

УДК 519.2+621.391

ПРИБЛИЖЁННЫЕ АЛГОРИТМЫ ДЛЯ НЕКОТОРЫХ ТРУДНОРЕШАЕМЫХ ЗАДАЧ ПОИСКА ПОДПОСЛЕДОВАТЕЛЬНОСТИ ВЕКТОРОВ *)

А. В. Кельманов, С. М. Романченко, С. А. Хамидуллин

Аннотация. Рассматриваются некоторые труднорешаемые задачи поиска подпоследовательности в последовательности векторов евклидова пространства, состоящей из конечного числа членов. Предполагается, что искомая подпоследовательность содержит фиксированное число векторов, близких между собой по критерию минимума суммы квадратов расстояний, причём поиск векторов подчинён условию: разность между номерами последующего и предыдущего искомым векторов ограничена сверху и снизу некоторыми константами. Предложены 2-приближённые эффективные алгоритмы решения этих задач.

Ключевые слова: поиск подпоследовательности векторов, минимум суммы квадратов расстояний, кластерный анализ, NP-трудность, эффективный приближённый алгоритм.

Введение

Предметом исследования являются труднорешаемые задачи, к которым сводится одна из актуальных проблем помехоустойчивого анализа данных. Цель исследования — обоснование приближённых эффективных алгоритмов решения этих задач.

Содержательная проблема анализа данных, которая приводит к решению указанных задач, состоит в следующем. Имеется таблица, содержащая упорядоченные по времени результаты измерения набора числовых информационно значимых характеристик для совокупности некоторых материальных объектов. Часть объектов в этой совокупности иден-

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 12-01-00090, 10-07-00195, 11-07-12083-офи-м), федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг. (гос. контракт 14.740.11.0362), а также целевой программы СО РАН (интеграционные проекты 7Б и 21А).

тичны и имеют одинаковые характеристики, число идентичных объектов известно. Остальные объекты различны и имеют отличающиеся характеристики. Известно, что временной интервал между двумя последовательными результатами измерения характеристик идентичных объектов ограничен сверху и снизу некоторыми константами. В каждом результате измерения, представленном в таблице, имеется ошибка, причём соответствие между объектом и набором неизвестно. Характеристики идентичных объектов в отличие от характеристик остальных объектов имеют принципиальную информационную ценность. Требуется, используя критерий минимума суммы квадратов расстояний, найти в таблице совокупность (подпоследовательность) наборов, соответствующих идентичным объектам, и оценить по результатам измерения набор характеристик этих объектов (учитывая, что данные содержат ошибку измерения).

Эту проблему можно трактовать как разновидность так называемой проблемы «обучения компьютера» (Machine learning) распознаванию образов (см., например, [8, 9]). Подобные этой содержательные проблемы с временными ограничениями на результаты измерения каких-либо информационно ценных характеристик весьма актуальны, в частности, при помехоустойчивой off-line обработке и распознавании числовых и векторных последовательностей (см., например, [1–4, 10, 11] и цитированные там работы), которые в приложениях трактуются как дискретные одномерные или многомерные сигналы. В этих проблемах наличие временных ограничений обусловлено имеющимися априорными данными о времени возможного появления принципиально значимой информации в обрабатываемом сигнале, последовательности, таблице и т. п.

Специальный случай сформулированной проблемы, когда временные ограничения отсутствуют, рассматривался в [5, 6]. В [5] показано, что индуцируемые этим случаем полиномиально эквивалентные оптимизационные задачи NP-трудны. В [6] предложены 2-приближённые эффективные алгоритмы решения этих задач. Полиномиальная приближённая схема для одной из таких задач предложена в [7]. Эту схему можно применять и для решения остальных задач в силу их полиномиальной эквивалентности.

Мотивацией данного исследования послужил тот факт, что до настоящего времени отсутствовали какие-либо эффективные алгоритмы с гарантированными оценками точности для решения экстремальных задач, к которым сводится общий случай сформулированной проблемы, когда имеются упомянутые выше временные ограничения. Результаты данной работы являются дополнением к результатам из [5, 6].

1. Модель анализа данных

Рассмотрим следующую структуру данных, представленных в виде совокупности векторов евклидова пространства.

