ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль—август 2015. Том 22, № 4. С. 50–62

УДК 519.16+519.85 DOI: 10.17377/daio.2015.22.463

ТОЧНЫЙ ПСЕВДОПОЛИНОМИАЛЬНЫЙ АЛГОРИТМ ДЛЯ ОДНОЙ ЗАДАЧИ ДВУХКЛАСТЕРНОГО РАЗБИЕНИЯ МНОЖЕСТВА ВЕКТОРОВ *)

А. В. Кельманов[1,2], *В. И. Хандеев*[1]
[1]Институт математики им. С. Л. Соболева,
пр. Коптюга, 4, 630090 Новосибирск, Россия
[2]Новосибирский гос. университет,
ул. Пирогова, 2, 630090 Новосибирск, Россия
e-mail: kelm@math.nsc.ru, khandeev@math.nsc.ru

Аннотация. Рассматривается евклидова NP-трудная в сильном смысле задача разбиения конечного множества векторов на два кластера заданных размеров по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров. Предполагается, что центр одного из искомых кластеров неизвестен и определяется как среднее значение по всем векторам, образующим этот кластер. Центр другого кластера задан в начале координат. Показано, что в случае фиксированной размерности пространства задача разрешима за полиномиальное время. Для случая фиксированной размерности пространства и целочисленных компонент векторов обоснован точный псевдополиномиальный алгоритм. Библиогр. 27.

Ключевые слова: разбиение, множество векторов, квадраты евклидовых расстояний, NP-трудность, точный псевдополиномиальный алгоритм.

Введение

Предметом исследования настоящей работы является труднорешаемая задача дискретной оптимизации. Цель работы — обоснование точного псевдополиномиального алгоритма для специального случая этой задачи.

Одной из известных труднорешаемых задач дискретной оптимизации является задача MSSC (Minimum Sum-of-Squares Clustering) [21, 26, 27]. В этой задаче требуется разбить заданное конечное множество векторов

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 15–01–00462 и 13–07–00070).

евклидова пространства на совокупность кластеров так, чтобы минимизировать по всем кластерам сумму внутрикластерных сумм квадратов расстояний от элементов кластеров до их центров. Под центром кластера понимается его геометрический центр, т. е. среднее значение векторов, входящих в этот кластер. Эта задача возникает, в частности, при решении проблем аппроксимации, анализа данных, распознавания образов, компьютерной геометрии, математической статистики. Вопросы построения алгоритмических решений задачи MSSC рассматриваются в сотнях публикаций (см., например, [3, 7, 21, 23–27] и цитированные там работы). Долгое время эта задача считалась NP-трудной. Однако корректное доказательство факта её труднорешаемости получено недавно [20] и инициировало исследования задач, близких к ней в постановочном плане (см., например, [1, 3, 8–18] и цитированные там работы). К их числу относится задача, рассматриваемая в настоящей работе.

Содержательная проблема, которая приводит к исследуемой задаче, состоит в следующем (см., например, [8, 9, 15]). Имеется таблица с результатами измерения набора характеристик некоторого объекта. Объект может находиться в одном из двух состояний: пассивном (значения всех характеристик равны нулю) и активном (значение хотя бы одной характеристики не равно нулю). При этом в каждом результате измерения имеется ошибка, а соответствие между результатом измерения и состоянием объекта неизвестно. Требуется найти подмножество наборов, соответствующих активному состоянию объекта, и оценить набор характеристик объекта в этом состоянии.

Сильная NP-трудность задачи, которая индуцируется сформулированной проблемой, установлена в [16]. Приближённые полиномиальные алгоритмы с оценками точности предложены в [8, 9, 19]. Характеристики этих алгоритмов приведены в разд. 1. Здесь лишь отметим, что из сильной NP-трудности задачи следует [22], что для неё не существует точных полиномиального и псевдополиномиального алгоритмов, если P≠NP. Поэтому представляет интерес поиск подклассов этой задачи, для которых возможно построение точных алгоритмов полиномиальной временно́й сложности.

В настоящей работе показана полиномиальная разрешимость задачи в случае, когда размерность пространства фиксирована. Для случая фиксированной размерности пространства и целочисленных компонент векторов построен точный псевдополиномиальный алгоритм.

