А. В. Кельманов, А. В. Пяткин, О сложности некоторых задач выбора подпоследовательности векторов,
Ж. вычисл. матем. и матем. физ., 2012, том 52, номер 12, 2284–2291

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2012, том 52, № 12, с. 2284–2291

УДК 519.7

О СЛОЖНОСТИ НЕКОТОРЫХ ЗАДАЧ ВЫБОРА ПОДПОСЛЕДОВАТЕЛЬНОСТИ ВЕКТОРОВ[1)]

© 2012 г. А. В. Кельманов, А. В. Пяткин

(630090 Новосибирск, пр-т Акад. Коптюга, 4, Институт математики им. С.Л. Соболева Сибирского отделения РАН)

e-mail: kelm@math.nsc.ru, artem@math.nsc.ru

Поступила в редакцию 01.08.2011 г.

Доказана NP-полнота некоторых задач выбора подпоследовательности из последовательности векторов евклидова пространства, состоящей из конечного числа членов. Предполагается, что искомая подпоследовательность содержит фиксированное число векторов, близких между собой по критерию минимума суммы квадратов расстояний, причем выбор векторов подчинен условию: разность между номерами последующего и предыдущего выбираемых векторов ограничена сверху и снизу некоторыми константами. Библ. 9.

Ключевые слова: выбор подпоследовательности векторов, минимум суммы квадратов расстояний, кластерный анализ, алгоритмическая сложность, NP-полнота.

ВВЕДЕНИЕ

Предметом исследования настоящей работы являются дискретные экстремальные задачи выбора из последовательности векторов евклидова пространства, состоящей из конечного числа членов, подпоследовательности элементов, близких по критерию минимума суммы квадратов расстояний при наличии ограничений на номера выбираемых векторов. Цель исследования – анализ алгоритмической сложности этих задач.

Представленные ниже результаты дополняют [1], где было установлено, что к числу NP-трудных задач относятся следующие тесно связанные между собой оптимизационные задачи кластерного анализа и выбора подмножества в конечном множестве векторов евклидова пространства.

Задача VS-1 (Vector Subset 1).

Дано: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральное число $M > 1$. *Найти*: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$F_1(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\left\|\sum_{\mathbf{y}\in\mathcal{C}}\mathbf{y}\right\|^2 + \sum_{\mathbf{y}\in\mathcal{Y}\setminus\mathcal{C}}\|\mathbf{y}\|^2$$

максимальна при ограничении $|\mathcal{C}| = M$ на мощность подмножества $\mathcal{C}$.

Задача VS-2 (Vector Subset 2).

Дано: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральное число $M > 1$. *Найти*: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$F_2(\mathcal{C}) = \sum_{\mathbf{y}\in\mathcal{C}}\|\mathbf{y} - \overline{\mathbf{y}}(\mathcal{C})\|^2,$$

где $\overline{\mathbf{y}}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\Sigma_{\mathbf{y}\in\mathcal{C}}\mathbf{y}$, минимальна при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

1) Работа выполнена при финансовой поддержке РФФИ (коды проектов 12-01-00090, 12-01-00093, 10-07-00195, 11-07-12083-офи-м и 12-01-33028), целевой программы СО РАН (интеграционные проекты № 7Б и № 21А), ФЦП “Научные и научно-педагогические кадры инновационной России” (гос. контракт № 14.740.11.0362).

Задача VS-3 (Vector Subset 3).

Дано: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральное число $M > 1$. *Найти*: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$F_3(\mathcal{C}) = \sum_{\mathbf{y}\in\mathcal{C}}\sum_{\mathbf{z}\in\mathcal{C}}\|\mathbf{y} - \mathbf{z}\|^2$$

минимальна при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Задача MSSC-Case (Minimum Sum-of-Squares Clustering, special Case).

Дано: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$, натуральное число $M > 1$. *Найти*: разбиение множества $\mathcal{Y}$ на $J = N - M + 1$ непустых кластеров $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J$ такое, что мощность одного из кластеров равна M и

$$F_4(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J) = \sum_{j=1}^{J}\sum_{\mathbf{y}\in\mathcal{C}_j}\|\mathbf{y} - \overline{\mathbf{y}}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{\mathbf{y}}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|}\Sigma_{\mathbf{y}\in\mathcal{C}_j}\mathbf{y}$, $j = 1, 2, \ldots, J$, – центр j-го кластера.

