ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Март—апрель 2011. Том 18, № 2. С. 29–40

УДК 519.2+621.391

ПРИБЛИЖЁННЫЙ АЛГОРИТМ РЕШЕНИЯ ОДНОЙ ЗАДАЧИ КЛАСТЕРНОГО АНАЛИЗА *)

А. В. Долгушев, А. В. Кельманов

Аннотация. Предложен 2-приближённый алгоритм для труднорешаемой задачи, к которой сводится одна из проблем разбиения множества векторов евклидова пространства на два подмножества (кластера) по критерию минимума суммы квадратов расстояний.

Ключевые слова: поиск подмножества векторов, кластерный анализ, NP-трудность, эффективный приближённый алгоритм.

Введение

Объектом исследования настоящей работы являются проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — труднорешаемая экстремальная задача, к которой сводится одна из проблем кластерного анализа данных. Цель работы — обоснование приближённого алгоритма для решения этой задачи.

Содержательная трактовка рассматриваемой проблемы анализа данных состоит в следующем. Имеется таблица, содержащая результаты измерения набора числовых информационно значимых характеристик некоторого материального объекта. Объект может находиться в одном из двух состояний: активном (включенном) и пассивном (выключенном). В пассивном состоянии значения всех измеряемых характеристик равны нулю, а в активном значение хотя бы одной характеристики не равно нулю. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие между результатом измерения и состояниями объекта неизвестно, а число активных состояний объекта известно.

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты № 09–01–00032, 10–07–00195), целевой программы № 2 Президиума РАН (проект № 227), целевой программы СО РАН (проект № 44), целевой программы АВЦП Рособразования (проект № 2.1.1/3235), а также федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» (гос. контракт № 14.740.11.0362).

Требуется, используя адекватный измеряемым характеристикам критерий, найти подмножество наборов, соответствующих активному состоянию объекта, и оценить по результатам измерения набор характеристик объекта в активном состоянии (учитывая, что данные содержат ошибку измерения). Эта содержательная проблема типична для многих приложений, связанных с компьютерным анализом данных и распознаванием образов (см., например, [3, 5, 6] и цитированные там работы).

В [3] дана формулировка этой содержательной проблемы как задачи поиска в множестве векторов евклидова пространства подмножества векторов фиксированной мощности такого, что векторы «близки» по критерию минимума суммы квадратов расстояний. Там же установлено, что решение этой (далее исходной) задачи эквивалентно отысканию в множестве векторов евклидова пространства такого подмножества фиксированной мощности, что длина суммы векторов из этого подмножества максимальна. Ниже, как и в [3], эта задача поиска подмножества на максимум имеет краткое название MLSVS. NP-трудность задачи MLSVS показана в [1, 3]. Точные и приближённые алгоритмы решения задачи MLSVS предложены в [1–4]. Характеристики этих алгоритмов приведены в следующем разделе.

Мотивацией исследований послужили следующие факты. Во-первых, алгоритм из [3] не имеет теоретических гарантий по точности. Во-вторых, приближённый алгоритм решения задачи MLSVS, предложенный в [1] и реализующий вполне полиномиальную аппроксимационную схему (FPTAS), не обеспечивает гарантированой оценки точности для исходной задачи анализа данных на минимум в неасимптотическом случае, хотя и позволяет находить асимптотически точное решение исходной задачи. В-третьих, временна́я сложность алгоритмов с оценками точности, обоснованных в [1, 2, 4], столь высока (см. следующий раздел), что эти алгоритмы практически непригодны для решения задач большой размерности.

В нашей статье рассмотрена одна из неизученных ранее задач кластерного анализа на минимум. Показано, что эта задача NP-трудна и эквивалентна исходной задаче. Для решения задачи предложен 2-приближённый алгоритм.

