ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 22 № 3 2016

УДК 519.16 + 519.85

ПРИБЛИЖЕННЫЙ АЛГОРИТМ ДЛЯ ЗАДАЧИ РАЗБИЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ НА КЛАСТЕРЫ С ОГРАНИЧЕНИЯМИ НА ИХ МОЩНОСТЬ[1]

А. В. Кельманов, Л. В. Михайлова, С. А. Хамидуллин, В. И. Хандеев

Рассматривается задача разбиения конечной последовательности точек евклидова пространства на заданное число кластеров (подпоследовательностей) по критерию минимума суммы по всем кластерам внутрикластерных сумм квадратов расстояний от элементов кластеров до их центров. Предполагается, что центр одного из искомых кластеров задан в начале координат, а центр каждого из остальных кластеров неизвестен и определяется как среднее значение по всем элементам, образующим этот кластер. При этом разбиение подчиненно структурным ограничениям на элементы последовательности, входящие в кластеры с неизвестными центрами: (1) конкатенация номеров элементов этих кластеров является возрастающей последовательностью, (2) разность между последующим и предыдущим номерами ограничена сверху и снизу заданными константами, (3) суммарная мощность кластеров с неизвестными центрами задана на входе. Показано, что задача NP-трудна в сильном смысле. Построен 2-приближенный алгоритм, полиномиальный при фиксированном числе кластеров.

Ключевые слова: разбиение, последовательность, евклидово пространство, минимум суммы квадратов расстояний, NP-трудность, приближенный алгоритм.

A. V. Kel'manov, L. V. Mikhailova, S. A. Khamidullin, V. I. Khandeev. An approximation algorithm for the problem of partitioning a sequence into clusters with constraints on their cardinalities.

We consider the problem of partitioning a finite sequence of points in Euclidean space into a given number of clusters (subsequences) minimizing the sum over all clusters of intracluster sums of squared distances from elements of the clusters to their centers. It is assumed that the center of one of the desired clusters is specified at the origin, while the centers of the other clusters are unknown. Very unknown cluster center is defined as the mean value of cluster elements. Additionally, there are a few structural constraints on the elements of the sequence that enter the clusters with unknown centers: (1) the concatenation of indices of elements of these clusters is an increasing sequence, (2) the difference between two consequent indices is bounded from below and above by prescribed constants, and (3) the total number of elements in these clusters is given as an input. It is shown that the problem is strongly NP-hard. A 2-approximation algorithm that is polynomial for a fixed number of clusters is proposed for this problem.

Keywords: partitioning, sequence, Euclidean space, minimum sum of squared distances, NP-hardness, approximation algorithm.

MSC: 68W25, 68Q25

DOI: 10.21538/0134-4889-2016-22-3-144-152

Введение

Предметом исследования является одна из задач разбиения конечной последовательности точек евклидова пространства на подпоследовательности. Цели исследования — выяснение сложностного статуса задачи и построение приближенного эффективного алгоритма с гарантированной оценкой точности.

Исследование мотивировано слабой изученностью задачи и ее актуальностью, в частности, для математических проблем аппроксимации, кластеризации и анализа последовательностей (временных рядов), а также для многих естественно-научных и технических приложений, в которых требуется классификация упорядоченных по времени данных численных экспериментов или результатов наблюдения за состояниями каких-либо материальных объектов (см.,

[1]Работа выполнена при финансовой поддержке Российского научного фонда (проект 16-11-10041).

например, [1–4] и цитированные там работы). Примеры приложений (истоков) исследуемой задачи приведены в разд. 1.

Настоящая работа является развитием результатов, полученных ранее в [5–7]. Каждая из цитируемых работ послужила необходимым элементом для предложенного ниже, первого на сегодняшний день, алгоритма решения задачи с гарантированной оценкой точности.

1. Формулировка задачи, ее истоки и сложность

Всюду далее $\mathbb{R}$ — множество вещественных чисел, $\|\cdot\|$ — евклидова норма в пространстве $\mathbb{R}^q$, $\langle\cdot,\cdot\rangle$ — скалярное произведение.

