

УДК 519.16+519.85

FP-TAS ДЛЯ ОДНОЙ ЗАДАЧИ ПОИСКА ПОДМНОЖЕСТВА ВЕКТОРОВ *)

А. В. Кельманов, С. М. Романченко

Аннотация. Рассматривается NP-трудная в сильном смысле задача поиска в конечном множестве векторов евклидова пространства подмножества заданной мощности, доставляющего минимум сумме квадратов расстояний от элементов подмножества до его центра. Центр искомого подмножества определяется как среднее значение вектора по всем элементам подмножества. Доказано, что если $P \neq NP$, то для общего случая этой задачи не существует полностью полиномиальной приближённой схемы (FP-TAS). Для специального случая этой задачи, когда размерность пространства фиксирована, такая схема обоснована.

Ключевые слова: поиск подмножества векторов, евклидово пространство, минимум суммы квадратов расстояний, NP-трудность, полностью полиномиальная приближённая схема.

Введение

Предметом исследования является труднорешаемая экстремальная задача, которая индуцируется, в частности, к одной из актуальных проблем классификации данных. Цель исследования — обоснование приближённого полиномиального алгоритма её решения.

Рассматриваемая задача в постановочном плане близка к классической [6, 9, 11] NP-трудной [5] задаче MSSC (Minimum Sum-of-Squares Clustering) анализа данных. Напомним, что в задаче MSSC требуется разбить заданное множество векторов евклидова пространства на семейство непустых подмножеств, называемых кластерами, так, чтобы минимизировать сумму по всем кластерам суммы квадратов расстояний от элементов кластера до его центра. Под центром кластера понимается среднее значение вектора по всем элементам кластера.

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 12-01-00090, 13-07-00070), а также целевой программы СО РАН (интеграционные проекты 7Б и 21А).

Задача MSSC моделирует (см., например, [6, 8, 9]) содержательную проблему, в которой требуется разбить таблицу с результатами многократных измерений набора характеристик для нескольких объектов на непустые подмножества, содержащие результаты измерений каждого из объектов, и оценить характеристики этих объектов в условиях, когда соответствие между объектом и результатом измерения отсутствует, а каждому результату измерения сопутствует измерительная ошибка.

В рассматриваемой ниже задаче требуется в конечном множестве векторов евклидова пространства найти только одно подмножество заданной мощности такое, что сумма квадратов расстояний от элементов искомого подмножества до его центра минимальна. Эта задача моделирует (см., например, [1, 2]) более простую в содержательном плане проблему, чем описана выше. В этой проблеме предполагается, что таблица с анализируемыми данными содержит результаты многократных измерений набора характеристик только одного, например, изучаемого (или важного по какой-либо причине) объекта, кроме того, эта таблица включает результаты однократных измерений произвольных (посторонних или случайных) объектов. При этом информационную ценность имеет только набор характеристик исследуемого объекта, но соответствие между результатами измерения и объектами отсутствует, а измерительная ошибка присутствует.

Формальная постановка задачи и полученные для неё результаты приведены в разд. 1. Здесь лишь отметим, что, несмотря на простоту формулировки, эта задача NP-трудна в сильном смысле [1]. Поэтому, как известно [7], для этой задачи (в предположении, что гипотеза $P \neq NP$ верна) не существует точных полиномиального и псевдополиномиального алгоритмов. Наконец, если $P \neq NP$, то для общего случая этой задачи не существует полностью полиномиальной приближённой схемы (FPTAS) (см. разд. 2). По этой причине представляется важным поиск подклассов и специальных случаев этой задачи, для которых указанная схема существует. Для одного из таких случаев (когда размерность пространства фиксирована) схема FPTAS обоснована ниже.

1. Формулировка задачи и известные результаты

Сформулированной во введении содержательной проблеме соответствует [1, 2]

Задача VS-2 (Vector Subset 2). ДАНО: множество $\mathcal{Y} = \{y_1, \dots, y_N\}$ векторов из \mathbb{R}^q и натуральное число $M > 1$. НАЙТИ: подмножество $\mathcal{C} \subseteq \mathcal{Y}$

мощности M такое, что

$$F(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 \rightarrow \min, \quad (1)$$

где $\bar{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ — центр подмножества \mathcal{C} .

