УДК 519.2+621.391

Точные алгоритмы поиска кластера наибольшего размера для двух целочисленных задач 2-кластеризации*

А.В. Кельманов[1,2], А.В. Панасенко[1,2], В.И. Хандеев[1,2]

[1]Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, просп. Акад. Коптюга, 4, Новосибирск, 630090

[2]Новосибирский национальный исследовательский государственный университет (НГУ), ул. Пирогова, 2, Новосибирск, 630090

E-mails: kelm@math.nsc.ru (Кельманов А.В.), a.v.panasenko@math.nsc.ru (Панасенко А.В.), khandeev@math.nsc.ru (Хандеев В.И.)

Кельманов А.В., Панасенко А.В., Хандеев В.И. Точные алгоритмы поиска кластера наибольшего размера для двух целочисленных задач 2-кластеризации // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск, 2019. — Т. 22, № 2. — С. 121–136.

Рассматриваются две родственные дискретные экстремальные задачи выбора (поиска) подмножества в конечном множестве точек евклидова пространства. Обе задачи индуцируются вариантами фундаментальной проблемы анализа данных — выбором в совокупности объектов подмножества похожих элементов. В обеих задачах требуется в заданном множестве точек найти кластер (подмножество) наибольшей мощности при ограничении на значение квадратичной кластеризационной функции. Совокупность точек входного множества вне искомого кластера соответствует второму (дополняющему) кластеру. В первой задаче кластеризационной функцией (из ограничения) является сумма по обоим кластерам внутрикластерных сумм квадратов расстояний между элементами кластеров и их центрами. Центр одного из кластеров неизвестен и определяется как центроид (геометрический центр), а центр другого фиксирован в заданной точке пространства (без ограничения общности в начале координат). Во второй задаче кластеризационной функцией является сумма по обоим кластерам мощностно-взвешенных внутрикластерных сумм квадратов расстояний между элементами кластеров и их центрами. Как и в первой задаче, центр одного из кластеров неизвестен и определяется как центроид, а центр другого фиксирован в начале координат. В настоящей работе установлено, что обе задачи NP-трудны в сильном смысле. Для вариантов задач, в которых точки входного множества имеют целочисленные координаты, предложены точные алгоритмы. Время работы алгоритмов псевдополиномиально, если размерность пространства ограничена константой.

DOI: 10.15372/SJNM20190201

Ключевые слова: евклидово пространство, 2-кластеризация, наибольшее подмножество, NP-трудность, целочисленная задача, псевдополиномиальная разрешимость.

Kel'manov A.V., Panasenko A.V., Khandeev V.I. Exact algorithms of searching for the largest size cluster in two integer 2-clustering problems // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. — Novosibirsk, 2019. — Vol. 22, № 2. — P. 121–136.

We consider two related discrete optimization problems of searching for a subset in a finite set of points in the Euclidean space. Both problems are induced by the versions of the fundamental problem in data analysis, namely, by selecting a subset of similar elements in a set of objects. In each problem, an input set and a positive real number are given, and it is required to find a cluster (i.e., a subset) of the largest size under

*Исследование задачи 1 поддержано грантом РНФ (проект № 16-11-10041). Исследование задачи 2 поддержано грантами РФФИ (проекты № 19-01-00308, 18-31-00398-мол-а), а также грантом РАН по программе фундаментальных научных исследований (проект № 0314-2016-0015) и Министерством образования и науки РФ в рамках программы 5-10.

constraints on the value of a quadratic clusterization function. The points in the input set which are outside the sought for subset are treated as the second (complementary) cluster. In the first problem, the function under the constraint is the sum over both clusters of the intracluster sums of the squared distances between the elements of the clusters and their centers. The center of the first (i.e., the sought) cluster is unknown and determined as the centroid, while the center of the second one is fixed at a given point in the Euclidean space (without loss of generality in the origin). In the second problem, the function under the constraint is the sum over both clusters of the weighted intracluster sums of the squared distances between the elements of the clusters and their centers. As in the first problem, the center of the first cluster is unknown and determined as the centroid, while the center of the second one is fixed in the origin. In this paper, we show that both problems are strongly NP-hard. Also, we present the exact algorithms for the cases of these problems in which the input points have integer components. If the space dimension is bounded by some constant, the algorithms are pseudopolynomial.

Keywords: *Euclidean space, 2-clustering, largest subset, NP-hardness, exact algorithm, pseudopolynomial-time solvability.*

Введение

Предметом исследования работы являются две близкие в постановочном плане оптимизационные задачи выбора (поиска) подмножества наибольшей мощности (размера) в конечном множестве точек евклидова пространства. Обе задачи моделируют одну из ключевых проблем анализа данных — выбор в конечном множестве объектов подмножества похожих элементов. Цель работы — анализ вычислительной сложности задач и построение алгоритмов с гарантированными оценками качества (точности и временной сложности), обеспечивающих эффективное решение этих задач.

Исследование мотивировано, с одной стороны, слабой изученностью задач в теоретическом плане, а именно — отсутствием опубликованных результатов об их статусе вычислительной сложности и отсутствием строго обоснованных алгоритмических решений. С другой стороны, исследование обусловлено важностью задач для ряда приложений (см. следующий пункт).

Статья имеет следующую структуру. В пункте 1 приведены формулировки задач, их интерпретации, близкие по постановке задачи, их отличительные черты и существующие алгоритмические результаты. В этом же пункте анонсированы результаты, полученные в настоящей работе. В следующем пункте анализируется вычислительная сложность задач. В п. 3 приведены вспомогательные результаты, необходимые для обоснования свойств предлагаемых алгоритмов. Наконец, п. 4 содержит пошаговую запись алгоритмов и обоснование их качественных свойств (точности и временной сложности).

1. Формулировки задач и их интерпретация, близкие по постановке задачи, известные и полученные результаты

Всюду далее: $\mathbb{R}$ — множество действительных чисел, d — размерность пространства, $\|\cdot\|$ — евклидова норма и $\langle\cdot,\cdot\rangle$ — скалярное произведение; *центроидом* (геометрическим центром) непустого конечного множества (кластера) $\mathcal{Y} \subset \mathbb{R}^d$ называется точка из $\mathbb{R}^d$, равная арифметическому среднему элементов из этого множества; *центром* кластера называется произвольная фиксированная точка $x \in \mathbb{R}^d$, относительно которой суммируются квадраты расстояний из этого кластера.

Первая из рассматриваемых задач близка (но не эквивалентна) по постановке к задаче *Minimum Sum-of-Squares Clustering* (MSSC), которая хорошо известна с прошлого века под другим названием k-Means [1–6]. Статус вычислительной сложности этой задачи анализировался в [7–10]. В [7] доказано, что даже простейший (базовый) двухкластерный вариант этой задачи — 2-MSSC (или 2-Means) — является NP-трудным в сильном смысле.