Рассматриваемая задача имеет следующую формулировку.

З а д а ч а 1. Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$, L и M. Найти непустые непересекающиеся подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_L$ множества $\mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ такие, что

$$F(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min, \tag{1.1}$$

где $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = (1/|\mathcal{M}_l|) \sum\limits_{j \in \mathcal{M}_l} y_j$ — центроид (геометрический центр) подмножества $\{y_j \mid j \in \mathcal{M}_l\}$ при ограничениях: (1) мощность объединенного множества $\mathcal{M}$ равна M, (2) в последовательности, образованной конкатенацией множеств $\mathcal{M}_1, \ldots, \mathcal{M}_L$, номера упорядочены по возрастанию при условии, что элементы каждого множества образуют возрастающую последовательность, (3) номера из объединенного набора $\mathcal{M} = \{n_1, \ldots, n_M\}$ связаны неравенствами

$$T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N, \qquad m = 2, \ldots, M. \tag{1.2}$$

Из приведенной формулировки видно, что задача 1 относится к классу задач кластеризации с квадратичным критерием. Кластерами являются искомые подмножества $\mathcal{M}_1$, …, $\mathcal{M}_L$, $\mathcal{N} \setminus \mathcal{M}$ номеров и соответствующие им подпоследовательности входной последовательности.

Одним из источников задачи 1 является следующая, общая для многих естественно-научных и технических приложений, содержательная проблема, характерная, в частности, для помехоустойчивого дистанционного мониторинга объектов, электронной разведки, анализа и распознавания биомедицинских и речевых сигналов и др.

Дана последовательность $\mathcal{Y}$, содержащая N упорядоченных по времени результатов $y_1, \ldots, y_N$ измерения набора y из q числовых характеристик некоторого объекта, который может находиться в $L+1$ состояниях. Среди этих состояний L активных и одно пассивное. В пассивном состоянии все элементы набора равны нулю, а в каждом из активных — хотя бы одна из компонент набора не равна нулю. Измерения сопровождаются инструментальной ошибкой. Известно, что объект некоторое время находится в одном из активных состояний, а затем переключается в другое активное состояние. При этом все активные состояния объекта сопровождаются переключениями в пассивное состояние на некоторое ограниченное сверху и снизу неизвестное время. Кроме того, известны (заданы) натуральные числа $T_{\min}$ и $T_{\max}$, которые соответствуют минимальному и максимальному интервалам времени между любыми двумя последовательными активными состояниями объекта. Соответствие элемента последовательности какому-либо состоянию объекта неизвестно. Требуется найти в последовательности все элементы, соответствующие активным состояниям объекта, и оценить характеристики объекта в каждом из активных состояний.

Формализация этой содержательной проблемы с использованием критерия минимума суммы квадратов уклонений индуцирует следующую задачу аппроксимации. Даны последовательность $y_1, \ldots, y_N$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$, L и M. Требуется найти

аппроксимирующую последовательность $z_1, \ldots, z_N$ вида

$$z_n = \begin{cases} x_1, & n \in \mathcal{M}_1, \\ \ldots & \ldots\ldots\ldots \\ x_L, & n \in \mathcal{M}_L, \\ 0, & n \in \mathcal{N} \backslash \mathcal{M}, \end{cases} \tag{1.3}$$

где $x_1, \ldots, x_L$ — произвольные неизвестные точки из $\mathbb{R}^q$, такую, что

$$\sum_{i \in \mathcal{N}} \|y_i - z_i\|^2 \longrightarrow \min \tag{1.4}$$

при тех же, что и в задаче 1, ограничениях на номера из подмножеств $\mathcal{M}_l$, $l = 1, \ldots, L$, и объединенного множества $\mathcal{M}$.