Для целевых функций задач VS-1, VS-2 и VS-3 выполняются следующие соотношения (см. [1]):

$$F_2(\mathcal{C}) = \sum_{\mathbf{y}\in\mathcal{Y}}\|\mathbf{y}\|^2 - F_1(\mathcal{C}) = \frac{1}{2|\mathcal{C}|}F_3(\mathcal{C}). \tag{1}$$

Поэтому задачи VS-1, VS-2 и VS-3 полиномиально эквивалентны. Кроме того, если считать, что в задаче MSSC-Case мощность, например, первого кластера $\mathcal{C}_1$ зафиксирована и равна M, то имеет место равенство (см. [1])

$$F_4(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J) = F_2(\mathcal{C}_1), \tag{2}$$

так как из условий задачи следует, что мощности кластеров из совокупности $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J\}\backslash\mathcal{C}_1$ равны 1. Поэтому задачи MSSC-Case и VS-2 эквивалентны. Доказано (см. [1]), что в форме верификации свойств сформулированные задачи NP-полны в сильном смысле.

Одна из возможных содержательных трактовок проблемы анализа данных, которая приводит к решению сформулированных задач, состоит в следующем (см. [1]). Имеется таблица, содержащая результаты измерения набора числовых информационно значимых характеристик для совокупности некоторых материальных объектов. Часть объектов из этой совокупности идентична и имеет одинаковые характеристики. Число идентичных объектов известно. Оставшиеся объекты различны и имеют отличающиеся характеристики. В каждом результате измерения, представленном в таблице, есть ошибка, причем соответствие между объектом и набором неизвестно. Характеристики идентичных объектов, в отличие от характеристик остальных объектов, имеют принципиальную информационную ценность. Требуется, используя критерий минимума суммы квадратов расстояний, найти подмножество наборов, соответствующих идентичным объектам, и оценить по результатам измерения набор характеристик этих объектов (учитывая, что данные содержат ошибку измерения).

Задачи, рассмотренные в настоящей работе, индуцируются близкой в содержательном плане проблемой. Отличие этой проблемы от сформулированной выше состоит лишь в том, что элементы таблицы упорядочены по времени, причем известно, что временно́й интервал между двумя последовательными результатами измерения характеристик идентичных объектов ограничен сверху и снизу некоторыми константами. Подобные этой содержательные проблемы с временны́ми ограничениями на результаты измерения каких-либо информационно значимых характеристик весьма актуальны, в частности, при помехоустойчивой off-line обработке числовых и векторных последовательностей (см., например, [2]–[7] 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 [1], рассмотренные ниже дискретные экстремальные задачи по своей сути являются аналогами приведенных выше задач. В рассматриваемых задачах предполагается, что входными данными являются не множества, а векторные последовательности, причем имеются ограничения на номера выбираемых векторов из входной последовательности. Эти ограничения соответствуют априорным данным о времени измерений характеристик идентичных объектов. Мотивацией исследований послужил тот факт, что статус сложности этих задач был неизвестен.

1. ЗАДАЧИ ВЫБОРА ПОДПОСЛЕДОВАТЕЛЬНОСТИ

Для учета ограничений на номера выбираемых векторов в приведенных ниже формулировках задач при записи целевых функций вместо суммирования по элементам множеств (см. предыдущий раздел) используется суммирование по номерам (индексам) элементов последовательности. Кроме того, с этой же целью в формулировки задач вводятся натуральные параметры $T_{\min}$ и $T_{\max}$.

Положим $\mathcal{N} = \{1, 2, \dots, N\}$. В форме верификации свойств задачи на последовательностях с ограничениями формулируются следующим образом.

Задача VSS1($T_{\min}, T_{\max}$) (Vector Subsequence in a Sequence 1).