Напомним, что в 2-MSSC требуется найти 2-разбиение конечного множества $\mathcal{Y} \subset \mathbb{R}^d$, минимизирующее сумму

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

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ — центроиды непустых непересекающихся подмножеств $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ соответственно.

Вторая из рассматриваемых задач в постановочном плане близка (но не эквивалентна) к известной задаче *Quadratic Min-Sum All-Pairs* 2*-Clustering* [11–18], в которой требуется найти 2-разбиение конечного множества $\mathcal{Y} \subset \mathbb{R}^d$, минимизирующее сумму

$$\sum_{y \in \mathcal{C}} \sum_{x \in \mathcal{C}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \sum_{x \in \mathcal{Y} \setminus \mathcal{C}} \|y - x\|^2.$$

Задача *Quadratic Min-Sum All-Pairs* 2*-Clustering* эквивалентна задаче *Cardinality-Weighted Minimum Sum-of-Squares* 2*-Clustering* или, по-другому, задаче *Cardinality-Weighted* 2-MSSC, в которой требуется найти 2-разбиение конечного множества $\mathcal{Y} \subset \mathbb{R}^d$, минимизирующее сумму

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

где, как и в задаче 2-MSSC, $\overline{y}(\mathcal{C})$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ — центроиды непустых непересекающихся подмножеств $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ соответственно. Эквивалентность указанных задач следует из того, что для любого непустого конечного множества $\mathcal{Y} \subset \mathbb{R}^d$ справедливо хорошо известное равенство $2|\mathcal{Y}| \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2 = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{Y}} \|y - x\|^2$, которое связывает мощностно взвешенный суммарный квадратичный разброс точек из множества $\mathcal{Y}$ относительно его центроида $\overline{y}(\mathcal{Y})$ и сумму квадратов попарных расстояний между элементами из этого множества. Гипотеза [12] о труднорешаемости этих задач доказана в [19, 20] вместе с доказательством сильной NP-трудности квадратичной евклидовой задачи о разрезе максимального веса (*Quadratic Euclidean Max-Cut*).

В целевых функциях (1) и (2) обе внутрикластерные суммы являются суммарными квадратичными разбросами точек из кластеров относительно их центроидов, т. е. в этих задачах оба центра неизвестны и определяются как центроиды. В каждой из сформулированных ниже задач в одной из внутрикластерных сумм фигурирует суммарный квадратичный разброс точек из кластера относительно фиксированной (заданной) точки $x \in \mathbb{R}^d$, в качестве которой без ограничения общности рассматривается начало координат. Центр другого кластера полагается равным его центроиду. Иными словами, в этих задачах неизвестным является только один центр, что отличает эти задачи от отмеченных выше.

Задачи 2-разбиения при одном неизвестном (и одном заданном) центре имеют следующие формулировки.

Задача 1 (2-MSSC *with given center*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d. *Найти*: 2-разбиение $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

В этой задаче требуется минимизировать сумму по обоим кластерам внутрикластерных сумм квадратов расстояний между элементами кластеров и их центрами. Центр кластера $\mathcal{C}$ неизвестен и определяется как центроид, а центр кластера $\mathcal{Y} \setminus \mathcal{C}$ фиксирован в начале координат. В задаче 2-MSSC, в отличие от задачи 1, оба центра неизвестны и определяются как центроиды. Сильная NP-трудность задачи 1 доказана в [21, 22].

Задача 2 (*Cardinality-Weighted* 2-*MSSC with given center*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d. *Найти*: 2-разбиение $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, такое, что

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

В этой задаче требуется минимизировать сумму по обоим кластерам мощностно взвешенных внутрикластерных сумм квадратов расстояний между элементами кластеров и их центрами. Как и в задаче 1, центр кластера $\mathcal{C}$ неизвестен и определяется как центроид, а центр кластера $\mathcal{Y} \setminus \mathcal{C}$ фиксирован в начале координат. Сильная NP-трудность задачи 2 установлена в [23, 24].

В настоящей работе исследуются следующие две задачи, которые близки по постановке к задачам 1 и 2.

Задача 3. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и число $\alpha \in (0, 1)$. *Найти*: подмножество $\mathcal{C} \subset \mathcal{Y}$ наибольшей мощности такое, что

$$F(\mathcal{C}) \leq \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{3}$$

В этой задаче требуется найти кластер $\mathcal{C}$ наибольшей мощности при ограничении на значение целевой функции $F(\mathcal{C})$ задачи 1. Это ограничение определяется правой частью неравенства (3), т. е. долей суммарного квадратичного разброса точек входного множества $\mathcal{Y}$ относительно его центроида $\overline{y}(\mathcal{Y})$.

Задача 4. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и число $\alpha \in (0, 1)$. *Найти*: подмножество $\mathcal{C} \subset \mathcal{Y}$ наибольшей мощности такое, что

$$G(\mathcal{C}) \leq \alpha N \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{4}$$

В этой задаче требуется найти кластер $\mathcal{C}$ наибольшей мощности при ограничении на значение целевой функции $G(\mathcal{C})$ задачи 2. Это ограничение определяется правой частью неравенства (4), т. е. долей суммарного мощностно взвешенного разброса точек входного множества $\mathcal{Y}$ относительно его центроида $\overline{y}(\mathcal{Y})$.

Сходство задач 1, 2 с задачами 3, 4, соответственно, и их отличие раскрывает следующее замечание. В задачах 3 и 4 функции $F(\mathcal{C})$ и $G(\mathcal{C})$ не являются целевыми. Они лишь определяют ограничения этих задач. Если $\mathcal{C}^*$ — оптимальное решение задачи 1

(или 2), то $F(\mathcal{C}^*) \leq F(\mathcal{C})$ в задаче 1 (или $G(\mathcal{C}^*) \leq G(\mathcal{C})$ в задаче 2) для любого $\mathcal{C} \subset \mathcal{Y}$. Поэтому в задачах 3 и 4 неравенства (3) и (4) определяют подмножества допустимых решений задач 1 и 2. Это значит, что в каждой из задач 3 и 4 требуется найти кластер наибольшего размера в подмножестве допустимых решений задач 1 и 2 соответственно.

