ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль—август 2013. Том 20, № 4. С. 36–45

УДК 519.2+621.391

ПОЛИНОМИАЛЬНЫЙ АЛГОРИТМ С ОЦЕНКОЙ ТОЧНОСТИ 2 ДЛЯ РЕШЕНИЯ ОДНОЙ ЗАДАЧИ КЛАСТЕРНОГО АНАЛИЗА *)

А. В. Кельманов, В. И. Хандеев

Аннотация. Предложен 2-приближённый полиномиальный алгоритм для труднорешаемой задачи, к которой сводится одна из проблем разбиения конечного множества векторов евклидова пространства на два подмножества (кластера) по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров. Центром первого кластера является среднее значение вектора в этом кластере, а центром второго — нуль-вектор.

Ключевые слова: кластерный анализ, поиск подмножества векторов, алгоритмическая сложность, полиномиальный приближённый алгоритм.

Введение

Объектом исследования настоящей работы являются проблемы оптимизации. Предмет исследования — труднорешаемая экстремальная задача, к которой сводится одна из проблем кластерного анализа данных. Цель работы — обоснование эффективного приближённого алгоритма для решения этой задачи.

Одной из самых известных [9–11, 14–16] в общем случае NP-трудных [8] задач анализа данных и распознавания образов является

Задача MSSC (Minimum Sum-of-Squares Clustering).
Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$. Найти: разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 12–01–00090, 12–01–33028 и 13–07–00070), а также целевой программы СО РАН (интеграционные проекты 7Б и 21А).

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, — центр j-го кластера.

В настоящей работе анализируется задача, близкая к этой в постановочном плане. Одна из возможных содержательных трактовок проблемы, которая приводит к решению этой задачи, состоит в следующем. Имеется таблица, содержащая многократные результаты измерения набора числовых информационно значимых характеристик некоторого материального объекта. Объект может находиться в одном из двух состояний: активном (включенном) и пассивном (выключенном). В пассивном состоянии значения всех измеряемых характеристик равны нулю, а в активном — значение хотя бы одной характеристики не равно нулю. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие между результатами измерения и состояниями объекта неизвестно. Требуется, используя адекватный измеряемым характеристикам критерий, найти подмножество наборов, соответствующих активному состоянию объекта, и оценить по результатам измерения набор характеристик объекта в активном состоянии (учитывая, что данные содержат ошибку измерения). Эта содержательная проблема типична для многих приложений, связанных с компьютерным анализом данных и распознаванием образов (см., например, [1, 4, 6, 7, 13] и цитированные там работы).

В [6, 7] дана формулировка этой содержательной проблемы анализа данных в виде оптимизационной модели (задачи), в которой наборам числовых характеристик соответствуют векторы евклидова пространства. Решение задачи — подмножество и оценка вектора (соответствующего искомому набору) — находится в результате минимизации суммы квадратов расстояний. В [6, 7] показано, что в рамках этой модели решение задачи сводится к поиску подмножества, максимизирующего среднее значение квадрата длины суммы векторов из входного множества. Ниже, как и в [6, 7], эта задача поиска подмножества на максимум имеет краткое название MALSSVS (Maximum of the Average value of the Length Square of the Sum of Vectors from a Subset). В [6, 7] показана NP-трудность задачи MALSSVS в сильном смысле. К задаче MALSSVS полиномиально сведена классическая труднорешаемая в сильном смысле задача 3-SAT [12]. Точные и приближённые алгоритмы решения задачи MALSSVS предложены в [2, 7]. Характеристики этих алгоритмов приведены в следующем разделе.

В [5, 7] установлено, что задаче MALSSVS на максимум полиномиально эквивалентна задача 1-MSSC-NF на минимум. Таким образом, одна и та же оптимизационная модель приводит к решению двух противо-

положных полиномиально эквивалентных задач: одна из них на максимум, а другая — на минимум. Задача 1-MSSC-NF состоит в разбиении входного множества векторов на два кластера по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров (как и в задаче MSSC, сформулированной выше) в случае, когда центр одного из двух кластеров определять не требуется: считается, что центр этого кластера фиксирован и равен нулю. Именно эта NP-трудная в сильном смысле задача рассматривается в настоящей работе.

