ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Сентябрь—октябрь 2008. Том 15, № 5. С. 20–34

УДК 519.2+621.391

ОБ ОДНОМ ВАРИАНТЕ ЗАДАЧИ ВЫБОРА ПОДМНОЖЕСТВА ВЕКТОРОВ*)

А. В. Кельманов, А. В. Пяткин

Аннотация. Доказана NP-полнота задачи выбора подмножества «похожих» векторов, к которой сводится один из вариантов проблемы апостериорного (off-line) помехоустойчивого обнаружения в числовой последовательности неизвестного повторяющегося вектора в случае, когда помеха аддитивна. Обоснован приближённый полиномиальный алгоритм решения этой задачи с гарантированной оценкой точности в случае фиксированной размерности пространства.

Ключевые слова: числовая векторная последовательность, апостериорная обработка, повторяющийся вектор, оптимальное помехоустойчивое обнаружение, сложность, NP-полнота, приближённый алгоритм.

Введение

В работе исследуется дискретная экстремальная задача, к которой сводится один из векторных вариантов [3, 4] проблемы апостериорного (off-line) обнаружения повторяющегося фрагмента в числовой последовательности в случае, когда помеха аддитивна. Одна из возможных содержательных трактовок рассматриваемой задачи состоит в следующем. Источник сообщений через канал передачи с помехой передаёт информацию об активном (включенном) или пассивном (выключенном) состоянии некоторого физического объекта или явления в виде числового набора — вектора — измеряемых информационно важных характеристик. В пассивном состоянии значения всех компонент этого вектора равны нулю, а в активном все измеряемые характеристики стабильны и значение хотя бы одной из них не равно нулю. На приёмную сторону поступает векторная последовательность результатов измерения состояний объекта, искажённая аддитивным шумом. Требуется обнаружить моменты времени, в которые объект находился в активном состоянии, и оценить значения измеряемых характеристик. Эта содержательная задача

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 06–01–00058, 07–07–00022 и 08–01–00516).

типична для многих приложений, связанных с компьютерным анализом данных и распознаванием образов (см., например, [9] и цитированные там работы). Традиционный и хорошо изученный подход к её решению опирается на on-line обработку данных. Он ориентирован на получение результата с наименьшими временными затратами. Корни этого подхода лежат в фундаментальной работе [12]. Напротив, потенциально более точный подход, который опирается на off-line обработку данных, до настоящего времени был не изучен. Его реализация сопряжена с решением специфической экстремальной задачи. Гипотеза о её NP-трудности оставалась неподтверждённой более 10 лет [3, 4]. В данной работе показано, что в форме верификации свойств эта задача NP-полна. Для оптимизационного варианта задачи предложен приближённый полиномиальный алгоритм решения с гарантированной оценкой точности в случае фиксированной размерности пространства. Показано, что рассматриваемую экстремальную задачу можно трактовать как задачу разбиения совокупности векторов на два подмножества (кластера) по критерию минимума суммы квадратов уклонений при условии, что центр одного из кластеров фиксирован и равен нулю.

1. Постановка задачи

Пусть векторная последовательность $x_n(\mathcal{M}, w) \in \mathbb{R}^q,\ n \in \mathcal{N}$, где $\mathcal{N} = \{1, 2, \ldots, N\}$, обладает свойством

$$x_n(\mathcal{M}, w) = \begin{cases} w, & n \in \mathcal{M}, \\ 0, & n \in \mathcal{N} \setminus \mathcal{M}, \end{cases} \tag{1}$$

где $\mathcal{M} \subseteq \mathcal{N}$. Положим $|\mathcal{M}| = M$. Вектор $w \in \mathbb{R}^q$ будем интерпретировать как информационно важный вектор, а M — как число его повторов. Рассмотрим аддитивную модель помехи (ошибок наблюдения). Доступной для обработки будем считать последовательность

$$v_n = x_n(\mathcal{M}, w) + e_n,\ n \in \mathcal{N},$$

где e_n — вектор помехи (ошибки), независимый от вектора $x_n(\mathcal{M}, w)$. Положим

$$S(\mathcal{M}, w) = \sum_{n \in \mathcal{N}} \|v_n - x_n(\mathcal{M}, w)\|^2, \tag{2}$$

допустим, что $0 < \|w\|^2 < \infty$, и рассмотрим следующую задачу среднеквадратического приближения.

Дано: последовательность $v_n \in \mathbb{R}^q, n \in \mathcal{N}$. Найти: подмножество $\mathcal{M} \subseteq \mathcal{N}$ и вектор $w \in \mathbb{R}^q$, минимизирующие $S(\mathcal{M}, w)$.