Следует заметить, что ранее в [4] был предложен точный псевдополиномиальный алгоритм для специального случая задачи поиска в конеч-

ном множестве векторов подмножества векторов с максимальной нормой их суммы. Этот же случай — фиксированная размерность пространства и целочисленность входов — анализируется в настоящей работе. Поскольку рассматриваемая в настоящей работе задача на минимум полиномиально эквивалентна рассмотренной в [4] задаче на максимум (см. разд. 2), ясно, что с помощью алгоритма, предложенного в [4], может быть найдено точное решение для частного случая рассматриваемой задачи. В настоящей работе для построения точного алгоритмического решения применяется подход, отличный от предложенного в [4]. Этот подход оказался менее продуктивным в плане быстродействия построенного алгоритма (см. разд. 2). Тем не менее, он представляется важным как иной эффективный инструмент решения сходных в постановочном плане задач.

1. Формулировка задачи и известные результаты

Рассматриваемая задача имеет следующую формулировку [8, 9, 16].

Задача 1-MSSC-F. *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число M. *Найти:* разбиение множества $\mathcal{Y}$ на два кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ — центр кластера $\mathcal{C}$, при ограничении $|\mathcal{C}| = M$.

В этой задаче входное множество $\mathcal{Y}$ соответствует таблице (см. содержательную трактовку во введении), содержащей N результатов измерения набора из q числовых характеристик некоторого объекта. Центр подмножества $\mathcal{C}$ неизвестен (является оптимизируемой переменной), а центр подмножества $\mathcal{Y} \setminus \mathcal{C}$ задан в начале координат. Число M соответствует числу измерений объекта в активном состоянии.

Задача 1-MSSC-F является частным, но NP-трудным случаем задачи J-MSSC-F (когда $J = 1$) [15]. В задаче J-MSSC-F требуется разбить входное множество на $J + 1$ кластер, причём центры J кластеров неизвестны, а центр одного из кластеров задан в начале координат [15]. Символы MSSC в кратком названии сформулированной задачи подчеркивают содержательное сходство с классической NP-трудной задачей MSSC, в которой центры всех искомых непересекающихся кластеров неизвестны. Последний символ F (от английского Fixed) указывает на то, что мощности кластеров фиксированы.

В [8] для задачи 1-MSSC-F построен 2-приближённый полиномиальный алгоритм, который находит решение за время $\mathcal{O}(qN^2)$. Полиномиальная приближённая схема (PTAS), имеющая временну́ю сложность $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — гарантированная оценка относительной погрешности, предложена в [9]. В [19] обоснован рандомизированный алгоритм, который при заданной относительной погрешности $\varepsilon > 0$ и фиксированной вероятности $\gamma \in (0,1)$ несрабатывания алгоритма для установленного значения параметра k находит $(1+\varepsilon)$-приближённое решение задачи за время $\mathcal{O}(2^k q(k+N))$, а также установлены условия, при которых алгоритм асимптотически точен и имеет трудоёмкость $\mathcal{O}(qN^2)$.

Ниже для случая, когда компоненты векторов целочисленны, предложен точный псевдополиномиальный алгоритм, который при фиксированной размерности пространства находит решение задачи за время $\mathcal{O}(N(MD)^q)$.

2. Точный псевдополиномиальный алгоритм

Покажем сначала, что при фиксированной размерности пространства задача 1-MSSC-F разрешима за полиномиальное время. Для этого нам потребуется следующая NP-трудная в сильном смысле [2, 5]

Задача LVS (Longest Vector Sum). *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число M. *Найти:* подмножество $\mathcal{C} \subseteq \mathcal{Y}$ мощности M такое, что $F(\mathcal{C}) = \big\| \sum\limits_{y \in \mathcal{C}} y \big\| \to \max$.

В [6] установлено, что задача LVS разрешима за время $\mathcal{O}(q^2 N^{2q})$, полиномиальное при фиксированной размерности q пространства.

Утверждение 1. *Задача* 1-MSSC-F *разрешима за время* $\mathcal{O}(q^2 N^{2q})$.

