

УДК 519.2+621.391

## О СЛОЖНОСТИ НЕКОТОРЫХ ЗАДАЧ КЛАСТЕРНОГО АНАЛИЗА ВЕКТОРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ \*)

*А. В. Кельманов, А. В. Пяткин*

**Аннотация.** Доказана NP-полнота двух задач кластеризации (разбиения) конечной последовательности векторов евклидова пространства. В оптимизационных вариантах обеих задач требуется разбить элементы последовательности на фиксированное число кластеров по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров. В одной из задач мощности кластеров заданы на входе, а в другой неизвестны (являются оптимизируемыми величинами). За исключением центра одного (специального) кластера центры остальных кластеров определяются как средние значения по всем векторам, образующим эти кластеры. Центр специального кластера полагается равным нулю. При этом разбиение подчинено условию: для всех векторов, не входящих в специальный кластер, разность между номерами последующего и предыдущего векторов, входящих в любой из этих кластеров, ограничена сверху и снизу заданными константами.

**Ключевые слова:** кластеризация, последовательность евклидовых векторов, минимум суммы квадратов расстояний, ограничение на номера векторов, алгоритмическая сложность.

### Введение

Предметом исследования настоящей работы являются дискретные экстремальные задачи, которые индуцируются актуальными проблемами кластерного анализа данных. Цель работы — анализ алгоритмической сложности этих задач.

Представленные ниже результаты дополняют [6–8], где установлено, что к числу NP-полных в сильном смысле задач относятся следующие задачи кластеризации конечного множества векторов евклидова пространства.

---

\*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты № 12-01-00093, № 12-01-00090, № 12-01-33028-мол-а-вед и № 13-07-00070), а также целевой программы СО РАН (интеграционные проекты № 7Б и № 21А).

**Задача  $J$ -MSSC-NF.** ДАНО: множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$  и положительное число  $A$ . ВОПРОС: существует ли разбиение множества  $\mathcal{Y}$  на непустые подмножества (кластеры)  $\mathcal{C}_1, \dots, \mathcal{C}_J$  и  $\mathcal{Y} \setminus \mathcal{C}$ , где  $\mathcal{C} = \mathcal{C}_1 \cup \dots \cup \mathcal{C}_J$ , такое, что

$$\sum_{j=1}^J \sum_{y \in \mathcal{C}_j} \|y - \bar{y}_j(\mathcal{C}_j)\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \leq A, \quad (1)$$

где  $\bar{y}_j(\mathcal{C}_j) = \sum_{y \in \mathcal{C}_j} y / |\mathcal{C}_j|$ ,  $j = 1, \dots, J$ , — центры кластеров?

**Задача  $J$ -MSSC-F.** ДАНО: множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$ , натуральные числа  $M_1, \dots, M_J$  и положительное число  $A$ . ВОПРОС: существует ли разбиение множества  $\mathcal{Y}$  на непустые подмножества  $\mathcal{C}_1, \dots, \mathcal{C}_J$  и  $\mathcal{Y} \setminus \mathcal{C}$ , где  $\mathcal{C} = \mathcal{C}_1 \cup \dots \cup \mathcal{C}_J$ , такое, что имеет место неравенство (1), при ограничениях  $|\mathcal{C}_j| = M_j$ ,  $j = 1, \dots, J$ , на мощности кластеров?

Последние два символа NF (Not Fixed) в названии первой задачи подчеркивают, что мощности кластеров не фиксированы, в отличие от второй задачи, где символ F означает, что они фиксированы. Символы MSSC подчеркивают сходство с задачей  $J$ -MSSC (Minimum Sum-of-Squares Clustering) — одной из известных [2, 11, 13, 15, 18–20] труднорешаемых [10] задач кластерного анализа.

