ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Январь—февраль 2014. Том 21, № 1. С. 53–66

УДК 519.16 + 519.85

ПРИБЛИЖЁННЫЙ ПОЛИНОМИАЛЬНЫЙ АЛГОРИТМ ДЛЯ ОДНОЙ ЗАДАЧИ РАЗБИЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ *)

А. В. Кельманов, С. А. Хамидуллин

Аннотация. Рассматривается NP-трудная в сильном смысле задача разбиения конечной последовательности векторов евклидова пространства на два кластера по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров. Предполагается, что мощности кластеров фиксированы. Центр одного из кластеров является оптимизируемой величиной и определяется как среднее значение по всем векторам, образующим этот кластер. Центр второго кластера полагается равным нулю. При этом разбиение подчинено условию: разность между номерами последующего и предыдущего векторов, входящих в первый кластер, ограничена сверху и снизу заданными константами. Предложен 2-приближённый полиномиальный алгоритм решения этой задачи.

Ключевые слова: последовательность евклидовых векторов, кластеризация, минимум суммы квадратов расстояний, NP-трудность, полиномиальный 2-приближённый алгоритм.

Введение

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

Пример содержательной трактовки проблемы, которая приводит к решению сфомулированной ниже задачи, состоит в следующем. Некоторый материальный объект может находиться в пассивном и конечном множестве уникальных активных состояний. Имеется таблица, содержащая

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты № 12–01–00090, № 12–01–33028–мол–а–вед, № 13–07–00070), а также целевой программы СО РАН (интеграционные проекты 7Б и 21А).

упорядоченные по времени результаты многократных измерений набора числовых информационно значимых характеристик этого объекта. В пассивном состоянии все числовые характеристики из набора равны нулю, а в любом активном — значение хотя бы одной характеристики не равно нулю. Число измерений, проведенных для каждого из активных состояний, известно. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие элементов таблицы какому-либо состоянию объекта неизвестно. Однако известно, что временно́й интервал между двумя последовательными активными состояниями объекта ограничен сверху и снизу некоторыми константами. Требуется, используя критерий минимума суммы квадратов расстояний, разбить таблицу на подмножества наборов, соответствующих пассивному и каждому активному состояниям объекта, а также оценить по результатам измерения наборы характеристик объекта в активных состояниях (учитывая, что данные содержат ошибку измерения).

Сформулированная содержательная проблема типична для многих приложений, связанных с анализом данных и распознаванием образов (см., например, [1–9] и цитированные там работы). В [1, 2] установлено, что эта проблема моделируется NP-трудной в сильном смысле задачей даже в простейшем случае, когда таблица содержит упорядоченные по времени данные только об одном активном и пассивном состояниях объекта и необходимо разбить эту таблицу на два кластера. Формальная постановка задачи кластеризации приведена в разд. 1. Здесь лишь отметим, что мотивацией данного исследования послужило отсутствие каких-либо эффективных алгоритмов с гарантированными оценками точности для решения экстремальной задачи, которая индуцируется сформулированной проблемой.

1. Формулировка задачи

Приведём математическую формулировку задачи, моделирующей содержательную проблему. Положим $\mathcal{N} = \{1, \ldots, N\}$.

Задача J-MSSCS-F.

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ векторов из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и $M_1, \ldots, M_J$.

Найти: разбиение множества $\mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ на непустые подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_J$ и $\mathcal{N} \setminus \mathcal{M}$, $\mathcal{M} = \mathcal{M}_1 \cup \ldots \cup \mathcal{M}_J = \{n_1, \ldots, n_M\}$, такое, что

$$\sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - \overline{y}(\mathcal{M}_j)\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}} \|y_n\|^2 \to \min, \tag{1}$$

$\overline{y}(\mathcal{M}_j) = \frac{1}{|\mathcal{M}_j|} \sum\limits_{n \in \mathcal{M}_j} y_n$, при условии, что $|\mathcal{M}_j| = M_j$, $j = 1, \ldots, J$, и ограничениях

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \quad m = 2, \ldots, M, \tag{2}$$

на элементы набора $\mathcal{M}$.

В этой задаче векторы последовательности $\mathcal{Y}$ с номерами из подмножества $\mathcal{M}_j$ образуют j-й кластер (подпоследовательность) с центром $\overline{y}(\mathcal{M}_j)$. Векторы с номерами из набора $\mathcal{N} \setminus \mathcal{M}$ образуют кластер с фиксированным в нуле центром.