Все сформулированные выше экстремальные задачи легко трактуются как задачи аппроксимации, комбинаторной геометрии, теории графов и статистики. Они имеют приложения, в частности, в проблемах Data science, Data mining, Pattern recognition и Machine learning. В этих приложениях и областях исследований алгоритмы кластеризации являются ключевыми инструментами, обеспечивающими решение проблем компьютерного анализа данных (см., например, [2–6, 25–31] и цитированные там работы). Приведем лишь один аргумент, поясняющий актуальность рассматриваемых задач. Как известно, проблемы Data mining и классические проблемы математической статистики близки по смыслу: и там, и там основная цель — выяснение структурных свойств множеств (данных или выборок). В классической статистике анализируются однородные выборки, в то время как в Data mining выборочные (экспериментальные) данные существенно неоднородны.

Хорошо известной статистической задачей является проверка гипотезы о том, что среднее выборочное значение совпадает с заданным значением, против альтернативы не совпадает. Существует несколько классических критериев, ориентированных на решение этой задачи.

Что делать в ситуации, когда есть основания полагать, что выборка неоднородна, состоит из элементов двух распределений, а соответствие между элементами выборки и распределением неизвестно? Эта ситуация (с неоднородностью выборочных данных) типична для Data mining и, в частности, для проблемы Big data. Ясно, что для применения классических статистических методов потребуется решить задачу разбиения данных на однородные множества (выборки). Сформулированные выше экстремальные задачи моделируют всего лишь несколько подобных слабо изученных проблем. Опыт исследования экстремальных задач, индуцированных сходными проблемами, показывает, что большинство из них относится к классу труднорешаемых задач. Выяснение вопросов о сложностном статусе индуцированных задач и вопросов их алгоритмической аппроксимируемости является первоочередной математической задачей.

К настоящему времени сложностной статус рассматриваемых задач 3 и 4 не установлен и какие-либо алгоритмические результаты отсутствуют. Поэтому ниже приведены существующие результаты только для наиболее близких к ним задач, а именно, для задач 1 и 2.

Напомним опубликованные результаты для задачи 1. Ряд результатов был получен для варианта задачи 1, в котором заданы мощности кластеров. Ниже этот вариант называется 2-*MSSC with given center and cluster cardinalities* (в п. 4 — задача 5). Сильная NP-трудность этого варианта задачи следует непосредственно из сильной NP-трудности задачи 1 (без ограничений на мощности кластеров), доказанной в [21, 22]. Однако первоначально факт труднорешаемости был установлен в [33–35] до получения результатов из [21, 22].

Всюду далее при описании существующих алгоритмических решений символом M обозначена заданная на входе задачи мощность кластера с неизвестным центром, а через D — максимальное абсолютное значение координат точек входного множества.

Из [36] следует, что задача разрешима за время $\mathcal{O}(d^2N^{2d})$, которое полиномиально, если размерность d пространства фиксирована (ограничена сверху константой). Ускоренный точный алгоритм с трудоемкостью $\mathcal{O}(dN^{d+1})$ предложен в [37]. Кроме того, в [38]

предложен точный алгоритм для случая задачи с целочисленными входами. Временная сложность этого алгоритма — $\mathcal{O}(dN(2MD+1)^d)$. Если размерность пространства фиксирована, то этот алгоритм псевдополиномиален.

В [39] обоснован 2-приближенный полиномиальный алгоритм с трудоемкостью $\mathcal{O}(dN^2)$.

Полиномиальная приближенная схема (PTAS) с трудоемкостью $\mathcal{O}(dN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — относительная ошибка, предложена в [40].

В [41] установлено, что если P $\neq$ NP, то для этой задачи не существует полностью полиномиальной аппроксимационной схемы (FPTAS). В этой же работе представлен алгоритм, который позволяет находить $(1+\varepsilon)$-приближенное решение за время $\mathcal{O}(dN^2(\sqrt{2q/\varepsilon}+2)^d)$ для заданного $\varepsilon \in (0,1)$. Если размерность d пространства фиксирована, то алгоритм имеет трудоемкость $\mathcal{O}(N^2(1/\varepsilon)^{q/2})$ и реализует схему FPTAS. Кроме того, в [42] обоснован улучшенный по быстродействию алгоритм с временем работы $\mathcal{O}\big(\sqrt{d}N^2\big(\frac{\pi e}{2}\big)^{d/2}\big(\sqrt{2/\varepsilon}+2\big)^d\big)$. Этот алгоритм реализует схему FPTAS с трудоемкостью $\mathcal{O}(N^2(1/\varepsilon)^{d/2})$, если размерность d пространства фиксирована, и остается полиномиальным в случае, когда $d=\mathcal{O}(\log N)$, т. е. когда размерность пространства — медленно растущая функция от мощности входного множества. В этом случае он реализует схему PTAS с трудоемкостью $\mathcal{O}\Big(N^{C\left(1.05+\log\left(2+\sqrt{2/\varepsilon}\right)\right)}\Big)$, где C — положительная константа.

В [43] предложен рандомизированный алгоритм. Если $M \geq \beta N$, где $\beta \in (0,1)$ — некоторая константа, то при заданных $\varepsilon > 0$ и $\gamma \in (0,1)$ алгоритм находит $(1+\varepsilon)$-приближенное решение задачи с вероятностью не менее $1-\gamma$ за время $\mathcal{O}(dN)$. В той же работе установлены условия, при которых алгоритм находит $(1+\varepsilon_N)$-приближенное решение задачи за время $\mathcal{O}(dN^2)$ с вероятностью не менее $1-\gamma_N$, где $\varepsilon_N \to 0$ и $\gamma_N \to 0$ при $N \to \infty$, т. е. условия, при которых алгоритм является асимптотически точным. Этот алгоритм имеет рекордные на сегодняшний день показатели качества, так как он обеспечивает отыскание приближенного решения за линейное по N и d время с вероятностными гарантиями, а также получение асимптотически точного решения за квадратичное по N и линейное по d время.

Полученные результаты можно использовать для решения основной задачи с неизвестными мощностями. Действительно, перебирая не более N возможных комбинаций мощностей двух кластеров, можно построить семейство из N решений варианта задачи с заданными мощностями и затем выбрать в полученном семействе наилучшее решение по значению целевой функции. Вместе с этим значительный интерес представляют приближенные алгоритмы решения задачи 1 без перебора вариантов мощностей, так как эти алгоритмы в $\mathcal{O}(N)$ раз быстрее. Такой полиномиальный приближенный алгоритм предложен в [32]. Он находит 2-приближенное решение за $\mathcal{O}(dN^2)$ операций (для сравнения: алгоритм из [39] с перебором по допустимым комбинациям мощностей — за $\mathcal{O}(dN^3)$ операций).

Напомним существующие результаты для задачи 2. Она была выявлена совсем недавно в [23, 24], поэтому имеющийся к настоящему времени набор алгоритмических решений для этой задачи меньше, чем для задачи 1, которая исследуется существенно дольше. В цитированных работах вместе с доказательством труднорешаемости было показано, что для задачи 2 не существует схемы FPTAS, если P $\neq$ NP.