**Задача  $J$ -MSSC.** ДАНО: множество  $\mathcal{Z} = \{z_1, \dots, z_K\}$  векторов из  $\mathbb{R}^q$  и положительное число  $A$ . ВОПРОС: существует ли разбиение множества  $\mathcal{Z}$  на непустые кластеры  $\mathcal{C}_1, \dots, \mathcal{C}_J$  такое, что

$$\sum_{j=1}^J \sum_{z \in \mathcal{C}_j} \|z - \bar{z}_j(\mathcal{C}_j)\|^2 \leq A, \quad (2)$$

где  $\bar{z}_j(\mathcal{C}_j) = \sum_{z \in \mathcal{C}_j} z / |\mathcal{C}_j|$ ,  $j = 1, \dots, J$ , — центры кластеров?

В этой задаче объединение непересекающихся кластеров совпадает с множеством  $\mathcal{Z}$ , т. е.  $\mathcal{Z} = \mathcal{C}_1 \cup \dots \cup \mathcal{C}_J$ .

Одна из возможных содержательных трактовок проблемы, которая приводит к решению сформулированных выше задач, состоит в следующем. Некоторый материальный объект может находиться в пассивном и конечном множестве активных состояний. Имеется таблица, содержащая результаты многократных измерений набора числовых информационно значимых характеристик этого объекта. В пассивном состоянии все

числовые характеристики из набора равны нулю, а в любом активном — значение хотя бы одной характеристики не равно нулю. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие элементов таблицы какому-либо состоянию объекта неизвестно. Требуется, используя критерий минимума суммы квадратов расстояний, разбить таблицу на подмножества наборов, соответствующих пассивному и каждому активному состояниям объекта, а также оценить по результатам измерения наборы характеристик объекта в активных состояниях (учитывая, что данные содержат ошибку измерения). Эта содержательная проблема типична для многих приложений, связанных с компьютерным анализом данных и распознаванием образов (см., например, [12] и цитированные там работы).

Задачи, рассмотренные в настоящей работе, порождаются близкой в содержательном плане проблемой. Отличие этой проблемы от сформулированной выше состоит лишь в том, что элементы таблицы упорядочены по времени, причём известно, что временной интервал между двумя последовательными активными состояниями объекта ограничен сверху и снизу некоторыми константами. Подобные содержательные проблемы с временными ограничениями на результаты измерения различных информационно значимых характеристик весьма актуальны, в частности, при помехоустойчивой обработке числовых и векторных последовательностей (см., например, [1, 3–5, 9, 16, 17] и цитированные там работы), которые в приложениях трактуются как дискретные одномерные или многомерные сигналы.

Поскольку модель анализируемых данных практически та же (за исключением дополнительных ограничений), что и в [7, 8], рассмотренные ниже дискретные экстремальные задачи по своей сути являются аналогами приведённых выше задач. В рассматриваемых задачах предполагается, что входными данными являются не множества, а векторные последовательности, причём на номера векторов, входящих в искомые кластеры, накладываются ограничения (см. следующий раздел). Эти ограничения соответствуют априорным данным о времени переключения объекта из пассивного в какое-либо активное состояние. Мотивацией исследований послужил тот факт, что статус сложности этих задач ранее не был установлен.

## 1. Задачи кластеризации последовательностей

Для учёта ограничений на номера векторов, которые являются потенциальными претендентами на включение в какой-либо кластер, в при-

ведённых ниже формулировках задач при записи целевых функций вместо суммирования по элементам множеств (см. предыдущий раздел) используется суммирование по номерам (индексам) элементов последовательности. С этой же целью в формулировки задач в качестве параметров вводятся натуральные константы  $T_{\min}$  и  $T_{\max}$ , удовлетворяющие условиям  $1 \leq T_{\min} \leq T_{\max}$ .

Положим  $\mathcal{N} = \{1, \dots, N\}$ . В форме верификации свойств задачи на последовательностях с ограничениями формулируются следующим образом.