1. Известные алгоритмические результаты

Напомним, что сформулированная в [3] содержательная проблема анализа данных сведена к решению задачи поиска в множестве $\mathcal{Y} =$

$\{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ подмножества $\mathcal{C} \subseteq \mathcal{Y}$ такого, что

$$\sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{\left\| \sum\limits_{y \in \mathcal{C}} y \right\|^2}{|\mathcal{C}|} \to \min, \tag{1}$$

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

Так как первый член в (1) является константой и мощность искомого подмножества $\mathcal{C}$ фиксирована, решение исходной задачи (1) на минимум можно получить, максимизируя норму в числителе второго члена функции (1). Поэтому в [3] рассмотрена

Задача MLSVS (**M**aximum of the **L**ength of **S**um of **V**ectors from a **S**ubset). *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$. *Найти*: подмножество $\mathcal{U} \subseteq \mathcal{Y}$ такое, что функция

$$F(\mathcal{U}) = \left\| \sum_{y \in \mathcal{U}} y \right\| \tag{2}$$

максимальна, при ограничении $|\mathcal{U}| = M$ на мощность искомого подмножества.

Приведём характеристики существующих алгоритмов решения этой задачи.

В [3] построен эффективный приближённый алгоритм, временна́я сложность которого равна $\mathcal{O}(qN)$. К сожалению, этот алгоритм не имеет гарантированной оценки точности. Он примечателен лишь тем, что в численных экспериментах на некоторых подклассах входных данных показывает точность, приемлемую для приложений [3].

В [1] предложен алгоритм с гарантированной оценкой относительной погрешности $\varepsilon = (q-1)/(8l^2)$, где l — целочисленный параметр алгоритма. Временна́я сложность алгоритма равна $\mathcal{O}[Nq^2(2l+1)^{q-1}]$. Фактически в [1] обоснована вполне полиномиальная аппроксимационная схема, устанавливающая полиномиальную относительно N и $1/\varepsilon$ оценку временно́й сложности алгоритма, для случая, когда размерность q пространства фиксирована. Для этого же случая в [4] конструктивно установлено, что задача разрешима за полиномиальное время $\mathcal{O}(q^2N^{2q})$.

Наконец, в [1, 2] построены псевдополиномиальные алгоритмы, гарантирующие оптимальность решения задачи для случая, когда компоненты векторов имеют целочисленные значения. Эти алгоритмы имеют временны́е сложности $\mathcal{O}(Nq^{q+1}(Ma)^{q-1})$ и $\mathcal{O}(qMN(2Ma)^{q-1})$ для [1] и [2]

соответственно, где a — максимальная абсолютная величина компоненты вектора в множестве $\mathcal{Y}$.

Относительно характеристик существующих алгоритмических решений заметим следующее. Во-первых, как для исходной задачи, так и для задачи MLSVS в настоящее время отсутствуют эффективные приближённые алгоритмы с константной оценкой точности. Следует отметить, что при фиксированной размерности пространства задача MLSVS принадлежит классу P. Факт принадлежности следует из существования полиномиального точного алгоритма, предложенного в [4], для решения этой задачи.

Во-вторых, опираясь на определение (2), для целевой функции (1) исходной задачи имеем

$$\sum_{y\in\mathcal{Y}} \|y\|^2 - \frac{\big\| \sum\limits_{y\in\mathcal{C}} y\big\|^2}{|\mathcal{C}|} = \sum_{y\in\mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|}F^2(\mathcal{C}). \tag{3}$$

Отсюда следует, что точное и асимптотически точное алгоритмическое решение исходной задачи может быть найдено с помощью известных алгоритмов [1, 2, 4] решения задачи MLSVS, максимизирующих $F(\mathcal{C})$. Однако получение оценки точности приближённого решения исходной задачи по оценке точности приближённого решения задачи MLSVS (и получение оценок в обратном порядке) проблематично из-за наличия в (3) аддитивной константы — суммы квадратов норм векторов из множества $\mathcal{Y}$.

В-третьих, анализируя приведённые выше оценки сложности алгоритмов, обоснованных в [1, 2, 4], легко заметить, что формулы этих оценок содержат в качестве множителя величины, в показателе степени которых фигурирует размерность пространства q. Во многих прикладных (естественно-научных и технических) задачах анализа данных размерность пространства составляет сотни и тысячи, а мощности анализируемых множеств ещё на несколько порядков выше. Для решения этих прикладных задач известные алгоритмы с оценками точности практически непригодны из-за высокой трудоёмкости. С другой стороны, малотрудоёмкий алгоритм, предложенный в [3], не имеет теоретических гарантий по точности.