Дано: последовательность (набор) $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ векторов из $\mathcal{R}^q$, натуральное число $M > 1$ и положительное число A. *Вопрос*: существует ли подмножество $\mathcal{M} = \{n_1, n_2, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов набора $\mathcal{Y}$ такое, что

$$f_1(\mathcal{M}) = \frac{1}{|\mathcal{M}|}\left\|\sum_{m=1}^{M} \mathbf{y}_{n_m}\right\|^2 + \sum_{m \notin \mathcal{M}} \|\mathbf{y}_m\|^2 \geq A,$$

при ограничениях

$$1 \leq T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N - 1, \quad m = 2, 3, \dots, M, \tag{1.1}$$

на элементы подмножества $\mathcal{M}$?

Задача VSS2($T_{\min}, T_{\max}$) (Vector Subsequence in a Sequence 2).

Дано: последовательность $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ векторов из $\mathcal{R}^q$, натуральное число $M > 1$ и положительное число B. *Вопрос*: существует ли подмножество $\mathcal{M} = \{n_1, n_2, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такое, что

$$f_2(\mathcal{M}) = \sum_{m=1}^{M} \|\mathbf{y}_{n_m} - \overline{\mathbf{y}}(\mathcal{M})\|^2 \leq B,$$

где $\overline{\mathbf{y}}(\mathcal{M}) = \frac{1}{|\mathcal{M}|}\Sigma_{n \in \mathcal{M}} \mathbf{y}_n$, при ограничениях (1.1) на элементы искомого подмножества $\mathcal{M}$?

Задача VSS3($T_{\min}, T_{\max}$) (Vector Subsequence in a Sequence 3).

Дано: последовательность $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ векторов из $\mathcal{R}^q$, натуральное число $M > 1$ и положительное число C. *Вопрос*: существует ли подмножество $\mathcal{M} = \{n_1, n_2, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такое, что

$$f_3(\mathcal{M}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \|\mathbf{y}_{n_i} - \mathbf{y}_{n_j}\|^2 \leq C,$$

при ограничениях (1.1) на элементы искомого подмножества $\mathcal{M}$?

Задача MSSC-Case-S($T_{\min}, T_{\max}$) (Minimum Sum-of-Squares Clustering, special Case for a Sequence).

Дано: последовательность $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ векторов из $\mathcal{R}^q$, натуральное число $M > 1$ и положительное число D. *Вопрос*: существует ли разбиение множества $\mathcal{N}$ на $J = N - M + 1$ непустых подмножеств $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_J$ такое, что $|\mathcal{M}_1| = M$ и

$$f_4(\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_J) = \sum_{j=1}^{J} \sum_{m \in \mathcal{M}_j} \|\mathbf{y}_m - \overline{\mathbf{y}}(\mathcal{M}_j)\|^2 \leq D,$$

где $\overline{\mathbf{y}}(\mathcal{M}_j) = \frac{1}{|\mathcal{M}_j|}\Sigma_{n \in \mathcal{M}_j} \mathbf{y}_n$, $j = 1, 2, \dots, J$, при ограничениях (1.1) на элементы подмножества $\mathcal{M}_1$?

2. АНАЛИЗ АЛГОРИТМИЧЕСКОЙ СЛОЖНОСТИ

Во-первых, отметим, что в случае $T_{\min} = T_{\max}$ все задачи, очевидно, решаются за полиномиальное время. Поэтому далее всюду будем считать, что $T_{\min} < T_{\max}$.

Во-вторых, заметим, что задачи в парах VS-1 и VSS1$(1, N-1)$, VS-2 и VSS2$(1, N-1)$, VS-3 и VSS3$(1, N-1)$, а также MSSC-Case и MSSC-Case-S$(1, N-1)$ эквивалентны. Поэтому NP-полнота задач VSS1$(1, N-1)$, VSS2$(1, N-1)$, VSS3$(1, N-1)$, а также MSSC-Case-S$(1, N-1)$ следует из результатов [1], где была показана NP-полнота задач VS1, VS2, VS3 и MSSC-Case.

