УДК 519.2+621.391

Аппроксимационная схема для задачи поиска подпоследовательности*

А.В. Кельманов[1,2], С.М. Романченко[1], С.А. Хамидуллин[1]

[1]Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, просп. Акад. В.А. Коптюга, 4, Новосибирск, 630090

[2]Новосибирский национальный исследовательский государственный университет, ул. Пирогова, 2, Новосибирск, 630090
E-mails: kelm@math.nsc.ru (Кельманов А.В.), rsm@math.nsc.ru (Романченко С.М.), kham@math.nsc.ru (Хамидуллин С.А.)

Кельманов А.В., Романченко С.М., Хамидуллин С.А. Аппроксимационная схема для задачи поиска подпоследовательности // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск, 2017. — Т. 20, № 4. — С. 379–392.

Рассматривается NP-трудная в сильном смысле задача поиска подпоследовательности с заданным числом элементов в конечной последовательности точек евклидова пространства. Критерием решения является минимум суммы квадратов расстояний от элементов искомой подпоследовательности до их геометрического центра. Выбор элементов подпоследовательности подчинен условию: разность между номерами последующего и предыдущего искомых элементов ограничена сверху и снизу заданными константами. Предложен приближенный алгоритм решения задачи. Доказано, что он является полностью полиномиальной аппроксимационной схемой (FPTAS), если размерность пространства ограничена константой.

DOI: 10.15372/SJNM20170403

Ключевые слова: *последовательность, евклидово пространство, минимум суммы квадратов расстояний, NP-трудность, FPTAS.*

Kelmanov A.V., Romanchenko S.M., Khamidullin S.A. An approximation scheme for a problem of finding a subsequence // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. — Novosibirsk, 2017. — Vol. 20, № 4. — P. 379–392.

We consider a strongly NP-hard Euclidean problem of finding a subsequence in a finite sequence. The criterion of the solution is a minimum sum of squared distances from the elements of a sought subsequence to its geometric center (centroid). It is assumed that a sought subsequence contains a given number of elements. In addition, a sought subsequence should satisfy the following condition: the difference between the indices of each previous and subsequent points is bounded with given lower and upper constants. We present an approximation algorithm of solving the problem and prove that it is a fully polynomial-time approximation scheme (FPTAS) when the space dimension is bounded by a constant.

Keywords: *euclidean space, sequence, minimum sum of squared distances, NP-hardness, FPTAS.*

Введение

Предметом исследования работы является NP-трудная в сильном смысле квадратичная евклидова задача поиска подпоследовательности, состоящей из заданного числа элементов, в конечной последовательности. Цель исследования — построение приближенного алгоритма с гарантированными оценками точности и трудоемкости.

*Работа выполнена при финансовой поддержке РФФИ (проекты № 15-01-00462, № 16-31-00186-мол-а, № 16-07-00168).

Исследование мотивировано слабой изученностью задачи и ее актуальностью, в частности, для математических проблем анализа временны́х рядов, проблем аппроксимации и проблем дискретной оптимизации (см., например, [1–3], цитированные там работы и следующий пункт), прикладных статистических проблем и междисциплинарной проблемы интерпретации данных (Data mining) (см., в частности, [4, 5] и следующий пункт). Кроме того, задача актуальна для многих естественно-научных и технических приложений, в которых требуется классификация упорядоченных по времени данных численных экспериментов, для помехоустойчивого дистанционного мониторинга объектов, электронной разведки, анализа и распознавания сигналов и др. (см., например, [1–7], цитированные там работы и следующий пункт).

Как известно, для NP-трудных задач не существует универсального полиномиального метода (алгоритма) решения, если гипотеза о несовпадении классов P и NP верна. Каждая из таких задач, в том числе рассматриваемая, требует разработки индивидуального подхода и вычислительного алгоритма. К тому же, рассматриваемая задача относится к числу сильно NP-трудных задач и для нее не существует (см. п. 2) полностью полиномиальной аппроксимационной схемы (FPTAS), если P$\neq$ NP. Поэтому значительный интерес представляет вопрос о построении схемы FPTAS для какого-либо специального случая (подкласса) задачи. Этот вопрос традиционен и важен для исследований как в области вычислительной математики, так и в области дискретной оптимизации. Ниже мы предлагаем схему FPTAS для одного из случаев рассматриваемой задачи. Предложенная схема основана на сеточной технике.