Большинство имеющихся алгоритмических решений получены с применением результативных техник построения алгоритмов для задачи 1. Для варианта задачи 2, в котором мощности кластеров заданы — *Cardinality-weighted 2-MSSC with given center and cluster cardinalities* (в п. 4 — задача 6), получен ряд алгоритмов. Оценки качества этих алгорит-

мов совпадают с оценками качества, приведенными выше для задачи 1. По этой причине мы не приводим их отдельно, а ограничиваемся их перечислением.

В [44] построен 2-приближенный алгоритм. Точный алгоритм для целочисленного варианта задачи предложен в [45]. В [46] предложен $(1+\varepsilon)$-приближенный алгоритм, реализующий схему FPTAS в случае, когда размерность пространства фиксирована. Улучшенная по быстродействию модификация этого алгоритма предложена в [42]. Эта модификация реализует схему PTAS, если размерность пространства является медленно растущей функцией от мощности входного множества ($d = \mathcal{O}(\log N)$). Рандомизированный алгоритм обоснован в [47].

В заключение обзора заметим, что остаются актуальными вопросы построения алгоритмов для задач 1 и 2 без перебора по N допустимым мощностям искомых кластеров.

В настоящей работе представлены первые на сегодняшний день результаты для новых задач 3 и 4 кластеризации данных. А именно, установлено, что обе задачи NP-трудны в сильном смысле, и обоснованы точные алгоритмы для случаев этих задач, в которых точки входного множества имеют целочисленные координаты. Если размерность пространства фиксирована, то оба предложенных алгоритма псевдополиномиальны.

2. Анализ вычислительной сложности

Перед анализом вычислительной сложности задач 3 и 4 заметим, что правые части (3) и (4) не зависят от искомого кластера $\mathcal{C}$ и для заданного входа являются константами. Положим

$$A = \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2, \quad B = \alpha N \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2. \tag{5}$$

Сформулируем задачи 3 и 4 в форме верификации свойств.

Задача 3А. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d, натуральное число M и действительное число $A > 0$. *Вопрос*: существует ли в $\mathcal{Y}$ подмножество $\mathcal{C}$ мощности не менее M такое, что

$$F(\mathcal{C}) \leq A \; ? \tag{6}$$

Задача 4А. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d, натуральное число M и действительное число $B > 0$. *Вопрос*: существует ли в $\mathcal{Y}$ подмножество $\mathcal{C}$ мощности не менее M такое, что

$$G(\mathcal{C}) \leq B \; ? \tag{7}$$

Сложностной статус этих задач устанавливает следующее

Утверждение. *Задачи 3А и 4А NP-полны в сильном смысле.*

Доказательство. Из выражений (3) и (4) легко видеть, что обе задачи 3А и 4А принадлежат классу NP.

Сформулируем оптимизационные задачи 1 и 2 в форме верификации свойств.

Задача 1А. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и действительное число $A > 0$. *Вопрос*: существует ли в $\mathcal{Y}$ подмножество $\mathcal{C}$ такое, что имеет место неравенство (6)?

Задача 2А. *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и действительное число $B > 0$. *Вопрос*: существует ли в $\mathcal{Y}$ подмножество $\mathcal{C}$ такое, что имеет место неравенство (7)?

Легко видеть, что при $M = 1$ в задачах 3А и 4А ответ положителен тогда и только тогда, когда ответ положителен в задачах 1А и 2А. Поэтому, справедливость утверждения следует из того, что NP-полные в сильном смысле задачи 1А и 2А являются специальными случаями (при $M = 1$) задач 3А и 4А соответственно. □

Из утверждения следует, что оптимизационные задачи 3 и 4 NP-трудны в сильном смысле.

3. Основы алгоритмов

Для построения алгоритмов решения задач 1, 2 и анализа их качественных свойств нам потребуются вспомогательные задачи, утверждения, множества и алгоритмы.

Во-первых, нам потребуются следующие отмеченные в п. 2 задачи.

Задача 5 (2-*MSSC with given center and cluster cardinalities*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и натуральное число M. *Найти*: подмножество $\mathcal{C} \subset \mathcal{Y}$ мощности M, минимизирующее значение функции $F(\mathcal{C})$.

Задача 6 (*Cardinality-Weighted 2-MSSC with given center and cluster cardinalities*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и натуральное число M. *Найти*: подмножество $\mathcal{C} \subset \mathcal{Y}$ мощности M, минимизирующее значение функции $G(\mathcal{C})$.

Вычислительной базой предлагаемых алгоритмов являются алгоритмы решения этих задач. В свою очередь, алгоритмы решения задач 5 и 6 основаны на двух приведенных ниже леммах 1 и 2, доказательства которых представлены в [39] и [45] соответственно.

Для произвольной точки $x \in \mathbb{R}^d$ положим:

$$r^x(y) = \langle y, x \rangle, \quad y \in \mathcal{Y}, \tag{8}$$

$$h^x(y) = (2M - N)\,\|y\|^2 - 2M\,\langle y, x \rangle, \quad y \in \mathcal{Y}, \tag{9}$$

и

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

$$g^x(\mathcal{C}) = M \sum_{y \in \mathcal{C}} \|y - x\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \quad \mathcal{C} \subseteq \mathcal{Y}. \tag{11}$$

Лемма 1. *Минимум функции* (10) *по всем подмножествам* $\mathcal{C} \subseteq \mathcal{Y}$ *мощности* M *достигается на подмножестве, состоящем из* M *векторов множества* $\mathcal{Y}$ *с наибольшими значениями функции* (8).

Лемма 2. *Минимум функции* (11) *по всем подмножествам* $\mathcal{C} \subseteq \mathcal{Y}$ *мощности* M *достигается на подмножестве, состоящем из* M *векторов множества* $\mathcal{Y}$ *с наименьшими значениями функции* (9).

Всюду далее полагаем, что точки множества $\mathcal{Y}$ имеют целочисленные координаты, т. е. будем рассматривать специальные (целочисленные) случаи задач. Определим вспомогательное множество в виде совокупности узлов (точек) равномерной решетки с рациональным шагом:

$$\mathcal{D} = \left\{ x \mid (x)^j = \frac{1}{M}(v)^j, \quad (v)^j \in \mathbb{Z}, \, |(v)^j| \leq MD, \quad j = 1, \ldots, d \right\}, \tag{12}$$

где

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1,\ldots,d\}} |(y)^j|, \tag{13}$$

а $(*)^j$ — j-я координата точки $*$.

Заметим, что

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

Ниже представлен алгоритм для целочисленного случая задачи 5.