Пусть векторная последовательность $x_n \in \mathbb{R}^q$, $n \in \mathcal{N} = \{1, \dots, N\}$, обладает свойством

$$x_n = \begin{cases} w, & n \in \mathcal{M}, \\ v_n, & n \in \mathcal{N} \setminus \mathcal{M}, \end{cases} \quad (1)$$

где $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$.

Допустим, что для обработки доступна последовательность

$$y_n = x_n + e_n, \quad n \in \mathcal{N}, \quad (2)$$

где e_n — вектор помехи (ошибки измерения), независимый от вектора x_n . Учитывая зависимость элементов последовательности (1) от множеств и векторов, положим

$$S(\mathcal{M}, w, \{v_i, i \in \mathcal{N} \setminus \mathcal{M}\}) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \quad (3)$$

и рассмотрим модель анализа данных в виде следующей экстремальной задачи.

Задача 1. ДАНО: последовательность y_n , $n \in \mathcal{N}$, векторов из \mathbb{R}^q и натуральные числа T_{\min} , T_{\max} и $M > 1$. НАЙТИ: непустое подмножество (набор) $\mathcal{M} \subseteq \mathcal{N}$ номеров элементов последовательности, вектор w и совокупность $\{v_i, i \in \mathcal{N} \setminus \mathcal{M}\}$ векторов, минимизирующих $S(\cdot)$, при условии, что структура последовательности описывается формулами (1) и (2), при следующих ограничениях на элементы набора \mathcal{M} :

$$1 \leq T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N - 1, \quad m = 2, \dots, M. \quad (4)$$

Совокупность элементов последовательности y_n , $n \in \mathcal{N}$, ассоциируется с набором наблюдаемых или имеющихся в распоряжении данных, в которых скрыта ненаблюдаемая последовательность x_n , $n \in \mathcal{N}$, включающая вектор w , повторяющийся M раз. Вектор w интерпретируется как информационно значимый набор характеристик идентичных объектов. Остальные элементы v_n , $n \in \mathcal{N} \setminus \mathcal{M}$, последовательности x_n , $n \in \mathcal{N}$, трактуются как возможный «мусор», который, как правило, имеется в совокупности обрабатываемых данных.

Отличие модели (1)–(4) от рассмотренной в [5] состоит лишь в дополнительных ограничениях (4). Если номера членов последовательностей интерпретировать как равномерные дискретные отсчёты времени, то параметры T_{\min} и T_{\max} из модели соответствуют минимальному и максимальному интервалам времени между двумя последовательными повторами неизвестного информационно значимого набора. Эти параметры на практике часто бывают априори известны. Величина $T_{\max} - T_{\min} + 1$ определяет степень аperiodичности повторов. Случай $T_{\min} = T_{\max}$ соответствует ситуации, когда повторы периодичны. В случае $T_{\min} < T_{\max}$ подслучай $T_{\min} = 1$ и $T_{\max} = N - 1$ соответствует другой ситуации, когда повторы в наибольшей степени аperiodичны.

Задача 1 по своей сути является задачей приближения последовательности (2) последовательностью (1) по критерию минимума суммы квадратов отклонений. Легко установить, что если e_n в формуле (2) есть выборка единичного объёма из q -мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия, то статистический подход к рассматриваемой проблеме анализа данных приводит к задаче минимизации функционала (3). Статистические аспекты проблемы находятся вне рамок данной работы.

2. Редуцированные задачи

Поскольку содержательная проблема и модель анализа данных практически те же (за исключением дополнительных ограничений (4)), что и в работе [5], сведение этой проблемы и задачи 1 к сформулированным ниже экстремальным задачам осуществляется точно так же, как и в [5]. По этой причине сформулированные ниже задачи по своей сути являются аналогами задач из [5, 6]. В этих задачах предполагается, что входными данными являются не множества, а структуры в виде векторных последовательностей, причём имеются ограничения на номера выбираемых векторов из входной последовательности. Для учёта этих ограничений в приведённых ниже формулировках задач при записи целевых функций вместо суммирования по элементам множеств (см. [5, 6]) используется суммирование по номерам (индексам) элементов последовательности.