**Задача  $J$ -MSSCS-NF( $T_{\min}, T_{\max}$ ).** ДАНО: последовательность  $\mathcal{Y} = (y_1, \dots, y_N)$  векторов из  $\mathbb{R}^q$  и положительное число  $A$ . ВОПРОС: существует ли разбиение множества  $\mathcal{N}$  номеров элементов последовательности  $\mathcal{Y}$  на непустые подмножества  $\mathcal{M}_1, \dots, \mathcal{M}_J$  и  $\mathcal{N} \setminus \mathcal{M}$ , где  $\mathcal{M} = \mathcal{M}_1 \cup \dots \cup \mathcal{M}_J = \{n_1, \dots, n_M\}$ , такое, что

$$\sum_{j=1}^J \sum_{n_m \in \mathcal{M}_j} \|y_{n_m} - \bar{y}(\mathcal{M}_j)\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}} \|y_n\|^2 \leq A, \quad (3)$$

где  $\bar{y}(\mathcal{M}_j) = \frac{1}{|\mathcal{M}_j|} \sum_{n \in \mathcal{M}_j} y_n$ , при ограничениях

$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max}, \quad m = 2, \dots, M, \quad (4)$$

на элементы набора  $\mathcal{M}$ ?

**Задача  $J$ -MSSCS-F( $T_{\min}, T_{\max}$ ).** ДАНО: последовательность  $\mathcal{Y} = (y_1, \dots, y_N)$  векторов из  $\mathbb{R}^q$ , натуральные числа  $M_1, \dots, M_J$  и положительное число  $A$ . ВОПРОС: существует ли разбиение множества  $\mathcal{N}$  номеров элементов последовательности  $\mathcal{Y}$  на непустые подмножества  $\mathcal{M}_1, \dots, \mathcal{M}_J$  и  $\mathcal{N} \setminus \mathcal{M}$ , где  $\mathcal{M} = \mathcal{M}_1 \cup \dots \cup \mathcal{M}_J = \{n_1, \dots, n_M\}$ , такое, что имеет место неравенство (3), при условии, что  $|\mathcal{M}_j| = M_j$ ,  $j = 1, \dots, J$ , и при ограничениях (4) на элементы набора  $\mathcal{M}$ ?

Набор символов MSSCS является аббревиатурой английского названия Minimum Sum-of-Squares Clustering for the case of Sequence.

Если номера членов последовательности  $\mathcal{Y}$  интерпретировать как равнономерные дискретные отсчёты времени, то элементы набора  $\mathcal{M} = \{n_1, \dots, n_M\}$  номеров этой последовательности в содержательной проблеме соответствуют моментам времени, в которые объект находится в каком-либо из активных состояний. Параметры  $T_{\min}$  и  $T_{\max}$  соответствуют минимальному и максимальному интервалам времени между двумя последовательными активными состояниями объекта. Эти параметры на практике часто бывают априори известны.

## 2. Анализ алгоритмической сложности

Сначала рассмотрим случай  $J = 1$ . Заметим, что при  $T_{\min} = T_{\max}$  задачи разрешимы за полиномиальное время. Действительно, при  $T_{\min} = T_{\max} = T$  в силу (4) выбираемый набор  $\mathcal{M} = \mathcal{M}_1$  полностью определяется своим первым элементом  $n_1$  и размером  $M$ . Поскольку существует не более  $\mathcal{O}(N)$  вариантов выбора первого элемента набора  $\mathcal{M}_1$  и не более  $\mathcal{O}(N)$  вариантов выбора его размера для задачи 1-MSSCS-NF( $T, T$ ), задачи 1-MSSCS-NF( $T, T$ ) и 1-MSSCS-F( $T, T$ ) решаются за время  $\mathcal{O}(qN^2)$  и  $\mathcal{O}(qN)$  соответственно.

Покажем, что в случае  $T_{\min} < T_{\max}$  обе задачи NP-полны в сильном смысле. Нам понадобится классическая NP-полная в сильном смысле [14]

**Задача MaxCut** (Максимальный разрез). Дано: граф  $G$  и натуральное число  $t$ . ВОПРОС: существует ли в этом графе такое разбиение множества вершин на два подмножества, что число рёбер с концами в разных подмножествах не меньше  $t$ ?