Наконец, в-третьих, легко проверить, что целевые функции задач выбора подпоследовательностей связаны формулами

$$f_2(\mathcal{M}) = \sum_{n\in\mathcal{M}} \|\mathbf{y}_n\|^2 - f_1(\mathcal{M}) = \frac{1}{2|\mathcal{M}|} f_3(\mathcal{M}), \tag{2.1}$$

$$f_4(\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J) = f_2(\mathcal{M}_1), \tag{2.2}$$

которые аналогичны (1) и (2). Следовательно, задачи VSS1$(T_{\min}, T_{\max})$, VSS2$(T_{\min}, T_{\max})$ и VSS3$(T_{\min}, T_{\max})$ полиномиально эквивалентны, а задача MSSC-Case-$(T_{\min}, T_{\max})$ эквивалентна задаче VSS2$(T_{\min}, T_{\max})$. Поэтому в случае, когда параметры $T_{\min}$ и $T_{\max}$ не являются частью входа, достаточно проанализировать сложность одной из этих задач. Докажем NP-полноту задачи VSS3$(T_{\min}, T_{\max})$.

Заметим, что доказательство существенно отличается для случаев $T_{\max} \geq 2T_{\min}$ и $T_{\max} \leq 2T_{\min} - 1$, а именно: к этим случаям VSS3$(T_{\min}, T_{\max})$ полиномиально сводятся две следующие разные NP-полные задачи.

Задача Clique in a regular graph (Клика в однородном графе).

Дано: однородный граф G степени $d \geq 2$ и натуральное число k. *Вопрос*: существует ли в этом графе такое подмножество вершин мощности k, что любые две вершины из этого подмножества связаны ребром?

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 ребер с концами в разных подмножествах не меньше t?

NP-полнота первой из этих задач доказана в [8]. Вторая задача относится к числу классических NP-полных задач (см. [9]).

NP-полнота задачи VSS3$(T_{\min}, T_{\max})$ вытекает из следующих двух лемм. Перед их доказательством заметим, что эта задача, очевидно, принадлежит классу NP для любых натуральных $T_{\min}$ и $T_{\max}$.

Лемма 1. *Если $T_{\max} \geq 2T_{\min}$, то задача* VSS3$(T_{\min}, T_{\max})$ *NP-полна в сильном смысле.*

Доказательство. Используя пример задачи Clique in a regular graph, т.е. однородный граф G степени d, содержащий p вершин и q ребер, а также число k, построим следующий пример задачи VSS3$(T_{\min}, T_{\max})$. Положим $M = k + p - 1$ и $C = 2(d-1)(k-1)k + 2kd(p-1)$.

Входную последовательность $\mathcal{Y}$ составим из $(q+1)$-мерных векторов трех типов: нулевых, основных и вспомогательных. Возьмем $p-1$ нулевых, p основных и $2(p-1)(T_{\min}-1)$ вспомогательных векторов, т.е. всего $N = 2(p-1)T_{\min} + 1$ членов.

Каждой вершине v_i графа G поставим в соответствие $(q+1)$-мерный основной вектор $\mathbf{u}_i$, в котором $(q+1)$-ая координата равна нулю, а j-я координата равна 1, если ребро e_j инцидентно вершине v_i, и 0 в противном случае $(i = 1, 2\ldots, p,\ j = 1, 2\ldots, q)$. Заметим, что у основных векторов d координат равны 1, а остальные равны 0 по построению.

Положим $X = \left\lceil \sqrt{C+1} \right\rceil$ и в качестве m-го вспомогательного вектора возьмем вектор, у которого $(q+1)$-я координата равна mX, $m = 1, 2, \ldots, 2(p-1)(T_{\min}-1)$, а все остальные координаты равны 0. В нулевом векторе все компоненты равны 0.

Последовательность $\mathcal{Y}$ векторов сформируем по следующему правилу. Сначала в произвольном порядке запишем все основные векторы. Затем в эту последовательность между основными векторами вставим нулевые векторы. Далее, в полученную последовательность между каждыми двумя последовательными векторами вставим $T_{\min} - 1$ вспомогательных векторов. Схематически построенная последовательность имеет следующий вид (для примера $T_{\min} = 4$):

$$\mathcal{Y} = \mathbf{y}_1\mathbf{y}_2\ldots\mathbf{y}_N = \mathbf{u}_1\mathbf{xxx0xxxu}_2\mathbf{xxx0xxxu}_3\ldots\mathbf{0xxxu}_p,$$

где $\mathbf{0}$ — нулевой вектор, $\mathbf{x}$ — вспомогательный вектор, а $\mathbf{u}_i$ — основной вектор.