Таким образом, характеристики существующих алгоритмов указывают на актуальность поиска таких приближённых алгоритмических решений исходной задачи и задачи MLSVS, сложность которых не зависит экспоненциально от размерности пространства.

2. Задача кластерного анализа на минимум

Рассмотрим следующую задачу поиска подмножества векторов.

Задача VS (**V**ector **S**ubset). *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$. *Найти:* подмножество $\mathcal{C} \subseteq \mathcal{Y}$ мощности M такое, что целевая функция

$$S(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \tag{4}$$

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$, минимальна.

Из (4) видно, что задачу VS можно трактовать как задачу кластерного анализа, в которой требуется разбить заданное множество $\mathcal{Y}$ векторов на два кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, мощности которых равны M и $N - M$ соответственно, таких, что целевая функция (4) минимальна. Справедлива следующая

Лемма 1. *Целевая функция* (4) *задачи* VS *минимальна тогда и только тогда, когда целевая функция* (2) *задачи* MLSVS *максимальна.*

Доказательство. Раскрывая первое слагаемое в правой части (4), установим связь между целевыми функциями задач VS и MLSVS:

$$\begin{aligned} S(\mathcal{C}) &= \sum_{y \in \mathcal{C}} \|y\|^2 - 2 \sum_{y \in \mathcal{C}} (y, \overline{y}(\mathcal{C})) + |\mathcal{C}| \cdot \|\overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\ &= \sum_{y \in \mathcal{Y}} \|y\|^2 - 2 \frac{\sum\limits_{y \in \mathcal{C}} \left(y, \sum\limits_{x \in \mathcal{C}} x\right)}{|\mathcal{C}|} + \frac{\left\| \sum\limits_{y \in \mathcal{C}} y \right\|^2}{|\mathcal{C}|} \\ &= \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{\left\| \sum\limits_{y \in \mathcal{C}} y \right\|^2}{|\mathcal{C}|} = \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} F^2(\mathcal{C}). \end{aligned} \tag{5}$$

Справедливость утверждения леммы следует из того, что первый член в правой части выражения (5) и положительный множитель во втором члене этого выражения являются константами. Лемма 1 доказана.

Теорема 1. *Задача* VS NP*-трудна.*

Доказательство. Справедливость утверждения теоремы следует из леммы 1 и NP-трудности [3] задачи MLSVS. Теорема 1 доказана.

Обратив внимание на предпоследнее равенство в цепочке равенств (5) и формулу (1), легко заметить, что задача VS эквивалентна исходной задаче (1).

Построим алгоритм решения задачи VS. Обозначим через $\mathcal{C}^*$ множество, доставляющее минимум S^* целевой функции S, и положим $y^* = \overline{y}(\mathcal{C}^*)$, где $\overline{y}(\cdot)$ — функция, определённая в формулировке задачи VS. Из леммы 1 вытекает

Следствие 1. *Для любых векторов $y \in \mathcal{C}^*$ и $z \in \mathcal{Y} \setminus \mathcal{C}^*$ справедливо неравенство*

$$(y,\, y^*) \geqslant (z,\, y^*).$$

Доказательство. Справедливость утверждения легко установить, если заметить, что из (5) следует равенство

$$S(\mathcal{C}^*) = \sum_{y \in \mathcal{Y}} \|y\|^2 - \sum_{y \in \mathcal{C}^*} (y, y^*).$$

Следствие 1 доказано.

Фактически следствие 1 показывает, что оптимальное подмножество $\mathcal{C}^*$ состоит из M векторов множества $\mathcal{Y}$, у которых проекции на направление вектора y^* имеют наибольшие значения. Это следствие указывает на изложенный ниже возможный подход к решению задачи VS.

3. Алгоритм решения задачи кластерного анализа

Идея подхода к приближённому решению задачи VS состоит в замене её решения решением более простой вспомогательной задачи и последующей оценкой точности этой замены. Рассмотрим следующую вспомогательную задачу.