**Теорема 1.** Для любых  $T_{\min} < T_{\max}$  задачи 1-MSSCS-NF( $T_{\min}, T_{\max}$ ) и 1-MSSCS-F( $T_{\min}, T_{\max}$ ) NP-полны в сильном смысле.

**ДОКАЗАТЕЛЬСТВО.** Сначала докажем NP-полноту более сложной задачи 1-MSSCS-NF( $T_{\min}, T_{\max}$ ) путём сведения к ней задачи MaxCut. Затем покажем, что аналогичное сведение можно использовать для анализа сложности задачи 1-MSSCS-F( $T_{\min}, T_{\max}$ ).

Нетрудно убедиться, что

$$\begin{aligned} \sum_{n_m \in \mathcal{M}_1} \|y_{n_m} - \bar{y}(\mathcal{M}_1)\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}_1} \|y_n\|^2 \\ = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \left\| \sum_{n_m \in \mathcal{M}_1} y_{n_m} \right\|^2 / |\mathcal{M}_1| \leq A \end{aligned} \quad (5)$$

тогда и только тогда, когда

$$f(\mathcal{M}_1) = \left\| \sum_{n_m \in \mathcal{M}_1} y_{n_m} \right\|^2 / |\mathcal{M}_1| \geq B, \quad (6)$$

где  $B = \sum_{n \in \mathcal{N}} \|y_n\|^2 - A$ . При сведении будем использовать целевую функцию (2) в форме (6).

Рассмотрим произвольный пример задачи MaxCut, т.е. граф  $G$  с  $p$  вершинами и  $q$  рёбрами, а также положительное целое число  $t$ . Положим

$$s = 8pq + 4q - 8tp, \quad T = \max\{T_{\min}, \lceil T_{\max}/2 \rceil\} - 1,$$

$$X = \lceil \sqrt{\max\{12q, (s + 4t)/(2p^2 - 1), (s - 8t)/(p + 1)^2\}} \rceil + 1.$$

Построим пример задачи 1-MSSCS-NF( $T_{\min}, T_{\max}$ ). Входную последовательность  $\mathcal{Y}$ , содержащую всего  $N = 2pT + 3p + 1$  членов, составим из  $(q + 1)$ -мерных векторов трёх типов: нулевых, вспомогательных и основных. Возьмём  $2pT$  нулевых,  $p + 1$  вспомогательных и  $2p$  основных векторов. В нулевом векторе все координаты равны 0. В каждом вспомогательном векторе  $(q + 1)$ -я координата равна  $X$ , а остальные компоненты равны 0.

Для определения основных векторов ориентируем все рёбра графа  $G$  произвольным образом. Каждой вершине  $v_i$ ,  $i = 1, \dots, p$ , ориентированного графа поставим в соответствие  $(q + 1)$ -мерный вектор  $u_i$ , в котором  $(q + 1)$ -я координата равна нулю, а  $j$ -я координата ( $j = 1, \dots, q$ ) равна 1, если дуга  $e_j$  исходит из вершины  $v_i$ ,  $-1$ , если дуга  $e_j$  входит в  $v_i$ , и 0, если дуга  $e_j$  неинцидентна  $v_i$ . Положим  $z_i = -u_i$ ,  $i = 1, \dots, p$ . В качестве основных векторов возьмём  $u_i$  и  $z_i$ ,  $i = 1, \dots, p$ .

