

УДК 519.2+621.391

К ВОПРОСУ ОБ АЛГОРИТМИЧЕСКОЙ СЛОЖНОСТИ ОДНОЙ ЗАДАЧИ КЛАСТЕРНОГО АНАЛИЗА *)

А. В. Долгушев, А. В. Кельманов

Аннотация. Доказана NP-полнота задачи MSSC — кластеризации множества векторов евклидова пространства по критерию минимума суммы квадратов — для случая, когда размерность пространства является, а число кластеров не является частью входа задачи.

Ключевые слова: кластерный анализ, задача MSSC, алгоритмическая сложность, NP-полнота.

Введение

Объект исследования настоящей статьи — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — задача кластеризации (разбиения) множества векторов евклидова пространства по критерию минимума суммы квадратов. Цель работы — анализ алгоритмической сложности этой задачи в случае, когда размерность пространства и число кластеров соответственно является и не является частью входа задачи.

Задача разбиения множества векторов евклидова пространства на подмножества (кластеры) по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров (центр кластера определяется как среднее значение вектора в кластере) известна в литературе как задача MSSC (от английского Minimum-Sum-of-Squares Clustering). Эта задача в некоторых публикациях фигурирует под названием k -means (k средних), которое соответствует названию одного из ранних алгоритмов [7], предложенных для её решения. К задаче MSSC сводится ряд типичных проблем анализа данных, возникающих в широком спектре приложений (см., например, [2, 4, 6–8] и цитированные там работы). Суть

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 09-01-00032 и 10-07-00195), гранта АБЦП Рособразования (проект № 2.1.1/3235), целевой программы СО РАН (интеграционный проект № 44), а также целевой программы № 2 Президиума РАН (проект № 227).

этих проблем состоит в том, чтобы по имеющейся совокупности, включающей несколько результатов измерения набора (вектора) каких-либо характеристик для элементов из некоторого множества материальных объектов, найти подмножества наборов, соответствующих каждому из этих объектов, или найти подмножества «похожих» объектов.

1. Анализ сложности

Задача MSSC в форме верификации свойств формулируется следующим образом.

Задача MSSC. *Дано:* множество $V = \{v_1, v_2, \dots, v_N\}$ векторов из \mathbb{R}^q и положительное число K . *Вопрос:* существует ли такое разбиение множества V на непустые подмножества (кластеры) C_1, C_2, \dots, C_J , что имеет место неравенство

$$\sum_{j=1}^J \sum_{v \in C_j} \|v - \bar{v}(C_j)\|^2 \leq K, \quad (1)$$

где $\bar{v}(C_j) = \sum_{v \in C_j} v / |C_j|$, $j = 1, 2, \dots, J$, — центр j -го кластера?

Напомним известные факты об алгоритмической сложности задачи MSSC. Одномерный вариант задачи разрешим за полиномиальное время [9]. Четыре возможных случая многомерного варианта этой задачи индуцируются комбинированием размерности пространства и числа кластеров как элементов, которые либо являются, либо не являются частью входа задачи. Относительно этих случаев известно следующее.

Если размерность пространства и число кластеров не являются частью входа, то задача решается точно за полиномиальное время [6]. Оставшиеся три случая на протяжении многих лет бездоказательно считались NP-трудными.

В [4] представлено доказательство труднорешаемости задачи MSSC для случая, когда размерность пространства является частью входа. Однако это доказательство оказалось ошибочным [2]. Корректное доказательство было опубликовано в [3] для подслучая, когда число кластеров фиксировано и равно двум. При доказательстве установлена полиномиальная сводимость NP-трудной задачи о максимально плотном разрезе графа к рассматриваемому подслучаю задачи MSSC [3]. Из неё следует NP-трудность задачи MSSC для случая, когда размерность пространства и число кластеров являются частью входа, так как задача с двумя кластерами является частным случаем задачи с числом кластеров не меньше двух.

Далее, в [8] установлена NP-трудность задачи MSSC в подслучае, когда число кластеров является частью входа, а размерность пространства фиксирована и равна двум. При доказательстве показана полиномиальная сводимость известной [5] NP-полной задачи 3-SAT к анализируемому подслучаю задачи MSSC. Из этого результата так же, как и из результата, полученного в [3], следует NP-трудность задачи MSSC в случае, когда число кластеров и размерность пространства являются частью входа, так как двумерная задача MSSC есть частный случай задачи MSSC большей размерности.

Наконец, нетрудно установить, что, когда число кластеров является частью входа, q -мерная задача MSSC полиномиально сводится к частному случаю $(q + 1)$ -мерной задачи путём введения нулевой $(q + 1)$ -й компоненты. Поэтому из результата, полученного в [8] для фиксированного $q = 2$, следует NP-трудность задачи MSSC в случае, когда число кластеров является, а размерность пространства не является частью входа задачи.

Ниже проанализирован последний из неизученных ранее случаев задачи. При анализе сложности применяется методика, сходная с методикой, использованной в [1].