Алгоритм $\mathcal{A}_1$.

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

Шаг 1. Найдем значение D и построим узлы решетки $\mathcal{D}$ по формулам (13) и (12).

Шаг 2. Для каждого $x \in \mathcal{D}$ построим множество $\mathcal{C}(x)$, состоящее из M точек $y \in \mathcal{Y}$, имеющих наибольшие значения функции (8). Вычислим $f^x(\mathcal{C}(x))$ по формуле (10).

Шаг 3. Найдем точку $x_A = \arg\min_{x \in \mathcal{D}} f^x(\mathcal{C}(x))$ и соответствующее ей подмножество $\mathcal{C}(x_A)$. В качестве решения задачи возьмем $\mathcal{C}^M_{A_1} = \mathcal{C}(x_A)$. Если решений несколько, то выберем любое из них.

Выход: множество $\mathcal{C}^M_{A_1}$.

Замечание 1. В [38] с использованием леммы 1 доказано, что если координаты всех точек входного множества $\mathcal{Y}$ целочисленны и лежат в интервале $[-D, D]$, то алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи 5 за время $\mathcal{O}(dN(2MD + 1)^d)$.

Наконец, нам нужен следующий алгоритм, который позволяет находить решение целочисленной задачи 6. Этот алгоритм отличается от алгоритма $\mathcal{A}_1$ лишь на шаге 2.

Алгоритм $\mathcal{A}_2$.

Вход: множество $\mathcal{Y}$ и положительное целое число M.

Шаг 1. Найдем значение D и построим узлы решетки $\mathcal{D}$ по формулам (13) и (12).

Шаг 2. Для каждого $x \in \mathcal{D}$ построим множество $\mathcal{C}(x)$, состоящее из M точек $y \in \mathcal{Y}$, имеющих наименьшие значения функции (9). Вычислим $g^x(\mathcal{C}(x))$ по формуле (11).

Шаг 3. Найдем точку $x_A = \arg\min_{x \in \mathcal{D}} g^x(\mathcal{C}(x))$ и соответствующее ей подмножество $\mathcal{C}(x_A)$. В качестве решения задачи возьмем $\mathcal{C}^M_{A_2} = \mathcal{C}(x_A)$. Если решений несколько, то выберем любое из них.

Выход: множество $\mathcal{C}^M_{A_2}$.

Замечание 2. В [45] с применением леммы 2 доказано, что если координаты всех точек входного множества $\mathcal{Y}$ целочисленны и лежат в интервале $[-D, D]$, то алгоритм $\mathcal{A}_2$ находит оптимальное решение задачи 6 за время $\mathcal{O}(dN(2MD + 1)^d)$.

4. Алгоритмы

Простая идея предлагаемых алгоритмов состоит в реализации сеточного подхода к аппроксимации неизвестного центроида искомого кластера наибольшего размера одним из узлов построенной равномерной решетки с рациональным шагом. Для каждого узла решетки с опорой на лемму 1 (при решении задачи 3) либо на лемму 2 (при решении задачи 4) с помощью алгоритмов $\mathcal{A}_1$ и $\mathcal{A}_2$ строится семейство допустимых решений — подмножеств. В построенном семействе допустимых подмножеств выбирается наибольшее по размеру и удовлетворяющее ограничению (3) (для задачи 3) либо ограничению (4) (для задачи 4).

Предлагается следующий алгоритм решения задачи 3.

Алгоритм $\mathcal{A}_3$.

Вход: множество $\mathcal{Y}$ и число α.

Шаг 1. Вычислим значение A по формуле (5).

Шаг 2. Для каждого $M = 1, \ldots, N$, используя алгоритм $\mathcal{A}_1$, найдем точное решение $\mathcal{C}_{A_1}^M$ задачи 5 и вычислим для этого решения значение целевой функции $F(\mathcal{C}_{A_1}^M)$.

Шаг 3. В семействе $\{\mathcal{C}_{A_1}^M, M = 1, \ldots, N\}$ множеств, полученных на шаге 2, найдем множество $\mathcal{C}_{A_1}$ наибольшей мощности, для которого $F(\mathcal{C}_{A_1}) \le A$.

Выход: множество $\mathcal{C}_{A_1}$.

Алгоритм решения задачи 4 аналогичен; основное отличие от алгоритма $\mathcal{A}_3$ заключается в построении допустимого решения задачи на шаге 2.

Алгоритм $\mathcal{A}_4$.

Вход: множество $\mathcal{Y}$ и число α.

Шаг 1. Вычислим значение B по формуле (5).

Шаг 2. Для каждого $M = 1, \ldots, N$, используя алгоритм $\mathcal{A}_2$, найдём точное решение $\mathcal{C}_{A_2}^M$ задачи 6 и вычислим для этого решения значение целевой функции $G(\mathcal{C}_{A_2}^M)$.

Шаг 3 В семействе $\{\mathcal{C}_{A_2}^M, M = 1, \ldots, N\}$ множеств, полученных на шаге 2, найдем множество $\mathcal{C}_{A_2}$ наибольшей мощности, для которого $G(\mathcal{C}_{A_2}) \le B$.

Выход: множество $\mathcal{C}_{A_2}$.

Справедлива следующая теорема.

Теорема. *Пусть точки входного множества $\mathcal{Y}$ имеют целочисленные координаты, лежащие в интервале $[-D, D]$. Тогда алгоритмы $\mathcal{A}_3$ и $\mathcal{A}_4$ находят точные решения задач 3 и 4 за время $\mathcal{O}(dN^2(2ND+1)^d)$.*

Доказательство. Докажем оптимальность решения, получаемого алгоритмом $\mathcal{A}_3$. Пусть $\mathcal{C}_1^*$ — оптимальное решение задачи 3, $M_1^* = |\mathcal{C}_1^*|$. Заметим, что алгоритм $\mathcal{A}_3$ находит допустимое решение задачи 5 при $M = M_1^*$. Поскольку $\mathcal{C}_{A_1}^{M^*}$ является оптимальным решением этой задачи,

$$F(\mathcal{C}_{A_1}^{M^*}) \le F(\mathcal{C}_1^*) \le A\,.$$

Таким образом, множество $\mathcal{C}_{A_1}^{M^*}$ было рассмотрено на шаге 3 работы алгоритма, и множество $\{\mathcal{C}_{A_1}^M, M = 1, \ldots, N \mid F(\mathcal{C}_{A_1}) \le A\}$ не пусто. Кроме того, из определения шага 3 получаем

$$|\mathcal{C}_{A_1}| \ge |\mathcal{C}_{A_1}^{M^*}| = M_1^* = |\mathcal{C}_1^*|.$$

С другой стороны, поскольку $\mathcal{C}_{A_1}$ является допустимым решением задачи 1, имеем $|\mathcal{C}_{A_1}| \leq |\mathcal{C}_1^*|$; таким образом, $|\mathcal{C}_{A_1}| = |\mathcal{C}_1^*|$.