Статья развивает результаты, полученные ранее в [1–3, 10–13], и имеет следующую структуру. В следующем пункте дана формулировка задачи и приведены примеры ее приложений (истоков). В п. 2 приведены известные результаты и анонсирован полученный результат. В п. 3 сформулированы и доказаны вспомогательные утверждения, обеспечивающие установление свойств предложенного алгоритма. Наконец, в п. 4 приведен приближенный алгоритм и доказано, что он реализует полностью полиномиальную аппроксимационную схему, если размерность пространства ограничена константой.

1. Формулировка задачи и ее истоки

Всюду далее используются следующие обозначения: $\mathbb{R}$ — множество вещественных чисел, $\mathbb{R}_+$ — множество положительных вещественных чисел, $\mathbb{Z}$ — множество целых чисел, $\|\cdot\|$ — евклидова норма, а $\mathcal{N}=\{1,\dots,N\}$ — подмножество натуральных чисел.

Рассматриваемая задача имеет следующую формулировку (см. [1–3]).

Задача 1 (*Finding a subsequence in a sequence*). *Дано*: последовательность $\mathcal{Y}=(y_1,\dots,y_N)$ точек 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

$$F(\mathcal{M})=\sum_{j\in\mathcal{M}}\|y_j-\overline{y}(\mathcal{M})\|^2\to\min, \tag{1}$$

где $\overline{y}(\mathcal{M})=\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}y_i$ — геометрический центр (центроид) мультимножества $\{y_i\in\mathcal{Y}\,|\,i\in\mathcal{M}\}$, состоящего из элементов искомой подпоследовательности, при ограничениях

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

на элементы набора $(n_1,\dots,n_M)$.

Сформулированная ниже простая содержательная проблема индуцирует задачу 1. Эта проблема ввиду своей содержательной простоты характерна для всех отмеченных во введении приложений и позволяет интерпретировать задачу 1 в терминах указанных приложений. Мы приведем лишь некоторые интерпретации. Заинтересованный читатель без особых усилий может привести другие трактовки.

Дана последовательность $\mathcal{Y}$, содержащая N упорядоченных по времени результатов $y_1, \ldots, y_N$ измерения набора y из q числовых характеристик некоторого объекта. Результаты измерений отображают состояния объекта. Среди этих состояний имеется M идентичных (повторяющихся). Каждый результат измерения сопровождается инструментальной ошибкой. Моменты времени $n_1, \ldots, n_M$ измерений идентичных состояний неизвестны. Однако известно, что временно́й интервал между двумя последовательными идентичными состояниями объекта ограничен сверху и снизу некоторыми константами $T_{\max}$ и $T_{\min}$. Характеристики повторяющегося состояния объекта в отличие от характеристик других состояний имеют информационную ценность.

Требуется найти подпоследовательность измерений, которая соответствует повторяющемуся состоянию объекта, используя критерий минимума суммы квадратов расстояний, и оценить набор характеристик объекта в этом состоянии, учитывая, что данные содержат ошибки измерения.

Формализация этой простой содержательной проблемы с использованием критерия минимума суммы квадратов уклонений индуцирует (см. [1–3]) следующую задачу аппроксимации. Даны последовательность $y_1, \ldots, y_N$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и M. Требуется найти аппроксимирующую последовательность $z_1, \ldots, z_N$ вида

$$z_n = \begin{cases} x, & n \in \mathcal{M}, \\ v_n, & n \in \mathcal{N} \setminus \mathcal{M}, \end{cases} \tag{3}$$

где x и v_n — произвольные неизвестные точки из $\mathbb{R}^q$, такую, что

$$\sum_{i \in \mathcal{N}} \|y_i - z_i\|^2 \longrightarrow \min, \tag{4}$$

при тех же, что и в задаче 1, ограничениях на номера из подмножества $\mathcal{M}$.