Аббревиатура MSSCS в кратком названии этой задачи образована от английского словосочетания Minimum Sum-of-Squares Clustering, for the case of Sequence и подчеркивает сходство с классической [6] трудноразрешимой [5] задачей MSSC (Minimum Sum-of-Squares Clustering). Литера J обозначает число кластеров, для которых требуется определить (оценить) центры (в задаче MSSC оцениваются центры всех кластеров). Буква F (от английского Fixed) указывает на то, что мощности искомых кластеров фиксированы.

Если номера членов последовательности $\mathcal{Y}$ интерпретировать как равномерные дискретные отсчёты времени, то элементы набора $\mathcal{M} = \{n_1, \ldots, n_M\}$ номеров этой последовательности в содержательной проблеме соответствуют моментам времени, в которые объект находился в каком-либо из активных состояний. При этом элементы последовательности соответствуют строкам таблицы. Номера из набора $\mathcal{N} \setminus \mathcal{M}$ соответствуют моментам времени, в которые объект находился в пассивном состоянии. Натуральные $T_{\min}$ и $T_{\max}$ в формулировке задачи соответствуют минимальному и максимальному интервалам времени между двумя последовательными активными состояниями объекта.

В [2] доказано, что эта задача NP-трудна в сильном смысле, когда $J \geqslant 2$ и $T_{\min} \leqslant T_{\max}$.

Простейшим случаем задачи J-MSSCS-F является

Задача 1-MSSCS-F. Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ векторов из $\mathbb{R}^q$ и натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$.
Найти: подмножество $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такое, что целевая функция

$$F(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2, \tag{3}$$

где $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum\limits_{i \in \mathcal{M}} y_i$, минимальна при ограничениях (2) на элементы искомого подмножества $\mathcal{M}$.

В [2] установлено, что эта задача NP-трудна в сильном смысле для любых $T_{\min} < T_{\max}$. В тривиальном случае, когда $T_{\min} = T_{\max}$, задача разрешима за полиномиальное время.

Поскольку сформулированные задачи NP-трудны в сильном смысле, в предположении справедливости гипотезы P≠NP для этих задач не существует [7] полиномиальных и псевдополиномиальных алгоритмов, гарантирующих отыскание точного решения, а также полностью полиномиальных аппроксимационных схем (FPTAS). Ниже для задачи 1-MSSCS-F предложен полиномиальный 2-приближённый алгоритм.

2. Алгоритм решения задачи

Суть предлагаемого подхода к построению приближённого алгоритма состоит в замене решения исходной задачи 1-MSSCS-F решением более простой вспомогательной задачи и последующей оценкой точности этой замены. Для построения алгоритма сформулируем свойства элементов из набора $\mathcal{M}$.

Лемма 1. *Пусть элементы набора $(n_1, \ldots, n_M)$ принадлежат множеству $\mathcal{N}$ и удовлетворяют системе ограничений* (2). *Тогда*

(i) *если $M \geqslant 2$, то её параметры связаны соотношением*

$$(M-1)T_{\min} \leqslant N-1; \tag{4}$$

(ii) *элемент n_m из набора $(n_1, \ldots, n_M)$ принадлежит множеству*

$$\omega_m = \big\{n \mid 1 + (m-1)T_{\min} \leqslant n \leqslant N - (M-m)T_{\min}\big\}; \tag{5}$$

(iii) *если в наборе $(n_1, \ldots, n_M)$ элемент n_m равен n, где $n \in \omega_m$, то n_{m-1} принадлежит множеству*

$$\gamma^{-}_{m-1}(n) = \big\{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \leqslant j \leqslant n - T_{\min}\big\} \tag{6}$$

для любого $m \in \{2, \ldots, M\}$.

Лемма 2. *Если $M \geqslant 2$ и параметры системы* (2) *связаны соотношением* (4), *то она совместна.*

Справедливость лемм 1 и 2 следует из [4]. Построим алгоритм решения следующей вспомогательной задачи.

Задача 1. Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ векторов из $\mathbb{R}^q$, вектор $b \in \mathbb{R}^q$ и натуральные числа $T_{\min}$, $T_{\max}$ и $M \geqslant 1$.
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$, доставляющий максимум целевой функции

$$G(\mathcal{M}) = \sum_{n \in \mathcal{M}} g(n), \tag{7}$$

где

$$g(n) = 2(y_n, b) - \|b\|^2, \quad n \in \mathcal{N}, \tag{8}$$

при дополнительных ограничениях (2) на элементы набора $\mathcal{M}$, если $M \neq 1$.