Цифра 2 в названии задачи обозначает порядковый номер в списке [1] полиномиально эквивалентных NP-трудных задач, которые индуцируются одной и той же содержательной проблемой.

Как известно, в статистических задачах, когда элементы некоторого непустого множества $\mathcal{Z} \subset \mathbb{R}^q$ суть выборочные случайные величины, значения $\bar{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ и $\frac{1}{|\mathcal{Z}|-1} F(\mathcal{Z})$ являются несмещёнными выборочными оценками математического ожидания и дисперсии соответственно. В детерминированных постановках задач эти величины определяют среднее значение по множеству и разброс элементов множества относительно среднего значения. Чем меньше разброс элементов множества, тем они более «компактны». Совместный поиск наиболее «компактного» подмножества элементов и вычисление среднего значения по элементам этого подмножества — суть задачи VS-2 в детерминированном случае. Если же элементы множества \mathcal{U} трактовать как случайные величины, то задачу VS-2 можно интерпретировать как совместное (одновременное) выделение из «засоренной» выборки \mathcal{U} подмножества \mathcal{C} элементов, которое соответствует выборочным данным (из некоторого распределения) с минимальной дисперсией, и оценивание математического ожидания этого распределения по подмножеству выделенных данных.

Напомним известные алгоритмические результаты. В [2] для общего случая задачи VS-2 построен 2-приближённый полиномиальный алгоритм, имеющий временную сложность $\mathcal{O}(qN^2)$. Полиномиальная приближённая схема (PTAS), позволяющая находить приближённое решение с относительной погрешностью $\varepsilon > 0$ за время $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, предложена в [4].

Как отмечено во введении, для общего случая задачи VS-2 не существует точного полиномиального и псевдополиномиального алгоритмов. Вместе с этим, в случае, когда мощность M искомого подмножества не является частью входа (фиксирована), задача разрешима за полиномиальное время. Действительно, для отыскания точного решения достаточно вычислить за время $\mathcal{O}(qM)$ значение целевой функции (1) для каждого допустимого подмножества мощности M множества \mathcal{U} и среди найденных $\mathcal{O}(N^M)$ значений выбрать наименьшее. Далее, введение

дополнительных ограничений на входные данные задачи позволяет построить точный псевдополиномиальный алгоритм [3] для случая, когда размерность q пространства фиксирована, а компоненты векторов входного множества имеют целочисленные значения. Временная сложность этого алгоритма есть величина $\mathcal{O}(qN(2MB+1)^q)$, где B — максимальное абсолютное значение координаты векторов входного множества.

Предложенный в настоящей работе алгоритм при фиксированной размерности q пространства позволяет находить $(1 + \varepsilon)$ -приближённое решение задачи для заданного $\varepsilon \in (0, 1)$ за время $\mathcal{O}(N^2(M/\varepsilon)^q)$. С учётом того, что $M \leq N$, этот алгоритм, очевидно, реализует схему FPTAS.

2. Геометрические основы алгоритма

Покажем сначала несуществование полностью полиномиальной приближённой схемы для общего случая задачи VS-2 в предположении справедливости гипотезы $P \neq NP$.

Теорема 1. *Если $P \neq NP$, то для задачи VS-2 схемы FPTAS не существует.*

ДОКАЗАТЕЛЬСТВО. В [1] показано, что к числу NP-трудных в сильном смысле относится задача VS-3 (с числовыми входами) отыскания в конечном множестве \mathcal{U} векторов из \mathbb{R}^q подмножества \mathcal{C} , доставляющего минимум функции

$$H(\mathcal{C}) = \sum_{x \in \mathcal{C}} \sum_{y \in \mathcal{C}} \|x - y\|^2,$$

при том же ограничении на мощность искомого множества \mathcal{C} , что и в задаче VS-2. Кроме того, к задаче VS-3 с бинарными векторами сведена NP-трудная в сильном смысле [10] задача отыскания в регулярном графе клики заданного размера. Поэтому задача VS-3 NP-трудна в сильном смысле для случая целочисленных входных данных.