Последовательность  $\mathcal{Y}$  векторов сформируем по следующему правилу. Сначала в произвольном порядке запишем все пары, состоящие из основных векторов, и получим последовательность  $u_1 z_1 u_2 z_2 \dots u_p z_p$ . Далее, между всеми парами  $u_i z_i$  основных векторов полученной последовательности вставим вспомогательный вектор  $x$ . Кроме того, добавим вспомогательный вектор в начало и в конец этой последовательности. Наконец, слева от каждого вектора  $u_i$  и справа от каждого вектора  $z_i$  вставим по  $T$  нулевых векторов ( $i = 1, \dots, p$ ). Схематически построенная последовательность имеет следующий вид (для примера  $T = 3$ , 0 — нулевой вектор,  $x$  — вспомогательный вектор):

$$\mathcal{Y} = y_1 y_2 \dots y_N = x000u_1 z_1 000x000u_2 z_2 000x000u_3 z_3 \dots 000x000u_p z_p 000x.$$

Положим  $B = ((p + 1)^2 X^2 + 4t)/(2p + 1)$ .

Покажем, что у построенной последовательности  $\mathcal{Y}$  существует набор  $\mathcal{M} = \mathcal{M}_1$  номеров, удовлетворяющий условиям (4) и (6), тогда и только тогда, когда в графе  $G$  (т. е. в задаче MaxCut) существует разрез, содержащий не менее  $t$  рёбер.

**НЕОБХОДИМОСТЬ.** Если в графе  $G$  есть разрез  $\{U, Z\}$ , где  $U \cap Z = \emptyset$ ,  $U \cup Z = V(G)$ , в котором не менее  $t$  рёбер имеют концы в разных подмножествах  $U$  и  $Z$ , то построим набор  $\mathcal{M}_1$  мощности  $2p + 1$  следующим образом. Включим в него номера всех  $2p + 1$  вспомогательных векторов, а также номера всех векторов  $u_i$ , для которых  $v_i \in U$ , и номера всех векторов  $z_i$ , для которых  $v_i \in Z$  (всего  $p$  таких векторов  $u_i$  и  $z_i$ ). Поскольку  $T_{\max} > T_{\min}$ , элементы построенного набора удовлетворяют условию (4).

Обозначим через  $h$  сумму векторов, номера которых принадлежат набору  $\mathcal{M}_1$ . Ясно, что  $(q+1)$ -я координата вектора  $h$  равна  $(p+1)X$ . Пусть  $e_l = v_i v_j$  —  $l$ -е ребро графа  $G$ . Нетрудно заметить, что если  $v_i$  и  $v_j$  лежат в одном и том же подмножестве ( $U$  или  $Z$ ), то  $l$ -я компонента вектора  $h$  равна 0. В противном случае она равна  $\pm 2$ . Поскольку таких компонент не менее  $t$ , имеем

$$f(\mathcal{M}_1) = \|h\|^2/|\mathcal{M}_1| \geq ((p+1)^2 X^2 + 4t)/(2p+1) = B,$$

т. е. условие (6) выполнено.

Таким образом, у последовательности  $\mathcal{Y}$  существует набор  $\mathcal{M} = \mathcal{M}_1$  номеров, удовлетворяющий условиям (4) и (6).

**ДОСТАТОЧНОСТЬ.** Пусть в задаче 1-MSSCS-NF( $T_{\min}, T_{\max}$ ) существует набор  $\mathcal{M}_1$  номеров последовательности  $\mathcal{Y}$ , удовлетворяющий условиям (4) и (6). Сначала покажем, что этот набор содержит номера всех вспомогательных векторов последовательности  $\mathcal{Y}$ .

Обозначим через  $f_i$  максимальное значение целевой функции задачи 1-MSSCS-NF( $T_{\min}, T_{\max}$ ) на решении, содержащем ровно  $i$  номеров вспомогательных векторов. Очевидно, что  $f_1 = X^2$ . Поскольку разность между номерами любых двух вспомогательных векторов в последовательности  $\mathcal{Y}$  больше  $T_{\max}$ , любое решение, содержащее ровно  $i$  номеров вспомогательных векторов, всего содержит не менее  $2i-1$  номеров векторов. Нетрудно заметить, что сумма квадратов первых  $q$  координат вектора  $h$ , равного сумме векторов с номерами из любого такого решения (набора), не превосходит  $4q$ . Отсюда  $f_i \leq (i^2 X^2 + 4q)/(2i-1)$ .