Доказательство. Цепочка равенств

$$S(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2$$
$$= \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} \Big\| \sum_{y \in \mathcal{C}} y \Big\|^2 = \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} (F(\mathcal{C}))^2$$

устанавливает связь между целевыми функциями задач 1-MSSC-F и LVS. Поскольку первый член в правой части полученного выражения — константа, а мощность множества $\mathcal{C}$ задана, минимизация $S(\mathcal{C})$ эквивалентна максимизации $F(\mathcal{C})$. Поэтому точный алгоритм решения задачи LVS можно применить для поиска оптимального решения задачи 1-MSSC-F. Утверждение 1 доказано.

Суть предлагаемого алгоритмического решения состоит в следующем. В области пространства, определяемой максимальным абсолютным значением координат входных векторов, строится многомерная равномерная по каждой координате сетка (решётка) с рациональным шагом. Шаг сетки выбирается так, чтобы один из её узлов совпал с геометрическим центром одного из оптимизируемых кластеров. Для каждого узла построенной сетки решается задача максимизации вспомогательной целевой функции. В результате решения находится набор векторов, доставляющий максимум этой функции. Найденный набор объявляется претендентом на решение. В качестве окончательного решения выбирается тот набор, для которого значение целевой функции исходной задачи минимально.

Для построения алгоритма потребуется вспомогательное утверждение. Положим

$$G(\mathcal{B}, b) = \sum_{y \in \mathcal{B}} \|y - b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2, \tag{2}$$

где $\mathcal{B} \subseteq \mathcal{Y}$, $|\mathcal{B}| = M$, $b \in \mathbb{R}^q$.

Лемма 1. *При любом фиксированном подмножестве $\mathcal{B} \subseteq \mathcal{Y}$ минимум функционала* (2) *доставляется вектором $\overline{y}(\mathcal{B}) = \frac{1}{M} \sum\limits_{y \in \mathcal{B}} y$ и равен $S(\mathcal{B})$. При любом фиксированном векторе $b \in \mathbb{R}^q$ минимум функционала* (2) *достигается на множестве, состоящем из M векторов множества $\mathcal{Y}$, имеющих наибольшие проекции на вектор b.*

Доказательство. Справедливость первого утверждения легко проверить аналитически (с помощью дифференцирования). Справедливость второго утверждения вытекает из следующей цепочки равенств:

$$\begin{aligned} G(\mathcal{B}, b) &= \sum_{y \in \mathcal{B}} \|y - b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2 \\ &= \sum_{y \in \mathcal{B}} (\|y - b\|^2 - \|y\|^2) + \sum_{y \in \mathcal{Y}} \|y\|^2 = \sum_{y \in \mathcal{Y}} \|y\|^2 + M \cdot \|b\|^2 - 2 \sum_{y \in \mathcal{B}} \langle y, b \rangle. \end{aligned}$$

Остаётся заметить, что первые два слагаемых в правой части полученного выражения являются константами. Лемма 1 доказана.

Допустим теперь, что векторы из множества $\mathcal{Y}$ имееют целочисленные компоненты. Положим

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1, \ldots, q\}} |(y)^j|, \tag{3}$$

где $(y)^j$ — j-я компонента вектора y.

Определим множество (многомерную решётку с равномерным рациональным шагом)

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

Заметим, что $|\mathcal{D}| = (2MD + 1)^q$.

Сформулируем следующий алгоритм решения задачи 1-MSSC-F.

Алгоритм $\mathcal{A}$

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

Шаг 1. Для каждого вектора $b \in \mathcal{D}$ построим множество $\mathcal{B}(b)$, состоящее из M векторов множества $\mathcal{Y}$, имеющих наибольшие проекции на вектор b. Вычислим значение $G(\mathcal{B}(b), b)$.

Шаг 2. Найдём вектор $b_A = \arg\min\limits_{b \in \mathcal{D}} G(\mathcal{B}(b), b)$ и соответствующее ему подмножество $\mathcal{B}(b_A)$. В качестве решения задачи возьмём подмножество $\mathcal{C}_A = \mathcal{B}(b_A)$. Если решений несколько, то берём любое из них.

Выход алгоритма: множество $\mathcal{C}_A$.