Легко видеть, что при целочисленных входных данных значение функции $H(\mathcal{C})$ всегда будет целочисленным и ограниченным полиномом от максимального абсолютного значения компонент входных векторов. Отсюда в соответствии с [7] следует несуществование схемы FPTAS (если $P \neq NP$) для задачи VS-3.

Напомним [1], что целевые функции задач VS-2 и VS-3 связаны соотношением $H(\mathcal{C}) = 2MF(\mathcal{C})$. Допустим теперь, что для задачи VS-2 существует схема FPTAS с алгоритмическим решением \mathcal{C}_A . Тогда в соответствии с определением схемы FPTAS для этого решения выполнено неравенство $F(\mathcal{C}_A) \leq (1 + \varepsilon)F(\mathcal{C}^*)$, где \mathcal{C}^* — оптимальное решение задачи.

Отсюда $H(\mathcal{C}_A) = 2MF(\mathcal{C}_A) \leq (1 + \varepsilon)2MF(\mathcal{C}^*) = (1 + \varepsilon)H(\mathcal{C}^*)$, что противоречит отсутствию схемы FPTAS для задачи VS-3. Теорема 1 доказана.

Для обоснования алгоритма сформулируем и докажем несколько вспомогательных утверждений.

Лемма 1. Для произвольного вектора $x \in \mathbb{R}^q$ и конечного множества $\mathcal{Z} \subset \mathbb{R}^q$ имеет место равенство

$$\sum_{z \in \mathcal{Z}} \|z - x\|^2 = \sum_{z \in \mathcal{Z}} \|z - \bar{z}\|^2 + |\mathcal{Z}| \cdot \|x - \bar{z}\|^2, \quad (2)$$

где $\bar{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ — центр множества \mathcal{Z} .

ДОКАЗАТЕЛЬСТВО. Просуммировав очевидные тождества

$$\|z - x\|^2 = \|z - \bar{z} + \bar{z} - x\|^2 = \|z - \bar{z}\|^2 + 2\langle z - \bar{z}, \bar{z} - x \rangle + \|\bar{z} - x\|^2,$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение векторов, по всем векторам $z \in \mathcal{Z}$, получим

$$\sum_{z \in \mathcal{Z}} \|z - x\|^2 = \sum_{z \in \mathcal{Z}} \|z - \bar{z}\|^2 + 2 \sum_{z \in \mathcal{Z}} \langle z - \bar{z}, \bar{z} - x \rangle + |\mathcal{Z}| \cdot \|\bar{z} - x\|^2. \quad (3)$$

Остаётся заметить, что сумма скалярных произведений в правой части формулы (3) равна нулю, так как $\sum_{z \in \mathcal{Z}} (z - \bar{z}) = 0$. Лемма 1 доказана.

Для непустого конечного множества \mathcal{Z} векторов из \mathbb{R}^q положим

$$D(\mathcal{Z}) = \max_{x, y \in \mathcal{Z}} \|x - y\|, \quad R(\mathcal{Z}) = \max_{z \in \mathcal{Z}} \|z - \bar{z}\|,$$

где \bar{z} — центр множества \mathcal{Z} , $D(\mathcal{Z})$ — диаметр множества \mathcal{Z} , а $R(\mathcal{Z})$ — максимальное расстояние от элементов этого множества до его центра. Взаимосвязь между $D(\mathcal{Z})$ и $R(\mathcal{Z})$ устанавливает

Лемма 2. Для непустого конечного множества векторов $\mathcal{Z} \subset \mathbb{R}^q$ справедливы неравенства

$$\frac{1}{2}D(\mathcal{Z}) \leq R(\mathcal{Z}) \leq D(\mathcal{Z}).$$

ДОКАЗАТЕЛЬСТВО. Из неравенства треугольника для произвольных $x, y \in \mathcal{Z}$ и \bar{z} следует оценка сверху

$$\|x - y\| \leq \|x - \bar{z}\| + \|y - \bar{z}\| \leq 2 \max_{z \in \mathcal{Z}} \|z - \bar{z}\| = 2R(\mathcal{Z}).$$

Поскольку эта оценка верна для произвольных $x, y \in \mathcal{Z}$, имеем

$$D(\mathcal{Z}) = \max_{x, y \in \mathcal{Z}} \|x - y\| \leq 2R(\mathcal{Z}),$$

что даёт оценку снизу для $R(\mathcal{Z})$.