Из условий задачи и (7), (8) следует, что

$$G(\mathcal{M}) = 2 \sum_{n \in \mathcal{M}} (y_n, b) - M\|b\|^2,$$

где $M\|b\|^2 = \text{const}$. Поэтому фактически в задаче 1 требуется найти подпоследовательность, состоящую из M векторов, максимизирующих сумму скалярных произведений этих векторов на заданный вектор b. Другими словами, искомые векторы должны быть максимально «похожи» (в смысле скалярного произведения) на заданный вектор.

Определим множество наборов номеров элементов последовательности $\mathcal{Y}$, допустимых в задаче 1:

$$\Psi_M = \begin{cases} \{(n_1) \mid n_1 \in \mathcal{N}\}, & \text{если } M = 1, \\ \{(n_1, \ldots, n_M) \mid n_i \in \mathcal{N},\ i = 1, \ldots, M; \\ 1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \\ \qquad\qquad m = 2, \ldots, M\}, & \text{если } 1 < M \leqslant N. \end{cases} \tag{9}$$

Заметим, что при $M = 1$ множество Ψ_M непусто при любых параметрах $T_{\min}$ и $T_{\max}$, входящих в определение (9). Для остальных допустимых значений M справедлива

Лемма 3. *Если $M \geqslant 2$, то множество Ψ_M непусто тогда и только тогда, когда имеет место неравенство* (4).

Справедливость утверждения следует из лемм 1 и 2.

Лемма 4. *Пусть $\Psi_M \neq \varnothing$ для некоторого значения $M \geqslant 1$. Тогда для этого M оптимальное значение $G_{\max} = \max\limits_{\mathcal{M}} G(\mathcal{M})$ целевой функции задачи* 1 *находится по формуле*

$$G_{\max} = \max_{n \in \omega_M} G_M(n), \tag{10}$$

а значения функции $G_M(n),\ n \in \omega_M$, *вычисляются по следующим рекуррентным формулам*:

$$G_m(n) = \begin{cases} g(n), & \text{если } n \in \omega_1,\ m = 1, \\ g(n) + \max\limits_{j \in \gamma^-_{m-1}(n)} G_{m-1}(j), & \text{если } n \in \omega_m,\ m = 2, \ldots, M, \end{cases} \quad (11)$$

где множества ω_m *и* $\gamma^-_{m-1}(n)$ *задаются формулами* (5) *и* (6).

Доказательство. Опираясь на лемму 1, запишем определение (9) в эквивалентном виде:

$$\Psi_M = \begin{cases} \{(n_1) \mid n_1 \in \omega_1\}, & \text{если } M = 1, \\ \{(n_1, \ldots, n_M) \mid n_M \in \omega_M; \\ \quad n_{m-1} \in \gamma^-_{m-1}(n_m),\ m = 2, \ldots, M\}, & \text{если } 1 < M \leqslant N. \end{cases} \quad (12)$$

При $M = 1$ формулы (10) и (11) очевидны. Пусть $M > 1$. Для каждого $n \in \omega_m$ определим множества

$$\psi_m(n) = \begin{cases} \{(n_1) \mid n_1 = n\}, & \text{если } m = 1, \\ \{(n_1, \ldots, n_m) \mid n_m = n, \\ n_{i-1} \in \gamma^-_{i-1}(n_i),\ i = 2, \ldots, m\}, & \text{если } m = 2, \ldots, M, \end{cases} \quad (13)$$

$$\psi^-_{m-1}(n) = \{(n_1, \ldots, n_{m-1}) \mid n_{m-1} \in \gamma^-_{m-1}(n);\ n_{i-1} \in \gamma^-_{i-1}(n_i), \\ i = 2, \ldots, m-1\},\ \ m = 2, \ldots, M. \quad (14)$$

Формула (13) определяет множество допустимых наборов размерности m, у которых последняя компонента n_m фиксирована и равна n, причём $n \in \omega_m$. Формула (14) задаёт множество допустимых поднаборов размерности $m-1$ набора $(n_1, \ldots, n_{m-1}, n_m)$ при условии, что $n_m = n$ и $n \in \omega_m$. Заметим, что эти множества непусты в силу лемм 2 и 3, так как $\Psi_M \neq \varnothing$ по предположению. Кроме того, заметим, что из (12)–(14) следуют равенства

$$\Psi_M = \bigcup_{n \in \omega_M} \psi_M(n), \quad (15)$$

$$\psi^-_{m-1}(n) = \bigcup_{j \in \gamma^-_{m-1}(n)} \psi_{m-1}(j), \quad n \in \omega_m,\ m = 2, \ldots, M. \quad (16)$$