Задача VSS1 (Vector Subsequence in a Sequence 1). ДАНО: последовательность (набор) $\mathcal{Y} = (y_1, \dots, y_N)$ векторов из \mathbb{R}^q , натуральные числа T_{\min}, T_{\max} и $M > 1$. НАЙТИ: подмножество $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов набора \mathcal{Y} такое, что целевая функция

$$f_1(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \left\| \sum_{i=1}^M y_{n_i} \right\|^2 + \sum_{i \notin \mathcal{M}} \|y_i\|^2$$

максимальна при ограничениях (4) на элементы \mathcal{M} .

Задача VSS2 (Vector Subsequence in a Sequence 2). ДАНО: последовательность $\mathcal{Y} = (y_1, \dots, y_N)$ векторов из \mathbb{R}^q , натуральные числа T_{\min}, T_{\max} и $M > 1$. НАЙТИ: подмножество $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности \mathcal{Y} такое, что целевая функция

$$f_2(\mathcal{M}) = \sum_{i=1}^M \|y_{n_i} - \bar{y}(\mathcal{M})\|^2, \quad (5)$$

где $\bar{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{n \in \mathcal{M}} y_n$, минимальна при ограничениях (4) на элементы \mathcal{M} .

Задача VSS3 (Vector Subsequence in a Sequence 3). ДАНО: последовательность $\mathcal{Y} = (y_1, \dots, y_N)$ векторов из \mathbb{R}^q , натуральные числа T_{\min}, T_{\max} и $M > 1$. НАЙТИ: подмножество $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности \mathcal{Y} такое, что целевая функция

$$f_3(\mathcal{M}) = \sum_{i=1}^M \sum_{j=1}^M \|y_{n_i} - y_{n_j}\|$$

минимальна при ограничениях (4) на элементы \mathcal{M} .

Задача MSSC-Case-S (Minimum Sum-of-Squares Clustering, special Case for a Sequence). ДАНО: последовательность $\mathcal{Y} = (y_1, \dots, y_N)$ векторов из \mathbb{R}^q , натуральные числа T_{\min}, T_{\max} и $M > 1$. НАЙТИ: разбиение множества \mathcal{N} на $J = N - M + 1$ непустых подмножеств $\mathcal{M}_1, \dots, \mathcal{M}_J$ такое, что $|\mathcal{M}_1| = M$ и целевая функция

$$f_4(\mathcal{M}) = \sum_{j=1}^J \sum_{m \in \mathcal{M}_j} \|y_m - \bar{y}(\mathcal{M}_j)\|^2$$

где $\bar{y}(\mathcal{M}_j) = \frac{1}{|\mathcal{M}_j|} \sum_{n \in \mathcal{M}_j} y_n$, $j = 1, \dots, J$, минимальна при ограничениях (4) на элементы \mathcal{M}_1 .

Во-первых, напомним [5], что вектор $\bar{y}(\mathcal{M})$ в формуле (5) является среднеквадратической оценкой для неизвестного вектора w в формулах (1) и (3).

Во-вторых, отметим, что в случае $T_{\min} = T_{\max}$ все задачи, очевидно, решаются за полиномиальное время. Поэтому далее будем рассматривать случай $T_{\min} < T_{\max}$.

В-третьих, заметим, что если T_{\min} и T_{\max} являются частью входа, то специальные случаи задач VSS1, VSS2, VSS3 и MSSC-Case-S поиска подпоследовательности, когда $T_{\min} = 1$ и $T_{\max} = N - 1$, эквивалентны соответствующим труднорешаемым задачам VS-1, VS-2, VS-3 и MSSC-Case поиска подмножества, изученным в [5, 6]. Поэтому сформулированные выше задачи NP-трудны.

Напомним, что в [6] для решения задач VS-2, VS-3 и MSSC-Case поиска подмножества предложены эффективные 2-приближённые алгоритмы, имеющие временную сложность $\mathcal{O}(qN^2)$. Эти алгоритмы, очевидно, можно применить для отыскания 2-приближённого решения специального случая задач VSS2, VSS3 и MSSC-Case-S поиска подпоследовательности, когда $T_{\min} = 1$ и $T_{\max} = N - 1$. Ниже обоснованы приближённые эффективные алгоритмы для общего случая сформулированных задач, т. е. для произвольных T_{\min} и T_{\max} таких, что $T_{\min} < T_{\max}$.

