exact pseudopolynomial algorithms in the case of integer components of the input points and fixed space dimension. Bibliogr. 24.

Keywords: Euclidean space, balanced clustering, NP-hardness, integer inputs, fixed space dimension, exact pseudopolynomial algorithm.

REFERENCES

1. **A. E. Baburin, E. 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**, No. 1, 32–42, 2007. Translated in *J. Appl. Ind. Math.,* **2**, No. 1, 32–38, 2008.
2. **N. Wirth,** *Algorithms + Data Structures = Programs,* Prentice Hall, Upper Saddle River, USA, 1976. Translated under the title *Algoritmy + struktury dannykh = programmy,* Mir, Moscow, 1985.
3. **E. Kh. Gimadi, Yu. V. Glazkov,** and **I. A. Rykov,** On two problems of choosing some subset of vectors with integer coordinates that has maximum norm of the sum of elements in Euclidean space, *Diskretn. Anal. Issled. Oper.,* **15**, No. 4, 30–43, 2008. Translated in *J. Appl. Ind. Math.,* **3**, No. 3, 343–352, 2009.

4. **E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova,** and **S. A. Khamidullin,** A posteriori detecting a quasiperiodic fragment with a given number of repetitions in a numerical sequence, *Sib. Zh. Ind. Mat.,* **9**, No. 1, 55–74, 2006.
5. **A. V. Dolgushev** and **A. V. Kel'manov,** An approximation algorithm for solving a problem of cluster analysis, *Diskretn. Anal. Issled. Oper.,* **18**, No. 2, 29–40, 2011. Translated in *J. Appl. Ind. Math.,* **5**, No. 4, 551–558, 2011.
6. **A. V. Dolgushev, A. V. Kel'manov,** and **V. V. Shenmaier,** Polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters, *Tr. Inst. Mat. Mekh.,* **21**, No. 3, 100–109, 2015.
7. **A. V. Kel'manov** and **A. V. Pyatkin,** NP-hardness of some quadratic Euclidean 2-clustering problems, *Dokl. Akad. Nauk,* **464**, No. 5, 535–538, 2015. Translated in *Dokl. Math.,* **92**, No. 2, 634–637, 2015.
8. **A. V. Kel'manov** and **A. V. Pyatkin,** On the complexity of some quadratic Euclidean 2-clustering problems, *Zh. Vychisl. Mat. Mat. Fiz.,* **56**, No. 3, 150–156, 2016. Translated in *Comput. Math. Math. Phys.,* **56**, No. 3, 491–497, 2016.
9. **A. V. Kel'manov** and **S. M. Romanchenko,** Pseudopolynomial algorithms for certain computationally hard vector subset and cluster analysis problems, *Avtom. Telemekh.,* No. 2, 156–162, 2012. Translated in *Autom. Remote Control,* **73**, No. 2, 349–354, 2012.
10. **A. V. Kel'manov** and **S. M. Romanchenko,** An FPTAS for a vector subset search problem, *Diskretn. Anal. Issled. Oper.,* **21**, No. 3, 41–52, 2014. Translated in *J. Appl. Ind. Math.,* **8**, No. 3, 329–336, 2014.
11. **A. V. Kel'manov** and **V. I. Khandeev,** A 2-approximation polynomial algorithm for a clustering problem, *Diskretn. Anal. Issled. Oper.,* **20**, No. 4, 36–45, 2013. Translated in *J. Appl. Ind. Math.,* **7**, No. 4, 515–521, 2013.
12. **A. V. Kel'manov** and **V. I. Khandeev,** A randomized algorithm for two-cluster partition of a set of vectors, *Zh. Vychisl. Mat. Mat. Fiz.,* **55**, No. 2, 335–344, 2015. Translated in *Comput. Math. Math. Phys.,* **55**, No. 2, 330–339, 2015.
13. **A. V. Kel'manov** and **V. I. Khandeev,** An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors, *Diskretn. Anal. Issled. Oper.,* **22**, No. 3, 36–48, 2015. Translated in *J. Appl. Ind. Math.,* **9**, No. 4, 497–502, 2015.
14. **D. Aloise, A. Deshpande, P. Hansen,** and **P. Popat,** NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.,* **75**, No. 2, 245–248, 2009.
15. **P. Brucker,** On the complexity of clustering problems, in *Optimization and Operations Research* (Proc. Workshop Held Univ. Bonn, Bonn, Germany, Oct. 2–8, 1977), pp. 45–54, Springer-Verlag, Heidelberg, 1978 (Lect. Notes Econ. Math. Syst., Vol. 157).

16. **W. F. de la Vega, M. Karpinski, C. Kenyon,** and **Y. Rabani,** Polynomial time approximation schemes for metric Min-Sum clustering, in *Electron. Colloq. Comput. Complexity,* Report No. 25, Hasso-Plattner-Institut Softwaresystemtechnik, Potsdam, 2002.
17. **W. F. de la Vega** and **C. Kenyon,** A randomized approximation scheme for metric Max-Cut, *J. Comput. Syst. Sci.,* **63**, 531–541, 2001.
18. **R. A. Fisher,** *Statistical methods and scientific inference*, Hafner Press, New York, 1959.
19. **M. R. Garey** and **D. S. Johnson,** *Computers and Intractability: A Guide to the Theory of NP-Completeness,* Freeman, San Francisco, 1979. Translated under the title *Vychislitel'nye mashiny i trudnoreshaemye zadachi,* Mir, Moscow, 1982.
20. **E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova,** and **S. A. Khamidullin,** A posteriori detecting a quasiperiodic fragment in a numerical sequence, *Pattern Recognit. Image Anal.,* **18**, No. 1, 30–42, 2008.
21. **S. Hasegawa, H. Imai, M. Inaba, N. Katoh,** and **J. Nakano,** Efficient algorithms for variance-based k-clustering, in *Computer Graphics and Applications* (Proc. 1st Pac. Conf. Comput. Gr. Appl., Seoul, Korea, Aug. 30 – Sept. 2, 1993), pp. 75–89, World Scientific, River Edge, NJ, USA, 1993.
22. **M. Inaba, N. Katoh,** and **H. Imai,** Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering, in *Proc. 10th Symp. Comput. Geom., Stony Brook, NY, USA, June 6–8, 1994,* pp. 332–339, ACM, New York, 1994.
23. **M. R. Rao,** Cluster analysis and mathematical programming, *J. Am. Stat. Assoc.,* **66**, 622–626, 1971.
24. **S. Sahni** and **T. Gonzalez,** P-complete approximation problems, *J. ACM,* **23**, 555–566, 1976.

Alexander V. Kel'manov,
Anna V. Motkova

Received
25 May 2016