

УДК 519.2+621.391

## NP-ПОЛНОТА НЕКОТОРЫХ ЗАДАЧ ВЫБОРА ПОДМНОЖЕСТВА ВЕКТОРОВ <sup>\*)</sup>

*А. В. Кельманов, А. В. Пяткин*

**Аннотация.** Доказана NP-полнота некоторых задач выбора подмножества векторов евклидова пространства. К решению таких задач сводится одна из проблем анализа данных. Предполагается, что искомое подмножество имеет фиксированную мощность и включает векторы, «близкие» между собой по критерию минимума суммы квадратов расстояний.

**Ключевые слова:** выбор подмножества векторов, кластерный анализ, алгоритмическая сложность, NP-полнота.

### Введение

В статье анализируется статус сложности дискретных оптимизационных задач, к которым сводится одна из проблем поиска подмножества векторов евклидова пространства. Содержательная трактовка рассматриваемой проблемы состоит в следующем.

Имеется таблица, содержащая результаты измерения набора числовых информационно значимых характеристик для совокупности некоторых материальных объектов. Часть объектов из этой совокупности идентичны и имеют одинаковые характеристики. Число идентичных объектов известно. Оставшиеся объекты различны и имеют отличающиеся характеристики. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие между объектом и набором неизвестно. Требуется, используя адекватный измеряемым характеристикам критерий, найти подмножество наборов, соответствующих идентичным объектам и оценить по результатам измерения набор характеристик этих объектов (учитывая, что данные содержат ошибку измерения).

---

<sup>\*)</sup>Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 09-01-00032, 08-01-00516 и 10-07-00195), целевой программы АВЦП Рособразования (проект 2.1.1/3235), целевых программ №2 Президиума РАН (проект 227) и Сибирского отделения РАН (интеграционный проект 44).

Мотивацией исследований послужил тот факт, что статус сложности экстремальных задач, к которым сводится эта проблема анализа числовых данных, в случае, когда критерием поиска подмножества является минимум суммы квадратов расстояний между наборами (или векторами) из искомого подмножества, был неизвестен. Сложностной статус оптимизационных задач, к которым сводятся сходные в содержательном плане проблемы анализа данных, связанные с поиском подмножеств векторов, изучался в [1–8, 12].

В первой части этих работ [1, 2, 4–6] рассматривалась модель анализа данных, в которой предполагалось, что анализируемое множество векторов наряду с «близкими» по критерию минимума суммы квадратов расстояний векторами может содержать векторы, «похожие» на централизованный около нуля шум. Проблема анализа данных состояла в поиске в заданном множестве векторов непустого подмножества «близких» между собой векторов. В цитируемых работах установлено, что экстремальные задачи, к которым сводится эта проблема, переформулированные в форме верификации свойств, относятся к классу NP-полных задач.

Во второй части упомянутых работ [3, 7, 8, 12] рассматривалась модель структурированных данных, в которой заданное множество векторов содержит непересекающиеся подмножества — кластеры, включающие «близкие» (по критерию минимума суммы квадратов расстояний) между собой векторы. Проблема анализа данных состояла в разбиении заданного множества на подмножества по указанному критерию. Оптимизационная задача, к которой сводится эта проблема анализа данных, широко известна под названием MSSC (Minimum-Sum-of-Squares Clustering). В некоторых публикациях эта же задача фигурирует под названием  $k$ -means [8, 11]. В [3, 7, 8, 12] показано, что задача MSSC в форме верификации свойств NP-полна.

Несмотря на содержательное сходство проблемы анализа данных, рассматриваемой в настоящей работе, с проблемами, изученными в [1–8, 12], оптимизационные задачи, к которым сводится эта проблема, отличны как от задач, исследованных в [1, 2, 4–6], так и от задачи MSSC. В настоящей работе установлено, что оптимизационные задачи, к которым сводится рассматриваемая проблема, относятся к классу труднорешаемых задач. Вместе с этим ниже показано, что анализируемую проблему можно интерпретировать как неизученную ранее разновидность проблемы кластерного анализа.

### 1. Модель анализа данных

Рассмотрим следующую структуру данных, представленных в виде совокупности векторов евклидова пространства. Пусть векторная последовательность  $x_n \in \mathbb{R}^q$ ,  $n \in \mathcal{N}$ , где  $\mathcal{N} = \{1, \dots, N\}$ , обладает свойством

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

где  $\mathcal{M} \subset \mathcal{N}$ ,  $\mathcal{M} \neq \emptyset$ .