Доказательство оптимальности решения, получаемого алгоритмом $\mathcal{A}_4$, строится аналогично доказательству оптимальности решения, получаемого алгоритмом $\mathcal{A}_3$.

Оценим временную сложность алгоритма $\mathcal{A}_3$. Шаг 1 выполняется за $\mathcal{O}(dN)$ операций. Наиболее трудоемкий шаг 2 требует $\mathcal{O}(dN^2(2ND+1)^d)$ операций, поскольку для каждого $M = 1, \ldots, N$ алгоритм $\mathcal{A}_1$ выполняется за $\mathcal{O}(dN(2MD+1)^d)$ операций. Наконец, шаг 3 выполняется за $\mathcal{O}(N)$ операций. Суммируя затраты на всех шагах, получаем представленную в теореме оценку временной сложности для алгоритма $\mathcal{A}_3$. Оценка временной сложности алгоритма $\mathcal{A}_4$ находится аналогично. □

Замечание 3. В случае когда размерность пространства фиксирована, алгоритмы $\mathcal{A}_3$ и $\mathcal{A}_4$ псевдополиномиальны, поскольку в этом случае время их выполнения можно оценить как $\mathcal{O}(N^2(ND)^d)$.

Заключение

В настоящей работе установлена сильная NP-трудность двух близких по постановке и важных для ряда приложений экстремальных задач кластеризации. Для решения этих задач предложены первые на сегодняшний день похожие по построению алгоритмы. Алгоритмы позволяют находить точные решения задач в случае, когда координаты входных точек целочисленны. Если размерность пространства ограничена сверху константой, то алгоритмы псевдополиномиальны. Иными словами, в работе показана псевдополиномиальная разрешимость целочисленных вариантов задач в случае, когда размерность пространства фиксирована (не является частью входа).

Ясно, что предложенные алгоритмы пригодны для решения практических задач лишь небольшой размерности. Тем не менее эти алгоритмы можно считать отправной точкой для получения улучшенных алгоритмических решений.

Важным направлением дальнейших исследований является обоснование эффективных приближенных алгоритмов с гарантированными оценками качества для выявленных труднорешаемых задач.

Литература

1. **MacQueen J.B.** Some methods for classification and analysis of multivariate observations // Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. — Berkeley: Univ. of California Press, 1967. — Vol. 1. — P. 281–297.
2. **Rao M.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. — 1971. — Vol. 66. — P. 622–626.
3. **Hansen P., Jaumard B., Mladenovich N.** Minimum sum of squares clustering in a low dimensional space // J. Classification. — 1998. — Vol. 15. — P. 37–55.
4. **Hansen P., Jaumard B.** Cluster analysis and mathematical programming // Mathematical Programming. — 1997. — Vol. 79. — P. 191–215.
5. **Fisher R.A.** Statistical Methods and Scientific Inference. — New York: Hafner, 1956.
6. **Jain A.K.** Data clustering: 50 years beyond k-means // Pattern Recognition Letters. — 2010. — Vol. 31, iss. 8. — P. 651–666.
7. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Machine Learning. — 2009. — Vol. 75, iss. 2. — P. 245–248.

8. **Drineas P., Frieze A., Kannan R., Vempala S., Vinay V.** Clustering large graphs via the singular value decomposition // Machine Learning. — 2004. — Vol. 56. — P. 9–33.

9. **Dolgushev A.V., Kel'manov A.V.** On the algorithmic complexity of a problem in cluster analysis // J. of Applied and Industrial Mathematics. — 2011. — Vol. 5, № 2. — P. 191–194.

10. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar k-means problem is NP-hard // Theoretical Computer Science. — 2012. — Vol. 442. — P. 3–21.

11. **Brucker P.** On the complexity of clustering problems // Lecture Notes in Economics and Mathematical Systems. — 1978. — Vol. 157. — P. 45–54.

12. **Bern M., Eppstein D.** Approximation algorithms for geometric problems // Approximation Algorithms for NP-Hard Problems. — Boston: PWS Publ., 1997. — P. 296–345.

13. **Indyk P.** A sublinear time approximation scheme for clustering in metric space // Proc. of the 40th Ann. IEEE Symp. on Foundations of Computer Science (FOCS). — 1999. — P. 154–159.

14. **de la Vega F., Kenyon C.** A randomized approximation scheme for metric max-cut // J. of Computer and System Sciences. — 2001. — Vol. 63. — P. 531–541.

15. **de la Vega F., Karpinski M., Kenyon C., Rabani Y.** Polynomial time approximation schemes for metric min-sum clustering // Electronic Colloquium on Computational Complexity (ECCC). — (Report № 25; 2002.)

16. **Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J.** Efficient algorithms for variance-based k-clustering // Proc. of the 1st Pacific Conference on Computer Graphics and Applications (Pacific Graphics'93, Seoul, Korea). — River Edge, NJ: World Scientific, 1993. — Vol. 1. — P. 75–89.

17. **Inaba M., Katoh N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: (extended abstract) // SCG'94 Proc. of the tenth annual symposium on Computational geometry. — Stony Brook, NY, USA, June 6–8, 1994. — P. 332–339. — (ACM, New York, 1994.)

18. **Sahni S., Gonzalez T.** P-complete approximation problems // J. of the ACM. — 1976. — Vol. 23. — P. 555–566.

19. **Ageev A.A., Kel'manov A.V., Pyatkin A.V.** NP-hardness of the Euclidean max-cut problem // Doklady Mathematics. — 2014. — Vol. 89, № 3. — P. 343–345.

20. **Ageev A.A., Kel'manov A.V., Pyatkin A.V.** Complexity of the weighted max-cut in Euclidean space // J. of Applied and Industrial Mathematics. — 2014. — Vol. 8, № 4. — P. 453–457.

21. **Kel'manov A.V., Pyatkin A.V.** On the complexity of a search for a subset of "similar" vectors // Doklady Mathematics. — 2008. — Vol. 78, № 1. — P. 574–575.

22. **Kel'manov A.V., Pyatkin A.V.** On a version of the problem of choosing a vector subset // J. of Applied and Industrial Mathematics. — 2009. — Vol. 3, № 4. — P. 447–455.

23. **Kel'manov A.V., Pyatkin A.V.** NP-hardness of some quadratic Euclidean 2-clustering problems // Doklady Mathematics. — 2015. — Vol. 92, № 2. — P. 634–637.