Рассмотрим задачу вычисления максимума целевой функции

$$G(n_1, \ldots, n_M \mid n_M = n) = g(n) + \sum_{i=1}^{M-1} g(n_i)$$

на множестве $\psi_M(n)$ наборов $(n_1, \ldots, n_{M-1}, n)$ при фиксированном n из ω_M. Обозначим эту задачу через $\langle M \mid n_M = n\rangle$ и для максимума функции положим

$$G_M(n) = \max_{\psi_M(n)} G(n_1, \ldots, n_M \mid n_M = n). \tag{17}$$

Погрузим задачу $\langle M \mid n_M = n\rangle$ в семейство подзадач $\langle m \mid n_m = n\rangle$, $n \in \omega_m$, $m = M, M-1, \ldots, 1$, вычисления максимума функции

$$G(n_1, \ldots, n_m \mid n_m = n) = g(n) + \sum_{i=1}^{m-1} g(n_i)$$

на множестве наборов (13). Положим $\eta_m = (n_1, \ldots, n_m)$ и

$$G_m(n) = \max_{\eta_m \in \psi_m(n)} G(n_1, \ldots, n_m \mid n_m = n). \tag{18}$$

Проанализируем случаи, указанные в формуле (11). Если $m = 1$ и $n \in \omega_1$, то из (7) и (5) получаем

$$G_1(n) = \max_{\eta_1 \in \psi_1(n)} \sum_{m=1}^{1} g(n_m) = \max_{(n_1) \in \psi_1(n)} g(n_1) = \max_{(n_1)=(n)} g(n_1) = g(n).$$

Пусть $m > 1$. Тогда, используя (13), (14), (16) и (18), для формулы (11) имеем

$$\begin{aligned} G_m(n) &= \max_{\eta_m \in \psi_m(n)} \left(g(n) + \sum_{i=1}^{m-1} g(n_i) \right) = g(n) + \max_{\eta_{m-1} \in \psi^-_{m-1}(n)} \sum_{i=1}^{m-1} g(n_i) \\ &= g(n) + \max_{n_{m-1} \in \gamma^-_{m-1}(n)} \max_{\eta_{m-1} \in \psi_{m-1}(n_{m-1})} \left(g(n_{m-1}) + \sum_{i=1}^{m-2} g(n_i) \right) \\ &= g(n) + \max_{j \in \gamma^-_{m-1}(n)} \left\{ \max_{\eta_{m-1} \in \psi_{m-1}(j)} \left(g(j) + \sum_{i=1}^{m-2} g(n_i) \right) \right\} \\ &= g(n) + \max_{j \in \gamma^-_{m-1}(n)} G_{m-1}(j). \end{aligned}$$

Наконец, принимая во внимание (15) и (17), формулу (10) получаем из цепочки равенств

$$\begin{aligned} G_{\max} = \max_{\eta_M \in \Psi_M} \sum_{m=1}^{M} g(n_m) &= \max_{n \in \omega_M} \left\{ \max_{\eta_M \in \psi_M(n)} \left(g(n) + \sum_{m=1}^{M-1} g(n_m) \right) \right\} \\ &= \max_{n \in \omega_M} G_M(n). \end{aligned}$$

Лемма 4 доказана.

Следствие 1. *Элементы* $\widehat{n}_1, \ldots, \widehat{n}_M$ *оптимального набора* $\widehat{\mathcal{M}} = \arg\max\limits_{\mathcal{M}} G(\mathcal{M})$ *находятся по следующим рекуррентным формулам*:

$$\widehat{n}_M = \arg\max_{n \in \omega_M} G_M(n), \tag{19}$$

$$\widehat{n}_{m-1} = \arg\max_{n \in \gamma^-_m(\widehat{n}_m)} G_m(n), \quad m = M, M-1, \ldots, 2. \tag{20}$$

Доказательство. Формула (19) следует непосредственно из (10). Справедливость (20) покажем по индукции. По определению оптимального набора имеем

$$G_{\max} = G(\widehat{n}_1, \ldots, \widehat{n}_M) = \sum_{m=1}^{M} g(\widehat{n}_m). \tag{21}$$

Предположим, что $\widehat{n}_M, \ldots, \widehat{n}_m$ найдены по формулам (19)–(20). Покажем, что формула (20) справедлива при вычислении $\widehat{n}_{m-1}$.