Мотивацией исследований послужили следующие факты. Во-первых, приближённый алгоритм решения задачи MALSSVS на максимум, предложенный в [7] и реализующий вполне полиномиальную аппроксимационную схему FPTAS, не обеспечивает гарантированой оценки точности для полиномиально эквивалентной задачи 1-MSSC-NF на минимум в неасимптотическом случае, хотя и позволяет находить асимптотически точное решение задачи 1-MSSC-NF. Во-вторых, временна́я сложность алгоритмов с оценками точности для задачи MALSSVS, обоснованных в [2, 7], столь высока (см. разд. 1), что эти алгоритмы практически непригодны для решения задач большой размерности.

В настоящей работе для решения задачи 1-MSSC-NF предложен приближённый алгоритм с оценкой точности 2, полиномиальный относительно размерности пространства и мощности входного множества.

1. Формулировка задачи и известные алгоритмические результаты

В [5, 7] показано, что формализация приведённой выше содержательной проблемы с использованием критерия минимума суммы квадратов расстояний приводит к следующим двум задачам.

Задача 1-MSSC-NF. Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ такое, что минимальна целевая функция

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

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

Задача MALSSVS. Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. Найти: подмножество $\mathcal{U} \subseteq \mathcal{Y}$ такое, что максимальна целевая функция

$$F(\mathcal{U}) = \frac{1}{|\mathcal{U}|} \Big\| \sum_{y \in \mathcal{U}} y \Big\|^2. \tag{2}$$

Задача 1-MSSC-NF близка в постановочном плане к задаче MSSC, но не эквивалентна ей. Символы MSSC в названии этой задачи подчеркивают сходство с задачей MSSC. Последние две литеры NF (Not Fixed) подчеркивают, что мощности кластеров не фиксированы.

Обобщение задачи 1-MSSC-NF на случай, когда число J кластеров не меньше единицы, а центр одного из кластеров определять не требуется, в [5] названо задачей J-MSSC-NF. Очевидно, что при $J \geqslant 1$ эта задача NP-трудна в сильном смысле, когда число кластеров J является частью входа задачи (как обобщение задачи 1-MSSC-NF). В [5] установлено, что задача J-MSSC-NF также NP-трудна в сильном смысле в случае, когда число кластеров не является частью входа.

Напомним, что имеет место равенство [5, 7]

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

Относительно характеристик существующих алгоритмических решений задач 1-MSSC-NF и MALSSVS заметим следующее.

Во-первых, в [7] для MALSSVS предложен алгоритм с гарантированной оценкой относительной погрешности $\varepsilon = (q-1)/(4l^2)$, где l — целочисленный параметр алгоритма. Временна́я сложность алгоритма есть величина $\mathcal{O}(Nq(q+\log N)(2l+1)^{q-1})$. Фактически, в [7] обоснована вполне полиномиальная аппроксимационная схема FPTAS, устанавливающая полиномиальную относительно N и $1/\varepsilon$ оценку временно́й сложности алгоритма для случая, когда размерность q пространства фиксирована. Для этого же случая в [2] конструктивно установлено, что задача разрешима за полиномиальное время $\mathcal{O}(q^2N^{2q})$.

Во-вторых, в [3] приведён 2-приближённый алгоритм с оценкой трудоёмкости $\mathcal{O}(qN^2)$ для задачи 1-MSSC-F, в которой мощность искомого кластера фиксирована. Очевидно, что с помощью этого алгоритма можно за время $\mathcal{O}(qN^3)$ получить приближённое решение задачи 1-MSSC-NF, перебирая возможные значения мощности кластера $\mathcal{C}$. В данной статье предложен менее трудоёмкий эффективный алгоритм, имеющий временну́ю сложность $\mathcal{O}(qN^2)$ при той же, что и в [3], оценке точности.