24. **Kel'manov A.V., Pyatkin A.V.** On the complexity of some quadratic Euclidean 2-clustering problems // Computational Mathematics and Mathematical Physics. — 2016. — Vol. 56, № 3. — P. 491–497.

25. **Bishop C.M.** Pattern Recognition and Machine Learning. — New York: Springer Science+Business Media, LLC, 2006.

26. **James G., Witten D., Hastie T., Tibshirani R.** An Introduction to Statistical Learning. — New York: Springer Science+Business Media, LLC, 2013.

27. **Hastie T., Tibshirani R., Friedman J.** The Elements of Statistical Learning (2nd edition). — Springer-Verlag, 2009.

28. **Aggarwal C.C.** Data Mining: The Textbook. — Springer International Publishing, 2015.

29. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series). — The MIT Press, 2017.

30. **Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T.** Big data clustering: a review // LNCS. — 2014. — Vol. 8583. — P. 707–720.

31. **Pach J., Agarwal P.K.** Combinatorial Geometry. — New York: Wiley, 1995.

32. **Kel'manov A.V., Khandeev V.I.** A 2-approximation polynomial algorithm for a clustering problem // J. of Applied and Industrial Mathematics. — 2013. — Vol. 7, № 4. — P. 515–521.

33. **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. матем. — 2006. — Т. 9, № 1. — С. 55–74.

34. **Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A.** A posteriori detecting a quasiperiodic fragment in a numerical sequence // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, № 1. — P. 30–42.

35. **Baburin A.E., Gimadi E.Kh., Glebov N.I., Pyatkin A.V.** The problem of finding a subset of vectors with the maximum total weight // J. of Applied and Industrial Mathematics. — 2008. — Vol. 2, № 1. — P. 32–38.

36. **Gimadi E.Kh., Pyatkin A.V., Rykov I.A.** On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension // J. of Applied and Industrial Mathematics. — 2010. — Vol. 4, № 1. — P. 48–53.

37. **Shenmaier V.V.** Solving some vector subset problems by Voronoi diagrams // J. of Applied and Industrial Mathematics. — 2016. — Vol. 10, № 4. — P. 560–566.

38. **Kel'manov A.V., Khandeev V.I.** An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors // J. of Applied and Industrial Mathematics. — 2015. — Vol. 9, № 4. — P. 497–502.

39. **Dolgushev A.V., Kel'manov A.V.** An approximation algorithm for solving a problem of cluster analysis // J. of Applied and Industrial Mathematics. — 2011. — Vol. 5, № 4. — P. 551–558.

40. **Dolgushev A.V., Kel'manov A.V., Shenmaier V.V.** Polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters // Proc. of the Steklov Institute of Mathematics. — 2016. — Vol. 295, supplement 1. — P. 47–56.

41. **Kel'manov A.V., Khandeev V.I.** Fully polynomial-time approximation scheme for a special case of a quadratic Euclidean 2-clustering problem // J. of Applied and Industrial Mathematics. — 2016. — Vol. 56, № 2. — P. 334–341.

42. **Kel'manov A.V., Motkova, A.V., Shenmaier V.V.** An approximation scheme for a weighted two-cluster partition problem // LNCS. — 2018. — Vol. 10716. — P. 323–333.

43. **Kel'manov A.V., Khandeev V.I.** A randomized algorithm for two-cluster partition of a set of vectors // Computational Mathematics and Mathematical Physics. — 2015. — Vol. 55, № 2. — P. 330–339.

44. **Kel'manov A.V., Motkova A.V.** Polynomial-time approximation algorithm for the problem of cardinality-weighted variance-based 2-clustering with a given center // Computational Mathematics and Mathematical Physics. — 2018. — Vol. 58, № 1. — P. 130–136.

45. **Kel'manov A.V., Motkova A.V.** Exact pseudopolynomial algorithms for a balanced 2-clustering problem // J. of Applied and Industrial Mathematics. — 2016. — Vol. 10, № 3. — P. 349–355.

46. **Kel'manov A.V., Motkova A.V.** A fully polynomial-time approximation scheme for a special Case of a balanced 2-clustering problem // LNCS. — 2016. — Vol. 9869. — P. 182–192.

47. **Kel'manov A.V., Khandeev V.I., Panasenko A.V.** Randomized algorithms for some clustering problems // Communications in Computer and Information Science. — 2018. — Vol. CCIS-871. — P. 109–119.

Поступила в редакцию 15 мая 2018 г.
После доработки 26 июня 2018 г.
Принята к публикации 21 января 2019 г.

Литература в транслитерации

1. **MacQueen J.B.** Some methods for classification and analysis of multivariate observations // Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. — Berkeley: Univ. of California Press, 1967. — Vol. 1. — P. 281–297.
2. **Rao M.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. — 1971. — Vol. 66. — P. 622–626.
3. **Hansen P., Jaumard B., Mladenovich N.** Minimum sum of squares clustering in a low dimensional space // J. Classification. — 1998. — Vol. 15. — P. 37–55.
4. **Hansen P., Jaumard B.** Cluster analysis and mathematical programming // Mathematical Programming. — 1997. — Vol. 79. — P. 191–215.
5. **Fisher R.A.** Statistical Methods and Scientific Inference. — New York: Hafner, 1956.
6. **Jain A.K.** Data clustering: 50 years beyond k-means // Pattern Recognition Letters. — 2010. — Vol. 31, iss. 8. — P. 651–666.
7. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Machine Learning. — 2009. — Vol. 75, iss. 2. — P. 245–248.
8. **Drineas P., Frieze A., Kannan R., Vempala S., Vinay V.** Clustering large graphs via the singular value decomposition // Machine Learning. — 2004. — Vol. 56. — P. 9–33.
9. **Dolgushev A.V., Kel'manov A.V.** On the algorithmic complexity of a problem in cluster analysis // J. of Applied and Industrial Mathematics. — 2011. — Vol. 5, № 2. — P. 191–194.
10. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar k-means problem is NP-hard // Theoretical Computer Science. — 2012. — Vol. 442. — P. 3–21.
11. **Brucker P.** On the complexity of clustering problems // Lecture Notes in Economics and Mathematical Systems. — 1978. — Vol. 157. — P. 45–54.
12. **Bern M., Eppstein D.** Approximation algorithms for geometric problems // Approximation Algorithms for NP-Hard Problems. — Boston: PWS Publ., 1997. — P. 296–345.
13. **Indyk P.** A sublinear time approximation scheme for clustering in metric space // Proc. of the 40th Ann. IEEE Symp. on Foundations of Computer Science (FOCS). — 1999. — P. 154–159.
14. **de la Vega F., Kenyon C.** A randomized approximation scheme for metric max-cut // J. of Computer and System Sciences. — 2001. — Vol. 63. — P. 531–541.
15. **de la Vega F., Karpinski M., Kenyon C., Rabani Y.** Polynomial time approximation schemes for metric min-sum clustering // Electronic Colloquium on Computational Complexity (ECCC). — (Report № 25; 2002.)
16. **Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J.** Efficient algorithms for variance-based k-clustering // Proc. of the 1st Pacific Conference on Computer Graphics and Applications (Pacific Graphics'93, Seoul, Korea). — River Edge, NJ: World Scientific, 1993. — Vol. 1. — P. 75–89.
17. **Inaba M., Katoh N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: (extended abstract) // SCG'94 Proc. of the tenth annual symposium on Computational geometry. — Stony Brook, NY, USA, June 6–8, 1994. — P. 332–339. — (ACM, New York, 1994.)