Из (18), (21) и леммы 4 следует, что

$$G_{\max} = \sum_{i=1}^{m} g(\widehat{n}_i) + \sum_{i=m+1}^{M} g(\widehat{n}_i) = G_m(\widehat{n}_m) + \sum_{i=m+1}^{M} g(\widehat{n}_i),$$

$$G_{\max} = \sum_{i=1}^{m-1} g(\widehat{n}_i) + \sum_{i=m}^{M} g(\widehat{n}_i) = G_{m-1}(\widehat{n}_{m-1}) + \sum_{i=m}^{M} g(\widehat{n}_i).$$

Отсюда

$$G_m(\widehat{n}_m) = G_{m-1}(\widehat{n}_{m-1}) + g(\widehat{n}_m). \tag{22}$$

Из (11) формально имеем

$$G_m(\widehat{n}_m) = g(\widehat{n}_m) + \max_{j \in \gamma^-_{m-1}(\widehat{n}_m)} G_{m-1}(j). \tag{23}$$

Комбинируя (22) и (23), получим $G_{m-1}(\widehat{n}_{m-1}) = \max\limits_{j \in \gamma^-_{m-1}(\widehat{n}_m)} G_{m-1}(j)$. Из этого равенства следует справедливость формулы (20). Следствие 1 доказано.

Таким образом, оптимальное решение задачи 1 можно найти с помощью следующего алгоритма, реализующего схему динамического программирования. Входами алгоритма являются $\mathcal{Y}$, b, $T_{\min}$, $T_{\max}$ и M.

Алгоритм $\mathcal{A}_1$

Шаг 1. Вычислим значения $g(n),\ n \in \mathcal{N}$, по формуле (8).

Шаг 2. Используя рекуррентные формулы (11), вычислим значения $G_m(n)$ для каждого $n \in \omega_m$ и $m = 1, \ldots, M$.

Шаг 3. Найдем значение $G_{\max}$ максимума целевой функции G по формуле (10) и оптимальный набор $\widehat{\mathcal{M}} = \{\widehat{n}_1, \ldots, \widehat{n}_M\}$ по формулам (19) и (20); выход.

Выходом алгоритма являются значения, зависящие от b, т. е. $G_{\max} = G_{\max}(b)$ и $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}(b) = \{\widehat{n}_1(b), \ldots, \widehat{n}_M(b)\}$.

Теорема 1. *Алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи* 1 *за время* $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$.

Доказательство. Оптимальность решения следует из леммы 4. Оценим временну́ю сложность алгоритма.

На первом шаге алгоритма требуется $\mathcal{O}(Nq)$ операций. Шаг 2 вносит основной вклад в трудоёмкость. Трудоёмкость этого шага определяется мощностью множеств ω_m и $\gamma^-_{m-1}(n)$, входящих в определение функции (11). Мощность первого из этих множеств не превосходит N, а мощность второго не больше $T_{\max} - T_{\min} + 1$. Вычисления по формуле (11) производятся для каждого $m = 1, \ldots, M$. Поэтому трудоёмкость шага 2 есть величина $\mathcal{O}(NM(T_{\max} - T_{\min} + 1))$. Из формул (10), (19) и (20) видно, что на шаге 3 требуется $\mathcal{O}(M(T_{\max} - T_{\min} + 1))$ операций. Суммируя затраты на всех шагах, получим оценку, приведённую в формулировке теоремы. Теорема 1 доказана.

Замечание 1. В оценке временно́й сложности алгоритма $\mathcal{A}_1$ множители M и $(T_{\max} - T_{\min} + 1)$ не превосходят N. Поэтому алгоритм полиномиален по N и по q, а его сложность можно оценить как $O(N(N^2 + q))$.

Изложим алгоритм решения задачи 1-MSSCS-F, в котором используется алгоритм $\mathcal{A}_1$. Входами алгоритма являются $\mathcal{Y}$, $T_{\min}$, $T_{\max}$ и M.

Алгоритм $\mathcal{A}$

Шаг 1. Положим $i = 0,\ \mathcal{M}_A = \varnothing,\ H = -\infty$.