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

$$\ldots v_{i-1} x v_{i+1} \ldots v_{k-1} x v_{k+1} \ldots v_{j-1} x v_{j+1} \ldots \ . \tag{5}$$

Здесь x — неизвестная точка, соответствующая повторяющемуся состоянию объекта в моменты времени $n_1, \ldots, n_M$, остальные элементы последовательности соответствуют произвольным состояниям. Число этих элементов между повторами точки x неизвестно и лежит в допустимом интервале от $T_{\min}-1$ до $T_{\max}-1$ в соответствии с ограничениями (2).

Раскрыв сумму (4) с учетом (3) и сгруппировав члены, легко проверить с помощью дифференцирования, что оптимальными в смысле (4) являются значения $x = \overline{y}(\mathcal{M})$ и $v_n = y_n$, $n \in \mathcal{N} \setminus \mathcal{M}$, а сформулированная задача аппроксимации индуцирует задачу 1 поиска набора $\mathcal{M}$. При этом в найденной оптимальной аппроксимирующей последовательности участок, соответствующий (5), как видно из формулировки задачи 1, имеет вид

$$\ldots v_{i-1} \overline{y}(\mathcal{M}) v_{i+1} \ldots v_{k-1} \overline{y}(\mathcal{M}) v_{k+1} \ldots v_{j-1} \overline{y}(\mathcal{M}) v_{j+1} \ldots \ .$$

В этой последовательности элемент $\overline{y}(\mathcal{M})$ — центроид мультимножества $\{y_j \,|\, j \in \mathcal{M}\}$ — определяется в результате решения задачи 1 и является оценкой для точки x.

Из приведенной выше схематичной строковой записи последовательностей видно, что их можно интерпретировать как последовательности, содержащие неизвестные квазипериодически (в силу ограничений (2)) повторяющиеся точки (элементы). Задачу аппроксимации можно трактовать как поиск квазипериодических повторов неизвестной точки совместно с оцениванием ее координат и отысканием положения этой точки в последовательности, т. е. как задачу совместного обнаружения и оценивания.

Кроме того, задача имеет приложения в междисциплинарной проблеме Data mining. Суть этой многогранной проблемы заключается в аппроксимации имеющихся данных некоторой математической моделью, позволяющей адекватно интерпретировать эти данные и правдоподобно объяснять их происхождение в терминах аппроксимирующей модели. В частности, среди вопросов о происхождении данных может быть статистическая гипотеза: верно ли, что входная последовательность $\mathcal{Y}$ является неоднородной выборкой из нескольких вероятностных распределений, причем M элементов этой последовательности являются выборкой из одного распределения с неизвестным средним (предполагается, что соответствие элементов выборки распределению неизвестно). Для проверки этой гипотезы нам потребуется сначала найти оптимальное решение задачи 1 (т. е. подпоследовательность). Лишь затем мы сможем воспользоваться классическими результатами из области статистической проверки гипотез.

Как известно, в руках исследователей-прикладников, изучающих и анализирующих данные, алгоритмы решения разнообразных задач поиска похожих или близких по некоторому критерию объектов являются основными математическими инструментами. Создание новых математических инструментов для решения проблемы Data mining обуславливает разработку эффективных алгоритмов с гарантированными оценками качества решения.

2. Известные и полученный результаты

Как отмечено выше, задача 1 относится к числу слабоизученных труднорешаемых проблем. Частный случай задачи 1, в котором $T_{\min} = 1$ и $T_{\max} = N$, эквивалентен (см. [1, 8]) NP-трудной в сильном смысле задаче поиска подмножества. В этой задаче (см. [8]) входом является не последовательность, а множество, и ограничения (2) на номера элементов, включаемых в подпоследовательность, отсутствуют.

Напомним сначала результаты, полученные для задачи поиска подмножества, так как она является частным случаем задачи 1. В общем случае эта задача NP-трудна в сильном смысле [8]. При этом в случае фиксированной размерности q пространства задача поиска подмножества полиномиально разрешима [9] за время $\mathcal{O}(N^{q+1})$.

Кроме того, к настоящему времени для задачи поиска подмножества построены следующие алгоритмы.

В [10] предложен 2-приближенный полиномиальный алгоритм, трудоемкость которого равна $\mathcal{O}(qN^2)$.