Задача SVSV (**S**earch for a **V**ector **S**ubset and **V**ector in the set). *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$. *Найти:* подмножество $\mathcal{B} \subseteq \mathcal{Y}$ мощности M и вектор $b \in \mathcal{Y}$ такие, что целевая функция

$$G(\mathcal{B}, b) = \sum_{y \in \mathcal{B}} \|y - b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2 \tag{6}$$

минимальна.

Построим алгоритм решения этой задачи. Обозначим через $\mathcal{B}^*$ и b^* подмножество и вектор, доставляющие минимум G^* целевой функции G. Для каждого фиксированного $b \in \mathcal{Y}$ определим функцию

$$h(y \mid b) = (y,\, b), \quad y \in \mathcal{Y}, \tag{7}$$

множество

$$\mathcal{H}(b) = \{h(y \mid b),\ y \in \mathcal{Y}\}, \tag{8}$$

совокупность

$$\mathcal{H}_M(b) = \{h_i \mid h_i \geqslant h_j,\ h_i \in \mathcal{H}(b),\ h_j \in \mathcal{H}(b), \\ i = 1, \ldots, M,\ j = M+1, \ldots, N\}, \tag{9}$$

состоящую из $M < N$ наибольших элементов множества $\mathcal{H}(b)$, и подмножество

$$\mathcal{B}(b) = \{y \mid y \in \mathcal{Y},\ h(y \mid b) \in \mathcal{H}_M(b)\} \tag{10}$$

множества $\mathcal{Y}$. При $M = N$ положим $\mathcal{H}_M(b) = \mathcal{H}(b)$.

Заметим, что в соответствии с определением (10) подмножество $\mathcal{B}(b)$ имеет структуру, аналогичную (см. следствие 1) структуре оптимального решения $\mathcal{C}^*$ задачи VS, а именно, подмножество $\mathcal{B}(b)$ состоит из вектора b и $M-1$ векторов, имеющих наибольшие проекции на направление вектора b.

Суть предлагаемого алгоритма решения задачи SVSV состоит в следующем. Сначала для каждого $b \in \mathcal{Y}$ вычисляется значение функции $G(\mathcal{B}(b), b)$. Затем в качестве решения задачи выбирается вектор b и соответствующее ему подмножество $\mathcal{B}(b)$ такие, что значение $G(\mathcal{B}(b), b)$ минимально.

Приведём алгоритм решения задачи SVSV в виде псевдокода.

Алгоритм $\mathcal{A}_1$

Шаг 1. Положим $b^* = 0$, $\mathcal{B}^* = \varnothing$, $G^* = +\infty$, $k := 0$.

Шаг 2. $k := k + 1$; положим $b = y_k$.

Шаг 3. Вычислим значение $h(y \mid b)$ для каждого $y \in \mathcal{Y}$ и сформируем множество $\mathcal{H}(b)$ по формулам (7), (8). Построим совокупность $\mathcal{H}_M(b)$ по формуле (9). Сформируем подмножество $\mathcal{B}(b)$ по формуле (10).

Шаг 4. Вычислим значение $G(\mathcal{B}(b), b)$ целевой функции (6).

Шаг 5. Если $G(\mathcal{B}(b), b) \leqslant G^*$, то положим $G^* = G(\mathcal{B}(b), b)$, $\mathcal{B}^* = \mathcal{B}(b)$, $b^* = b$; иначе переходим к следующему шагу.

Шаг 6. Если $k < N$, то переходим на шаг 2; иначе — к следующему шагу.

Шаг 7. Выход. Подмножество $\mathcal{B}^*$, вектор b^* и значение G^* целевой функции объявляем результатом работы алгоритма.