Теорема 1. *Задача MSSC NP-полна в случае, когда размерность пространства является, а число кластеров не является частью входа задачи соответственно.*

Доказательство. Принадлежность задачи MSSC классу NP очевидна. Пусть число кластеров J не является частью входа задачи. Покажем, что для любого $J \geq 2$ задача J -MSSC полиномиально сводится к частному случаю задачи $(J + 1)$ -MSSC. Тогда результат будет следовать из NP-полноты [3] задачи 2-MSSC.

По произвольной индивидуальной задаче J -MSSC построим следующую индивидуальную задачу $(J + 1)$ -MSSC. Положим

$$a = \max_{i=1,\dots,N} \max_{j=1,\dots,q} |v_i^j|,$$

где v_i^j — j -я компонента вектора $v_i = (v_i^1, \dots, v_i^q) \in V$. Без ограничения общности будем считать, что $a \neq 0$. Далее, в задаче $(J + 1)$ -MSSC положим $\tilde{K} = K$ и $\tilde{V} = V \cup \{x\}$, где $x = (2a + K + 1, \dots, 2a + K + 1) \in \mathbb{R}^q$.

Покажем, что для того, чтобы в задаче J -MSSC существовало разбиение множества V на кластеры C_1, C_2, \dots, C_J , удовлетворяющее условию (1), необходимо и достаточно, чтобы в задаче $(J + 1)$ -MSSC существовало разбиение множества \tilde{V} на кластеры B_1, B_2, \dots, B_{J+1} такое,

что

$$\sum_{j=1}^{J+1} \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 \leq \tilde{K}. \quad (2)$$

НЕОБХОДИМОСТЬ. Пусть в задаче J -MSSC существует разбиение множества V на кластеры C_1, C_2, \dots, C_J такое, что выполняется неравенство (1). Тогда нетрудно убедиться, что в задаче $(J+1)$ -MSSC существует разбиение множества \tilde{V} на кластеры $B_1 = C_1, B_2 = C_2, \dots, B_J = C_J, B_{J+1} = \{x\}$ такое, что выполнено (2). Действительно,

$$\begin{aligned} \sum_{j=1}^{J+1} \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 &= \sum_{j=1}^J \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 + \sum_{v \in B_{J+1}} \|v - \bar{v}(B_{J+1})\|^2 \\ &= \sum_{j=1}^J \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 + \|x - x\|^2 \\ &= \sum_{j=1}^J \sum_{v \in C_j} \|v - \bar{v}(C_j)\|^2 \leq K = \tilde{K}. \end{aligned} \quad (3)$$

ДОСТАТОЧНОСТЬ. Заметим сначала, что для любого подмножества $X \subseteq V$ такого, что $|X| > 1$, в составе которого имеется некоторый вектор y , справедлива следующая цепочка равенств:

$$\begin{aligned} \sum_{v \in X} \|v - \bar{v}(X)\|^2 &= \sum_{v \in X} \|v\|^2 - 2 \sum_{v \in X} (v, \bar{v}(X)) + |X| \|\bar{v}(X)\|^2 \\ &= \sum_{v \in X} \|v\|^2 - 2 \frac{\sum_{v \in X} (v, \sum_{u \in X} u)}{|X|} + \frac{\|\sum_{v \in X} v\|^2}{|X|} = \sum_{v \in X} \|v\|^2 - \frac{\|\sum_{v \in X} v\|^2}{|X|} \\ &= \|y\|^2 + \sum_{u \in X \setminus \{y\}} \|u\|^2 - \frac{\|y\|^2}{|X|} - \frac{2 \left(y, \sum_{u \in X \setminus \{y\}} u \right)}{|X|} - \frac{\|\sum_{u \in X \setminus \{y\}} u\|^2}{|X|} \\ &= \frac{|X| - 1}{|X|} \|y\|^2 + \left(\sum_{u \in X \setminus \{y\}} \|u\|^2 - \frac{\|\sum_{u \in X \setminus \{y\}} u\|^2}{|X| - 1} \right) \end{aligned}$$

$$\begin{aligned}
 & + \frac{\left\| \sum_{u \in X \setminus \{y\}} u \right\|^2}{|X| - 1} - \frac{\left\| \sum_{u \in X \setminus \{y\}} u \right\|^2}{|X|} - \frac{2(y, \sum_{u \in X \setminus \{y\}} u)}{|X|} \\
 & = \frac{|X| - 1}{|X|} \|y\|^2 - \frac{2 \sum_{u \in X \setminus \{y\}} (u, y)}{|X|} \\
 & + \sum_{v \in X \setminus \{y\}} \|v - \bar{v}(X \setminus \{y\})\|^2 + \frac{\left\| \sum_{u \in X \setminus \{y\}} u \right\|^2}{(|X| - 1)|X|}. \quad (4)
 \end{aligned}$$

Допустим теперь, что в задаче $(J + 1)$ -MSSC существует разбиение множества \tilde{V} на кластеры B_1, B_2, \dots, B_{J+1} , удовлетворяющее неравенству (2). Не ограничивая общность, будем считать, что $x \in B_{J+1}$.