Полиномиальная приближенная схема (PTAS) обоснована в [11]. Эта схема позволяет решать задачу за время $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$ с произвольной относительной погрешностью ε.

Для случая фиксированной размерности q пространства и целочисленных координат точек построен [12] точный псевдополиномиальный алгоритм, имеющий трудоемкость $\mathcal{O}(N(MD)^q)$, где D — максимальное абсолютное значение координат входных точек. В [13] установлено, что для задачи поиска подмножества, имеющей числовые входы, не существует схемы FPTAS, если P$\neq$ NP, и там же такая схема обоснована для специального случая, в котором размерность q пространства фиксирована. Эта схема позволяет решать задачу с произвольной относительной погрешностью ε за время $\mathcal{O}(N^2(M/\varepsilon)^q)$.

К настоящему времени для рассматриваемой задачи 1 были получены следующие результаты. Заметим сначала, что поскольку задача 1 является обобщением NP-трудной в сильном смысле задачи поиска подмножества, для нее, как и для задачи поиска подмножества, не существует ни точного полиномиального, ни точного псевдополиномиального алгоритмов, ни схемы FPTAS, если P$\neq$ NP.

Случай, когда $T_{\min}$ и $T_{\max}$ являются параметрами задачи 1, анализировался в [1]. В этой работе установлено, что задача 1 NP-трудна в сильном смысле для любых $T_{\min} < T_{\max}$. В тривиальном случае, когда $T_{\min} = T_{\max}$, эта задача разрешима за полиномиальное время.

В работе [2] предложен 2-приближенный полиномиальный алгоритм, временная сложность которого есть величина $\mathcal{O}(N^2(MN + q))$. Для случая задачи 1, в котором компоненты элементов последовательности целочисленны, а размерность q пространства фиксирована, в [3] обоснован точный псевдополиномиальный алгоритм, который находит решение задачи за время $\mathcal{O}(N^3(MD)^q)$, где D — максимальное абсолютное значение координат входных точек.

Основной результат настоящей работы — приближенный алгоритм, который при заданной относительной погрешности ε находит $(1+\varepsilon)$-приближенное решение задачи 1 за время $\mathcal{O}\Big(N^2(M(T_{\max} - T_{\min} + 1) + q)\Big(\sqrt{\frac{2q}{\varepsilon}} + 2\Big)^q\Big)$. Показано, что при фиксированной размерности q пространства время работы алгоритма есть величина $\mathcal{O}\big(MN^3(1/\varepsilon)^{q/2}\big)$, и он реализует схему FPTAS.

3. Основы алгоритма

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

Геометрической основой алгоритма являются следующие утверждения.

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

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

где $\overline{z}$ — центроид множества $\mathcal{Z}$.

Лемма 2. *Пусть выполнены условия леммы 1. Тогда если некоторая точка $u \in \mathbb{R}^q$ является ближайшей (в смысле расстояния) к центроиду $\overline{z}$ множества $\mathcal{Z}$ из всех точек этого множества, то*

$$\sum_{z\in\mathcal{Z}} \|z - u\|^2 \le 2\sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2.$$

Обе леммы относятся к числу хорошо известных. Их доказательство представлено во множестве публикаций (см., например, [12, 13]). Легко проверить, что утверждения обеих лемм справедливы в случае, когда $\mathcal{Z}$ — мультимножество.

Лемма 3. *Пусть*

$$S(\mathcal{M}, x) = \sum_{n \in \mathcal{M}} \|y_n - x\|^2, \quad x \in \mathbb{R}^q, \quad \mathcal{M} \subseteq \mathcal{N}, \tag{6}$$

где y_n — элемент последовательности $\mathcal{Y}$, а элементы набора $\mathcal{M} = \{n_1, \ldots, n_M\}$ удовлетворяют ограничениям (2). *Тогда для любого фиксированного набора $\mathcal{M}$ минимум функции $S(\mathcal{M}, x)$ по x достигается в точке $x = \overline{y}(\mathcal{M})$ и равен $F(\mathcal{M})$.*

Доказательство. Справедливость утверждения леммы легко проверяется дифференцированием и следует также из леммы 1.