Допустим, что для обработки доступна последовательность

$$y_n = x_n + e_n, \quad n \in \mathcal{N}, \quad (2)$$

где  $e_n$  — вектор помехи (ошибки), независимый от вектора  $x_n$ . Учитывая зависимость элементов последовательности (1) от множеств, положим

$$S(\mathcal{M}, w, \{v_i, i \in \mathcal{N} \setminus \mathcal{M}\}) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \quad (3)$$

и рассмотрим модель анализа данных в виде следующей экстремальной задачи.

**Задача SVS** (Searching for Vector Subsets). *Дано*: последовательность  $y_n$ ,  $n \in \mathcal{N}$ , векторов из  $\mathbb{R}^q$  и натуральное число  $M > 1$ . *Найти*: подмножество  $\mathcal{M} \subset \mathcal{N}$ , вектор  $w$  и совокупность  $\{v_i, i \in \mathcal{N} \setminus \mathcal{M}\}$  векторов, минимизирующих  $S(\cdot)$ , при ограничении  $|\mathcal{M}| = M$  на мощность подмножества  $\mathcal{M}$  и при условии, что структура последовательности описывается формулами (1) и (2).

Эта задача соответствует сформулированной выше содержательной проблеме. В модели анализа данных вектор  $w$  можно интерпретировать как набор искомых характеристик идентичных объектов, а векторы  $v_i$ ,  $i \in \mathcal{N} \setminus \mathcal{M}$ , — как наборы характеристик остальных объектов совокупности. Мощность подмножества  $\mathcal{M}$  соответствует числу идентичных объектов.

Задачу SVS можно трактовать как поиск наилучшего варианта приближения по критерию минимума суммы квадратов отклонений последовательности (2) от последовательности (1), которая включает повторяющийся вектор, перемежающийся с произвольными векторами евклидова пространства. Нетрудно установить, что к минимизации функционала (3) приводит статистическая формулировка проблемы, если считать,

что вектор  $e_n$  есть выборка единичного объёма из  $q$ -мерного нормального распределения с параметрами  $(0, \sigma^2 I)$ , где  $I$  — единичная матрица, а в качестве критерия решения использовать максимум функционала правдоподобия.

Проанализируем возможные варианты оптимизационных задач, к которым сводится сформулированная задача SVS анализа данных.

## 2. Задачи выбора подмножества векторов

Раскрывая сумму квадратов в правой части (3) с учётом (1), получим

$$S = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \sum_{j \in \mathcal{M}} \{2(y_j, w) - \|w\|^2\} - \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \{2(y_i, v_i) - \|v_i\|^2\}. \quad (4)$$

Минимум функционала  $S(\cdot)$  по неизвестным векторам  $w$  и  $v_i$ ,  $i \in \mathcal{N} \setminus \mathcal{M}$ , находится аналитически. Используя (4), нетрудно убедиться, что для любого непустого подмножества  $\mathcal{M} \subset \mathcal{N}$  этот минимум доставляется векторами  $\bar{w} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$  и  $\bar{v}_j = y_j$ ,  $j \in \mathcal{N} \setminus \mathcal{M}$ , и равен

$$S_{\min}(\mathcal{M}) = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \left( \frac{1}{|\mathcal{M}|} \left\| \sum_{i \in \mathcal{M}} y_i \right\|^2 + \sum_{j \in \mathcal{N} \setminus \mathcal{M}} \|y_j\|^2 \right). \quad (5)$$

Заметим, что первый член в правой части этого равенства является константой. Следовательно, задача минимизации функционала  $S(\cdot)$ , сформулированная выше, сводится к задаче максимизации выражения в скобках в правой части (5). Сформулируем эту задачу максимизации в форме верификации свойств. Перед формулировкой задачи положим  $\mathcal{Y} = \{y_n \mid n \in \mathcal{N}\}$ ,  $\mathcal{C} = \{y_j \mid j \in \mathcal{M}\}$  и заменим в выражении (5) суммирование по индексам на суммирование по элементам множеств.

**Задача VS-1 (Vector Subset 1).** *Дано:* множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$ , натуральное число  $M > 1$  и положительное число  $A$ . *Вопрос:* существует ли такое подмножество  $\mathcal{C} \subset \mathcal{Y}$ , что имеет место неравенство

$$\frac{1}{M} \left\| \sum_{y \in \mathcal{C}} y \right\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \geq A, \quad (6)$$

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

Кроме того, заметим, что для правой части (5) в терминах суммирования по элементам множеств справедлива цепочка равенств

$$\begin{aligned} \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} \left\| \sum_{y \in \mathcal{C}} y \right\|^2 - \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 &= \sum_{y \in \mathcal{C}} \|y\|^2 - \frac{1}{|\mathcal{C}|} \left\| \sum_{y \in \mathcal{C}} y \right\|^2 \\ &= \sum_{y \in \mathcal{C}} \|y\|^2 - \frac{2}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} \left( y, \sum_{u \in \mathcal{C}} u \right) + \frac{1}{|\mathcal{C}|} \left\| \sum_{y \in \mathcal{C}} y \right\|^2 \\ &= \sum_{y \in \mathcal{C}} \|y\|^2 - 2 \sum_{y \in \mathcal{C}} (y, \bar{w}) + |\mathcal{C}| \|\bar{w}\|^2 = \sum_{y \in \mathcal{C}} \|y - \bar{w}\|^2, \quad (7) \end{aligned}$$

где  $\bar{w} = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ . Поэтому минимизация функционала  $S(\cdot)$  сводится к минимизации правой части (7). Отсюда получаем следующую задачу выбора подмножества.