Шаг 2. Положим $i := i + 1,\ b = y_i$.

Шаг 3. Для фиксированного вектора $b \in \mathcal{Y}$ найдём оптимальное решение $\widehat{\mathcal{M}}(b)$ и значение $G_{\max}(b)$ целевой функции задачи 1 с помощью алгоритма $\mathcal{A}_1$.

Шаг 4. Если $H < G_{\max}(b)$, то положим $b_A = b$, $H = G_{\max}(b)$, $\mathcal{M}_A = \widehat{\mathcal{M}}(b)$.

Шаг 5. Если $i < N$, то переходим на шаг 2, иначе — к следующему шагу.

Шаг 6. Вычислим вектор $\overline{y}(\mathcal{M}_A) = \frac{1}{M} \sum\limits_{n \in \mathcal{M}_A} y_n$ и значение $F(\mathcal{M}_A)$ целевой функции по формуле (3); положим $G_{\max}(b_A) = H$; выход.

Выходом алгоритма (решением задачи) объявляем набор $\mathcal{M}_A$, значение $F(\mathcal{M}_A)$, а также векторы $\overline{y}(\mathcal{M}_A)$ и b_A. Если максимуму $G_{\max}(b_A)$ соответствует несколько наборов $\widehat{\mathcal{M}}_A$, то выбираем любой из них.

Суть алгоритма $\mathcal{A}$ состоит в решении задачи 1 с помощью алгоритма $\mathcal{A}_1$ для каждого вектора последовательности $\mathcal{Y}$ и последующего выбора из найденных решений (наборов) наилучшего набора $\widehat{\mathcal{M}}_A$, которому соответствует наибольшее значение $G_{\max}(b_A)$ целевой функции задачи 1.

Для обоснования точности алгоритмического решения потребуется

Лемма 5. *Пусть $\mathcal{Z}$ — непустое конечное множество векторов из $\mathbb{R}^q$, а $\overline{z} = \frac{1}{|\mathcal{Z}|} \sum\limits_{z \in \mathcal{Z}} z$ — центр этого множества. Если вектор $t \in \mathbb{R}^q$ удовлетворяет условиям $\|t - \overline{z}\| \leqslant \|z - \overline{z}\|$ для любого $z \in \mathcal{Z}$, то имеет место неравенство*

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

Справедливость леммы 5 установлена в [3].

Теорема 2. *Алгоритм $\mathcal{A}$ находит 2-приближённое решение задачи* 1-MSSCS-F *за время $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q))$. Оценка 2 точности алгоритма асимптотически достижима.*

Доказательство. Пусть $\mathcal{M}^*$ — оптимальное решение задачи 1-MSSCS-F, $\mathcal{C}^* = \{y_n \mid n \in \mathcal{M}^*\}$ — подмножество векторов из $\mathcal{Y}$, соответствующих оптимальному набору $\mathcal{M}^*$, $\overline{y}(\mathcal{M}^*) = \frac{1}{|\mathcal{M}^*|} \sum\limits_{n \in \mathcal{M}^*} y_n$ — центр подмножества $\mathcal{C}^* \subseteq \mathcal{Y}$, а $u = \arg\min\limits_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{M}^*)\|$ — вектор, ближайший к центру подмножества $\mathcal{C}^*$.

Согласно пошаговой записи алгоритм находит вектор

$$b_A = \arg\max_{b \in \mathcal{Y}} G_{\max}(b) \tag{24}$$

из множества $\mathcal{Y}$, набор $\mathcal{M}_A = \widehat{\mathcal{M}}(b_A) = \{\widehat{n}_1(b_A), \ldots, \widehat{n}_M(b_A)\}$, вектор $\overline{y}(\mathcal{M}_A)$, значение $F(\mathcal{M}_A)$ целевой функции задачи 1-MSSCS-F, а также максимум

$$G_{\max}(b_A) = \sum_{n \in \mathcal{M}_A} \{2(y_n, b_A) - \|b_A\|^2\} = \max_{b \in \mathcal{Y}} \max_{\mathcal{M}} \sum_{n \in \mathcal{M}} \{2(y_n, b) - \|b\|^2\} \tag{25}$$

функции $G_{\max}(b)$, $b \in \mathcal{Y}$.