Покажем, что у построенной последовательности $\mathcal{Y}$ (т.е. в примере задачи VSS3($T_{\min}, T_{\max}$)) существует набор $\mathcal{M}$ из M номеров, компоненты которого удовлетворяют условиям (1.1) и $f_3(\mathcal{M}) \le C$, тогда и только тогда, когда в графе G (т.е. в примере задачи Клика в однородном графе) есть клика мощности k.

Сначала вычислим квадраты расстояний между векторами. Очевидно, что квадрат расстояния от вспомогательного вектора до любого другого вектора не меньше $X^2 \ge C + 1$. Квадрат расстояния от нулевого до любого основного вектора равен d, так как у основного вектора ровно d компонент равны 1, а остальные равны 0. Наконец, квадрат расстояния между основными векторами равен $2d - 2$, если соответствующие им вершины в графе G смежны, и $2d$, если они не смежны.

Необходимость. Если в G имеется клика размера k, то, выбрав в качестве совокупности $\mathcal{M}$ номера всех нулевых векторов из последовательности $\mathcal{Y}$, а также номера тех основных векторов этой последовательности, которые соответствуют вершинам клики, получим $f_3(\mathcal{M}) = C$. Выполнение условия (1.1) следует из построения и неравенства $T_{\max} \ge 2T_{\min}$.

Таким образом, у последовательности $\mathcal{Y}$ существует набор $\mathcal{M}$ номеров, компоненты которого удовлетворяют условиям задачи VSS3($T_{\min}, T_{\max}$).

Достаточность. Предположим, что у последовательности $\mathcal{Y}$ существует набор $\mathcal{M}$ номеров членов этой последовательности, удовлетворяющих условиям (1.1) и $f_3(\mathcal{M}) \le C$. Тогда из последнего неравенства следует, что набор $\mathcal{M}$ не содержит номеров вспомогательных векторов.

Обозначим через a число номеров основных векторов в наборе $\mathcal{M}$. Тогда $\mathcal{M}$ содержит $k + p - 1 - a$ номеров нулевых векторов, так как $|\mathcal{M}| = M = k + p - 1$ и $\mathcal{M}$ не содержит номеров вспомогательных векторов.

Из построения следует, что суммарное число нулевых векторов в последовательности не превосходит $p - 1$. Поэтому $k + p - 1 - a \le p - 1$. Отсюда вытекает неравенство $a \ge k$. С другой стороны, из построения следует, что число a основных векторов в последовательности не превосходит p. Таким образом, $a \in \{k, k+1, \ldots, p\}$.

Учитывая, что квадрат расстояния от нулевого вектора до любого основного равен d, а минимальный квадрат расстояния между основными векторами равен $2d - 2$, для целевой функции задачи имеем оценку

$$f_3(\mathcal{M}) \ge 2d(k + p - 1 - a)a + (2d - 2)(a - 1)a.$$

Нетрудно проверить, что правая часть этого неравенства, как функция $f(a)$, строго возрастает при $d \ge 2$ и $a \in [k, p]$. Следовательно, эта функция принимает наименьшее значение на левой границе отрезка $[k, p]$, т.е. в точке k, причем это значение $f(k) = C$. Поэтому для целевой функции имеем оценку $f_3(\mathcal{M}) \ge C$. Но по нашему предположению справедливо условие $f_3(\mathcal{M}) \le C$. Следовательно, для выполнения этого условия требуется, чтобы набор $\mathcal{M}$ содержал ровно k номеров основных векторов, квадрат расстояния между любыми двумя из которых равен $2d - 2$. Но тогда в графе G (из задачи Клика в однородном графе) соответствующие этим векторам вершины образуют клику. Лемма 1 доказана.

Поскольку $T_{\max} > T_{\min}$, из леммы 1 следует NP-полнота задачи VSS3($1, T_{\max}$) для любого $T_{\max} > 1$.

Замечание 1. NP-полнота в сильном смысле рассмотренного случая задачи VSS3($T_{\min}, T_{\max}$) вытекает из того, что числовые значения входных параметров этого случая задачи ограничены полиномом от размера входа задачи Clique in a regular graph.