С другой стороны, очевидно, что  $f_{i+1} \geq (i+1)^2 X^2/(2i+1)$ . Следовательно,

$$f_{i+1} - f_i \geq \frac{(i+1)^2 X^2}{2i+1} - \frac{i^2 X^2 + 4q}{2i-1} = \frac{(2i^2 - 1)X^2 - 4q(2i+1)}{4i^2 - 1} > 0,$$

так как  $X^2 > 12q$ , а максимум выражения  $(2i+1)/(2i^2-1)$  равен 3 при  $i=1$ .

Таким образом, достаточно показать, что  $f_p < B$ . Действительно,

$$\begin{aligned} f_p - B &\leq \frac{p^2 X^2 + 4q}{2p-1} - \frac{(p+1)^2 X^2 + 4t}{2p+1} \\ &= \frac{8pq + 4q + 4t - 8pt - (2p^2 - 1)X^2}{4p^2 - 1} < 0 \end{aligned}$$

по выбору  $X$ . Значит, набор  $\mathcal{M}_1$  содержит номера всех вспомогательных векторов последовательности  $\mathcal{Y}$ .

Покажем, что  $|\mathcal{M}_1| = 2p + 1$ . Действительно, в противном случае имеем

$$\begin{aligned} f(\mathcal{M}_1) - B &\leq \frac{(p+1)^2 X^2 + 4q}{2p+2} - \frac{(p+1)^2 X^2 + 4t}{2p+1} \\ &= \frac{8pq + 4q - 8pt - 8t - (p+1)^2 X^2}{(2p+1)(2p+2)} < 0 \quad (7) \end{aligned}$$

по выбору  $X$ ; противоречие.

Обозначим через  $\mathcal{Y}(\mathcal{M}_1)$  подпоследовательность векторов, номера которых принадлежат набору  $\mathcal{M}_1$ . Положим  $U' = \{v_i \in G \mid u_i \in \mathcal{Y}(\mathcal{M}_1)\}$ ,  $Z' = \{v_i \in G \mid z_i \in \mathcal{Y}(\mathcal{M}_1)\}$  и  $W = \{v_i \in G \mid u_i, z_i \notin \mathcal{Y}(\mathcal{M}_1)\}$ . Пусть  $e_l = v_i v_j$  —  $l$ -е ребро графа  $G$ , а  $h$  — сумма векторов, номера которых принадлежат набору  $\mathcal{M}_1$ .

Нетрудно заметить, что  $l$ -я координата вектора  $h$  равна 0, если  $v_i$  и  $v_j$  лежат в одном и том же подмножестве ( $U'$ ,  $Z'$  или  $W$ ),  $\pm 1$ , если одна из этих вершин лежит в  $W$ , а вторая — в  $U'$  или  $Z'$ , и  $\pm 2$ , если одна из этих вершин лежит в  $U'$ , а вторая — в  $Z'$ .

Обозначим через  $t_1$  и  $t_2$  число компонент вектора  $h$ , равных  $\pm 1$  и  $\pm 2$ , соответственно. Тогда для значения целевой функции имеем

$$f(\mathcal{M}_1) = \frac{(p+1)^2 X^2 + 4t_2 + t_1}{2p+1} \geq B,$$

откуда следует, что  $4t_2 + t_1 \geq 4t$ .

Преобразуем множества  $U'$ ,  $Z'$  и  $W$  следующим образом: каждую вершину  $w \in W$  добавим к  $U'$ , если у неё больше соседей в  $Z'$ , чем в  $U'$ , и к  $Z'$  — в противном случае. Обозначим полученные множества через  $U$  и  $Z$ . Ясно, что мощность полученного разреза  $\{U, Z\}$ , где  $U \cap Z = \emptyset$ ,  $U \cup Z = V(G)$ , не меньше  $t_2 + t_1/2 \geq t_2 + t_1/4 \geq t$ .