Всюду далее мы используем $f^u(v)$ для обозначения некоторой функции $f(u, v)$ при условии, что аргумент u у этой функции фиксирован. Для произвольной фиксированной точки $x \in \mathbb{R}^q$ положим

$$\mathcal{M}^x = \arg \min_{\mathcal{M} \subseteq \mathcal{N}} S^x(\mathcal{M}). \tag{7}$$

Нетрудно заметить, что набор $\mathcal{M}^x$ является допустимым решением задачи 1, порожденным точкой x. □

Лемма 4. *Пусть $\mathcal{M}^*$ — оптимальное решение задачи 1, $\overline{y}(\mathcal{M}^*)$ — центроид мультимножества $\{y_i \mid i \in \mathcal{M}^*\}$. Тогда при любой фиксированной точке $x \in \mathbb{R}^q$ для значения целевой функции* (1) *на наборе $\mathcal{M}^x$ справедлива оценка*

$$F(\mathcal{M}^x) \le F(\mathcal{M}^*) + M\|x - \overline{y}(\mathcal{M}^*)\|^2. \tag{8}$$

Доказательство. Пусть $\overline{y}(\mathcal{M}^x)$ — центроид мультимножества $\{y_i \in \mathcal{Y} \mid i \in \mathcal{M}^x\}$. Тогда из определений (1), (6) и первого утверждения леммы 3 имеем

$$F(\mathcal{M}^x) = \sum_{i \in \mathcal{M}^x} \|y_i - \overline{y}(\mathcal{M}^x)\|^2 \le \sum_{i \in \mathcal{M}^x} \|y_i - x\|^2 = S^x(\mathcal{M}^x). \tag{9}$$

Кроме того, в соответствии с определением (7) набора $\mathcal{M}^x$ имеем неравенство

$$S^x(\mathcal{M}^x) \le S^x(\mathcal{M}^*). \tag{10}$$

Далее, применив лемму 1 к точке x и мультимножеству $\{y_i \,|\, i \in \mathcal{M}^*\}$, для правой части (10) найдем

$$S^x(\mathcal{M}^*) = \sum_{i \in \mathcal{M}^*} \|y_i - x\|^2 = \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 + |\mathcal{M}^*| \, \|x - \overline{y}(\mathcal{M}^*)\|^2. \tag{11}$$

Наконец, объединив (9)–(11), получим

$$\begin{aligned} F(\mathcal{M}^x) &\le S^x(\mathcal{M}^x) \le S^x(\mathcal{M}^*) \\ &= \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 + |\mathcal{M}^*| \, \|x - \overline{y}(\mathcal{M}^*)\|^2 = F(\mathcal{M}^*) + M\|x - \overline{y}(\mathcal{M}^*)\|^2. \end{aligned}$$ □

Лемма 4 показывает, что точность допустимого решения $\mathcal{M}^x$, полученного по некоторой точке $x \in \mathbb{R}^q$, можно оценить через расстояние от этой точки до (неизвестного) оптимального центроида $\overline{y}(\mathcal{M}^*)$. Чем ближе точка x к оптимальному центроиду, тем меньше абсолютная ошибка аппроксимации значения $F(\mathcal{M}^*)$ значением $F(\mathcal{M}^x)$. При этом для точности аппроксимации из (8) имеем

$$\frac{F(\mathcal{M}^x)}{F(\mathcal{M}^*)} \le 1 + \frac{M\|x - \overline{y}(\mathcal{M}^*)\|^2}{F(\mathcal{M}^*)}.$$

Лемма 5. *Пусть $\mathcal{M}^*$ — оптимальное решение задачи 1, а*

$$t = \arg \min_{y \in \{y_i \mid i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\| \tag{12}$$

есть точка из мультимножества $\{y_i \mid i \in \mathcal{M}^*\}$, ближайшая к его центроиду $\overline{y}(\mathcal{M}^*)$. Тогда справедлива оценка

$$\|t - \overline{y}(\mathcal{M}^*)\|^2 \le \frac{1}{M} F(\mathcal{M}^t), \tag{13}$$

где $\mathcal{M}^t$ — подмножество, доставляющее минимум функции $S^t(\mathcal{M})$ по всем $\mathcal{M} \subseteq \mathcal{N}$ при ограничениях (2) на элементы $\mathcal{M}$.