2. Алгоритм

Суть алгоритмического решения задачи 1-MSSC-NF состоит в следующем. Для каждого вектора из исходного множества строится гиперплоскость, перпендикулярная этому вектору и проходящая через его середину. Подмножество векторов, лежащих в полупространстве, не включающем начало координат, объявляется претендентом на решение. Для

каждого подмножества-претендента вычисляется значение вспомогательной оценочной функции. В качестве решения выбирается подмножество, для которого значение этой функции минимально.

Обозначим через $\mathcal{C}^*$ множество, доставляющее минимум функционалу $S(\cdot)$, и положим $y^* = \overline{y}(\mathcal{C}^*) = \frac{1}{|\mathcal{C}^*|} \sum\limits_{y \in \mathcal{C}^*} y$.

Лемма 1. *Для любых векторов $x \in \mathcal{C}^*$ и $z \in \mathcal{Y} \setminus \mathcal{C}^*$ имеют место неравенства* (i) $\|x - y^*\| \leqslant \|x\|$ *и* (ii) $\|z - y^*\| \geqslant \|z\|$.

Доказательство. Докажем неравенство (i). Допустим, напротив, что существует вектор $u \in \mathcal{C}^*$ такой, что $\|u - y^*\| > \|u\|$. Тогда

$$\begin{aligned} S(\mathcal{C}^*) &= \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \\ &> \sum_{y \in \mathcal{C}^* \setminus \{u\}} \|y - y^*\|^2 + \|u\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \\ &= \sum_{y \in \mathcal{C}^* \setminus \{u\}} \|y - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}^* \setminus \{u\})} \|y\|^2 \geqslant S(\mathcal{C}^* \setminus \{u\}), \end{aligned}$$

что противоречит оптимальности $\mathcal{C}^*$. Последнее неравенство следует из того, что для любого непустого конечного множества $\mathcal{Z}$ векторов из $\mathbb{R}^q$ минимум суммы $\sum\limits_{z \in \mathcal{Z}} \|z - x\|^2$ по x доставляется вектором $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum\limits_{z \in \mathcal{Z}} z$.

Аналогично, для доказательства неравенства (ii) допустим, что существует вектор $v \in \mathcal{Y} \setminus \mathcal{C}^*$ такой, что $\|v - y^*\| < \|v\|$. Тогда

$$\begin{aligned} S(\mathcal{C}^*) &= \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \\ &= \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + \|v\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}^* \cup \{v\})} \|y\|^2 \\ &> \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + \|v - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}^* \cup \{v\})} \|y\|^2 \\ &= \sum_{y \in \mathcal{C}^* \cup \{v\}} \|y - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}^* \cup \{v\})} \|y\|^2 \geqslant S(\mathcal{C}^* \cup \{v\}); \end{aligned}$$

противоречие. Лемма 1 доказана.

Лемма 1 имеет наглядный геометрический смысл. Действительно, равенство $\|y - y^*\| = \|y\|$ равносильно равенству $2(y, y^*) = \|y^*\|^2$, которое определяет гиперплоскость, перпендикулярную y^* и проходящую через

его середину. При этом оптимальное множество лежит в полупространстве, не включающем начало координат.

Рассмотрим вспомогательную задачу.

Задача SVSV-NF (Search for a Vector Subset and a Vector in the set, the case when the subset cardinality is Not Fixed).
Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. Найти: подмножество $\mathcal{B} \subseteq \mathcal{Y}$ и вектор $b \in \mathcal{Y}$ такие, что минимальна целевая функция

$$G(\mathcal{B}, b) = \sum_{y \in \mathcal{B}} \|y - b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2. \tag{4}$$

Для фиксированного вектора $b \in \mathcal{Y}$ положим $\mathcal{B}^*(b) = \arg\min\limits_{\mathcal{B} \subseteq \mathcal{Y}} G(\mathcal{B}, b)$.