**Задача VS-2 (Vector Subset 2).** *Дано:* множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$ , натуральное число  $M > 1$  и положительное число  $B$ . *Вопрос:* существует ли такое подмножество  $\mathcal{C} \subset \mathcal{Y}$ , что имеет место неравенство

$$\sum_{y \in \mathcal{C}} \|y - \bar{w}\|^2 \leq B, \quad (8)$$

где  $\bar{w} = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ , при ограничении  $|\mathcal{C}| = M$  на мощность подмножества  $\mathcal{C}$ ?

Наконец, заметим, что для любого множества  $\mathcal{Z}$  векторов евклидова пространства имеет место равенство [9]:

$$\sum_{z \in \mathcal{Z}} \|z - \bar{z}(\mathcal{Z})\|^2 = \frac{1}{2|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Z}} \|z - y\|^2, \quad (9)$$

где  $\bar{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ .

Учитывая, что мощность подмножества  $\mathcal{C}$  фиксирована, из (8) и (9) получим ещё одну задачу выбора подмножества.

**Задача VS-3 (Vector Subset 3).** *Дано:* множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$ , натуральное число  $M > 1$  и положительное число  $D$ . *Вопрос:* существует ли такое подмножество  $\mathcal{C} \subset \mathcal{Y}$ , что имеет место неравенство

$$\sum_{y \in \mathcal{C}} \sum_{z \in \mathcal{C}} \|y - z\|^2 \leq D, \quad (10)$$

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

Основным результатом настоящей работы является установление статуса NP-полноты сформулированных выше задач.

### 3. Анализ сложности

Напомним следующую вспомогательную NP-полную [10] задачу.

**Задача Клика.** *Дано:* граф  $G$  и натуральное число  $M$ . *Вопрос:* существует ли в этом графе такое подмножество вершин мощности  $M$ , что любые две вершины из этого подмножества связаны ребром?

Специальный случай этой задачи для однородных графов, степень которых не фиксирована, также относится к числу NP-полных задач [13]. Покажем, что задача Клика полиномиально сводится к задаче VS-3.

**Теорема 1.** *Задача VS-3 NP-полна в сильном смысле.*

**ДОКАЗАТЕЛЬСТВО.** Задача VS-3, очевидно, принадлежит классу NP.

Пусть  $G$  — однородный граф степени  $d$  с  $N$  вершинами и  $q$  рёбрами. Построим следующий пример задачи VS-3. Каждой вершине графа  $G$  сопоставим  $q$ -мерный вектор  $y$ , в котором  $i$ -я координата равна 1, если ребро  $i$  инцидентно этой вершине, и равна 0 в противном случае. Тогда для пары векторов  $y, z$  из множества  $\mathcal{Y} = \{y_1, \dots, y_N\}$ , очевидно, имеет место равенство  $\|y - z\|^2 = 2d$ , если соответствующие этим векторам вершины в графе  $G$  не смежны, и  $\|y - z\|^2 = 2d - 2$ , если эти вершины смежны. Следовательно, целевая функция (10) задачи VS-3 не превосходит числа  $D = 2(d - 1)M(M - 1)$  тогда и только тогда, когда граф  $G$  содержит клику на  $M$  вершинах.

Нетрудно установить, что построенное сведение полиномиально. NP-полнота задачи в сильном смысле следует из того, что при доказательстве NP-полная в сильном смысле задача Клика была сведена к частному случаю задачи VS-3 с бинарными векторами. Теорема 1 доказана.

**Теорема 2.** *Задача VS-2 NP-полна в сильном смысле.*

**ДОКАЗАТЕЛЬСТВО.** Ясно, что задача VS-2 принадлежит классу NP. Покажем, что к ней полиномиально сводится NP-полная задача VS-3.

Действительно, из (9) легко видеть, что в задаче VS-3 подмножество векторов, удовлетворяющее неравенству (10), существует тогда и только тогда, когда в задаче VS-2 при  $B = \frac{D}{2M}$  существует соответствующее подмножество векторов, удовлетворяющее неравенству (8). Теорема 2 доказана.