Справедливость утверждения теоремы 2 вытекает из следующей цепочки равенств и неравенств:

$$
\begin{aligned}
F(\mathcal{M}_A) &= \sum_{n\in\mathcal{M}_A} \|y_n - \overline{y}(\mathcal{M}_A)\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}_A} \|y_n\|^2 \\
&\leqslant_{(1)} \sum_{n\in\mathcal{M}_A} \|y_n - b_A\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}_A} \|y_n\|^2 = \sum_{n\in\mathcal{N}} \|y_n\|^2 - \sum_{n\in\mathcal{M}_A} \{2(y_n, b_A) - \|b_A\|^2\} \\
&=_{(2)} \sum_{n\in\mathcal{N}} \|y_n\|^2 - \max_{b\in\mathcal{Y}} \max_{\mathcal{M}} \sum_{n\in\mathcal{M}} \{2(y_n, b) - \|b\|^2\} \\
&=_{(3)} \min_{b\in\mathcal{Y}} \min_{\mathcal{M}} \Big(\sum_{n\in\mathcal{N}} \|y_n\|^2 - \sum_{n\in\mathcal{M}} \{2(y_n, b) - \|b\|^2\} \Big) \\
&= \min_{b\in\mathcal{Y}} \min_{\mathcal{M}} \Big(\sum_{n\in\mathcal{M}} \|y_n - b\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}} \|y_n\|^2 \Big) \leqslant_{(4)} \sum_{n\in\mathcal{M}^*} \|y_n - u\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}^*} \|y_n\|^2 \\
&\leqslant_{(5)} 2 \sum_{n\in\mathcal{M}^*} \|y_n - \overline{y}(\mathcal{M}^*)\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}^*} \|y_n\|^2 \\
&\leqslant 2\Big(\sum_{n\in\mathcal{M}^*} \|y_n - \overline{y}(\mathcal{M}^*)\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}^*} \|y_n\|^2 \Big) = 2F(\mathcal{M}^*). \quad (26)
\end{aligned}
$$

В этой цепочке справедливость непомеченных знаков равенств и неравенств очевидна. Неравенство 1 справедливо, так как для любого конечного множества $\mathcal{Z}$ векторов из $\mathbb{R}^q$ (в частности, для $\mathcal{Z} = \{y_n \mid n \in \mathcal{M}_A\}$) минимум суммы квадратов $\sum\limits_{z\in\mathcal{Z}} \|z - c\|^2$ по c достигается в точке $c = \overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum\limits_{z\in\mathcal{Z}} z$ (т. е. в точке $c = \overline{y}(\mathcal{M}_A)$). Равенство 2 следует из теоремы 1 и формул (24), (25). Равенство 3 справедливо, так как $\sum\limits_{n\in\mathcal{N}} \|y_n\|^2$ — константа. Неравенство 4 следует из того, что подмножество $\mathcal{M}^*$ и вектор $u \in \mathcal{Y}$ — допустимое решение задачи $\min\limits_{b\in\mathcal{Y}} \min\limits_{\mathcal{M}}(\cdot)$. Справедливость неравенства 5 вытекает из леммы 5.

Таким образом, из (26) следует, что $F(\mathcal{M}_A)/F(\mathcal{M}^*) \leqslant 2$, т. е. алгоритм $\mathcal{A}$ находит 2-приближённое решение.

Покажем, что оценка 2 точности алгоритма асимптотически достижима. Для этого приведём пример, показывающий существование таких входных данных задачи, для которых отношение $F(\mathcal{M}_A)/F(\mathcal{M}^*)$ может быть сколь угодно близко к 2.

Пусть $q = 2$, $N = 4$, $M = 2$, $T_{\min} = 1$, $T_{\max} = 3$, $\mathcal{Y} = (y_1, y_2, y_3, y_4)$, где $y_1 = (0, 0)$, $y_2 = (0, 1)$, $y_3 = (1, 0)$, $y_4 = (\alpha, 0)$ и $\alpha < -1$.

Легко видеть, что в этом примере оптимальным решением задачи 1-MSSCS-F является набор $\mathcal{M}^* = \{2, 4\}$, которому соответствуют векторы y_2 и y_4 последовательности $\mathcal{Y}$, причём $F(\mathcal{M}^*) = \frac{3+\alpha^2}{2}$.