Оценку сложности и точности алгоритма устанавливает

Лемма 2. *Алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи* SVSV *за время $\mathcal{O}(qN^2)$.*

Доказательство. Из определения (10) множества $\mathcal{B}(b)$ следует, что при каждом фиксированном $b \in \mathcal{Y}$ справедливо очевидное равенство

$$\max_{\mathcal{B}\subseteq\mathcal{Y}} \sum_{y\in\mathcal{B}} (y,b) = \sum_{y\in\mathcal{B}(b)} (y,b).$$

Поэтому для оптимального значения целевой функции (6) задачи SVSV имеем

$$\begin{aligned}
G^* = G(\mathcal{B}^*, b^*) &= \min_{b\in\mathcal{Y},\ \mathcal{B}\subseteq\mathcal{Y}} G(\mathcal{B}, b) = \min_{b\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} G(\mathcal{B}, b) \\
&= \min_{b\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} \Big(\sum_{y\in\mathcal{B}} \|y-b\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}} \|y\|^2 \Big) \\
&= \min_{b\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} \Big(\sum_{y\in\mathcal{B}} (\|y-b\|^2 - \|y\|^2) + \sum_{y\in\mathcal{Y}} \|y\|^2 \Big) \\
= \sum_{y\in\mathcal{Y}} \|y\|^2 + \min_{b\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} \sum_{y\in\mathcal{B}} (-2(y,b)) &= \sum_{y\in\mathcal{Y}} \|y\|^2 + \min_{b\in\mathcal{Y}} \Big\{ -2 \max_{\mathcal{B}\subseteq\mathcal{Y}} \sum_{y\in\mathcal{B}} (y,b) \Big\} \\
= \sum_{y\in\mathcal{Y}} \|y\|^2 + \min_{b\in\mathcal{Y}} \sum_{y\in\mathcal{B}(b)} (-2(y,b)) &= \min_{b\in\mathcal{Y}} \Big(\sum_{y\in\mathcal{B}(b)} (\|y-b\|^2 - \|y\|^2) + \sum_{y\in\mathcal{Y}} \|y\|^2 \Big) \\
&= \min_{b\in\mathcal{Y}} \Big(\sum_{y\in\mathcal{B}(b)} \|y-b\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}(b)} \|y\|^2 \Big) = \min_{b\in\mathcal{Y}} G(\mathcal{B}(b), b).
\end{aligned}$$

Отсюда следует, что минимум G^* целевой функции $G(\mathcal{B}, b)$, оптимальное подмножество $\mathcal{B}^* \subseteq \mathcal{Y}$ и вектор $b^* \in \mathcal{Y}$ можно найти по значениям $G(\mathcal{B}(b), b)$ целевой функции, вычисленным для каждого $b \in \mathcal{Y}$, выбрав среди этих значений наименьшее и соответствующие этому значению вектор b и подмножество $\mathcal{B}(b)$.

Оценим временну́ю сложность алгоритма. Основной вклад в трудоёмкость алгоритма дают шаги 3 и 4.

Для вычисления значений $h(y \mid b)$, $y \in \mathcal{Y}$, и формирования множества $\mathcal{H}(b)$ на шаге 3 потребуется $\mathcal{O}(qN)$ операций. Построение совокупности $\mathcal{H}_M(b) \subseteq \mathcal{H}(b)$ можно осуществить за время $\mathcal{O}(N)$ (например, с помощью алгоритма [7] поиска M-го наибольшего числа в массиве из N чисел). Число операций, необходимых для формирования подмножества $\mathcal{B}(b)$, не превышает $\mathcal{O}(N)$. Вычисление значения целевой функции G на шаге 4 потребует $\mathcal{O}(qN)$ операций. Суммируя перечисленные затраты, устанавливаем, что на шагах 3 и 4 требуется $\mathcal{O}(qN)$ операций.

На каждом из шагов 1, 2, 5–7 число операций равно константе, не зависящей от размера входа задачи. Поскольку шаги 3 и 4 выполняются N раз, итоговая временна́я сложность алгоритма равна $\mathcal{O}(qN^2)$. Лемма 2 доказана.

Докажем вспомогательное утверждение.