Лемма 2. *Пусть $b \in \mathcal{Y}$. Тогда для любых $x \in \mathcal{B}^*(b)$ и $z \in \mathcal{Y} \setminus \mathcal{B}^*(b)$ имеют место неравенства* (i) $\|x - b\| \leqslant \|x\|$ *и* (ii) $\|z - b\| \geqslant \|z\|$.

Доказательство аналогично доказательству леммы 1.

Заметим, что равенство $\|y - b\| = \|y\|$ определяет гиперплоскость $2(y, b) = \|b\|^2$, перпендикулярную b и проходящую через его середину. При этом оптимальное множество $\mathcal{B}^*(b)$ лежит в полупространстве, не включающем начало координат.

Кроме того, для целевой функции (4) задачи SVSV-NF справедливо

Свойство. *Пусть $b \in \mathcal{Y}$, $\mathcal{B} \subseteq \mathcal{Y}$. Тогда если $u \in \mathcal{B}$ и $\|u - b\| = \|u\|$, то $G(\mathcal{B}, b) = G(\mathcal{B} \setminus \{u\}, b)$.*

В соответствии с леммой 2 определим множества

$$\mathcal{B}(b) = \{y \in \mathcal{Y} \mid 2(y, b) > \|b\|^2\}, \tag{5}$$

$$\mathcal{B}'(b) = \{y \in \mathcal{Y} \mid 2(y, b) \geqslant \|b\|^2\}. \tag{6}$$

Из леммы 2 и сформулированного свойства функции $G(\mathcal{B}, b)$ вытекает

Следствие. *Для любого вектора $b \in \mathcal{Y}$ множество $\mathcal{B}^*$ доставляет минимум функции $G(\mathcal{B}, b)$ тогда и только тогда, когда $\mathcal{B}(b) \subseteq \mathcal{B}^* \subseteq \mathcal{B}'(b)$.*

Это утверждение позволяет сформулировать следующий алгоритм решения задачи SVSV-NF.

Алгоритм $\mathcal{A}_1$.

Шаг 1. Для каждого $b \in \mathcal{Y}$ формируем множество $\mathcal{B}(b)$.

Шаг 2. Для каждого $b \in \mathcal{Y}$ вычислим значение $G(\mathcal{B}(b), b)$ целевой функции (4).

Шаг 3. В качестве решения задачи выберем вектор b^* и соответствующее ему множество $\mathcal{B}^* = \mathcal{B}(b^*)$ такие, что значение функции $G(\mathcal{B}^*, b^*)$ минимально. Если оптимальных значений несколько, то возьмём любое из них.

Лемма 3. *Оптимальное решение задачи* SVSV-NF *алгоритм* $\mathcal{A}_1$ *находит за время* $\mathcal{O}(qN^2)$.

Доказательство. Оптимальность вытекает из следствия и цепочки равенств

$$\min_{\mathcal{B}\subseteq\mathcal{Y}, b\in\mathcal{Y}} G(\mathcal{B}, b) = \min_{b\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} G(\mathcal{B}, b) = \min_{b\in\mathcal{Y}} G(\mathcal{B}(b), b).$$

Оценим временну́ю сложность алгоритма. Для фиксированного $b \in \mathcal{Y}$ формирование множества $\mathcal{B}(b)$ требует $\mathcal{O}(qN)$ операций, как и вычисление функции $G(\mathcal{B}(b), b)$. Поэтому трудоёмкость шагов 1 и 2 составляет $\mathcal{O}(qN^2)$ операций. Шаг 3 — поиск наименьшего элемента — требует не более $\mathcal{O}(N)$ операций. Таким образом, временна́я сложность алгоритма есть $\mathcal{O}(qN^2)$. Лемма 3 доказана.

Замечание. В алгоритме $\mathcal{A}_1$ можно вместо множеств $\mathcal{B}(b)$ взять множества $\mathcal{B}'(b)$. Ввиду приведённого следствия получаемое алгоритмом решение, очевидно, тоже будет оптимальным.