Схематически участок последовательности z_n, $n \in \mathcal{N}$, можно представить в виде

$$\ldots 0 x_{l-1} 0 \ldots 0 x_{l-1} 0 \ldots\ldots 0 x_l 0 \ldots 0 x_l 0 \ldots . \tag{1.5}$$

Здесь $x_{l-1},\ x_l \in \mathbb{R}^q$ — неизвестные ненулевые точки (соответствующие $(l-1)$-у и l-у активным состояниям объекта), 0 — начало координат (соответствующее пассивному состоянию), а число нулей между ненулевыми точками неизвестно и лежит в допустимом интервале от $T_{\min} - 1$ до $T_{\max} - 1$ в соответствии с ограничениями (1.2).

Раскрыв сумму (1.4) с учетом (1.3) и сгруппировав члены, легко проверить с помощью дифференцирования, что оптимальными в смысле (1.4) являются значения $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$, а сформулированная задача аппроксимации индуцирует задачу 1. При этом в найденной оптимальной аппроксимирующей последовательности участок, соответствующий (1.5), как видно из формулировки задачи 1, имеет вид:

$$\ldots 0 \overline{y}(\mathcal{M}_{l-1}) 0 \ldots 0 \overline{y}(\mathcal{M}_{l-1}) 0 \ldots\ldots 0 \overline{y}(\mathcal{M}_l) 0 \ldots 0 \overline{y}(\mathcal{M}_l) 0 \ldots .$$

В этой последовательности для всех $l = 1, \ldots, L$ номера из набора $\mathcal{M}_l$, кластер $\{y_j \mid j \in \mathcal{M}_l\}$ и его центроид $\overline{y}(\mathcal{M}_l)$ определяются в результате решения задачи 1. Центроид $\overline{y}(\mathcal{M}_l)$ является оценкой для точки x_l.

Из приведенной выше схематичной строковой записи последовательностей видно, что их можно интерпретировать как последовательности, содержащие участки с серийными квазипериодическими (в силу ограничений (1.2)) повторами. Если условиться о границах серий, например, по первому (или по последнему) повтору, то все рассмотренные выше задачи можно трактовать как задачи разбиения последовательности на серийные участки с квазипериодическими повторами неизвестных точек совместно с оцениванием точек и отысканием их положения в последовательности.

Сложностной статус задачи 1 устанавливает следующее

Утверждение. *Задача* 1 *NP-трудна в сильном смысле.*

Доказательство утверждения следует из того, что частный случай задачи 1, в котором $L = 1$, является [5] NP-трудной в сильном смысле задачей.

Из этого утверждения следует, что сформулированная выше характерная для многих приложений содержательная задача, а также задача аппроксимации относится к числу труднорешаемых задач.

2. Известные и полученные результаты

Задача 1 относится к числу слабоизученных проблем дискретной оптимизации. Близкой в постановочном плане является задача (см. [7]), в которой входная последовательность $\mathcal{Y}$ одномерна, т. е. $q = 1$. Точки из набора $(x_1, \ldots, x_L)$ заданы на входе и принадлежат $\mathbb{R}^d$, где

$d \geq 1$, причем $T_{\min} \geq d$ в ограничениях (1.2). В целевой функции этой задачи вместо центроидов $\overline{y}(\mathcal{M}_1), \ldots, \overline{y}(\mathcal{M}_L)$ искомых подмножеств в формуле (1.1) фигурируют элементы заданного набора $(x_1, \ldots, x_L)$. Искомыми переменными являются множества $\mathcal{M}_1, \ldots, \mathcal{M}_L$. Эту задачу можно трактовать как задачу поиска в последовательности серийных участков с квазипериодическими повторами точек из заданного набора (шаблона) совместно с отысканием положения этих точек в последовательности. В [7] показано, что эта задача разрешима за полиномиальное время с помощью построенной схемы динамического программирования. Ниже мы применяем упрощенную модификацию этой схемы для построения предлагаемого алгоритма.

В настоящее время для задачи 1, за исключением ее частного случая, когда $L = 1$ в формуле (1.1), а также двух подслучаев этого случая отсутствуют какие-либо эффективные алгоритмы с оценками точности. Для указанного частного случая задачи получены следующие результаты.