Эта задача соответствует сформулированной содержательной задаче обнаружения по критерию минимума суммы квадратов уклонений ненулевого неизвестного информационно-важного вектора, повторяющегося в ненаблюдаемой последовательности (1). Задачу можно также трактовать как поиск подмножества векторов, «похожих» в среднеквадратическом. Легко убедиться, что к аналогичной формулировке можно прийти, если считать, что вектор e_n есть выборка из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения использовать традиционный для статистики максимум функционала правдоподобия.

Заметим, что в частном случае, когда множество $\mathcal{M}$ известно, эта задача вырождается в классическую задачу оценивания среднего значения вектора. В другом частном случае, когда вектор w задан, задача решается за полиномиальное время, как при известном, так и при неизвестном числе M повторов [5, 10].

В рассматриваемом общем случае, когда повторяющийся вектор w и множество $\mathcal{M}$ неизвестны, минимум функционала (2) по вектору w легко находится аналитически [2]. Этот минимум доставляется вектором $\overline{w} = \sum\limits_{n \in \mathcal{M}} v_n / M$ и равен

$$S_{\min}(\mathcal{M}) = \sum_{n \in \mathcal{N}} \|v_n\|^2 - \frac{1}{M} \Big\| \sum_{n \in \mathcal{M}} v_n \Big\|^2. \tag{3}$$

Первый член в правой части выражения (3) — константа. Поэтому в оптимизационной форме имеем следующую редуцированную дискретную экстремальную задачу.

Дано: последовательность $v_1, v_2, \ldots, v_N$ векторов из $\mathbb{R}^q$. Найти: подмножество $\mathcal{M} \subseteq \mathcal{N}$, максимизирующее $\Big\| \sum\limits_{n \in \mathcal{M}} v_n \Big\|^2 / M$.

Сформулируем её в форме задачи верификации свойств. Будем называть эту задачу «Среднее значение квадрата длины суммы векторов из подмножества» и использовать для её обозначения англоязычную аббревиатуру ALSSVS (от Average value of a Length Square of the Sum of Vectors from a Subset)

Задача ALSSVS. Дано: семейство V, состоящее из N векторов евклидова пространства $\mathbb{R}^q$, и положительное число K. Вопрос: существует ли такое непустое $U \subseteq V$, что имеет место неравенство

$$\Big\| \sum_{v \in U} v \Big\|^2 \Big/ |U| \geqslant K? \tag{4}$$

Эта задача NP-трудна [1, 2], когда мощность $|U|$ фиксирована.

Перед анализом сложности задачи ALSSVS в случае нефиксированной мощности $|U|$ рассмотрим следующую близкую к ней содержательную задачу. Она также возникает в приложениях, связанных с анализом данных и распознаванием образов, в случае, когда вместо векторной последовательности вида (1) анализируется последовательность

$$x_n(\mathcal{M}, w_1, w_2) = \begin{cases} w_1, \ n \in \mathcal{M}, \\ w_2, \ n \in \mathcal{N} \setminus \mathcal{M}, \end{cases}$$

и рассматривается следующая задача среднеквадратического приближения.

Дано: последовательность $v_n \in \mathbb{R}^q, n \in \mathcal{N}$. Найти: подмножество $\mathcal{M} \subseteq \mathcal{N}$ и векторы $w_1 \in \mathbb{R}^q$ и $w_2 \in \mathbb{R}^q$, минимизирующие

$$S(\mathcal{M}, w_1, w_2) = \sum_{n \in \mathcal{N}} \|v_n - x_n(\mathcal{M}, w_1, w_2)\|^2.$$

Нетрудно установить, что для любых непустых $\mathcal{M}$ и $\mathcal{N} \setminus \mathcal{M}$ минимальное значение целевой функции этой задачи достигается при $\overline{w}_1 = \sum\limits_{n \in \mathcal{M}} v_n/M$ и $\overline{w}_2 = \sum\limits_{n \in \mathcal{N} \setminus \mathcal{M}} v_n/(N - M)$. Отсюда имеем экстремальную задачу, которая обычно называется кластеризацией (разбиением) множества векторов на два кластера или подмножества (в общем случае число кластеров может быть больше двух) по критерию минимума суммы квадратов. Эта задача имеет краткое название MSSC (от Minimum Sum-of-Squares Clustering). Встречается также и другое название этой задачи — k-Means Clustering Problem, которое было дано в [11] по названию приближённого алгоритма для её решения. В форме верификации свойств эта задача формулируется в следующем виде.

Задача MSSC. Дано: семейство V, состоящее из N векторов евклидова пространства $\mathbb{R}^q$, и положительное число $\widetilde{K}$. Вопрос: существует ли такое разбиение множества V на непустые подмножества (кластеры) B_1 и B_2, что имеет место неравенство

$$\sum_{v \in B_1} \|v - \overline{w}_1\|^2 + \sum_{v \in B_2} \|v - \overline{w}_2\|^2 \leqslant \widetilde{K},$$

где $\overline{w}_i = \sum\limits_{v \in B_i} v/|B_i|, \ i = 1, 2,$ — центры кластеров.