Пусть $|B_{J+1}| = 1$. Тогда $B_{J+1} = \{x\}$ и в задаче J -MSSC существует разбиение множества V на кластеры $C_1 = B_1, C_2 = B_2, \dots, C_J = B_J$ такое, что справедливо (1). В самом деле,

$$\begin{aligned}
 \sum_{j=1}^J \sum_{v \in C_j} \|v - \bar{v}(C_j)\|^2 &= \sum_{j=1}^J \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 + \|x - x\|^2 \\
 &= \sum_{j=1}^J \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 + \sum_{v \in B_{J+1}} \|v - \bar{v}(B_{J+1})\|^2 \leq \tilde{K} = K. \quad (5)
 \end{aligned}$$

Пусть $|B_{J+1}| > 1$. Заметим, что цепочка равенств (4) справедлива при подстановке $X = B_{J+1}$ и $y = x$. Сделав эту подстановку, найдём

$$\begin{aligned}
 \sum_{v \in B_{J+1}} \|v - \bar{v}(B_{J+1})\|^2 &= \frac{|B_{J+1}| - 1}{|B_{J+1}|} \|x\|^2 - \frac{2 \sum_{u \in B_{J+1} \setminus \{x\}} (u, x)}{|B_{J+1}|} \\
 &+ \sum_{v \in B_{J+1} \setminus \{x\}} \|v - \bar{v}(B_{J+1} \setminus \{x\})\|^2 + \frac{\left\| \sum_{u \in B_{J+1} \setminus \{x\}} u \right\|^2}{(|B_{J+1}| - 1)|B_{J+1}|}. \quad (6)
 \end{aligned}$$

В правой части этого выражения последние два члена неотрицательны. Напомним, что $x = (2a + K + 1, \dots, 2a + K + 1)$, и заметим, что $\|u\| \leq a\sqrt{q}$ для всех $u \in B_{J+1}$ по построению. Используя свойства скалярного произведения, для разности первых двух членов в правой части (6) получим

оценку

$$\begin{aligned} & \frac{|B_{J+1}| - 1}{|B_{J+1}|} \|x\|^2 - \frac{2 \sum_{u \in B_{J+1} \setminus \{x\}} (u, x)}{|B_{J+1}|} \\ & \geq \frac{|B_{J+1}| - 1}{|B_{J+1}|} \{q(2a + K + 1)^2 - 2aq(2a + K + 1)\} \\ & = \left(1 - \frac{1}{|B_{J+1}|}\right) q(2a + K + 1)(K + 1) > qK/2 \geq K = \tilde{K}, \end{aligned}$$

так как $q \geq 2$. Отсюда для задачи $(J + 1)$ -MSSC имеем

$$\sum_{j=1}^{J+1} \sum_{v \in B_j} \|v - \bar{v}(B_{J+1})\|^2 = \sum_{j=1}^J \sum_{v \in B_j} \|v - \bar{v}(B_j)\|^2 + \sum_{v \in B_{J+1}} \|v - \bar{v}(B_{J+1})\|^2 > \tilde{K},$$

что противоречит условию (2). Следовательно, $|B_{J+1}| = 1$, а для этого случая существование разбиения множества V , удовлетворяющего условию (1), на кластеры C_1, C_2, \dots, C_J было показано выше. Теорема 1 доказана.

Таким образом, установлена NP-полнота задачи MSSC в случае, когда размерность пространства и число кластеров соответственно является и не является частью входа задачи. Установленный факт дополняет известные результаты и закрывает вопрос о сложности статусе возможных случаев этой задачи.

ЛИТЕРАТУРА

1. **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.
2. **Aloise D., Hansen P.** On the complexity of minimum sum-of-squares clustering // Les Cahiers du GERAD, G-2007-50. — 2007. — 12 p.
3. **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.
4. **Drineas P., Frieze A., Kannan R., Vempala S., Vinay V.** Clustering large graphs via the singular value decomposition // Machine Learning. — 2004. — Vol. 56. — P. 9–33.
5. **Garey M.R., Johnson D.S.** Computers and intractability: a guide to the theory of NP-completeness. — San Francisco, CA: Freeman, 1979. — 338 p.
6. **Inaba M., Katch N., Imai H.** Applications of weighted Voronoi diagrams and randomization to variance-based clustering // Proc. Ann. Symp. Comput. Geom. — NY, USA: Stony Brook, 1994. — P. 332–339.

7. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Stat. Probab. Vol. 1.— Berkley: University of California Press, 1967. — P. 281–297.
8. **Mahajan M., Nimbhorkar P., Varadarajan K.** The planar k-means problem is NP-hard // Lect. Notes Comput. Sci. — 2009. — Vol. 5431. — P. 284–285.
9. **Rao M.** Cluster analysis and mathematical programming // J. Am. Stat. Assoc. — 1971. — Vol. 66. — P. 622–626.

Долгушев Алексей Владимирович,
e-mail: dolgushev@math.nsc.ru
Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru

Статья поступила
1 декабря 2009 г.
Переработанный вариант —
17 декабря 2009 г.