В [5] анализировался вариант задачи, в котором $T_{\min}$ и $T_{\max}$ — параметры. В цитируемой работе установлено, что в случае $L = 1$ параметрический вариант задачи 1 NP-труден в сильном смысле для любых $T_{\min} < T_{\max}$. В тривиальном случае, когда $T_{\min} = T_{\max}$, задача разрешима за полиномиальное время.

В [6] для этого же случая (когда $L = 1$) задачи 1 предложен 2-приближенный полиномиальный алгоритм, временная сложность которого оценивается величиной $\mathcal{O}(N^2(MN + q))$.

Кроме того, в [8] и [9] исследованы два подслучая этого же случая задачи, в которых размерность q пространства фиксирована. Для подслучая с целочисленными входами задачи в [8] предложен точный псевдополиномиальный алгоритм, трудоемкость которого есть величина $\mathcal{O}(MN^2(MD)^q)$, где D — максимальное абсолютное значение координат входных точек. Для подслучая с вещественными входами в [9] обоснована полностью полиномиальная приближенная схема. Предложенный в этой работе алгоритм при заданной относительной погрешности ε позволяет находить $(1 + \varepsilon)$-приближенное решение задачи 1 за время $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$.

В настоящей работе для задачи 1 предложен алгоритм, позволяющий находить 2-приближенное решение за время $\mathcal{O}(LN^{L+1}(MN + q))$, полиномиальное при фиксированном числе L кластеров.

3. Основы алгоритма

Для построения алгоритма нам потребуются несколько базовых утверждений, вспомогательная задача и точный полиномиальный алгоритм ее решения.

Геометрической базой алгоритма являются следующие утверждения.

Лемма 1. *Для произвольной точки $u \in \mathbb{R}^q$ и конечного непустого множества $\mathcal{Z} \subset \mathbb{R}^q$ имеет место равенство*

$$\sum_{z \in \mathcal{Z}} \|z - u\|^2 = \sum_{z \in \mathcal{Z}} \|z - \overline{z}\|^2 + |\mathcal{Z}| \, \|u - \overline{z}\|^2, \tag{3.1}$$

где $\overline{z} = 1/|\mathcal{Z}| \sum\limits_{z \in \mathcal{Z}} z$ — центроид множества $\mathcal{Z}$.

Эта лемма доказывается весьма просто и относится к общеизвестным результатам. Ее доказательство представлено во множестве публикаций, в частности в [10].

Лемма 2. *Пусть выполнены условия леммы* 1. *Тогда если некоторая точка $u \in \mathbb{R}^q$ лежит ближе (в смысле расстояния) к центроиду $\overline{z}$ множества $\mathcal{Z}$, чем все точки этого множества, то справедливо неравенство $\sum\limits_{z \in \mathcal{Z}} \|z - u\|^2 \leq 2 \sum\limits_{z \in \mathcal{Z}} \|z - \overline{z}\|^2$.*

Справедливость леммы следует из (3.1), так как по условию леммы $\|u - \overline{z}\| \le \|z - \overline{z}\|$ для всех $z \in \mathcal{Z}$.

Всюду далее будем использовать обозначение $f^x(y)$ для функции $f(x, y)$ при фиксированном аргументе x.

Лемма 3. *Пусть*

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - x_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2, \tag{3.2}$$

$$G(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l \rangle - \|x_l\|^2),$$

где $x_1, \ldots, x_L$ — точки из $\mathbb{R}^q$, а элементы множеств $\mathcal{M}_l, \ldots, \mathcal{M}_L$ и $\mathcal{M}$ удовлетворяют ограничениям задачи 1. *Тогда для условных оптимумов функции* (3.2) *справедливы следующие утверждения*: 1) *для любых непустых фиксированных подмножеств $\mathcal{M}_1, \ldots, \mathcal{M}_L$ минимум функции* (3.2) *по переменным $x_1, \ldots, x_L$ достигается в точках $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$, и равен $F(\mathcal{M}_1, \ldots, \mathcal{M}_L)$*; 2) *для любого набора $x = (x_1, \ldots, x_L)$ фиксированных точек из $\mathbb{R}^q$ минимум функции $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ по подмножествам $\mathcal{M}_1, \ldots, \mathcal{M}_L$ достигается на подмножествах $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ номеров элементов $\{y_i \mid i \in \bigcup_{l=1}^{L} \mathcal{M}_l^x\}$ последовательности $\mathcal{Y}$, для которых максимальна функция $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$.*