Лемма 2. *Пусть в условиях задачи* 1-MSSC-F *векторы из множества $\mathcal{Y}$ имеют целочисленные компоненты из интервала $[-D, D]$, $\mathcal{C}^*$ — оптимальное решение задачи* 1-MSSC-F*, а $\mathcal{C}_A$ — множество, полученное в результате работы алгоритма $\mathcal{A}$. Тогда $S(\mathcal{C}^*) = S(\mathcal{C}_A)$.*

Доказательство. Из (3) следует, что центр $\overline{y}(\mathcal{C}) = \frac{1}{M} \sum\limits_{y \in \mathcal{C}} y$ любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M лежит в $\mathcal{D}$. Стало быть, и центр $y^* = \overline{y}(\mathcal{C}^*)$ оптимального подмножества $\mathcal{C}^*$ лежит в этом же множестве.

В силу шага 2 имеем

$$G(\mathcal{C}_A, b_A) \leqslant G(\mathcal{B}(y^*), y^*). \tag{4}$$

Из первого утверждения леммы 1 следует оценка

$$S(\mathcal{C}_A) \leqslant G(\mathcal{C}_A, b_A), \tag{5}$$

а из второго — равенство

$$G(\mathcal{B}(y^*), y^*) = S(\mathcal{C}^*). \tag{6}$$

Объединяя (4)–(6), получаем оценку $S(\mathcal{C}_A) \leqslant S(\mathcal{C}^*)$.

С другой стороны, так как множество $\mathcal{C}_A$ является допустимым решением задачи 1-MSSC-F, справедливо неравенство $S(\mathcal{C}^*) \leqslant S(\mathcal{C}_A)$, что устанавливает равенство значений $S(\mathcal{C}^*)$ и $S(\mathcal{C}_A)$. Лемма 2 доказана.

Теорема 1. *Если выполнены условия леммы* 2, *то алгоритм* $\mathcal{A}$ *находит оптимальное решение задачи* 1-MSSC-F *за время* $\mathcal{O}(qN(2MD+1)^q)$.

Доказательство. Оптимальность решения следует из леммы 2.

Оценим временну́ю сложность алгоритма. Шаг 1 выполняется $|\mathcal{D}|$ раз. При этом для каждого вектора $b \in \mathcal{D}$ вычисление проекций на этот вектор требует $\mathcal{O}(qN)$ операций, а выбор M векторов, имеющих наибольшие проекции — $\mathcal{O}(N)$ операций. Затраты на вычисление значения функции $G(\mathcal{B}(b), b)$ составляют $\mathcal{O}(qN)$ операций. Поскольку $|\mathcal{D}| = (2MD+1)^q$, трудоёмкость шага 1 оценивается величиной $\mathcal{O}(qN(2MD+1)^q)$. Шаг 2 — поиск наименьшего элемента — требует $\mathcal{O}((2MD+1)^q)$ операций. Таким образом, итоговая временна́я сложность алгоритма есть величина $\mathcal{O}(qN(2MD+1)^q)$. Теорема 1 доказана.

Покажем, что в случае фиксированной размерности пространства алгоритм псевдополиномиален. Действительно, поскольку $MD \geqslant \frac{1}{2}$, то

$$(2MD+1)^q = 2^q\left(MD+\frac{1}{2}\right)^q \leqslant 4^q(MD)^q.$$

Отсюда следует, что при указанных условиях время работы алгоритма оценивается величиной $\mathcal{O}(N(MD)^q)$.

Время работы известного алгоритма [6], гарантирующего оптимальное решение общего случая задачи, есть величина $\mathcal{O}(N^{2q})$. Поэтому при $MD < N^{2-\frac{1}{q}}$ предложенный для частного случая псевдополиномиальный алгоритм $\mathcal{A}$ более эффективен по сравнению с точным алгоритмом, ориентированным на общий случай.