18. **Sahni S., Gonzalez T.** P-complete approximation problems // J. of the ACM. — 1976. — Vol. 23. — P. 555–566.

19. **Ageev A.A., Kel'manov A.V., Pyatkin A.V.** NP-hardness of the Euclidean max-cut problem // Doklady Mathematics. — 2014. — Vol. 89, № 3. — P. 343–345.

20. **Ageev A.A., Kel'manov A.V., Pyatkin A.V.** Complexity of the weighted max-cut in Euclidean space // J. of Applied and Industrial Mathematics. — 2014. — Vol. 8, № 4. — P. 453–457.

21. **Kel'manov A.V., Pyatkin A.V.** On the complexity of a search for a subset of "similar" vectors // Doklady Mathematics. — 2008. — Vol. 78, № 1. — P. 574–575.

22. **Kel'manov A.V., Pyatkin A.V.** On a version of the problem of choosing a vector subset // J. of Applied and Industrial Mathematics. — 2009. — Vol. 3, № 4. — P. 447–455.

23. **Kel'manov A.V., Pyatkin A.V.** NP-hardness of some quadratic Euclidean 2-clustering problems // Doklady Mathematics. — 2015. — Vol. 92, № 2. — P. 634–637.

24. **Kel'manov A.V., Pyatkin A.V.** On the complexity of some quadratic Euclidean 2-clustering problems // Computational Mathematics and Mathematical Physics. — 2016. — Vol. 56, № 3. — P. 491–497.

25. **Bishop C.M.** Pattern Recognition and Machine Learning. — New York: Springer Science+Business Media, LLC, 2006.

26. **James G., Witten D., Hastie T., Tibshirani R.** An Introduction to Statistical Learning. — New York: Springer Science+Business Media, LLC, 2013.

27. **Hastie T., Tibshirani R., Friedman J.** The Elements of Statistical Learning (2nd edition). — Springer-Verlag, 2009.

28. **Aggarwal C.C.** Data Mining: The Textbook. — Springer International Publishing, 2015.

29. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series). — The MIT Press, 2017.

30. **Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T.** Big data clustering: a review // LNCS. — 2014. — Vol. 8583. — P. 707–720.

31. **Pach J., Agarwal P.K.** Combinatorial Geometry. — New York: Wiley, 1995.

32. **Kel'manov A.V., Khandeev V.I.** A 2-approximation polynomial algorithm for a clustering problem // J. of Applied and Industrial Mathematics. — 2013. — Vol. 7, № 4. — P. 515–521.

33. **Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A.** Aposteriornoe obnaruzhenie v chislovoy posledovatel'nosti kvaziperiodicheskogo fragmenta pri zadannom chisle povtorov // Sib. zhurn. industr. matem. — 2006. — T. 9, № 1. — S. 55–74.

34. **Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A.** A posteriori detecting a quasiperiodic fragment in a numerical sequence // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, № 1. — P. 30–42.

35. **Baburin A.E., Gimadi E.Kh., Glebov N.I., Pyatkin A.V.** The problem of finding a subset of vectors with the maximum total weight // J. of Applied and Industrial Mathematics. — 2008. — Vol. 2, № 1. — P. 32–38.

36. **Gimadi E.Kh., Pyatkin A.V., Rykov I.A.** On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension // J. of Applied and Industrial Mathematics. — 2010. — Vol. 4, № 1. — P. 48–53.

37. **Shenmaier V.V.** Solving some vector subset problems by Voronoi diagrams // J. of Applied and Industrial Mathematics. — 2016. — Vol. 10, № 4. — P. 560–566.

38. **Kel'manov A.V., Khandeev V.I.** An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors // J. of Applied and Industrial Mathematics. — 2015. — Vol. 9, № 4. — P. 497–502.

39. **Dolgushev A.V., Kel'manov A.V.** An approximation algorithm for solving a problem of cluster analysis // J. of Applied and Industrial Mathematics. — 2011. — Vol. 5, № 4. — P. 551–558.

40. **Dolgushev A.V., Kel'manov A.V., Shenmaier V.V.** Polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters // Proc. of the Steklov Institute of Mathematics. — 2016. — Vol. 295, supplement 1. — P. 47–56.

41. **Kel'manov A.V., Khandeev V.I.** Fully polynomial-time approximation scheme for a special case of a quadratic Euclidean 2-clustering problem // J. of Applied and Industrial Mathematics. — 2016. — Vol. 56, № 2. — P. 334–341.

42. **Kel'manov A.V., Motkova, A.V., Shenmaier V.V.** An approximation scheme for a weighted two-cluster partition problem // LNCS. — 2018. — Vol. 10716. — P. 323–333.

43. **Kel'manov A.V., Khandeev V.I.** A randomized algorithm for two-cluster partition of a set of vectors // Computational Mathematics and Mathematical Physics. — 2015. — Vol. 55, № 2. — P. 330–339.

44. **Kel'manov A.V., Motkova A.V.** Polynomial-time approximation algorithm for the problem of cardinality-weighted variance-based 2-clustering with a given center // Computational Mathematics and Mathematical Physics. — 2018. — Vol. 58, № 1. — P. 130–136.

45. **Kel'manov A.V., Motkova A.V.** Exact pseudopolynomial algorithms for a balanced 2-clustering problem // J. of Applied and Industrial Mathematics. — 2016. — Vol. 10, № 3. — P. 349–355.

46. **Kel'manov A.V., Motkova A.V.** A fully polynomial-time approximation scheme for a special Case of a balanced 2-clustering problem // LNCS. — 2016. — Vol. 9869. — P. 182–192.

47. **Kel'manov A.V., Khandeev V.I., Panasenko A.V.** Randomized algorithms for some clustering problems // Communications in Computer and Information Science. — 2018. — Vol. CCIS-871. — P. 109–119.