Лемма 2. *Если $T_{\max} \le 2T_{\min} - 1$ и $T_{\min} \ge 2$, то задача* VSS3($T_{\min}, T_{\max}$) NP-*полна в сильном смысле.*

Доказательство. Построим следующий пример задачи VSS3($T_{\min}, T_{\max}$), используя пример задачи MaxCut, т.е. граф G с p вершинами и q ребрами, а также положительное целое число t. Обозначим через d_i, $i = 1, 2, \ldots, p$, степень i-й вершины графа G. Положим $M = 2p - 1$ и $C = 8pq - 4q - 8t$.

Входную последовательность $\mathcal{Y}$, содержащую всего $N = 2(p-1)T_{\min} + p + 1$ членов, составим из векторов трех типов: нулевых, вспомогательных и основных. Возьмем $p - 1$ нулевых, $N = 2(p-1)(T_{\min} - 1)$ вспомогательных и $2p$ основных векторов.

В качестве вспомогательных и нулевых используем те же векторы, что и при доказательстве леммы 1. Напомним, что в нулевом векторе все компоненты равны 0, а в m-м вспомогательном векторе $(q+1)$-ая координата равна mX, $m = 1, 2, \ldots, 2(p-1)(T_{\min} - 1)$, где $X = \left\lceil \sqrt{C+1} \right\rceil$, а остальные координаты равны 0.

Ориентируем все ребра графа G произвольным образом. Каждой вершине v_i, $i = 1, 2 \ldots, p$, ориентированного графа поставим в соответствие $(q+1)$-мерный вектор $\mathbf{u}_i$, в котором $(q+1)$-ая координата равна нулю, а j-я координата ($j = 1, 2 \ldots, q$) равна 1, если дуга e_j исходит из вершины v_i, равна -1, если дуга e_j входит в v_i, и равна 0, если дуга e_j неинцидентна v_i. Положим $\mathbf{z}_i = -\mathbf{u}_i$, $i = 1, 2 \ldots, p$. В качестве основных векторов возьмем векторы $\mathbf{u}_i$ и $\mathbf{z}_i$, $i = 1, 2 \ldots, p$.

Последовательность $\mathcal{Y}$ векторов сформируем по следующему правилу. Сначала в произвольном порядке запишем все пары, состоящие из основных векторов, и получим следующую последовательность: $\mathbf{u}_1\mathbf{z}_1\mathbf{u}_2\mathbf{z}_2\ldots\mathbf{u}_p\mathbf{z}_p$. Далее, в каждую третью позицию этой последовательности вставим нулевой вектор. Наконец, слева и справа от каждого нулевого вектора вставим по $T_{\min} - 1$ вспомогательных векторов. Схематически построенная последовательность имеет следующий вид (для примера $T_{\min} = 4$):

$$\mathcal{Y} = \mathbf{y}_1\mathbf{y}_2\ldots\mathbf{y}_N = \mathbf{u}_1\mathbf{z}_1\mathbf{xxx0xxxu}_2\mathbf{z}_2\mathbf{xxx0xxxu}_3\mathbf{z}_3\ldots\mathbf{0xxxu}_p\mathbf{z}_p.$$

Покажем, что у построенной последовательности $\mathcal{Y}$ существует набор $\mathcal{M}$ из M номеров, удовлетворяющий условиям (1.1) и $f_3(\mathcal{M}) \le C$, тогда и только тогда, когда в графе G (т.е. в задаче MaxCut) существует разрез, содержащий не менее t ребер.