Однако для этого же частного случая задачи LVS в [4] обоснован алгоритм, имеющий трудоёмкость $\mathcal{O}(NM(MD)^{q-1})$. Поскольку задачи LVS и 1-MSSC-F полиномиально эквивалентны, этот алгоритм гарантирует отыскание точного решения рассматриваемой в настоящей работе задачи в D раз быстрее, чем предложенный алгоритм $\mathcal{A}$. Тем не менее, изложенный подход к построению алгоритма может оказаться полезным как ещё один эффективный инструмент решения сходных в постановочном плане задач. В частности, этот — по своей сути сеточный — подход более привлекателен в плане распараллеливания алгоритма. Кроме того, этот подход может быть использован при построении полностью полиномиальной приближённой схемы (FPTAS) для частного случая задачи, в котором размерность пространства фиксирована.

Заключение

В работе рассмотрена одна из актуальных NP-трудных в сильном смысле задач разбиения конечного множества векторов евклидова пространства. Установлено, что в случае фиксированной размерности пространства задача разрешима за полиномиальное время. Обоснован точный псевдополиномиальный алгоритм, позволяющий находить оптимальное решение задачи в случае, когда размерность пространства фиксирована, а компоненты векторов целочисленны.

Сеточный подход к построению алгоритмического решения, применённый в настоящей работе, по сути указывает следующий путь к построению схемы FPTAS для случая фиксированной размерности пространства — аппроксимация центра неизвестного кластера узлом специально построенной решетки. Обоснование такой схемы является важным направлением дальнейших исследований.

Ещё одним важным направлением представляется построение алгоритмов с оценками качества для обобщения рассмотренной задачи на случай, в котором число кластеров с неизвестными центрами больше единицы.

ЛИТЕРАТУРА

1. **Агеев А. А., Кельманов А. В., Пяткин А. В.** Труднорешаемость задачи о разрезе максимального веса в евклидовом пространстве // Докл. АН. 2014. Т. 456, № 5. С. 511–513.
2. **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.
3. **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.
4. **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. операций. 2008. Т. 15, № 4. С. 30–43.
5. **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.
6. **Гимади Э. Х., Пяткин А. В., Рыков И. А.** О полиномиальной разрешимости некоторых задач выбора подмножеств векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. операций. 2008. Т. 15, № 6. С. 11–19.

7. **Долгушев А. В., Кельманов А. В.** К вопросу об алгоритмической сложности одной задачи кластерного анализа // Дискрет. анализ и исслед. операций. 2010. Т. 17, № 2. С. 39–45.
8. **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.
9. **Долгушев А. В., Кельманов А. В., Шенмайер В. В.** Приближенная полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: Сб. докл. 9-й междунар. конф. (Республика Черногория, г. Будва, 16–22 сентября 2012 г.). М.: Торус Пресс, 2012. С. 242–244.
10. **Еремин И. И., Гимади Э. Х., Кельманов А. В., Пяткин А. В., Хачай М. Ю.** 2-Приближенный алгоритм поиска клики с минимальным весом вершин и ребер // Тр. Ин-та математики и механики УрО РАН. 2013. Т. 19, № 2. С. 134–143.
11. **Кельманов А. В.** Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // Тр. Ин-та математики и механики УрО РАН. 2008. Т. 14, № 2. С. 81–88.
12. **Кельманов А. В.** О сложности некоторых задач анализа данных // Журн. вычисл. математики и мат. физики. 2010. Т. 50, № 11. С. 2045–2051.
13. **Кельманов А. В.** О сложности некоторых задач кластерного анализа // Журн. вычисл. математики и мат. физики. 2011. Т. 51, № 11. С. 2106–2112.
14. **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.
15. **Кельманов А. В., Пяткин А. В.** О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. физики. 2009. Т. 49, № 11. С. 2059–2067.
16. **Кельманов А. В., Пяткин А. В.** О сложности некоторых задач кластерного анализа векторных последовательностей // Дискрет. анализ и исслед. операций. 2013. Т. 20, № 2. С. 47–57.
17. **Кельманов А. В., Романченко С. М., Хамидуллин С. А.** Точные псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подпоследовательности векторов // Журн. вычисл. математики и мат. физики. 2013. Т. 53, № 1. С. 143–153.
18. **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.
19. **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.
20. **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.