Наконец, заметим, что целевые функции задач выбора подпоследовательностей связаны формулами

$$f_2(\mathcal{M}) = \sum_{n \in \mathcal{N}} \|y_n\|^2 - f_1(\mathcal{M}) = \frac{1}{2|\mathcal{M}|} f_3(\mathcal{M}), \quad (6)$$

$$f_4(\mathcal{M}_1, \dots, \mathcal{M}_J) = f_2(\mathcal{M}_1), \quad (7)$$

аналогичными формулам, связывающим целевые функции задач VS-1, VS-2, VS-3 и MSSC-Case [5]. Поэтому при помощи (6) и (7) по найденному решению одной из задач легко находятся решения остальных задач и соответствующие значения целевых функций. В качестве базовой рассмотрим задачу VSS2. Построим алгоритм её решения.

Суть подхода к построению алгоритма состоит в замене решения исходной задачи решением более простой вспомогательной задачи и последующей оценкой точности этой замены.

3. Алгоритмы решения задач

Опираясь на результаты из [4], построим алгоритм решения следующей вспомогательной задачи.

Задача SVN (Subsequence of Vectors which are Nearest to given Vector). ДАНО: последовательность $\mathcal{Y} = (y_1, \dots, y_N)$ векторов из \mathbb{R}^q , вектор $b \in \mathbb{R}^q$, натуральные числа T_{\min} , T_{\max} и $M > 1$. НАЙТИ: набор

$\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности \mathcal{Y} такой, что целевая функция $G(\mathcal{M}) = \sum_{m=1}^M \|y_{n_m} - b\|^2$ минимальна при ограничениях (4) на элементы \mathcal{M} .

Положим

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

Следуя [4], определим функцию

$$G_m(n) = \begin{cases} g(n), & \text{если } n \in \omega_1, \quad m = 1, \\ g(n) + \min_{j \in \gamma_{m-1}^-(n)} G_{m-1}(j), & \text{если } n \in \omega_m, \quad m = 2, \dots, M, \end{cases} \quad (9)$$

где $\omega_m = \{n \mid 1 + (m-1)T_{\min} \leq n \leq N - (M-m)T_{\min}\}$, $m = 1, \dots, M$, — область допустимых значений переменной n_m , а

$$\gamma_{m-1}^-(n) = \{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \leq j \leq n - T_{\min}\}, \\ n \in \omega_m, \quad m = 2, \dots, M,$$

— область допустимых значений переменной n_{m-1} при условии, что значение переменной n_m фиксировано и равно n .

По своему смыслу $G_m(n)$ есть условно-оптимальное значение целевой функции G при условии, что из последовательности выбрано m векторов, последний из которых совпадает с y_n .

Лемма 1. Оптимальное значение $G_{\min} = \min_{\mathcal{M}} G(\mathcal{M})$ целевой функции задачи SVNВ находится по формуле

$$G_{\min} = \min_{n \in \omega_M} G_M(n), \quad (10)$$

а значения функции $G_M(n)$, $n \in \omega_M$, вычисляются по формуле (9).

Справедливость этой леммы и приведённого далее утверждения следует из [4].

Следствие 1. Элементы оптимального набора $\widehat{\mathcal{M}} = \arg \min_{\mathcal{M}} G(\mathcal{M})$ находятся по следующим рекуррентным формулам:

$$\widehat{n}_M = \arg \min_{n \in \omega_M} G_M(n), \quad (11)$$

$$\widehat{n}_{m-1} = \arg \min_{n \in \gamma_{m-1}^-(\widehat{n}_m)} G_{m-1}(n), \quad m = M, M-1, \dots, 2. \quad (12)$$

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

АЛГОРИТМ \mathcal{A}_1

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

ШАГ 2. Вычислим значения $G_m(n)$ для каждого $n \in \omega_m$ и $m = 1, \dots, M$, используя рекуррентные формулы (9).

ШАГ 3. Найдём значение G_{\min} минимума целевой функции G и оптимальный набор $\widehat{\mathcal{M}} = \{\widehat{n}_1, \dots, \widehat{n}_M\}$ по формулам (10)–(12).