Сначала вычислим квадраты расстояний между векторами. От вспомогательного вектора до любого другого вектора квадрат расстояния не меньше $X^2 \ge C + 1$. Квадрат расстояния от нулевого вектора до векторов $\mathbf{u}_i$ и $\mathbf{z}_i$ равен степени d_i вершины v_i графа G. Легко проверить, что если вершины v_i и v_k не смежны, то

$$\|\mathbf{u}_i - \mathbf{u}_k\|^2 = \|\mathbf{z}_i - \mathbf{z}_k\|^2 = \|\mathbf{u}_i - \mathbf{z}_k\|^2 = \|\mathbf{z}_i - \mathbf{u}_k\|^2 = d_i + d_k, \tag{2.3}$$

а если вершины v_i и v_k смежны, то

$$\|\mathbf{u}_i - \mathbf{u}_k\|^2 = \|\mathbf{z}_i - \mathbf{z}_k\|^2 = d_i + d_k + 2, \tag{2.4}$$

$$\|\mathbf{u}_i - \mathbf{z}_k\|^2 = \|\mathbf{z}_i - \mathbf{u}_k\|^2 = d_i + d_k - 2. \tag{2.5}$$

Необходимость. Если граф G содержит разрез $\{V_1, V_2\}$, где $V_1 \cap V_2 = \varnothing$, $V_1 \cup V_2 = V(G)$, с $t_0 \ge t$ ребрами, то выберем в набор $\mathcal{M}$ номера всех нулевых векторов и номера всех векторов $\mathbf{u}_i$, для которых $v_i \in V_1$, а также номера всех векторов $\mathbf{z}_i$, для которых $v_i \in V_2$. Нетрудно заметить, что разность номеров соседних элементов в полученном наборе $\mathcal{M}$ равна $T_{\min}$ или $T_{\min} + 1$, откуда следует выполнение условия (1.1).

Для i-й вершины графа G через a_i обозначим число инцидентных ей ребер разреза. Тогда мощность разреза равна $t_0 = \Sigma_{i=1}^p a_i/2$.

Вычислим значение функции $f_3(\mathcal{M}) = \Sigma_{n_m \in \mathcal{M}} \Sigma_{n_k \in \mathcal{M}} \|\mathbf{y}_{n_m} - \mathbf{y}_{n_k}\|^2$. Сначала заметим, что если $\mathbf{y}_{n_m} = \mathbf{0}$, то

$$\sum_{n_k \in \mathcal{M}} \|\mathbf{y}_{n_m} - \mathbf{y}_{n_k}\|^2 = \sum_{i=1}^{p} d_i = 2q,$$

так как квадрат расстояния от нулевого до основного вектора равен d_i, набор $\mathcal{M}$ содержит номера векторов последовательности $\mathcal{Y}$, соответствующих всем вершинам графа G, а сумма степеней всех вершин графа равна удвоенному числу ребер.

Если $\mathbf{y}_{n_m} = \mathbf{u}_i$ или $\mathbf{y}_{n_m} = \mathbf{z}_i$, то сумма $\Sigma_{n_k\in\mathcal{M}}\|\mathbf{y}_{n_m} - \mathbf{y}_{n_k}\|^2$ содержит $M = 2p-1$ слагаемых, среди которых: 1) $p-1$, слагаемых, равных d_i (это квадраты расстояний до нулевых векторов), 2) одно слагаемое, равное 0 (при $n_k = n_m$), 3) $p-1-d_i$ слагаемых вида (2.3) (это квадраты расстояний до векторов, соответствующих несмежным с v_i вершинам v_k), 4) $d_i - a_i$ слагаемых вида (2.4), 5) a_i слагаемых вида (2.5). Поэтому, используя несложные преобразования и очевидное равенство $\Sigma_{k=1}^{p} d_k = 2q$, находим

$$\sum_{n_k\in\mathcal{M}}\|\mathbf{y}_{n_m} - \mathbf{y}_{n_k}\|^2 = (p-1)d_i + \sum_{k=1,\,k\neq i}^{p}(d_i + d_k) - 2a_i + 2(d_i - a_i) = (2p-1)d_i + 2q - 4a_i.$$

Суммируя найденные значения по всем номерам $n_m \in \mathcal{M}$, получаем

$$f_3(\mathcal{M}) = 2q(p-1) + \sum_{i=1}^{p}((2p-1)d_i + 2q - 4a_i) =$$
$$= 2q(p-1) + 2q(2p-1) + 2pq - 8t_0 = 8pq - 4q - 8t_0 \le C$$

при $t_0 \ge t$.

Таким образом, у последовательности $\mathcal{Y}$ существует набор $\mathcal{M}$ номеров, удовлетворяющий условиям (1.1) и $f_3(\mathcal{M}) \le C$.