Отметим, что сложность задачи MSSC на данный момент не установлена (хотя на протяжении нескольких десятков лет эта задача в оптимизационном варианте бездоказательно считается NP-трудной). В опубликованном недавно [7] доказательстве её NP-полноты впоследствии была обнаружена ошибка [6].

Покажем, что задача ALSSVS получается из MSSC, если в последней центр одного из кластеров зафиксирован и равен 0 (при этом допускается, что этот кластер может быть пустым). Действительно, пусть для определённости $\overline{w}_2 = 0$. Тогда

$$\sum_{v\in B_1}\|v-\overline{w}_1\|^2+\sum_{v\in B_2}\|v\|^2=\sum_{v\in B_1}\|v\|^2+|B_1|\|\overline{w}_1\|^2-2\sum_{v\in B_1}(v,\overline{w}_1)+\sum_{v\in B_2}\|v\|^2$$

$$=\sum_{v\in V}\|v\|^2+\frac{\Big\|\sum\limits_{v\in B_1}v\Big\|^2}{|B_1|}-2\frac{\sum\limits_{v\in B_1}\Big(v,\sum\limits_{u\in B_1}u\Big)}{|B_1|}=\sum_{v\in V}\|v\|^2-\frac{\Big\|\sum\limits_{v\in B_1}v\Big\|^2}{|B_1|}.$$

Таким образом, в специальном случае задачи MSSC, когда центр одного из кластеров, например B_2, равен нулю, разбиение семейства векторов V на кластеры B_1 и B_2 существует тогда и только тогда, когда в задаче ALSSVS при $K=\sum\limits_{v\in V}\|v\|^2-\widetilde{K}$ существует соответствующее подмножество векторов U.

2. Анализ комбинаторной сложности задачи

Статус сложности задачи ALSSVS устанавливает следующая

Теорема 1. *Задача* ALSSVS NP-*полна в сильном смысле.*

Доказательство. Принадлежность задачи ALSSVS к классу NP очевидна. Для доказательства её NP-полноты используем NP-полную задачу 3-ВЫПОЛНИМОСТЬ [8].

Задача 3-SAT. Дано: m дизъюнкций $C_1, C_2, \ldots, C_m$ над множеством из n переменных, причём каждая из дизъюнкций содержит по 3 литерала (под литералом понимается переменная или её отрицание). Вопрос: можно ли назначить этим переменным такие значения истинности, чтобы каждая дизъюнкция содержала по крайней мере один истинный литерал?

Построим полиномиальное сведение задачи 3-SAT к задаче ALSSVS. Рассмотрим произвольный пример задачи 3-SAT с n переменными и m дизъюнкциями, каждая из которых содержит по три литерала.

Пусть

$$c = n(m+n)^2/2 - m - n + 2,$$

$$b = \lceil \sqrt{2n(c+2n+3m-1)^2 - n(c+m+n-1)^2/(m+n) + 1}\ \rceil,$$

$$a = \lceil b\sqrt{(3m+2n)(2m+n+1)/6}\ \rceil.$$

Положим

$$K = 6a^2 + (m+n)b^2 + n(c+m+n-1)^2/(m+n).$$

Обозначим k-ю координату вектора v_i через $v_i(k)$. Сконструируем семейство V, состоящее из $N = 2n+3m$ векторов размерности $q = 4n+3m+1$, по следующим правилам. Положим

$$v_i(2n+3m+1) = b, \quad i = 1, 2, \ldots, 2n+3m,$$

и будем называть *средней* координату с номером $2n+3m+1$. Координаты $1, 2, \ldots, 2n+3m$ всех векторов будем называть *левыми*, а остальные — *правыми*.

Для каждого $i = 1, 2, \ldots, n$ зададим

$$v_{2i-1}(2i-1) = a\sqrt{3}, \quad v_{2i-1}(2i) = -a\sqrt{3},$$

$$v_{2i}(2i-1) = -a\sqrt{3}, \quad v_{2i}(2i) = a\sqrt{3},$$

а все остальные левые координаты векторов v_{2i-1} и v_{2i} положим равными 0. Пусть

$$v_{2i-1}(2n+3m+2i) = v_{2i}(2n+3m+2i+1) = c,$$

а все остальные правые координаты этих векторов равны 1. Кроме того, положим

$$v_{2n+3j-2}(2n+3j-2) = v_{2n+3j-1}(2n+3j-1) = v_{2n+3j}(2n+3j) = 2a,$$

$$v_{2n+3j-2}(2n+3j-1) = v_{2n+3j-2}(2n+3j) = v_{2n+3j-1}(2n+3j-2)$$

$$= v_{2n+3j-1}(2n+3j) = v_{2n+3j}(2n+3j-2) = v_{2n+3j}(2n+3j-1) = -a$$

для каждого $j = 1, 2, \ldots, m$, а все остальные левые координаты этих векторов зададим равными 0.