Приведём вспомогательное утверждение.

Лемма 4 [3]. *Пусть* $\mathcal{Z}$ *— непустое конечное множество векторов из* $\mathbb{R}^q$, *а* $\overline{z}(\mathcal{Z}) = \frac{1}{\mathcal{Z}} \sum\limits_{z\in\mathcal{Z}} z$. *Тогда если* $x \in \mathbb{R}^q$ *удовлетворяет условиям* $\|x - \overline{z}\| \leqslant \|z - \overline{z}\|$ *для любого* $z \in \mathcal{Z}$, *то имеет место неравенство*

$$\sum_{z\in\mathcal{Z}} \|z - x\|^2 \leqslant 2 \sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2.$$

Лемма 5. *Пусть* $\mathcal{B}^*, b^*$ *— оптимальное решение вспомогательной задачи* SVSV-NF, *а* $\mathcal{C}^*$ *— оптимальное решение задачи* 1-MSSC-NF. *Тогда имеет место оценка* $S(\mathcal{B}^*) \leqslant 2S(\mathcal{C}^*)$.

Доказательство. Рассмотрим множество $\mathcal{B}^*$ и вектор b^*. Поскольку для любого непустого конечного множества $\mathcal{Z}$ векторов из $\mathbb{R}^q$ минимум суммы $\sum\limits_{z\in\mathcal{Z}} \|z - x\|^2$ по x доставляется вектором $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum\limits_{z\in\mathcal{Z}} z$, имеет место оценка

$$S(\mathcal{B}^*) = \sum_{y\in\mathcal{B}^*} \|y - \overline{y}(\mathcal{B}^*)\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}^*} \|y\|^2$$

$$\leqslant \sum_{y\in\mathcal{B}^*} \|y - b^*\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}^*} \|y\|^2 = G(\mathcal{B}^*, b^*). \quad (7)$$

Далее, рассмотрим множество $\mathcal{C}^*$ и вектор $t = \arg\min\limits_{y\in\mathcal{C}^*} \|y - y^*\|$, где $y^* = \overline{y}(\mathcal{C}^*) = \frac{1}{|\mathcal{C}^*|} \sum\limits_{y\in\mathcal{C}^*} y$ (вектор t — ближайший к y^* вектор в оптимальном множестве $\mathcal{C}^*$). Так как они удовлетворяют условиям леммы 4, имеем

$$\sum_{y\in\mathcal{C}^*} \|y - t\|^2 \leqslant 2 \sum_{y\in\mathcal{C}^*} \|y - y^*\|^2,$$

следовательно,

$$G(\mathcal{C}^*, t) = \sum_{y\in\mathcal{C}^*} \|y - t\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 \leqslant 2 \sum_{y\in\mathcal{C}^*} \|y - y^*\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2$$
$$\leqslant 2 \sum_{y\in\mathcal{C}^*} \|y - y^*\|^2 + 2 \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 = 2S(\mathcal{C}^*). \quad (8)$$

Кроме того, заметим, что $\mathcal{C}^*$, t — допустимое решение вспомогательной задачи SVSV-NF, а $\mathcal{B}^*$, b^* — её оптимальное решение. Поэтому

$$G(\mathcal{B}^*, b^*) \leqslant G(\mathcal{C}^*, t). \quad (9)$$

Объединяя (7)–(9), получим оценку

$$S(\mathcal{B}^*) \leqslant G(\mathcal{B}^*, b^*) \leqslant G(\mathcal{C}^*, t) \leqslant 2S(\mathcal{C}^*).$$

Лемма 5 доказана.

В силу леммы 5 решение задачи 1-MSSC-NF даёт

Алгоритм $\mathcal{A}$.

Шаг 1. По заданному множеству $\mathcal{Y}$ находим оптимальное решение $\mathcal{B}^*$, b^* вспомогательной задачи SVSV-NF с помощью алгоритма $\mathcal{A}_1$.