Доказательство. Первое утверждение леммы легко проверяется дифференцированием и следует также из леммы 1. Для доказательства второго утверждения достаточно заметить, что справедливо следующее легко проверяемое равенство

$$S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{j \in \mathcal{N}} \|y_j\|^2 - G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L), \tag{3.3}$$

в правой части которого сумма не зависит от $\mathcal{M}_1, \ldots, \mathcal{M}_L$.

Лемма доказана.

Вычислительной базой предлагаемого алгоритма является точный полиномиальный алгоритм решения следующей вспомогательной задачи.

Задача 2. Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ и набор $x = (x_1, \ldots, x_L)$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и M. Найти непустые непересекающиеся подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_L$ множества $\mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такие, что целевая функция $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ максимальна при тех же, что и в задаче 1, ограничениях на искомые переменные.

Для изложения алгоритма решения вспомогательной задачи определим функцию

$$g_l^x(n) = 2\langle y_n, x_l \rangle - \|x_l\|^2, \qquad n \in \mathcal{N}, \qquad l = 1, \ldots, L, \tag{3.4}$$

где x_l — точка из набора x, y_n — элемент последовательности $\mathcal{Y}$. В соответствии с этим определением для целевой функции $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ имеем $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{n \in \mathcal{M}_l} g_l^x(n)$. Кроме того, заметим, что утверждение 2) леммы 3 означает равенства

$$(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x) = \arg \min_{\mathcal{M}_1, \ldots, \mathcal{M}_L} S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \arg \max_{\mathcal{M}_1, \ldots, \mathcal{M}_L} G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L). \tag{3.5}$$

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ задачи 2 и (согласно приведенным выше равенствам (3.5)) оптимального решения задачи минимизации функции $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$. Схема следует из результатов работы [7].

Лемма 4. *Пусть выполнены условия задачи* 2. *Тогда для любых натуральных* L *и* M *таких, что* $(M-1)T_{\min} < N$ *и* $L \le M$, *оптимальное значение* $G^x_{\max}$ *целевой функции этой задачи находится по формуле*

$$G^x_{\max} = \max_{n\in\{1+(M-1)T_{\min},\dots,N\}} G^x_{L,M}(n), \tag{3.6}$$

а значения функции $G^x_{L,M}(n)$ *вычисляются по следующим рекуррентным формулам*

$$G^x_{l,m}(n) = g^x_l(n) + \begin{cases} 0, \quad \textit{если} \quad l=1, \quad m=1, \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{1,m-1}(j), \quad \textit{если} \quad l=1, \quad m=2,\dots,M-(L-1), \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), \quad \textit{если} \quad l=2,\dots,L, \quad m=l, \\ \max\{\max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j), \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j)\}, \\ \qquad \textit{если} \quad l=2,\dots,L, \quad m=l+1,\dots,M-(L-l), \end{cases} \tag{3.7}$$

где

$$\gamma_{m-1}(n) = \bigl\{ j \mid \max\{1+(m-2)T_{\min}, n-T_{\max}\} \le j \le n-T_{\min} \bigr\}, \qquad m=2,\dots,M \tag{3.8}$$

при каждом $n = 1+(m-1)T_{\min},\dots,N-(M-m)T_{\min}$.