**Теорема 3.** *Задача VS-1 NP-полна в сильном смысле.*

ДОКАЗАТЕЛЬСТВО. Задача VS-1, очевидно, принадлежит классу NP, и к ней полиномиально сводится NP-полная задача VS-2.

В самом деле, из (7) следует, что в задаче VS-2 подмножество векторов, удовлетворяющее неравенству (8), существует тогда и только тогда, когда в задаче VS-1 при  $A = \sum_{y \in \mathcal{Y}} \|y\|^2 - B$  существует соответствующее подмножество векторов, удовлетворяющее неравенству (6). Теорема 3 доказана.

Сформулируем упомянутую во введении задачу MSSC кластерного анализа и покажем её связь с рассмотренными задачами.

**Задача MSSC.** *Дано:* множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$  и положительное число  $B$ . *Вопрос:* существует ли такое разбиение множества  $\mathcal{Y}$  на непустые подмножества (кластеры)  $\mathcal{C}_1, \dots, \mathcal{C}_J$ , что имеет место неравенство

$$\sum_{j=1}^J \sum_{y \in \mathcal{C}_j} \|y - \bar{w}_j\|^2 \leq B, \quad (11)$$

где  $\bar{w}_j = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$ ,  $j = 1, \dots, J$ , — центр  $j$ -го кластера?

Рассмотрим следующую модификацию этой задачи.

**Задача MSSC-Case.** *Дано:* множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  векторов из  $\mathbb{R}^q$ , натуральное число  $M > 1$  и положительное число  $B$ . *Вопрос:* существует ли такое разбиение множества  $\mathcal{Y}$  на  $J = N - M + 1$  непустых кластеров  $\mathcal{C}_1, \dots, \mathcal{C}_{N-M+1}$ , что мощность одного из кластеров равна  $M$  и имеет место неравенство (11)?

Пусть, например, мощность кластера  $\mathcal{C}_1$  равна  $M$ . Тогда из условий задачи следует, что мощность оставшихся  $N - M$  кластеров равна 1. Поскольку центр кластера, имеющего мощность, равную 1, совпадает с единственным элементом — вектором из этого кластера, для целевой функции задачи MSSC-Case из (11) имеем

$$\sum_{j=1}^J \sum_{y \in \mathcal{C}_j} \|y - \bar{w}_j\|^2 = \sum_{y \in \mathcal{C}_1} \|y - \bar{w}_1\|^2 + \sum_{j=2}^{N-M} \sum_{y \in \mathcal{C}_j} \|y - \bar{w}_j\|^2 = \sum_{y \in \mathcal{C}_1} \|y - \bar{w}_1\|^2.$$

Отсюда следует эквивалентность задач VS-2 и MSSC-Case. Поэтому рассмотренную проблему анализа данных можно трактовать как одну из труднорешаемых проблем кластерного анализа.

### Заключение

Показана NP-полнота оптимизационных задач, которые индуцируются проблемой поиска в множестве векторов евклидова пространства такого подмножества векторов, что оно имеет заданную (фиксированную) мощность и включает векторы, «близкие» между собой по критерию минимума суммы квадратов расстояний. Из полученного результата следует труднорешаемость соответствующей проблемы анализа данных. Остаётся заметить, что эффективные алгоритмы с гарантированными оценками точности для решения рассмотренных задач в настоящее время неизвестны.

### 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. 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.
4. 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.
5. Кельманов А. В., Пяткин А. В. Об одном варианте задачи выбора подмножества векторов // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 5. — С. 25–40.
6. 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.
7. 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.
8. Dasgupta S. The hardness of  $k$ -means clustering // Technical report CS-2007-0890. University of California, 2007. — 6 p.
9. Edwards A. W. F., Cavalli-Sforza L. L. A method for cluster analysis // Biometrics. — 1965. — Vol. 21. — P. 362–375.
10. Garey M. R., Johnson D. S. Computers and intractability: a guide to the theory of NP-completeness. — San Francisco: Freeman, 1979. — 314 p.



11. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Statistics Probab. — 1967. — Vol. 1. — P. 281–297.
12. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar  $k$ -means problem is NP-hard // Proc. of 3rd Annual Workshop on Algorithms and Computation (WALCOM). — Berlin, Heidelberg, New York: Springer-Verl., 2009. — P. 274–285. (Lect. Notes Comput. Sci.; Vol. 5431).
13. **Papadimitriou C. H.** Computational complexity. — New York: Addison-Wesley, 1994. — 523 p.

*Кельманов Александр Васильевич,*

e-mail: kelm@math.nsc.ru

*Пяткин Артем Валерьевич,*

e-mail: artem@math.nsc.ru

Статья поступила

12 апреля 2010 г.

Переработанный вариант —

6 июля 2010 г.