Лемма 3. *Пусть $\mathcal{Z}$ — непустое конечное множество векторов из $\mathbb{R}^q$, а $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|}\sum\limits_{z\in\mathcal{Z}} z$. Тогда если вектор $x \in \mathbb{R}^q$ удовлетворяет условиям*

$$\|x - \overline{z}\| \leqslant \|z - \overline{z}\| \quad \forall z \in \mathcal{Z}, \tag{11}$$

то имеет место неравенство

$$\sum_{z\in\mathcal{Z}} \|z - x\|^2 \leqslant 2\sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2.$$

Доказательство. Пусть некоторый вектор $x \in \mathbb{R}^q$ удовлетворяет условиям (11). Тогда для каждого вектора $z \in \mathcal{Z}$ имеем

$$\begin{aligned}\|z - x\|^2 &= \|z - \overline{z} + \overline{z} - x\|^2 \\ &= \|z - \overline{z}\|^2 + \|\overline{z} - x\|^2 + 2(z - \overline{z}, \overline{z} - x) \\ &\leqslant 2\|z - \overline{z}\|^2 + 2(z - \overline{z}, \overline{z} - x).\end{aligned}$$

Суммируя полученное неравенство по всем векторам $z \in \mathcal{Z}$, найдём

$$\begin{aligned}\sum_{z\in\mathcal{Z}} \|z - x\|^2 &\leqslant 2\sum_{z\in\mathcal{Z}} (\|z - \overline{z}\|^2 + \langle z - \overline{z}, \overline{z} - x\rangle) \\ &= 2\sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2 + 2\Big\langle \sum_{z\in\mathcal{Z}} z - |\mathcal{Z}|\overline{z}, \overline{z} - x\Big\rangle = 2\sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2,\end{aligned}$$

где $\langle\cdot,\cdot\rangle$ — скалярное произведение векторов. Лемма 3 доказана.

Лемма 4. *Пусть $\mathcal{B}^*$, b^* — оптимальное решение вспомогательной задачи* SVSV, *а $\mathcal{C}^*$ — оптимальное решение задачи* VS. *Тогда имеет место оценка $S(\mathcal{B}^*) \leqslant 2S(\mathcal{C}^*)$.*

Доказательство. Для любого непустого конечного множества $\mathcal{Z}$ векторов из $\mathbb{R}^q$ минимум суммы $\sum\limits_{z\in\mathcal{Z}} \|z - x\|^2$ по x доставляется вектором $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|}\sum\limits_{z\in\mathcal{Z}} z$. Поэтому для множества $\mathcal{B}^*$ и векторов $\overline{y}(\mathcal{B}^*) = \frac{1}{|\mathcal{B}^*|}\sum\limits_{y\in\mathcal{B}^*} y$ и b^* имеем неравенство

$$\sum_{y\in\mathcal{B}^*} \|y - \overline{y}(\mathcal{B}^*)\|^2 \leqslant \sum_{y\in\mathcal{B}^*} \|y - b^*\|^2.$$

Из этого неравенства и определений функций S и G следует оценка

$$S(\mathcal{B}^*) = \sum_{y\in\mathcal{B}^*} \|y - \overline{y}(\mathcal{B}^*)\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}^*} \|y\|^2 \leqslant \sum_{y\in\mathcal{B}^*} \|y - b^*\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}^*} \|y\|^2 = G(\mathcal{B}^*, b^*). \quad (12)$$

Далее, рассмотрим вектор $t = \arg\min\limits_{y\in\mathcal{C}^*} \|y - y^*\|$, где $y^* = \overline{y}(\mathcal{C}^*) = \frac{1}{|\mathcal{C}^*|} \sum\limits_{y\in\mathcal{C}^*} y$. Этот ближайший к y^* вектор в оптимальном множестве $\mathcal{C}^*$ и само множество $\mathcal{C}^*$ удовлетворяют условиям леммы 4. Следовательно,

$$\sum_{y\in\mathcal{C}^*} \|y - t\|^2 \leqslant 2 \sum_{y\in\mathcal{C}^*} \|y - y^*\|^2. \quad (13)$$

Кроме того, заметим, что $|\mathcal{C}^*| = |\mathcal{B}^*| = M$, $t \in \mathcal{Y}$ и $\mathcal{C}^* \subseteq \mathcal{Y}$. Поэтому $\mathcal{C}^*, t$ — допустимое решение вспомогательной задачи SVSV. Следовательно,

$$G(\mathcal{B}^*, b^*) \leqslant G(\mathcal{C}^*, t). \quad (14)$$

Объединяя (12)–(14) и опираясь на определение функции G, получим оценку

$$S(\mathcal{B}^*) \leqslant G(\mathcal{B}^*, b^*) \leqslant G(\mathcal{C}^*, t) = \sum_{y\in\mathcal{C}^*} \|y - t\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2$$
$$\leqslant 2 \sum_{y\in\mathcal{C}^*} \|y-y^*\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 \leqslant 2 \sum_{y\in\mathcal{C}^*} \|y-y^*\|^2 + 2 \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 = 2S(\mathcal{C}^*).$$

Лемма 4 доказана.