21. **Jain A. K.** Data clustering: 50 years beyond K-means // Pattern Recognit. Lett. 2010. Vol. 31, No. 8. P. 651–666.
22. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. San Francisco, CA: Freeman, 1979. 314 p.
23. **Hansen P., Jaumard B.** Cluster analysis and mathematical programming // Math. Program., Ser. A. 1997. Vol. 79. P. 191–215.
24. **Hansen P., Jaumard B., Mladenovich N.** Minimum sum of squares clustering in a low dimensional space // J. Classif. 1998. Vol. 15, No. 1. P. 37–55.
25. **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, June 6–8, 1994). New York: ACM, 1995. P. 332–339.
26. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Stat. Probab. Vol. 1. Berkeley, CA: Univ. California Press, 1967. P. 281–297.
27. **Rao M. R.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. 1971. Vol. 66. P. 622–626.

Кельманов Александр Васильевич,
Хандеев Владимир Ильич

Статья поступила
16 сентября 2014 г.
Исправленный вариант —
22 февраля 2015 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
July–August 2015. Volume 22, No. 4. P. 50–62

UDC 519.16+519.85 DOI: 10.17377/daio.2015.22.463

AN EXACT PSEUDOPOLYNOMIAL ALGORITHM FOR A BI-PARTITIONING PROBLEM

A. V. Kel'manov[1,2], *V. I. Khandeev*[1]
[1]Sobolev Institute of Mathematics,
4 Koptyug Ave., 630090 Novosibirsk, Russia
[2]Novosibirsk State University,
2 Pirogov St., 630090 Novosibirsk, Russia
e-mail: kelm@math.nsc.ru, khandeev@math.nsc.ru

Abstract. We consider the strongly NP-hard problem of partitioning a set of Euclidean vectors into two sets (clusters) under the criterion of minimum sum-of-squared distances from the elements of clusters to their centers. The center of the first cluster is the average value of the vectors in the cluster, and the center of the second one is the origin. We prove that the problem is solvable in polynomial time in the case of fixed space dimension. We also present a pseudopolynomial algorithm which finds an optimal solution in the case of integer values of the components of the vectors in the input set and fixed space dimension. Bibliogr. 27.

Keywords: bi-partitioning, vector subset, squared Euclidean distances, NP-hardness, exact pseudopolynomial algorithm.

REFERENCES

1. **A. A. Ageev, A. V. Kel'manov,** and **A. V. Pyatkin,** NP-hardness of the Euclidean max-cut problem, *Dokl. Akad. Nauk,* **456**, No. 5, 511–513, 2014. Translated in *Dokl. Math.,* **89**, No. 3, 343–345, 2014.
2. **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.
3. **A. E. Galashov** and **A. V. Kel'manov,** A 2-approximate algorithm to solve one problem of the family of disjoint vector subsets, *Avtom. Telemekh.,* No. 4, 5–19, 2014. Translated in *Autom. Remote Control,* **75**, No. 4, 595–606, 2014.
4. **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 an Euclidean space, *Diskretn. Anal. Issled. Oper.,* **15**, No. 4, 30–43, 2008. Translated in *J. Appl. Ind. Math.,* **3**, No. 3, 343–352, 2009.