Пусть в условиях леммы 1 $x \in \mathcal{Z}$. Тогда из (2) получаем

$$\sum_{z \in \mathcal{Z}} \|z - \bar{z}\|^2 = \sum_{z \in \mathcal{Z}} \|z - x\|^2 - |\mathcal{Z}| \cdot \|x - \bar{z}\|^2 = \sum_{z \in \mathcal{Z}} (\|z - x\|^2 - \|x - \bar{z}\|^2) \geq 0.$$

Из этого неравенства следует, что по крайней мере одно из слагаемых в скобках неотрицательно. Стало быть, для любого вектора $x \in \mathcal{Z}$ найдётся вектор $z_x \in \mathcal{Z}$ такой, что

$$\|x - \bar{z}\| \leq \|x - z_x\| \leq D(\mathcal{Z}).$$

Так как это неравенство справедливо для произвольного $x \in \mathcal{Z}$, имеем оценку сверху

$$\max_{x \in \mathcal{Z}} \|x - \bar{z}\| = R(\mathcal{Z}) \leq D(\mathcal{Z}).$$

Лемма 2 доказана.

Для конечного множества \mathcal{Z} векторов из \mathbb{R}^q , натурального $M \leq |\mathcal{Z}|$ и произвольного вектора $x \in \mathbb{R}^q$ определим множество

$$\mathcal{Z}_M(x, \mathcal{Z}) = \{z_i \mid \|z_i - x\| \leq \|z_j - x\|; z_i, z_j \in \mathcal{Z}, j \neq i, i \leq M\}, \quad (4)$$

состоящее из M ближайших (по расстоянию) к x элементов множества \mathcal{Z} , и максимальное расстояние

$$r_M(x, \mathcal{Z}) = \max_{z \in \mathcal{Z}_M(x, \mathcal{Z})} \|z - x\| \quad (5)$$

от этого вектора до векторов из $\mathcal{Z}_M(x, \mathcal{Z})$.

Непосредственно из определений (4) и (5) вытекает

Лемма 3. Пусть \mathcal{Z} — конечное множество векторов из \mathbb{R}^q , а \mathcal{X} — подмножество мощности M множества \mathcal{Z} . Тогда

(i) для любого вектора $x \in \mathbb{R}^q$ справедливо неравенство

$$\sum_{z \in \mathcal{Z}_M(x, \mathcal{Z})} \|z - x\|^2 \leq \sum_{z \in \mathcal{X}} \|z - x\|^2;$$

(ii) для любого $z \in \mathcal{X}$ верно неравенство

$$r_M(z, \mathcal{Z}) \leq D(\mathcal{X}). \quad (6)$$

Обоснуем подход к получению приближённого решения задачи VS-2. Из определения (4) следует, что подмножество $\mathcal{Z}_M(x, \mathcal{Y}) \subseteq \mathcal{Y}$ является допустимым решением задачи VS-2 для любого $x \in \mathbb{R}^q$. Опираясь на леммы 1–3, сформулируем важное свойство этого подмножества.

Лемма 4. Пусть \mathcal{C}^* — оптимальное решение задачи VS-2, $\bar{y}(\mathcal{C}^*) = \frac{1}{M} \sum_{y \in \mathcal{C}^*} y$ — центр подмножества \mathcal{C}^* , а $\mathcal{Z}_M(x, \mathcal{Y})$ — допустимое решение задачи VS-2, порождённое некоторым вектором $x \in \mathbb{R}^q$. Тогда справедлива оценка

$$F(\mathcal{Z}_M(x, \mathcal{Y})) \leq F(\mathcal{C}^*) + M\|x - \bar{y}(\mathcal{C}^*)\|^2. \quad (7)$$