Опираясь на лемму 4, представим алгоритм решения задачи VS.

Алгоритм $\mathcal{A}$.

Шаг 1. По заданному множеству $\mathcal{Y}$ и числу M находим оптимальное решение $\mathcal{B}^*, b^*$ вспомогательной задачи SVSV с помощью алгоритма $\mathcal{A}_1$.

Шаг 2. Подмножество $\mathcal{B}^*$ объявляем решением задачи VS.

Теорема 2. *Алгоритм $\mathcal{A}$ находит 2-приближённое решение задачи* VS *за время $\mathcal{O}(qN^2)$. Оценка 2 точности алгоритма достижима и неулучшаема.*

Доказательство. Справедливость утверждения теоремы следует из лемм 2–4 и приведённого ниже примера, показывающего существование входных данных задачи, для которых отношение $S(\mathcal{B}^*)/S(\mathcal{C}^*)$ может быть сколь угодно близко к 2 и равно 2.

Пусть $q = 2$, $N = 5$, $M = 4$, $y_1 = (0, 0)$, $y_2 = (10, 0)$, $y_3 = (11, 0)$, $y_4 = (10, 1 + \alpha)$, $y_5 = (11, 1)$. Тогда если $-\sqrt{807} - 1 < \alpha < \sqrt{807} - 1$, то $\mathcal{B}^* = \mathcal{C}^* = \{y_2, y_3, y_4, y_5\}$, $b^* = y_5$, $S(\mathcal{B}^*) = 4 + \alpha^2$, $S(\mathcal{C}^*) = 2 + \alpha + 3\alpha^2/4$ и

$$S(\mathcal{B}^*)/S(\mathcal{C}^*) = (4 + \alpha^2)/(2 + \alpha + 3\alpha^2/4).$$

Отсюда видно, что $S(\mathcal{B}^*)/S(\mathcal{C}^*) = 2$ при $\alpha = 0$, и существуют такие данные, что это отношение сколько угодно близко к 2. Теорема 2 доказана.

Заключение

Рассмотрена одна из неизученных ранее задач разбиения множества векторов евклидова пространства на два кластера фиксированной мощности по критерию минимума суммы квадратов расстояний. Установлена связь этой задачи с изучавшимися ранее труднорешаемыми задачами поиска подмножества векторов. Показано, что рассматриваемая задача NP-трудна. Построен эффективный приближённый алгоритм с константной оценкой точности 2 для её решения.

Делом ближайшей перспективы является изучение вопросов аппроксимируемости задачи, а также построение более эффективных приближённых алгоритмов. Представляют интерес алгоритмы рандомизированного типа, а также алгоритмы, ориентированные на решение задачи со случайными входами.

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., Глазков Ю. В., Рыков И. А.** О двух задачах выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммой разности // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 5. — С. 30–43.
3. **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.

4. **Гимади Э. Х., Пяткин А. В., Рыков И. А.** О полиномиальной разрешимости некоторых задач выбора подмножества векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 6. — С. 11–19.
5. **Кельманов А. В.** Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // Тр. Ин-та математики и механики УрО РАН. — 2008. — Т. 14, № 2. — С. 81–88.
6. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Trans. Signal Processing. — 2004. — Vol. 52, N 3. — P. 1–12.
7. **Wirth H.** Algorithms + data structures = programs. — New Jersey: Prentice Hall, 1976. — 366 p.

Долгушев Алексей Владимирович,
e-mail: dolgushev@math.nsc.ru

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

Статья поступила
26 декабря 2010 г.

Переработанный вариант —
18 января 2011 г.