После несложных вычислений с использованием рекуррентных формул (19), (20), (10), (11) динамического программирования и формул (24), (25) устанавливаем, что максимуму $G_{\max}(b_A) = 0$ соответствует пять равноправных алгоритмических решений $\mathcal{M}_A^{(1)} = \{1, 2\}$, $\mathcal{M}_A^{(2)} = \{1, 3\}$, $\mathcal{M}_A^{(3)} = \{1, 4\}$, $\mathcal{M}_A^{(4)} = \{2, 3\}$, $\mathcal{M}_A^{(5)} = \{2, 4\}$. Для этих решений имеем следующие значения целевой функции F:

$$F\big(\mathcal{M}_A^{(1)}\big) = F\big(\mathcal{M}_A^{(2)}\big) = \frac{3}{2} + \alpha^2,$$

$$F\big(\mathcal{M}_A^{(3)}\big) = 2 + \frac{\alpha^2}{2}, \quad F\big(\mathcal{M}_A^{(4)}\big) = 1 + \alpha^2, \quad F\big(\mathcal{M}_A^{(5)}\big) = \frac{3+\alpha^2}{2}.$$

Отношение

$$\frac{F\big(\mathcal{M}_A^{(1)}\big)}{F\big(\mathcal{M}^*\big)} = \frac{F\big(\mathcal{M}_A^{(2)}\big)}{F\big(\mathcal{M}^*\big)} = \frac{3+2\alpha^2}{3+\alpha^2}$$

может быть сколь угодно близко к 2 при $\alpha \to -\infty$, т. е. оценка 2 точности алгоритма асимптотически достижима.

Оценим временну́ю сложность алгоритма. Время вычислений определяется трудоёмкостью шага 3. На этом шаге N раз решается вспомогательная задача 1 с помощью алгоритма $\mathcal{A}_1$, трудоёмкость которого оценена в теореме 1. Отсюда следует оценка сложности. Теорема 2 доказана.

Замечание 2. В оценке временно́й сложности алгоритма $\mathcal{A}$ множители M и $(T_{\max} - T_{\min} + 1)$ не превосходят N. Поэтому алгоритм полиномиален по N и q, а его сложность можно оценить как $O(N^2(N^2 + q))$.

Остаётся заметить, что алгоритм можно применять и для решения обобщения задачи на случай, когда центр одного из кластеров зафиксирован не в нуле, а в произвольной заданной точке (векторе) евклидова пространства. Действительно, в этом случае достаточно переместить начало координат в заданную точку, пересчитать координаты векторов и применить изложенный алгоритм.

Заключение

Обоснован 2-приближённый полиномиальный алгоритм для решения NP-трудной в сильном смысле задачи разбиения последовательности на

два кластера, которая индуцируется, в частности, оптимизационной моделью одной из актуальных проблем анализа данных.

Рассмотренная задача относится к числу практически неизученных в алгоритмическом плане. Поэтому исследование вопросов её аппроксимируемости, а также обоснование алгоритмов другого типа (асимптотически точных, рандомизированных и др.) для её решения представляется делом ближайшей перспективы.

Важным направлением исследований является поиск подклассов этой задачи, для которых возможно построение полностью полиномиальной приближённой схемы, а также точного полиномиального и псевдополиномиального алгоритмов.

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.
   Kel'manov A. V., Pyatkin A. V. Complexity of certain problems of searching for subsets of vectors and cluster analysis // Comput. Math. Math. Physics. — 2009. — Vol. 49, N 11. — P. 1966–1971.
2. **Кельманов А. В., Пяткин А. В.** О сложности некоторых задач кластерного анализа векторных последовательностей // Дискрет. анализ и исслед. операций. — 2013. — Т. 20, № 2. — С. 47–57.
   Kel'manov A. V., Pyatkin A. V. On complexity of some problems of cluster analysis of vector sequences // J. Appl. Industr. Math. — 2013. — Vol. 7, N 3. — P. 363–369.
3. **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.
4. **Кельманов А. В., Хамидуллин С. А.** Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Журн. вычисл. математики и мат. физики. — 2001. — Т. 41, № 5. — С. 807–820.
   Kel'manov A. V., Khamidullin S. A. Posterior detection of a given number of identical subsequences in a quasi-periodic sequence // Comput. Math. Math. Physics. — 2001. — Vol. 41, N 5. — P. 762–774.
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 Recognit. 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. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Trans. Signal Process. — 2004. — Vol. 52, N 3. — P. 645–656.

Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru
Хамидуллин Сергей Асгадуллович,
e-mail: kham@math.nsc.ru

Статья поступила
1 марта 2013 г.

Переработанный вариант —
13 мая 2013 г.