ДОКАЗАТЕЛЬСТВО. Обозначив центр допустимого решения задачи VS-2 через $\bar{y}(\mathcal{Z}_M(x, \mathcal{Y})) = \frac{1}{M} \sum_{y \in \mathcal{Z}_M(x, \mathcal{Y})} y$, из определений (1) и (4) получим следующую цепочку равенств и неравенств:

$$\begin{aligned} F(\mathcal{Z}_M(x, \mathcal{Y})) &= \sum_{y \in \mathcal{Z}_M(x, \mathcal{Y})} \|y - \bar{y}(\mathcal{Z}_M(x, \mathcal{Y}))\|^2 \leq \sum_{y \in \mathcal{Z}_M(x, \mathcal{Y})} \|y - x\|^2 \\ &\leq \sum_{y \in \mathcal{C}^*} \|y - x\|^2 = \sum_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|^2 + M\|x - \bar{y}(\mathcal{C}^*)\|^2 = F(\mathcal{C}^*) + M\|x - \bar{y}(\mathcal{C}^*)\|^2. \end{aligned}$$

В этой цепочке справедливость первого неравенства и предпоследнего равенства следует из леммы 1, а второго неравенства — из леммы 3. Лемма 4 доказана.

Лемма 4 показывает, что качество допустимого решения $\mathcal{Z}_M(x, \mathcal{Y})$, полученного по некоторому вектору $x \in \mathbb{R}^q$, можно оценить через расстояние от этого вектора до центра $\bar{y}(\mathcal{C}^*)$ оптимального решения. Чем ближе x к $\bar{y}(\mathcal{C}^*)$, тем точнее решение $\mathcal{Z}_M(x, \mathcal{Y})$.

Суть предлагаемого подхода состоит в построении многомерной сетки (решётки), среди узлов которой найдётся вектор, близкий к центру оптимального кластера. Достаточное условие на размер шага сетки устанавливает

Лемма 5. Пусть выполнены условия леммы 4, $x \in \mathbb{R}^q$ и $y \in \mathcal{C}^*$. Тогда для того чтобы при фиксированном $\varepsilon > 0$ множество $\mathcal{Z}_M(x, \mathcal{Y})$ было $(1 + \varepsilon)$ -приближённым решением задачи VS-2, достаточно, чтобы вектор x удовлетворял неравенству

$$\|x - \bar{y}(\mathcal{C}^*)\|^2 \leq \frac{\varepsilon}{4M} r_M^2(y, \mathcal{Y}).$$

ДОКАЗАТЕЛЬСТВО. Из леммы 2 для множества \mathcal{C}^* следует, что

$$\frac{1}{4}D^2(\mathcal{C}^*) \leq R^2(\mathcal{C}^*) = \max_{u \in \mathcal{C}^*} \|u - \bar{y}(\mathcal{C}^*)\|^2 \leq \sum_{u \in \mathcal{C}^*} \|u - \bar{y}(\mathcal{C}^*)\|^2 = F(\mathcal{C}^*). \quad (8)$$

Кроме того, из неравенства (6) для вектора x и множеств \mathcal{C}^* и \mathcal{Y} имеем

$$\|x - \bar{y}(\mathcal{C}^*)\|^2 \leq \frac{\varepsilon}{4M} r_M^2(y, \mathcal{Y}) \leq \frac{\varepsilon}{4M} D^2(\mathcal{C}^*). \quad (9)$$

Комбинируя (8) и (9), получим

$$M\|x - \bar{y}(\mathcal{C}^*)\|^2 \leq \varepsilon F(\mathcal{C}^*). \quad (10)$$

Применив (10) к правой части неравенства (7), приходим к оценке

$$F(\mathcal{Z}_M(x, \mathcal{Y})) \leq (1 + \varepsilon)F(\mathcal{C}^*).$$

Значит, множество $\mathcal{Z}_M(x, \mathcal{Y})$ — $(1 + \varepsilon)$ -приближённое решение задачи VS-2. Лемма 5 доказана.