5. **E. 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**, No. 1, 55–74, 2006.
6. **E. Kh. Gimadi, A. V. Pyatkin,** and **I. A. Rykov,** On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension, *Diskretn. Anal. Issled. Oper.,* **15**, No. 6, 11–19, 2008. Translated in *J. Appl. Ind. Math.,* **4**, No. 1, 48–53, 2010.
7. **A. V. Dolgushev** and **A. V. Kel'manov,** On the algorithmic complexity of a problem in cluster analysis, *Diskretn. Anal. Issled. Oper.,* **17**, No. 2, 39–45, 2010. Translated in *J. Appl. Ind. Math.,* **5**, No. 2, 191–194, 2011.
8. **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.
9. **A. V. Dolgushev, A. V. Kel'manov,** and **V. V. Shenmaier,** A polynomial approximation scheme for a problem of cluster analysis, in *Doklady 9-i mezhdunarodnoi konferentsii "Intellektualizatsiya obrabotki informatsii"* (Doklady 9th Int. Conf. "Intellectualization of Information Processing"), *Budva, Montenegro, Sept. 16–22, 2012,* pp. 242–244, Torus Press, Moscow, 2012.
10. **I. I. Eremin, E. Kh. Gimadi, A. V. Kel'manov, A. V. Pyatkin,** and **M. Yu. Khachai,** 2-Approximation algorithm for finding a clique with minimum weight of vertices and edges, *Tr. Inst. Mat. Mekh.,* **19**, No. 2, 134–143, 2013. Translated in *Proc. Steklov Inst. Math.,* **284**, Suppl. 1, S87–S95, 2014.
11. **A. V. Kel'manov,** Off-line detection of a quasi-periodically recurring fragment in a numerical sequence, *Tr. Inst. Mat. Mekh.,* **14**, No. 2, 81–88, 2008. Translated in *Proc. Steklov Inst. Math.,* **263**, Suppl. 2, S84–S92, 2008.
12. **A. V. Kel'manov,** On the complexity of some data analysis problems, *Zh. Vychisl. Mat. Mat. Fiz.,* **50**, No. 11, 2045–2051, 2010. Translated in *Comput. Math. Math. Phys.,* **50**, No. 11, 1941–1947, 2010.
13. **A. V. Kel'manov,** On the complexity of some cluster analysis problems, *Zh. Vychisl. Mat. Mat. Fiz.,* **51**, No. 11, 2106–2112, 2011. Translated in *Comput. Math. Math. Phys.,* **51**, No. 11, 1983–1988, 2011.
14. **A. V. Kel'manov** and **A. V. Pyatkin,** On the complexity of a search for a subset of "similar" vectors, *Dokl. Akad. Nauk,* **421**, No. 5, 590–592, 2008. Translated in *Dokl. Math.,* **78**, No. 1, 574–575, 2008.
15. **A. V. Kel'manov** and **A. V. Pyatkin,** Complexity of certain problems of searching for subsets of vectors and cluster analysis, *Zh. Vychisl. Mat. Mat. Fiz.,* **49**, No. 11, 2059–2067, 2009. Translated in *Comput. Math. Math. Phys.,* **49**, No. 11, 1966–1971, 2009.
16. **A. V. Kel'manov** and **A. V. Pyatkin,** On complexity of some problems of

cluster analysis of vector sequences, *Diskretn. Anal. Issled. Oper.,* **20**, No. 2, 47–57, 2013. Translated in *J. Appl. Ind. Math.,* **7**, No. 3, 363–369, 2013.

17. **A. V. Kel'manov, S. M. Romanchenko,** and **S. A. Khamidullin,** Accurate pseudopolynomial-time algorithms for some NP-hard problems of searching for a vector subsequence, *Zh. Vychisl. Mat. Mat. Fiz.,* **53**, No. 1, 143–153, 2013.
18. **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.
19. **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.
20. **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.
21. **A. K. Jain,** Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett.,* **31**, No. 8, 651–666, 2010.
22. **M. R. Garey** and **D. S. Johnson,** *Computers and Intractability: A Guide to the Theory of NP-Completeness,* Freeman, San Francisco, 1979.
23. **P. Hansen** and **B. Jaumard,** Cluster analysis and mathematical programming, *Math. Program., Ser. A,* **79**, No. 1–3, 191–215, 1997.
24. **P. Hansen, B. Jaumard,** and **N. Mladenovic,** Minimum sum of squares clustering in a low dimensional space, *J. Classif.,* **15**, No. 1, 37–55, 1998.
25. **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.
26. **J. B. MacQueen,** Some methods for classification and analysis of multivariate observations, in L. M. Le Cam and J. Neyman, eds., *Proc. 5th Berkeley Symp. Math. Stat. Probab., Berkeley, USA, June 21 – July 18, 1965 and Dec. 27, 1965 – Jan. 7, 1966,* Vol. 1, pp. 281–297, Univ. of California Press, Berkeley, 1967.
27. **M. R. Rao,** Cluster analysis and mathematical programming, *J. Am. Stat. Assoc.,* **66**, 622–626, 1971.

Alexander V. Kel'manov,
Vladimir I. Khandeev

Received
16 September 2014
Revised
22 February 2015