Достаточность. Пусть имеется набор $\mathcal{M}$ номеров последовательности $\mathcal{Y}$, удовлетворяющих условиям (1.1) и $f_3(\mathcal{M}) \le C$. Тогда из последнего неравенства следует, что набор $\mathcal{M}$ не содержит номеров вспомогательных векторов.

Далее, заметим, что ни для какого $i = 1,2\ldots,p$ набор $\mathcal{M}$ не может содержать одновременно номера векторов $\mathbf{u}_i$ и $\mathbf{z}_i$, так как $T_{\min} \ge 2$. Отсюда, а также из условий $M = 2p-1$ и $T_{\max} \le 2T_{\min} - 1$ следует, что набор $\mathcal{M}$ содержит номера всех нулевых векторов и ровно по одному номеру векторов $\mathbf{u}_i$ или $\mathbf{z}_i$ для каждого i.

Рассмотрим разрез вершин графа G на два множества $V_1 = \{v_i\,|\mathbf{y}_{n_m} = \mathbf{u}_i,\, n_m \in \mathcal{M}\}$ и $V_2 = \{v_i\,|\mathbf{y}_{n_m} = \mathbf{z}_i,\, n_m \in \mathcal{M}\}$. Используя приведенные выше вычисления, нетрудно убедиться, что $f_3(\mathcal{M}) = 8pq - 4q - 8t_0$, где t_0 — мощность разреза. Поскольку $f_3(\mathcal{M}) \le C$, этот разрез имеет мощность не менее t. Лемма 2 доказана.

Замечание 2. NP-полнота в сильном смысле рассмотренного случая задачи VSS3($T_{\min}, T_{\max}$) следует из того, что числовые значения входных параметров этого случая задачи ограничены полиномом от размера входа задачи MaxCut.

Из лемм 1 и 2 вытекает

Теорема 1. *Для любых натуральных $T_{\min}$ и $T_{\max}$, удовлетворяющих неравенству $T_{\min} < T_{\max}$, задача* VSS3($T_{\min}, T_{\max}$) NP-*полна в сильном смысле*.

Из этой теоремы и формул (2.1), (2.2) вытекает

Следствие 1. Для любых натуральных $T_{\min}$ и $T_{\max}$, удовлетворяющих неравенству $T_{\min} < T_{\max}$, задачи VSS1($T_{\min}, T_{\max}$), VSS2($T_{\min}, T_{\max}$) и MSSC-Case-S($T_{\min}, T_{\max}$) NP-полны в сильном смысле.

ЗАКЛЮЧЕНИЕ

Показана NP-полнота оптимизационных задач, которые индуцируются проблемой поиска в последовательности векторов евклидова пространства, содержащей конечное число членов, такой подпоследовательности, что она имеет фиксированное число элементов и включает векторы, близкие между собой по критерию минимума суммы квадратов расстояний. Из полученного результата следует труднорешаемость соответствующей проблемы анализа упорядоченных по времени табличных данных. Остается заметить, что эффективные алгоритмы с гарантированными оценками точности для решения рассмотренных задач в настоящее время неизвестны.

СПИСОК ЛИТЕРАТУРЫ

1. *Кельманов А.В., Пяткин А.В.* NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. 2010. Т. 17. № 5. С. 37–45.
2. *Кельманов А.В., Хамидуллин С.А.* Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Ж. вычисл. матем. и матем. физ. 2001. Т. 41. № 5. С. 807–820.
3. *Гимади Э.Х., Кельманов А.В., Кельманова М.А., Хамидуллин С.А.* Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сибирск. ж. индустр. матем. 2006. Т. 9. № 1(25). С. 55–74.
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 // Ж. вычисл. матем. и матем. физ. 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 // Дискретный анализ и исследование операций. 2009. Т. 16. № 4. С. 31–46.
6. *Kel'manov A.V., Jeon B.* A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // IEEE Trans. Sign. Proc. 2004. V. 52. № 3. P. 645–656.
7. *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 and Image Analysis. 2010. V. 20. № 4. P. 451–458.
8. *Papadimitriou C.H.* Computational Complexity. New-York: Addison-Wesley, 1994.
9. *Garey M.R., Johnson D.S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman, 1979.