Следствие. *Пусть выполнены условия леммы* 4. *Пусть, кроме того,*

$$r^x_{l,m}(n) = \begin{cases} 1, \quad \textit{если} \quad l=1, \quad m=2,\dots,M-(L-1), \\ l-1, \quad \textit{если} \quad l=2,\dots,L, \quad m=l, \\ l-1, \quad \textit{если} \quad \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j) < \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), \\ \qquad l=2,\dots,L, \quad m=l+1,\dots,M-(L-l), \\ l, \quad \textit{если} \quad \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j) \ge \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,m-1}(j), \\ \qquad l=2,\dots,L, \quad m=l+1,\dots,M-(L-l); \end{cases}$$

$$I^x_{l,m}(n) = \arg\max_{j\in\gamma_{m-1}(n)} G^x_{l,m-1}(j), \qquad l=1,\dots,L, \qquad m=l+1,\dots,M-(L-l)$$

при каждом $n = 1+(m-1)T_{\min},\dots,N-(M-m)T_{\min}$;

$$n^x(m) = \begin{cases} \arg\max\limits_{n\in\{1+(M-1)T_{\min},\dots,N\}} G^x_{L,M}(n), & \textit{если} \quad m=M, \\ I^x_{k^x(m),m+1}(n^x(m+1)), & \textit{если} \quad m=M-1,\dots,1; \end{cases}$$

$$k^x(m) = \begin{cases} L, & \textit{если} \quad m=M, \\ r^x_{k^x(m+1),m+1}(n^x(m+1)), & \textit{если} \quad m=M-1,\dots,1; \end{cases}$$

$$J^x(l) = \begin{cases} 0, & \textit{если} \quad l=0, \\ \bigl|\{m\in\{1,\dots,M\} \mid k^x(m)\le l\}\bigr|, & \textit{если} \quad l=1,\dots,L. \end{cases}$$

Тогда множества $\mathcal{M}^x_1,\dots,\mathcal{M}^x_L$ *определяются по правилу*

$$\mathcal{M}^x_l = \{n \mid n = n^x(m),\ m = J^x(l-1)+1,\dots,J^x(l)\} \tag{3.9}$$

при каждом $l=1,\dots,L$.

Запишем алгоритм, реализующий приведенную схему, в пошаговом виде.

А л г о р и т м $\mathcal{A}_1$.

Вход алгоритма: последовательность $\mathcal{Y}$, набор $(x_1, \ldots, x_L)$ точек, числа $T_{\min}$, $T_{\max}$ и M.

Ш а г 1. Вычислим значения $g_l^x(n)$ для $l = 1, \ldots, L$, $n = 1 + (l-1)T_{\min}, \ldots, N - (L-l)T_{\min}$ по формуле (3.4).

Ш а г 2. Используя рекуррентные формулы (3.7) и (3.8), вычислим значения $G_{l,m}^x(n)$ для каждого $l = 1, \ldots, L$, $m = l, \ldots, M - (L-l)$, $n = 1 + (m-1)T_{\min}, \ldots, N - (M-m)T_{\min}$.

Ш а г 3. Найдем значение $G_{\max}^x$ максимума целевой функции G^x по формуле (3.6) и оптимальные подмножества $\mathcal{M}_l^x$ по формуле (3.9).

Выход алгоритма: набор подмножеств $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$.

З а м е ч а н и е 1. Перед началом работы алгоритма требуется проверка справедливости двух условий леммы 4. Эти необходимые условия обеспечивают совместность ограничений в задачах 1 и 2, а также корректность входных данных алгоритма.

З а м е ч а н и е 2. В [7] установлено, что алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи 2 за время $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$. В этом выражении значение $T_{\max} - T_{\min} + 1$ не превосходит N. Поэтому время работы алгоритма оценивается величиной $\mathcal{O}(LN(MN + q))$.

4. Приближенный алгоритм

Суть подхода к поиску решения заключается в следующем. Для каждого упорядоченного набора, содержащего L элементов последовательности $\mathcal{Y}$, находим точное решение вспомогательной задачи 2 — набор подмножеств номеров элементов последовательности, который является допустимым решением исходной задачи 1. Найденный набор подмножеств объявляется претендентом на решение исходной задачи и включается в семейство допустимых решений. В качестве окончательного решения из построенного семейства выбирается набор подмножеств, доставляющий наибольшее значение целевой функции задачи 2.