Замечание 1. Временная сложность алгоритма \mathcal{A}_1 равна

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

Действительно, на первом шаге алгоритма требуется $\mathcal{O}(Nq)$ операций. Основной вклад в трудоёмкость вносит шаг 2. Трудоёмкость этого шага определяется мощностью множеств ω_m и $\gamma_{m-1}^-(n)$, входящих в определение функции (9). Мощность первого из этих множеств не превосходит N , а мощность второго не больше $T_{\max} - T_{\min} + 1$. Вычисления по формуле (9) производятся для каждого $m = 1, \dots, M$. Поэтому трудоёмкость шага 2 есть величина $\mathcal{O}(NM(T_{\max} - T_{\min} + 1))$. Из формул (10)–(12) видно, что на шаге 3 требуется $\mathcal{O}(M(T_{\max} - T_{\min} + 1))$ операций. Суммируя затраты на всех шагах, получим приведённую оценку.

Изложим алгоритм решения задачи VSS2 с использованием алгоритма \mathcal{A}_1 . Входами алгоритма являются \mathcal{U} , T_{\min} , T_{\max} и M .

АЛГОРИТМ \mathcal{A}

ШАГ 1. Положим $i = 0$, $\widehat{\mathcal{M}} = \emptyset$, $G^* = +\infty$.

ШАГ 2. $i := i + 1$, $b = y_i$.

ШАГ 3. Для фиксированного вектора $b \in \mathcal{U}$ найдём оптимальное решение $\widehat{\mathcal{M}}(b) = \{\widehat{n}_1(b), \dots, \widehat{n}_M(b)\}$ и значение $G_{\min}(b)$ целевой функции задачи SVNВ с помощью алгоритма \mathcal{A}_1 .

ШАГ 4. Если $G^* > G_{\min}(b)$, то положим $G^* = G_{\min}(b)$, $\widehat{b} = b$, $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}(b)$.

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

ШАГ 6. Вычислим вектор $\bar{y}(\widehat{\mathcal{M}}) = \frac{1}{M} \sum_{n \in \widehat{\mathcal{M}}} y_n$ и значение $f_2(\widehat{\mathcal{M}})$ целевой функции по формуле (5); выход.

Решением задачи объявляем набор $\widehat{\mathcal{M}}$, значение $f_2(\widehat{\mathcal{M}})$ и вектор $\bar{y}(\widehat{\mathcal{M}})$.

Для обоснования точности алгоритмического решения потребуется

Лемма 2 [6]. Пусть \mathcal{Z} — непустое конечное множество векторов из \mathbb{R}^q , а $\bar{z} = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$ — центр этого множества. Тогда если вектор $t \in \mathbb{R}^q$ удовлетворяет условиям $\|t - \bar{z}\| \leq \|z - \bar{z}\|$ для любого $z \in \mathcal{Z}$, то имеет место неравенство

$$\sum_{z \in \mathcal{Z}} \|z - t\|^2 \leq 2 \sum_{z \in \mathcal{Z}} \|z - \bar{z}\|^2.$$

Теорема 1. Алгоритм \mathcal{A} находит 2-приближённое решение задачи VSS2 за время $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q))$. Оценка 2 точности алгоритма достижима.

ДОКАЗАТЕЛЬСТВО. Пусть \mathcal{M}^* — оптимальное решение задачи VSS2, $\mathcal{C}^* = \{y_n \mid n \in \mathcal{M}^*\}$ — подмножество векторов из \mathcal{Y} , соответствующих оптимальному набору \mathcal{M}^* , и $\bar{y}(\mathcal{M}^*) = \frac{1}{|\mathcal{M}^*|} \sum_{n \in \mathcal{M}^*} y_n$ — центр множества \mathcal{C}^* .