Пусть $j \in \{1, 2, \ldots, m\}$. Допустим, что k-й литерал ($k = 1, 2, 3$), входящий в дизъюнкцию C_j, равен z_i или $\bar{z}_i$, где $i \in \{1, 2, \ldots, n\}$. Если имеет место первая альтернатива, то пусть $v_{2n+3j-k+1}(2n+3m+2i) = 0$, а все

остальные правые координаты вектора $v_{2n+3j-k+1}$ равны 1. В противном случае считаем $v_{2n+3j-k+1}(2n+3m+2i+1)=0$, а все остальные правые координаты этого вектора положим равными 1. Далее, для всех $i=1,2,\ldots,n$ и $j=1,2,\ldots,m$ условимся, что векторы v_{2i-1} и v_{2i} соответствуют переменной z_i и её отрицанию $\bar{z}_i$, а векторы $v_{2n+3j-2}$, $v_{2n+3j-1}$ и v_{2n+3j} — третьему, второму и первому литералам, входящим в дизъюнкцию C_j. Кроме того, условимся, что координаты $2i-1$ и $2i$ соответствуют переменной z_i, координаты $2n+3j-2$, $2n+3j-1$ и $2n+3j$ — дизъюнкции C_j, а координаты $2n+3m+2i$, $2n+3m+2i+1$ — литералам z_i и $\bar{z}_i$.

Ниже приведён пример задачи ALSSVS, соответствующей примеру задачи 3-SAT при $n=2$ и $m=2$ с переменными z_1, z_2 и дизъюнкциями $z_1 \vee z_2 \vee \bar{z}_2$ и $z_1 \vee \bar{z}_1 \vee z_2$ (вектор v_i совпадает с i-й строкой матрицы, $i=1,2,\ldots,10$):

$$\begin{pmatrix}
a\sqrt{3} & -a\sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b & c & 1 & 1 & 1 \\
-a\sqrt{3} & a\sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b & 1 & c & 1 & 1 \\
0 & 0 & a\sqrt{3} & -a\sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & b & 1 & 1 & c & 1 \\
0 & 0 & -a\sqrt{3} & a\sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & b & 1 & 1 & 1 & c \\
0 & 0 & 0 & 0 & 2a & -a & -a & 0 & 0 & 0 & b & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & -a & 2a & -a & 0 & 0 & 0 & b & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & -a & -a & 2a & 0 & 0 & 0 & b & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2a & -a & -a & b & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -a & 2a & -a & b & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -a & -a & 2a & b & 0 & 1 & 1 & 1
\end{pmatrix}.$$

Покажем, что подмножество векторов $U \subseteq V$, удовлетворяющее условию (4), существует тогда и только тогда, когда найдётся такое назначение истинности переменных, что каждая из дизъюнкций содержит истинный литерал.

Предположим, что набор дизъюнкций выполним. Для каждого $i=1,2,\ldots,n$ включим в U вектор v_{2i-1}, если переменная z_i ложна, и вектор v_{2i}, если она истинна. Кроме того, добавим в U вектор $v_{2n+3j-k_j+1}$, где k_j, $j=1,2,\ldots,m$, — номер первого истинного литерала, входящего в дизъюнкцию C_j. Таким образом, множество U состоит из $m+n$ векторов. Обозначим их сумму через u, k-ю координату вектора u — через $u(k)$ и оценим $\|u\|^2$.

Поскольку для каждого $i=1,2,\ldots,n$ ровно один из векторов v_{2i-1}, v_{2i} входит в U, то $u(2i-1)=\pm a\sqrt{3}$, $u(2i)=\mp a\sqrt{3}$, а сумма квадратов этих координат равна $6a^2$. Так как для каждого $j=1,2,\ldots,m$ ровно один из векторов $v_{2n+3j-2}$, $v_{2n+3j-1}$, v_{2n+3j} входит в U, то

$$u^2(2n+3j-2)+u^2(2n+3j-1)+u^2(2n+3j)=6a^2.$$

Очевидно, что $u(2n+3m+1)=(m+n)b$. По определению, если z_i истинна,

то $u(2n+3m+2i) \geqslant c$, а если она ложна, то $u(2n+3m+2i+1) \geqslant c$. В любом случае правая координата вектора u не меньше c тогда и только тогда, когда соответствующий ей литерал ложен. Но отсюда следует, что эта координата не равна 0 во всех векторах из U.