Свойство векторов из оптимального решения устанавливает

Лемма 6. Пусть выполнены условия леммы 4. Тогда для любого вектора $y \in \mathcal{C}^*$ справедлива оценка

$$\|y - \bar{y}(\mathcal{C}^*)\|^2 \leq M r_M^2(y, \mathcal{Y}). \quad (11)$$

ДОКАЗАТЕЛЬСТВО. Из оптимальности множества \mathcal{C}^* следует, что

$$\|y - \bar{y}(\mathcal{C}^*)\|^2 \leq \sum_{x \in \mathcal{C}^*} \|x - \bar{y}(\mathcal{C}^*)\|^2 = F(\mathcal{C}^*) \leq F(\mathcal{Z}_M(y, \mathcal{Y})). \quad (12)$$

С другой стороны, в силу определений (4) и (5) множества $\mathcal{Z}_M(y, \mathcal{Y})$ и величины $r_M(y, \mathcal{Y})$ имеет место цепочка неравенств

$$F(\mathcal{Z}_M(y, \mathcal{Y})) \leq \sum_{x \in \mathcal{Z}_M(y, \mathcal{Y})} \|x - y\|^2 \leq M r_M^2(y, \mathcal{Y}). \quad (13)$$

Объединив неравенства (12) и (13), получим утверждение леммы. Лемма 6 доказана.

Фактически лемма 5 определяет область (шар с центром в $\bar{y}(\mathcal{C}^*)$), любой элемент из которой позволяет гарантированно получить приближённое решение с требуемой точностью. В то же время лемма 6 определяет область поиска, содержащую (неизвестный) центр оптимального решения.

3. Алгоритм решения задачи

Для произвольного $y \in \mathcal{Y}$ и положительных чисел h, H определим векторное множество

$$\mathcal{B}(y, h, H) = \{b \mid b = y + h(j_1, \dots, j_q), j_i \in \mathbb{Z}, |h * j_i| \leq H, i = 1, \dots, q\}. \quad (14)$$

Элементы из этого множества соответствуют узлам равномерной сетки с центром в y и шагом h . Параметр H определяет геометрические размеры сетки. Число точек сетки по каждой координате, очевидно, не превосходит $2 \lfloor \frac{H}{h} \rfloor + 1 \leq 2 \frac{H}{h} + 1$ и не зависит от вектора y .

Замечание 1. Для любого $x \in \mathbb{R}^q$ такого, что $\|y - x\| \leq H$, где y — центр сетки $\mathcal{B}(y, h, H)$, расстояние от вектора x до ближайшего узла сетки $\mathcal{B}(y, h, H)$ не превосходит $\sqrt{q}h/2$, так как по каждой из q координат это расстояние, очевидно, не превосходит $h/2$.

Для произвольного $\varepsilon > 0$ и $y \in \mathcal{Y}$ положим

$$h = \frac{\varepsilon}{\sqrt{qM}} r_M(y, \mathcal{Y}), \quad (15)$$

$$H = \sqrt{M} r_M(y, \mathcal{Y}), \quad (16)$$

где $r_M(y, \mathcal{Y})$ определяется формулой (5).

Построим алгоритм решения задачи, основанный на поиске вектора, близкого к центру оптимального решения.

АЛГОРИТМ \mathcal{A} .

ВХОД. Множество \mathcal{Y} и числа M и ε .

Для каждого вектора $y \in \mathcal{Y}$ выполним шаги 1–4.

ШАГ 1. Вычислим $r_M(y, \mathcal{Y})$, h и H по формулам (5), (15) и (16) соответственно.

ШАГ 2. Если $r_M(y, \mathcal{Y}) = 0$, то построим множество $\mathcal{Z}_M(y, \mathcal{Y})$ по формуле (4) и объявим его результатом работы алгоритма. Иначе переходим к следующему шагу.

ШАГ 3. Построим множество $\mathcal{B}(y, h, H)$ в соответствии с (14).

ШАГ 4. Для каждого вектора $b \in \mathcal{B}(y, h, H)$ построим множество $\mathcal{Z}_M(b, \mathcal{Y})$ по формуле (4) и вычислим значение $F(\mathcal{Z}_M(b, \mathcal{Y}))$.

ШАГ 5. Результатом работы алгоритма объявим множество $\mathcal{Z}_M(b, \mathcal{Y})$, у которого значение целевой функции $F(\mathcal{Z}_M(b, \mathcal{Y}))$ минимально.

ВЫХОД.