Таким образом, в графе  $G$  существует разрез мощности не менее  $t$ .

Для доказательства NP-полноты задачи 1-MSSCS-F( $T_{\min}, T_{\max}$ ) используем то же самое сведение. Отличие состоит лишь в том, что при доказательстве следует положить  $|\mathcal{M}_1| = 2p + 1$ . Остальные выкладки и рассуждения проводятся аналогично. Теорема 1 доказана.

В случае произвольного  $J$  обе задачи оказываются NP-полными даже при  $T_{\min} = T_{\max}$ .

**Теорема 2.** Задачи  $J$ -MSSCS-NF( $T_{\min}, T_{\max}$ ) и  $J$ -MSSCS-F( $T_{\min}, T_{\max}$ ) NP-полны в сильном смысле для любых  $T_{\min} \leq T_{\max}$  и  $J \geq 2$ .



**ДОКАЗАТЕЛЬСТВО.** Покажем, что задача  $J\text{-MSSCS-NF}(T_{\min}, T_{\max})$  NP-полна. Отсюда следует NP-полнота задачи  $J\text{-MSSCS-F}(T_{\min}, T_{\max})$ , поскольку для решения первой задачи достаточно решить не более  $\mathcal{O}(N^J)$  задач  $J\text{-MSSCS-F}(T_{\min}, T_{\max})$ .

Рассмотрим задачу  $J\text{-MSSC}$  (см. введение). Напомним, что входом этой задачи является множество  $\mathcal{Z} = \{z_1, \dots, z_K\}$  векторов из  $\mathbb{R}^q$  и положительное число  $A$ . Построим пример последовательности  $\mathcal{Y}$  для задачи  $J\text{-MSSCS-NF}(T_{\min}, T_{\max})$ .

Положим  $X = \lceil \sqrt{A+1} \rceil$ . Для каждого  $k = 1, \dots, K$  образуем вектор  $y_k$  из вектора  $z_k$  добавлением  $(q+1)$ -й координаты, равной  $X$ . Запишем полученные векторы в произвольном порядке и вставим между векторами  $y_k$  и  $y_{k+1}$  по  $T_{\min} - 1$  нулевых векторов ( $k = 1, \dots, K-1$ ).

Покажем, что у построенной векторной последовательности  $\mathcal{Y}$  длины  $K + (K-1)(T_{\min} - 1)$  существует разбиение множества  $\mathcal{N}$  номеров на подмножества  $\mathcal{M}_1, \dots, \mathcal{M}_J$  и  $\mathcal{N} \setminus \mathcal{M}$ , где  $\mathcal{M} = \mathcal{M}_1 \cup \dots \cup \mathcal{M}_J$ , удовлетворяющее условиям (3) и (4), тогда и только тогда, когда в задаче  $J\text{-MSSC}$  существует разбиение на кластеры  $\mathcal{C}_1, \dots, \mathcal{C}_J$ , удовлетворяющее условию (2).

**НЕОБХОДИМОСТЬ.** Рассмотрим разбиение  $\{\mathcal{M}_1, \dots, \mathcal{M}_J, \mathcal{N} \setminus \mathcal{M}\}$  номеров последовательности  $\mathcal{Y}$  в задаче  $J\text{-MSSCS-NF}(T_{\min}, T_{\max})$ , удовлетворяющее условиям (3) и (4). Поскольку для любого  $k = 1, \dots, K$  только за счёт  $(q+1)$ -й координаты имеем  $\|y_k\|^2 > A$ , множество  $\bigcup_{i=1}^J \mathcal{M}_i$  содержит номера всех ненулевых векторов. Но тогда из условия (4) следует, что ни одного номера нулевого вектора это множество содержать не может.