Действительно, правая координата вектора из V равна 0 лишь если она соответствует входящему в дизъюнкцию литералу. Но для каждого $j = 1, 2, \ldots, m$ множество U содержит вектор, соответствующий истинному литералу, входящему в дизъюнкцию C_j. Следовательно, все правые координаты этого вектора, соответствующие ложным литералам, должны быть равны 1. Таким образом, в u найдутся n правых координат, равных $c+m+n-1$. Имеем

$$\frac{\|u\|^2}{|U|} \geqslant \frac{6(m+n)a^2 + (m+n)^2b^2 + n(c+m+n-1)^2}{m+n}$$
$$= 6a^2 + (m+n)b^2 + n(c+m+n-1)^2/(m+n) = K,$$

что и требовалось.

Допустим теперь, что существует подмножество $U \subseteq V$, удовлетворяющее условию (4), и пусть $t = |U|$. Как и прежде, через u обозначим сумму всех векторов из U. Очевидно, что $u(2n+3m+1) = tb$ и $u(k) \leqslant c + 2n + 3m - 1$ для каждой правой координаты k. Заметим, что сумма квадратов левых координат вектора u, соответствующих переменной z_i, равна $6a^2$, если ровно один из векторов v_{2i-1}, v_{2i} принадлежит U, и равна 0 в противном случае. Аналогично сумма квадратов левых координат вектора u, соответствующих дизъюнкции C_j, равна $6a^2$, если один или два из векторов $v_{2n+3j-2}$, $v_{2n+3j-1}$, v_{2n+3j} принадлежат U, и равна 0 в противном случае. Покажем, что U содержит ровно $m+n$ векторов.

Предположим, что $1 \leqslant t < m+n$. Тогда

$$\|u\|^2/t \leqslant (6ta^2 + (tb)^2 + 2n(c+2n+3m-1)^2)/t$$
$$= 6a^2 + tb^2 + 2n(c+2n+3m-1)^2/t \leqslant 6a^2 + (m+n)b^2 + 2n(c+2n+3m-1)^2 - b^2$$
$$= K - (b^2 - 2n(c+2n+3m-1)^2 + n(c+m+n-1)^2/(m+n)) < K$$

по выбору b. Таким образом, U не удовлетворяет условию (4), что противоречит нашему предположению.

Допустим теперь, что $3m+2n \geqslant t > m+n$. Тогда

$$\|u\|^2/t \leqslant (6(m+n)a^2 + (tb)^2 + 2n(c+2n+3m-1)^2)/t$$
$$= 6a^2 - (t-m-n)6a^2/t + tb^2 + 2n(c+2n+3m-1)^2/t$$

$$\begin{aligned}
&\leqslant 6a^2 + (3m+2n)b^2 + 2n(c+2n+3m-1)^2 - 6a^2/(3m+2n) \\
&\quad = K - (6a^2/(3m+2n) - (2m+n)b^2 - 2n(c+2n+3m-1)^2 \\
&\qquad + n(c+m+n-1)^2/(m+n)) \\
&\qquad = K - (6a^2/(3m+2n) - (2m+n+1)b^2 + 1) < K
\end{aligned}$$

в силу выбора a. Снова имеем противоречие с выбором U.

Тем самым U содержит ровно $m+n$ векторов. Если хотя бы одна из левых координат вектора u равна 0, то

$$\begin{aligned}
\|u\|^2/(m+n) &\leqslant (6(m+n-1)a^2 + ((m+n)b)^2 + 2n(c+2n+3m-1)^2)/(m+n) \\
&\leqslant K - 6a^2/(m+n) - n(c+m+n+1)^2/(m+n) \\
&\qquad + 2n(c+2n+3m-1)^2)/(m+n) < K,
\end{aligned}$$

противоречие. Следовательно, все левые координаты вектора u ненулевые, что с учётом условия $t = m+n$ может быть лишь в том и только в том случае, когда для любых $i = 1, 2, \ldots, n$ и $j = 1, 2, \ldots, m$ множество U содержит ровно один из векторов v_{2i-1}, v_{2i} и ровно один из векторов $v_{2n+3j-2}$, $v_{2n+3j-1}$, v_{2n+3j}. Таким образом, сумма u содержит ровно n правых координат, больших или равных c. Причём если хотя бы одна из них меньше $c+m+n-1$, то

$$\begin{aligned}
\|u\|^2/(m+n) &\leqslant (6(m+n)a^2 + ((m+n)b)^2 + (n-1)(c+n+m-1)^2 \\
&\qquad + (c+n+m-2)^2 + n(m+n)^2)/(m+n) \\
&\leqslant K + ((c+n+m-2)^2 - (c+n+m-1)^2)/(m+n) + n(m+n) \\
&= K + n(m+n) - (2c+2n+2m-3)/(m+n) < K
\end{aligned}$$

по выбору c, чего быть не может, так как U удовлетворяет условию (4).