Сформулируем алгоритм решения задачи 1, реализующий описанный подход. В приведенной ниже пошаговой записи предполагается, что входные натуральные числа заданы в соответствии с ограничениями задачи и условиями леммы 4 (см. замечание 1).

А л г о р и т м $\mathcal{A}$.

Вход алгоритма: последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$, M и L.

Ш а г 1. Для каждого набора $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$, сформированного из элементов последовательности $\mathcal{Y}$, с помощью алгоритма $\mathcal{A}_1$ найдем оптимальное решение $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ задачи 2.

Ш а г 2. Найдем наборы $x(A) = \arg\max\limits_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ и $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$. Если оптимальных решений несколько, то выберем любое из них.

Выход алгоритма: набор $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$.

Лемма 5. *Пусть $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ — оптимальное решение задачи 1, а $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ — решение, полученное в результате работы алгоритма $\mathcal{A}$. Тогда*

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*).$$

Д о к а з а т е л ь с т в о. Оптимальному решению $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ задачи 1 соответствует набор $\{\overline{y}(\mathcal{M}_1^*), \ldots, \overline{y}(\mathcal{M}_L^*)\}$ центроидов. Рассмотрим точку $t_l = \arg\min\limits_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|$, $l = 1, \ldots, L$, из подмножества $\mathcal{M}_l^*$, ближайшую к центроиду этого подмножества. Эта точка и само подмножество $\mathcal{M}_l^*$ удовлетворяют условиям леммы 2. Поэтому, применяя неравенство этой

леммы к каждому из подмножеств $\mathcal{M}_l^*$, $l = 1, \ldots, L$, вычислим оценку для величины

$$S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) = \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - t_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2$$

$$\leq 2 \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2$$

$$\leq 2 \sum_{l=1}^{L} \sum_{y \in \mathcal{M}_l^*} \|y - \overline{y}(\mathcal{M}_l^*)\|^2 + 2 \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 = 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*), \tag{4.1}$$

где $\mathcal{M}^* = \bigcup_{l=1}^{L} \mathcal{M}_l^*$.

С другой стороны, заметим, что набор $t = (t_1, \ldots, t_L)$ точек, ближайших к центроидам $\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*$, является одним из наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$, рассмотренных на шаге 1 алгоритма $\mathcal{A}$. Пусть $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ — оптимальное решение задачи 2 при $x = t$, полученное на шаге 1 алгоритма $\mathcal{A}$. Тогда в соответствии с утверждением 2) леммы 3, т. е. согласно (3.5), набор $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ доставляет минимум функции $S^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ при $x = t$. Поэтому

$$S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \leq S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L). \tag{4.2}$$

Далее, по определению шага 2 в соответствии с (3.3) справедлива оценка

$$S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \leq S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L), \tag{4.3}$$

где $(x_1^A, \ldots, x_L^A) = x(A)$. Кроме того, из первого утверждения леммы 3 следует неравенство

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A). \tag{4.4}$$

Наконец, объединяя (4.1)–(4.4), получим цепочку оценочных неравенств

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq S(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \leq S(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L)$$
$$\leq S(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) \leq 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*),$$

которая завершает доказательство леммы.

Свойства изложенного алгоритмического решения устанавливает

Теорема. *Алгоритм $\mathcal{A}$ находит 2-приближенное решение задачи* 1 *за время*

$$\mathcal{O}(LN^{L+1}(M(T_{\max} - T_{\min} + 1) + q)).$$

Оценка 2 *точности алгоритма достижима.*

Доказательство. Оценка 2 точности алгоритма следует из леммы 5. Оценка трудоемкости алгоритма следует из того, что на шаге 1 для каждого из N^L наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$ алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи 2 за время $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$, а на шаге 2 выбор наименьшего элемента осуществляется за $\mathcal{O}(N^L)$ операций. Достижимость оценки точности алгоритма $\mathcal{A}$ следует из достижимости оценки точности 2-приближенного алгоритма для частного случая (когда $L = 1$) задачи 1 (см. [6]).