Положив  $\mathcal{C}_j = \{y_i \mid i \in \mathcal{M}_j\}$  для всех  $j = 1, \dots, J$ , получим допустимое решение задачи  $J\text{-MSSC}$  с тем же самым значением целевой функции. Таким образом, в задаче  $J\text{-MSSC}$  разбиение на кластеры, удовлетворяющее условию (2), существует.

**ДОСТАТОЧНОСТЬ.** С другой стороны, любое решение  $\{\mathcal{C}_1, \dots, \mathcal{C}_J\}$  задачи  $J\text{-MSSC}$  порождает допустимое решение  $\{\mathcal{M}_1, \dots, \mathcal{M}_J, \mathcal{N} \setminus \mathcal{M}\}$  задачи  $J\text{-MSSCS-NF}(T_{\min}, T_{\max})$ , в котором наборы  $\mathcal{M}_j$  выбираются из условия  $\mathcal{C}_j = \{y_i \mid i \in \mathcal{M}_j\}$ , с тем же самым значением целевой функции. Таким образом, задача  $J\text{-MSSCS-NF}(T_{\min}, T_{\max})$  NP-полна. Теорема 2 доказана.

### Заключение

Установлена NP-трудность в сильном смысле актуальных задач кластерного анализа векторных последовательностей с дополнительными ограничениями на номера выбираемых векторов. Наличие таких ограничений не упрощает аналогов этих задач с алгоритмической точки зрения.

Вопрос обоснования эффективных приближённых алгоритмов с оценками точности для решения этих задач остаётся открытым.

### 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. 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.
3. 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.
4. Кельманов А. В., Михайлова Л. В. Совместное обнаружение в квазипериодической последовательности заданного числа фрагментов из эталонного набора и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. математики и мат. физики. — 2006. — Т. 46, № 1. — С. 172–189.
5. Кельманов А. В., Михайлова Л. В., Хамидуллин С. А. Об одной задаче поиска упорядоченных наборов фрагментов в числовой последовательности // Дискрет. анализ и исслед. операций. — 2009. — Т. 16, № 4. — С. 31–46.
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. 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.
9. Кельманов А. В., Хамидуллин С. А. Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Журн. вычисл. математики и мат. физики. — 2001. — Т. 41, № 5. — С. 807–820.

10. **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.
11. **Aloise D., Hansen P.** On the complexity of minimum sum-of-squares clustering // Les Cahiers du GERAD, G-2007-50. — 2007. — 12 p.
12. **Anil K., Jain K.** Data clustering: 50 years beyond  $k$ -means // Pattern Recognit. Lett. — 2010. — Vol. 31. — P. 651–666.
13. **Edwards A. W. F., Cavalli-Sforza L. L.** A method for cluster analysis // Biometrics. — 1965. — Vol. 21. — P. 362–375.
14. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. — San Francisco: Freeman, 1979. — 314 p.
15. **Inaba M., Katch N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based clustering // Proc. 10th Ann. Symp. Comput. Geom. (Stony Brook, New York, June 06–08, 1994). — Stony Brook, New York, NY: ACM Press, 1994. — P. 332–339.
16. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Trans. Signal Processing. — 2004. — Vol. 52, N 3. — P. 645–656.
17. **Kel'manov A. V., Khamidullin S. A.** An algorithm for recognition of a vector alphabet generating a sequence with a quasi-periodic structure // Pattern Recognition Image Anal. — 2010. — Vol. 20, N 4. — P. 451–458.
18. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar  $k$ -means problem is NP-hard // Theor. Comput. Sci. — 2012. Vol. 442. — P. 13–21.
19. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Statist. Probab. (California, Berkeley, June 21–July 18, 1965; December 27, 1965–January 7, 1966). Vol. 1. — Berkeley, California: University of California Press, 1967. — P. 281–297.
20. **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

Пяткин Артём Валерьевич,

e-mail: artem@math.nsc.ru

Статья поступила

27 июня 2012 г.

Переработанный вариант —

11 октября 2012 г.