Пусть для каждого $i = 1, 2, \ldots, n$ переменная z_i истинна, если множество U содержит вектор v_{2i}, и ложна, если оно содержит v_{2i-1}. Поскольку u имеет n правых координат, равных $c+n+m-1$, каждая дизъюнкция содержит по крайней мере один истинный литерал.

Остаётся заметить, что сводимость полиномиальна. Действительно, поскольку фигурируемые при сведении иррациональные числа могут быть записаны лишь приблизительно, следует так выбрать точность приближения, чтобы погрешность не повлияла на справедливость приведённых в доказательстве рассуждений и при этом сводимость осталась полиномиальной. В первую очередь это относится к числам $a\sqrt{3}$, которые присутствуют в записи матрицы. Заметим, что если выбрать $\varepsilon \leqslant$

$1/(2(n+m+1)a^2\sqrt{3})$, то $a^2(\sqrt{3}\pm\varepsilon)^2-3a^2=\pm 2a^2\sqrt{3}\varepsilon+a^2\varepsilon^2\leqslant 1/(n+m)$. Таким образом, суммарная погрешность не превысит 1, что никак не повлияет на справедливость приведённых выкладок. С другой стороны, так как a ограничено полиномом от m и n, то сводимость является полиномиальной.

Так как элементы матрицы ограничены полиномом от m и n, то задача является NP-полной в сильном смысле. Теорема 1 доказана.

3. Приближённый алгоритм решения задачи

Для произвольного непустого конечного множества векторов U определим функцию

$$F(U)=\frac{\Big\|\sum\limits_{v\in U} v\Big\|^2}{|U|}.$$

Покажем, что в случае $q=1$ задача ALSSVS разрешима за время $O(N\log N)$. К множеству $V\subset\mathbb{R}$, $|V|=N$, применим

Алгоритм $\mathcal{A}_1$.

Шаг 1. Разобьём множество V на два подмножества

$$V^+=\{v\in V\mid v>0\},\ \ V^-=\{v\in V\mid v<0\}.$$

Упорядочим их элементы по невозрастанию абсолютных величин так, что

$$V^+=\{v_1^+\geqslant v_2^+\geqslant\ldots\geqslant v_{|V^+|}^+>0\},\ \ V^-=\{v_1^-\leqslant v_2^-\leqslant\ldots\leqslant v_{|V^-|}^-<0\}.$$

Шаг 2. Вычислим значения

$$F_i^+=F\big(\{v_1^+,v_2^+,\ldots,v_i^+\}\big),\ \ i=1,\ldots,|V^+|;$$

положим

$$F^+=\max\big\{F_1^+,\ldots,F_{|V^+|}^+\big\},\ \ U^+=\big\{\{v_1^+,v_2^+,\ldots,v_i^+\}\mid F_i^+=F^+\big\}.$$

Шаг 3. Вычислим значения

$$F_i^-=F\big(\{v_1^-,v_2^-,\ldots,v_i^-\}\big),\ \ i=1,\ldots,|V^-|;$$

положим

$$F^-=\max\big\{F_1^-,\ldots,F_{|V^-|}^-\big\},\ \ U^-=\big\{\{v_1^-,v_2^-,\ldots,v_i^-\}\mid F_i^-=F^-\big\}.$$

Шаг 4. Положим $\widehat{F} = \max\{F^+, F^-\}$, а также $\widehat{U} = U^+$, если $F^+ \geqslant F^-$, и $\widehat{U} = U^-$, если $F^+ < F^-$. Подмножество $\widehat{U}$ объявляем результатом работы алгоритма.

Алгоритм $\mathcal{A}_1$, очевидно, находит оптимальное решение — подмножество $\widehat{U} \subseteq V$ — одномерной задачи ALSSVS, так как оптимальное подмножество не может содержать ни чисел разных знаков, ни меньшее по абсолютной величине число при отсутствии большего. Трудоёмкость алгоритма $\mathcal{A}_1$ определяется шагом 1, для выполнения которого необходимо $O(N \log N)$ операций.

Описанный ниже алгоритм $\mathcal{A}$ находит приближённое решение задачи ALSSVS — подмножество $U_A \subseteq V$ — в пространстве $\mathbb{R}^q$. По своей структуре и идее поиска решения с использованием вспомогательного конечного множества эталонных (опорных) векторов он сходен с алгоритмами, изложенными в [1, 5, 10] и ориентированными на решение близких по смыслу задач поиска подмножеств векторов.

Алгоритм $\mathcal{A}$.

На входе алгоритма $\mathcal{A}$ — множество $V = \{v_1, \ldots, v_N\}$ векторов из $\mathbb{R}^q$ и натуральный параметр $L \geqslant \sqrt{(q-1)/8}$. Алгоритм $\mathcal{A}$ строит специальное семейство решений. Затем из этого семейства выбирается наилучший элемент.