Теорема доказана.

Замечание 3. Согласно замечанию 2 время работы алгоритма $\mathcal{A}$ оценивается величиной $\mathcal{O}(LN^{L+1}(MN + q))$, полиномиальной при фиксированном числе L кластеров.

Заключение

В работе показано, что к числу NP-трудных в сильном смысле проблем относится одна из актуальных задач разбиения конечной последовательности точек евклидова пространства

на заданное число кластеров при ограничениях на их мощность. Для этой задачи предложен алгоритм, который позволяет находить 2-приближенное решение задачи за полиномиальное время при фиксированном числе кластеров. На наш взгляд, представленный в работе алгоритм решения задачи будет полезен как одно из инструментальных средств решения проблем в области приложений, связанных с анализом и распознаванием временных рядов (сигналов).

Значительный интерес представляет обоснование более быстрых приближенных алгоритмов с гарантированными оценками точности, а также поиск подклассов рассмотренной задачи, для которых возможно построение таких алгоритмов.

СПИСОК ЛИТЕРАТУРЫ

1. **Tak-chung Fu.** A review on time series data mining // Engineering Applications of Artificial Intelligence. 2011. Vol. 24, no. 1. P. 164–181.
2. Remote sensing time series: Revealing Land Surface Dynamics / eds. C. Kuenzer, S. Dech, W. Wagner. New York etc.: Springer International Publishing, 2015. 441 p. (Remote Sensing and Digital Image Processing; vol. 22.)
3. **T. Warren Liao.** Clustering of time series data — a survey // Pattern Recognition. 2005. Vol. 38, no. 11. P. 1857–1874.
4. **Aggarwal C. C.** Data mining: The textbook. New York etc.: Springer International Publishing, 2015. 734 p.
5. **Кельманов А. В., Пяткин А. В.** О сложности некоторых задач кластерного анализа векторных последовательностей // Дискретный анализ и исследование операций. 2013. Т. 20, № 2. С. 47–57.
6. **Кельманов А. В., Хамидуллин С. А.** Приближенный алгоритм для одной задачи разбиения последовательности // Дискретный анализ и исследование операций. 2014. Т. 21, № 1. С. 53–66.
7. **Кельманов А. В., Михайлова Л. В.** Совместное обнаружение в квазипериодической последовательности заданного числа фрагментов из эталонного набора и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. математики и мат. физики. 2006. Т. 46, № 1. С. 172–189.
8. **Кельманов А. В., Хамидуллин С. А., Хандеев В. И.** Точный псевдополиномиальный алгоритм для одной задачи бикластеризации последовательности // XV Всеросс. конф. “Математическое программирование и приложения”: тез. докл. / ИММ УрО РАН. Екатеринбург, 2015. С. 139.
9. **Кельманов А. В., Хамидуллин С. А., Хандеев В. И.** Полностью полиномиальная аппроксимационная схема для одной задачи двухкластерного разбиения последовательности // Дискретный анализ и исследование операций. 2016. Т. 23, № 2. С. 21–40.
10. **Кельманов А. В., Романченко С. М.** FPTAS для одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. 2014. Т. 21, № 3. С. 41–52.

Кельманов Александр Васильевич
д-р физ.-мат. наук, зав. лабораторией
Институт математики им. С. Л. Соболева СО РАН
Новосибирский государственный университет
e-mail: kelm@math.nsc.ru

Михайлова Людмила Викторовна
канд. физ.-мат. наук, старший науч. сотрудник
Институт математики им. С. Л. Соболева СО РАН
e-mail: mikh@math.nsc.ru

Хамидуллин Сергей Асгадуллович
канд. тех. наук, старший науч. сотрудник
Институт математики им. С. Л. Соболева СО РАН
e-mail: kham@math.nsc.ru

Хандеев Владимир Ильич
аспирант
Институт математики им. С. Л. Соболева СО РАН
e-mail: khandeev@math.nsc.ru

Поступила 30.05.2016