Шаг 2. Подмножество $\mathcal{B}^*$ объявляем решением задачи 1-MSSC-NF.

Теорема. *Алгоритм $\mathcal{A}$ находит 2-приближённое решение задачи* 1-MSSC-NF *за время $\mathcal{O}(qN^2)$. Оценка 2 точности алгоритма достижима.*

Доказательство. Справедливость утверждения теоремы следует из лемм 3, 5 и следующего примера.

Пусть $q = 2,\ N = 2,\ y_1 = (0, \alpha),\ y_2 = (1, \alpha)$. Тогда если $0 < \alpha < 1$, то $\mathcal{B}^* = \{y_2\}$, $\mathcal{C}^* = \{y_1, y_2\}$, $S(\mathcal{B}^*) = \alpha^2$, $S(\mathcal{C}^*) = 1/2$. Таким образом,

отношение $S(\mathcal{B}^*)/S(\mathcal{C}^*) = 2\alpha^2$ может быть сколь угодно близко к 2 при $\alpha \to 1$.

Если же $\alpha = 1$, то имеем два оптимальных решения: либо $\mathcal{B}^* = \{y_1, y_2\}$, либо $\mathcal{B}^* = \{y_2\}$. При этом для второго решения $S(\mathcal{B}^*) = 1$, $\mathcal{C}^* = \{y_1, y_2\}$, $S(\mathcal{C}^*) = 1/2$ и $S(\mathcal{B}^*)/S(\mathcal{C}^*) = 2$, т. е. оценка точности алгоритма достижима. Теорема доказана.

Заключение

В работе построен 2-приближённый эффективный алгоритм решения NP-трудной задачи, к которой сводится оптимизационная модель одной из актуальных проблем анализа данных.

Поскольку рассмотренная задача относится к числу слабо изученных в алгоритмическом плане, важными направлениями дальнейших исследований являются: 1) обоснование эффективного рандомизированного алгоритма; 2) построение схемы PTAS; 3) поиск специальных случаев задачи, для которых возможно построение точных и приближённых полиномиальных алгоритмов; 4) разработка эффективных алгоритмов с оценками точности для обобщения этой задачи на случай нескольких кластеров — задачи J-MSSC-NF.

ЛИТЕРАТУРА

1. **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.
2. **Гимади Э. Х., Пяткин А. В., Рыков И. А.** О полиномиальной разрешимости некоторых задач выбора подмножества векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 6. — С. 11–19.
3. **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.
4. **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.
5. **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.
6. **Кельманов А. В., Пяткин А. В.** О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Докл. РАН. — 2008. — Т. 421, № 5. — С. 590–592.

7. **Кельманов А. В., Пяткин А. В.** Об одном варианте задачи выбора подмножества векторов // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 5. — С. 20–34.
8. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Les Cahiers du GERAD, G-2008-33. — 2008. — 4 p.
9. **Aloise D., Hansen P.** On the complexity of minimum sum-of-squares clustering // Les Cahiers du GERAD, G-2007-50. — 2007. — 12 p.
10. **Anil K., Jain K.** Data clustering: 50 years beyond k-means // Pattern Recognit. Lett. — 2010. — Vol. 31. —P. 651–666.
11. **Edwards A. W. F., Cavalli-Sforza L. L.** A method for cluster analysis // Biometrics. — 1965. — Vol. 21. — P. 362–375.
12. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. — San Francisco: Freeman, 1979. — 314 p.
13. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Trans. Signal Process. — 2004. — Vol. 52, N 3. — P. 645–656.
14. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Stat. Probab. Vol. 1. — 1967. — P. 281–297.
15. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar k-means problem is NP-hard // Lect. Notes Comput. Sci. — 2009. — Vol. 5431. — P. 284–285.
16. **Rao M.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. — 1971. — Vol. 66. — P. 622–626.

Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru
Хандеев Владимир Ильич,
e-mail: vladimir.handeev@gmail.com

Статья поступила
12 июня 2012 г.

Переработанный вариант —
21 октября 2012 г.