Перед формальным описанием алгоритма определим вспомогательное множество

$$H(L) = \{h \in \mathbb{Z}^q \mid \max\{|h(1)|, |h(2)|, \ldots, |h(q)|\} = L\},$$

где $\mathbb{Z}$ — множество целых чисел, а $h(k)$, $k = 1, \ldots, q$, — k-я компонента вектора h. Для мощности этого множества справедлива оценка

$$|H(L)| \leqslant 2q(2L+1)^{q-1},$$

которая следует из неравенства

$$|H(L)| \leqslant \sum_{k=1}^{q} |H_k(L)|,$$

где

$$H_k(L) = \{h \in \mathbb{Z}^q \mid |h(k)| = L;\ |h(m)| \leqslant L,\ m \in \{1, \ldots, q\} \setminus \{k\}\}.$$

Пусть множество $H(L)$ упорядочено, например, в лексикографическом порядке. Обозначим i-й вектор в этом порядке через h_i.

Шаг 1. Положим $i = 1$.

Шаг 2. Вычислим $u_j^i = (v_j, h_i)/\|h_i\|$, $j = 1, 2, \ldots, N$. Для множества $\{u_1^i, \ldots, u_N^i\}$ с помощью алгоритма $\mathcal{A}_1$ найдём оптимальное решение одномерной задачи ALSSVS — подмножество $\widehat{U}_i \subseteq \{u_1^i, \ldots, u_N^i\}$. Найдём подмножество

$$V_i = \{v \mid v \in V,\ (v, h_i)/\|h_i\| = u_j^i,\ u_j^i \in \widehat{U}_i,\ j = 1, \ldots, N\}$$

множества V по элементам подмножества $\widehat{U}_i$. Вычислим $\widehat{F}_i = F(V_i)$.

Шаг 3. Если $i < |H(L)|$, то полагаем $i := i + 1$ и выполняем шаги 2 и 3. Иначе переходим к шагу 4.

Шаг 4. Из семейства $\{V_i \mid i = 1, 2, \ldots, |H(L)|\}$ решений, полученного на шагах 2 и 3, выбираем такой элемент U_A, что

$$F_A = F(U_A) = \max\{\widehat{F}_i \mid i = 1, 2, \ldots, |H(L)|\}.$$

Подмножество $U_A \subseteq V$ объявляем результатом работы алгоритма $\mathcal{A}$.

Оценку сложности и точности алгоритма даёт следующая

Теорема 2. *Алгоритм $\mathcal{A}$ имеет временну́ю сложность*

$$O(Nq(q + \log N)(2L + 1)^{q-1})$$

и находит решение задачи ALSSVS *с гарантированной относительной погрешностью, не превышающей* $(q - 1)/(4L^2)$.

Доказательство. Шаг 2, определяющий трудоёмкость алгоритма, выполняется $|H(L)|$ раз. Для вычисления скалярных произведений N входных векторов на вектор h_i требуется $O(Nq)$ операций. Трудоёмкость нахождения подмножеств $\widehat{U}_i$ и V_i есть величина порядка $O(N \log N)$. Поэтому шаг 2 выполняется за $O(Nq + N \log N)$ операций, а временна́я сложность всего алгоритма есть величина порядка

$$O(|H(L)|N(q + \log N)) = O(Nq(2L + 1)^{q-1}(q + \log N)).$$

Пусть U^* — оптимальное решение задачи ALSSVS и $u^* = \sum\limits_{v \in U^*} v$. Тогда для оптимального значения $F^* = F(U^*)$ целевой функции F имеем $F^* = \|u^*\|^2/|U^*|$.

Пусть h — вектор, максимально близкий по углу к вектору u^* среди всех векторов из множества $H(L)$, а φ — угол между u^* и h. Положим

$$s = \frac{u^* L}{\max\limits_{k=1,\ldots,q} |u^*(k)|}$$

и рассмотрим вектор $\widetilde{u} = (\lfloor s(1) \rceil, \lfloor s(2) \rceil, \ldots, \lfloor s(q) \rceil)$, где $\lfloor x \rceil$ — ближайшее к x целое число. Угол между $\widetilde{u}$ и u^* (или, что то же самое, между $\widetilde{u}$ и s) обозначим через ψ. Очевидно, что $\widetilde{u} \in H(L)$ и $\varphi \leqslant \psi$.

Применяя теорему косинусов к треугольнику, образованному векторами $\widetilde{u}$ и s, получим оценку

$$\cos\psi = -\frac{\|\widetilde{u}-s\|^2 - \|s\|^2 - \|\widetilde{u}\|^2}{2\|s\|\|\widetilde{u}\|} = 1 - \frac{\|\widetilde{u}-s\|^2 - (\|s\| - \|\widetilde{u}\|)^2}{2\|s\|\|\widetilde{u}\|}$$
$$\geqslant 1 - \frac{\|\widetilde{u}-s\|^2}{2\|s\|\|\widetilde{u}\|} \geqslant 1 - \frac{q-1}{8(\min\{\|\widetilde{u}\|, \|s\|\})^2} \geqslant 1 - \frac{q-1}{8L^2},$$

поскольку каждая координата вектора $\widetilde{u} - s$ по абсолютной величине не превосходит $1/2$ и при этом хотя бы одна из них равна 0. Заметим, что $\cos\psi \geqslant 0$ по выбору L.