Теорема 2. Для любого фиксированного $\varepsilon > 0$ алгоритм \mathcal{A} находит $(1 + \varepsilon)$ -приближённое решение задачи VS-2 за время

$$\mathcal{O}(qN^2(2\sqrt{q}M/\varepsilon + 1)^q).$$

ДОКАЗАТЕЛЬСТВО. Согласно пошаговой записи алгоритма на шаге 2 возможны два случая. В случае, когда $r_M(y, \mathcal{Y}) = 0$, из леммы 5 получаем равенство $\|y - \bar{y}(\mathcal{C}^*)\| = 0$, которое вместе с леммой 4 доказывает оптимальность множества $\mathcal{Z}_M(y, \mathcal{Y})$.

Проанализируем случай, когда $r_M(y, \mathcal{Y}) > 0$. Очевидно, что в ходе работы алгоритма будет выбран каждый вектор $y \in \mathcal{C}^*$. В соответствии с леммой 6 для всех векторов $y \in \mathcal{C}^*$ выполнено неравенство (11). Из этого неравенства и (16) следует, что $\|y - \bar{y}(\mathcal{C}^*)\| \leq H$, т. е. центр оптимального кластера лежит в области сетки $\mathcal{B}(y, h, H)$.

Рассмотрим вектор $b^* = \arg \min_{b \in \mathcal{B}(y, h, H)} \|b - \bar{y}(\mathcal{C}^*)\|$, ближайший к центру множества \mathcal{C}^* . Согласно замечанию 1 из (15) имеем

$$\|b^* - \bar{y}(\mathcal{C}^*)\|^2 \leq \frac{qh^2}{4} = \frac{\varepsilon}{4M} r_M^2(y, \mathcal{Y}).$$

Тем самым вектор b^* удовлетворяет условиям леммы 5, следовательно, множество $\mathcal{Z}_M(b^*, \mathcal{Y})$ — $(1 + \varepsilon)$ -приближённое решение задачи VS-2.

Оценим время работы алгоритма. На шагах 1 и 2 для вычисления $r_M(y, \mathcal{Y})$ и построения множества $\mathcal{Z}_M(y, \mathcal{Y})$ достаточно найти M ближайших к y векторов из \mathcal{Y} . Это можно осуществить за $\mathcal{O}(N)$ операций над векторами размерности q , т. е. за $\mathcal{O}(qN)$ элементарных операций [12].

Время построения множества $\mathcal{B}(y, h, H)$ на шаге 3 определяется его мощностью. Из формул (15) и (16) получаем

$$|\mathcal{B}(y, h, H)| \leq (2H/h + 1)^q = (2\sqrt{q}M/\varepsilon + 1)^q.$$

Таким образом, третий шаг потребует $\mathcal{O}(q(2\sqrt{q}M/\varepsilon + 1)^q)$ элементарных операций. На шаге 4 для каждого вектора из множества \mathcal{B} вычисляются M ближайших к нему элементов множества \mathcal{Y} . Эти вычисления потребуют не более $\mathcal{O}(qN)$ элементарных операций для каждого вектора из \mathcal{B} и $\mathcal{O}(qN(2\sqrt{q}M/\varepsilon + 1)^q)$ операций для всех векторов из \mathcal{B} [12]. Поэтому суммарная сложность шагов 1–4 для каждого вектора $y \in \mathcal{Y}$ равна $\mathcal{O}(qN(2\sqrt{q}M/\varepsilon + 1)^q)$.

Шаги 1–4 выполняются для каждого вектора $y \in \mathcal{Y}$, т. е. всего N раз. Отсюда следует окончательная оценка временной сложности алгоритма. Теорема 2 доказана.