Согласно пошаговой записи алгоритм находит вектор

$$\widehat{b} = \arg \min_{b \in \mathcal{Y}} G_{\min}(b) = \arg \min_{b \in \mathcal{Y}} \min_{\mathcal{M}} \sum_{n \in \mathcal{M}} \|y_n - b\|^2$$

из множества \mathcal{Y} , набор $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}(\widehat{b}) = \{\widehat{n}_1(\widehat{b}), \dots, \widehat{n}_M(\widehat{b})\}$ и

$$G^* = \sum_{n \in \widehat{\mathcal{M}}} \|y_n - \widehat{b}\|^2 = \min_{b \in \mathcal{Y}} G_{\min}(b) = \min_{b \in \mathcal{Y}} \min_{\mathcal{M}} \sum_{n \in \mathcal{M}} \|y_n - b\|^2. \quad (13)$$

Возьмём вектор $u = \arg \min_{c \in \mathcal{C}^*} \|c - \bar{y}(\mathcal{M}^*)\|$, ближайший к центру подмножества $\mathcal{C}^* \subseteq \mathcal{Y}$. Для этого вектора из леммы 2 следует оценка

$$\sum_{n \in \mathcal{M}^*} \|y_n - u\|^2 \leq 2 \sum_{n \in \mathcal{M}^*} \|y_n - \bar{y}(\mathcal{M}^*)\|^2 = 2f_2(\mathcal{M}^*). \quad (14)$$

С другой стороны, для левой части (14) имеем очевидную оценку снизу

$$\sum_{n \in \mathcal{M}^*} \|y_n - u\|^2 \geq \min_{b \in \mathcal{Y}} \min_{\mathcal{M}} \sum_{n \in \mathcal{M}} \|y_n - b\|^2 = \sum_{n \in \widehat{\mathcal{M}}} \|y_n - \widehat{b}\|^2. \quad (15)$$

Кроме того, справедлива оценка

$$f_2(\widehat{\mathcal{M}}) = \sum_{n \in \widehat{\mathcal{M}}} \|y_n - \bar{y}(\widehat{\mathcal{M}})\|^2 \leq \sum_{n \in \widehat{\mathcal{M}}} \|y_n - \widehat{b}\|^2, \quad (16)$$

так как для любого конечного множества \mathcal{Z} векторов из \mathbb{R}^q минимум суммы квадратов $\sum_{z \in \mathcal{Z}} \|z - c\|^2$ по c достигается в точке $\bar{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} z$.

Из (13)–(16) следует, что для решения $\widehat{\mathcal{M}}$, полученного с помощью алгоритма \mathcal{A} , выполнено неравенство $f_2(\widehat{\mathcal{M}}) \leq 2f_2(\mathcal{M}^*)$, т. е. решение 2-приближённое.

Покажем, что оценка 2 точности алгоритма достижима. Для этого, опираясь на результаты [6], приведём пример существования входных данных задачи таких, что отношение $f_2(\widehat{\mathcal{M}})/f_2(\mathcal{M}^*)$ может быть сколь угодно близко к 2 и равно 2.

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

Легко видеть, что в этом примере оптимальным решением задачи является набор $\mathcal{M}^* = \{1, 2, 4\}$, причём $f_2(\mathcal{M}^*) = 1 + \frac{\alpha^2 - 3}{6}$. После несложных вычислений с использованием рекуррентных формул (9)–(12) динамического программирования для алгоритмического решения найдём $\widehat{\mathcal{M}} = \{1, 2, 3\}$, $f_2(\widehat{\mathcal{M}}) = 2$. Отношение $f_2(\widehat{\mathcal{M}})/f_2(\mathcal{M}^*) = 2/(1 + \frac{\alpha^2 - 3}{6})$ может быть сколь угодно близко к 2 при $\alpha \rightarrow \sqrt{3}$.

Если входные данные таковы, что $\alpha^2 = 3$, то оптимальным решением задачи является набор $\mathcal{M}^* = \{1, 2, 4\}$, причём $f_2(\mathcal{M}^*) = 1$. Возможны два равноправных алгоритмических решения: $\widehat{\mathcal{M}}_1 = \{1, 2, 3\}$ и $\widehat{\mathcal{M}}_2 = \{1, 2, 4\}$, так как $G(\widehat{\mathcal{M}}_1) = G(\widehat{\mathcal{M}}_2) = 2$ при $\widehat{b} = y_1$, причём $f_2(\widehat{\mathcal{M}}_1) = 2$, а $f_2(\widehat{\mathcal{M}}_2) = 1$. Для первого решения имеет место равенство $f_2(\widehat{\mathcal{M}}_1)/f_2(\mathcal{M}^*) = 2$, т. е. оценка 2 точности алгоритма достижима.