Далее, для приближённого решения U_A, найденного алгоритмом $\mathcal{A}$, имеем следующую оценку значения целевой функции:

$$F(U_A) \geqslant \widehat{F}_i = F(V_i) = \frac{\big\|\sum\limits_{v\in V_i} v\big\|^2}{|V_i|} \geqslant \frac{\big[\big(\sum\limits_{v\in V_i} v, h\big)\big]^2}{|V_i|\cdot\|h\|^2}$$
$$\geqslant \frac{\big[\big(\sum\limits_{v\in U^*} v, h\big)\big]^2}{|U^*|\cdot\|h\|^2} = \frac{(u^*,h)^2}{|U^*|\cdot\|h\|^2} = \frac{\|u^*\|^2\cos^2\varphi}{|U^*|} = F^*\cos^2\varphi.$$

Отсюда, учитывая, что $\varphi \leqslant \psi$, устанавливаем оценку

$$\frac{F^* - F(U_A)}{F^*} \leqslant 1 - \cos^2\varphi \leqslant 1 - \cos^2\psi \leqslant \frac{q-1}{4L^2} - \frac{(q-1)^2}{64L^4} \leqslant \frac{q-1}{4L^2}$$

относительной погрешности алгоритма. Теорема 2 доказана.

4. Заключение

Установлено, что рассмотренная задача ALSSVS не тождественна известной задаче MSSC евклидовой кластеризации множества векторов на два кластера по критерию минимума суммы квадратов. Её можно трактовать как специальный случай задачи MSSC, когда центр одного из кластеров фиксирован (известен) и равен нулю. Обоснован приближённый полиномиальный алгоритм решения задачи с гарантированной оценкой точности в случае фиксированной размерности пространства. Полученный результат, как базовый, имеет важное значение для установления

статуса комбинаторной сложности других труднорешаемых задач из целого семейства близких к рассмотренной (в содержательном плане) задач помехоустойчивого off-line анализа данных и распознавания образов, представленного в [3, 4]. Это семейство на сегодняшний день включает несколько сотен неизученных задач [13], для которых какие-либо алгоритмы с доказуемыми оценками точности неизвестны. Что касается задачи MSSC, то установление статуса её сложности представляется важным делом ближайшей перспективы.

ЛИТЕРАТУРА

1. **Бабурин А. Е., Гимади Э. Х., Глебов Н. И., Пяткин А. В.** Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. операций. Сер. 2. — 2007. — Т. 14, № 1. — С. 32–42.
2. **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.
3. **Кельманов А.В.** О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Сб. докл. 13-й Всеросс. конф. «Математические методы распознавания образов» (Зеленогорск, 30 сентября–6 октября 2007). — М.: МАКС Пресс, 2007. — С. 261–264.
4. **Кельманов А. В.** Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Российской конф. «Дискретная оптимизация и исследование операций» (Владивосток, 7–14 сентября 2007). — Новосибирск: Изд-во Ин-та математики, 2007. — С. 46–50. (http://math.nsc.ru/conference/door07/DOOR_abstracts.pdf).
5. **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.
6. **Aloise D., Hansen P.** On the complexity of minimum sum-of-squares clustering // Les Cahiers du GERAD, G-2007-50. — 2007. — 12 p.
7. **Drineas P., Frieze A., Kannan R., Vempala S., Vinay V.** Clustering large graphs via the singular value decomposition // Machine Learning, 56. — 2004. — P. 9–33.
8. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. — San Francisco, CA: Freeman, 1979. — 340 p.
9. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Transactions on Signal Processing. — 2004. — Vol. 52, N 3. — P. 1–12.

10. Kel'manov A. V., Khamidullin S. A., Okol'nishnikova L. V. A posteriori detection of identical subsequences in a quasiperiodic sequence // Pattern Recognition and Image Analysis. — 2002. — Vol. 12, N 4. — P. 438–447.

11. MacQueen J. Some methods for classification and analysis of multivariate observations // Proc. Fifth Berkeley Symp. Math. Statistics and Probability. — 1967. — Vol. 1. — P. 281–296.

12. Wald A. Sequential analysis. — New York: John Wiley, 1947. — 230 p.

13. http://math.nsc.ru/∼serge/qpsl/

Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru

Пяткин Артём Валерьевич,
e-mail: artem@math.nsc.ru

Статья поступила
1 апреля 2008 г.