Покажем, что при фиксированной размерности q предложенный алгоритм реализует схему FPTAS. Действительно, если $\varepsilon \in (0, 1)$, то для одного из сомножителей в оценке вычислительной сложности алгоритма имеем

$$\begin{aligned} \left(\frac{2\sqrt{q}M}{\varepsilon} + 1 \right)^q &= \left(2\sqrt{q} \left(\frac{M}{\varepsilon} + \frac{1}{2\sqrt{q}} \right) \right)^q \leq (2\sqrt{q})^q \left(\frac{M}{\varepsilon} + 1 \right)^q \\ &\leq (2\sqrt{q})^q 2^q \left(\frac{M}{\varepsilon} \right)^q = 2^{2q} q^{q/2} \left(\frac{M}{\varepsilon} \right)^q = \mathcal{O}((M/\varepsilon)^q), \end{aligned}$$

так как $\frac{M}{\varepsilon} + 1 \leq \frac{2M}{\varepsilon}$. Отсюда следует, что при указанных условиях время работы алгоритма \mathcal{A} есть величина $\mathcal{O}(N^2(M/\varepsilon)^q)$. В силу того, что $M \leq N$, вычислительная сложность алгоритма \mathcal{A} ограничена полиномом от размера входа задачи и $1/\varepsilon$. Это значит, что построенный алгоритм реализует схему FPTAS [7].

Заключение

Обоснована схема FPTAS для специального случая NP-трудной в сильном смысле задачи поиска в конечном множестве векторов евклидова пространства подмножества векторов заданной мощности, доставляющего минимум сумме квадратов расстояний от элементов искомого подмножества до его геометрического центра. Рассмотренный случай фиксированной размерности пространства актуален в приложениях. Показано, что при условии $P \neq NP$ для общего случая этой задачи схемы FPTAS не существует. Важным направлением исследований является обоснование рандомизированного алгоритма для общего случая рассмотренной задачи.

ЛИТЕРАТУРА

1. Кельманов А. В., Пяткин А. В. NP-полнота некоторых задач выбора подмножества векторов // Дискрет. анализ и исслед. операций. — 2010. — Т. 17, № 5. — С. 37–45.
Kel'manov A. V., Pyatkin A. V. NP-completeness of some problems of choosing a vector subset // J. Appl. Industr. Math. — 2011. — Vol. 5, N 3. — P. 352–357.
2. Кельманов А. В., Романченко С. М. Приближённый алгоритм для решения одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. — 2011. — Т. 18, № 1. — С. 61–69.
Kel'manov A. V., Romanchenko S. M. An approximation algorithm for solving a problem of search for a vector subset // J. Appl. Industr. Math. — 2012. — Vol. 6, N 1. — P. 90–96.

3. **Кельманов А. В., Романченко С. М.** Псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. — 2012. — № 2. — С. 156–162.
Kel'manov A. V., Romanchenko S. M. Pseudopolynomial algorithms for certain computationally hard vector subset and cluster analysis problems // Automation and Remote Control. — 2012. — Vol. 73, N 2. — P. 349–354.
4. **Шенмайер В. В.** Аппроксимационная схема для одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. — 2012. — Т. 19, № 2. — С. 92–100.
Shenmaier V. V. An approximation scheme for a problem of search for a vector subset // J. Appl. Industr. Math. — 2012. — Vol. 6, N 3. — P. 381–386.
5. **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.
6. **Anil K., Jain K.** Data clustering: 50 years beyond k -means // Pattern Recogn. Lett. — 2010. — Vol. 31. — P. 651–666.
7. **Garey M. R., Johnson D. S.** Computers and intractability: a guide to the theory of NP-completeness. — San Francisco: Freeman, 1979. — 314 p.
8. **Hastie T., Tibshirani R., Friedman J.** The elements of statistical learning: data mining, inference, and prediction. — New York: Springer-Verl., 2001. — 533 p.
9. **MacQueen J. B.** Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Statist. Probab. (Berkeley, June 21–July 18, 1965; December 27, 1965–January 7, 1966). Vol. 1. — Berkeley: University of California Press, 1967. — P. 281–297.
10. **Papadimitriou C. H.** Computational complexity. — New York: Addison-Wesley, 1994. — 523 p.
11. **Rao M.** Cluster analysis and mathematical programming // J. Amer. Stat. Assoc. — 1971. — Vol. 66. — P. 622–626.
12. **Wirth H.** Algorithms + data structures = programs. — New Jersey: Prentice Hall, 1976. — 366 p.

Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru
Романченко Семён Михайлович,
e-mail: rsm@math.nsc.ru

Статья поступила
11 ноября 2013 г.
Переработанный вариант —
29 января 2014 г.