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

Замечание 2. В оценку временной сложности алгоритма входят числовые параметры M и $T_{\max} - T_{\min} + 1$, ограниченные размером N входа задачи, поэтому алгоритм \mathcal{A} полиномиален.

Алгоритм решения задачи VSS3 состоит в следующем. Сначала находим решение $\widehat{\mathcal{M}}$ задачи VSS2 и значение целевой функции $f_2(\widehat{\mathcal{M}})$ с по-

мощью алгоритма \mathcal{A} . Затем, опираясь на (6), вычисляем значение целевой функции $f_3(\widehat{\mathcal{M}}) = 2M f_2(\widehat{\mathcal{M}})$. Решением задачи объявляем набор $\widehat{\mathcal{M}}$ и значение $f_3(\widehat{\mathcal{M}})$. Очевидно, что это решение будет 2-приближённым.

Алгоритм решения задачи MSSC-Case-S аналогичен. В качестве искомого набора $\widehat{\mathcal{M}}_1$ берём набор $\widehat{\mathcal{M}}$, найденный с помощью алгоритма \mathcal{A} , и совокупность $\mathcal{N} \setminus \widehat{\mathcal{M}}_1$ одноэлементных наборов. Ясно, что это решение тоже 2-приближённое.

В силу (6) алгоритм \mathcal{A} можно применить для отыскания приближённого эффективного решения задачи VSS1. Следует, однако, заметить, что найденное решение не будет иметь каких-либо теоретических гарантий по точности из-за связи целевых функций задач VSS1 и VSS2 через аддитивную константу (см. (6)) — сумму квадратов длин векторов последовательности \mathcal{U} .

Остаётся заметить, что при решении специального случая рассмотренных задач, когда $T_{\min} = 1$ и $T_{\max} = N - 1$, лучше применять менее трудоёмкие алгоритмы, предложенные в [6]. Действительно, в этом случае трудоёмкость предложенных в настоящей работе алгоритмов в соответствии с замечанием 2 есть величина $\mathcal{O}(N^2(N^2 + q))$, в то время как алгоритмы, предложенные в [6], имеют трудоёмкость $\mathcal{O}(N^2q)$.

Заключение

В работе обоснованы 2-приближённые эффективные алгоритмы для решения задач, к которым сводится оптимизационная модель одной из актуальных проблем анализа данных.

Поскольку рассмотренные задачи относятся к числу практически неизученных в алгоритмическом плане, исследование вопросов их аппроксимируемости, а также обоснование алгоритмов другого типа (асимптотически точных, рандомизированных и др.) для их решения представляется делом ближайшей перспективы. Интерес представляют алгоритмические решения задачи VSS1, так как на сегодняшний день для этой задачи какие-либо эффективные алгоритмы с оценками точности неизвестны.

ЛИТЕРАТУРА

1. 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.

2. 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.
3. Кельманов А. В., Михайлова Л. В., Хамидуллин С. А. Об одной задаче поиска упорядоченных наборов фрагментов в числовой последовательности // Дискрет. анализ и исслед. операций. — 2009. — Т. 16, № 4. — С. 31–46.
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.
5. 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.
6. 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.
7. Шенмайер В. В. Аппроксимационная схема для одной задачи поиска подмножества векторов // Математические методы распознавания образов: 15-я Всерос. конф. (ММРО-15). Сб. докл. — М.: МАКС Пресс, 2011. — С. 284–286.
8. Anil K., Jain K. Data clustering: 50 years beyond k -means // Pattern Recognit. Lett. — 2010. — Vol. 31. — P. 651–666.
9. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. — New York: Springer-Verl., 2001. — 533 p.
10. 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.
11. Kel'manov A. V., Khamidullin S. A. An algorithm for recognition of a vector alphabet generating a sequence with a quasi-periodic structure // Pattern Recognit. Image Anal. — 2010. — Vol. 20, N 4. — P. 451–458.

Кельманов Александр Васильевич,
e-mail: kelm@math.nsc.ru
Романченко Семён Михайлович,
e-mail: semenr@bk.ru
Хамидуллин Сергей Асгадуллович,
e-mail: kham@math.nsc.ru

Статья поступила
11 августа 2011 г.
Переработанный вариант —
7 ноября 2011 г.