Доказательство. Из определения (12) точки t следует, что для любого $i \in \mathcal{M}^*$ справедливо неравенство

$$\|t - \overline{y}(\mathcal{M}^*)\|^2 \le \|y_i - \overline{y}(\mathcal{M}^*)\|^2.$$

Просуммировав обе части этого неравенства по $i \in \mathcal{M}^*$, получим

$$M\|t - \overline{y}(\mathcal{M}^*)\|^2 \le \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2. \tag{14}$$

Поскольку $\mathcal{M}^t$ — допустимое решение задачи 1, а $\mathcal{M}^*$ — ее оптимальное решение, имеем неравенство

$$F(\mathcal{M}^*) \le F(\mathcal{M}^t). \tag{15}$$

Объединив (14) и (15), получим оценку

$$M\|t - \overline{y}(\mathcal{M}^*)\|^2 \le \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 = F(\mathcal{M}^*) \le F(\mathcal{M}^t),$$

которая устанавливает справедливость неравенства (13). □

Лемма 5 дает оценку сверху на расстояние от центроида $\overline{y}(\mathcal{M}^*)$ оптимального решения $\mathcal{M}^*$ задачи 1 до ближайшей к нему точки из входной последовательности $\mathcal{Y}$.

Лемма 6. *Пусть выполнены условия леммы 5. Тогда для того, чтобы при заданном $\varepsilon > 0$ для некоторой точки $x \in \mathbb{R}^q$ множество $\mathcal{M}^x$ было $(1+\varepsilon)$-приближенным решением задачи 1, достаточно, чтобы точка x удовлетворяла неравенству*

$$\|x - \overline{y}(\mathcal{M}^*)\|^2 \le \frac{\varepsilon}{4M} F(\mathcal{M}^t). \tag{16}$$

Доказательство. Пусть $\overline{y}(\mathcal{M}^t)$ — центроид мультимножества $\{y_i \mid i \in \mathcal{M}^t\}$. Поскольку $\mathcal{M}^t = \arg\min_{\mathcal{M}} S^t(\mathcal{M})$, из определений (1), (6) функций F и S следует неравенство

$$F(\mathcal{M}^t) \le S^t(\mathcal{M}^t). \tag{17}$$

Кроме того, в соответствии с определением (7) имеем неравенство

$$S^t(\mathcal{M}^t) \le S^t(\mathcal{M}^*). \tag{18}$$

Далее, применяя лемму 2 к точке t и мультимножеству $\{y_i \mid i \in \mathcal{M}^*\}$, найдем оценку для правой части (18):

$$S^t(\mathcal{M}^*) = \sum_{i\in\mathcal{M}^*} \|y_i - t\|^2 \le 2 \sum_{i\in\mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 = 2F(\mathcal{M}^*). \tag{19}$$

Объединив (16)–(19), получим

$$\|x - \overline{y}(\mathcal{M}^*)\|^2 \le \frac{\varepsilon}{2M} F\big(\mathcal{M}^t\big) \le \frac{\varepsilon}{2M} S^t\big(\mathcal{M}^t\big) \le \frac{\varepsilon}{2M} S^t(\mathcal{M}^*) \le \frac{\varepsilon}{M} F(\mathcal{M}^*). \tag{20}$$

Наконец, применив (20) к правой части неравенства (8), получим оценку

$$F(\mathcal{M}^x) \le (1+\varepsilon) F(\mathcal{M}^*),$$

из которой следует справедливость леммы. □

Лемма 6 показывает, насколько близко к центроиду оптимального решения должна находиться точка x, чтобы порожденное этой точкой допустимое решение $\mathcal{M}^x$ гарантировало получение $(1+\varepsilon)$-приближенного решения задачи 1.

Вычислительной базой алгоритма является точный полиномиальный алгоритм поиска набора $\mathcal{M}^x$, т. е. алгоритм решения следующей вспомогательной задачи.

Задача 2. *Дано*: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$, точка $x \in \mathbb{R}^q$, 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

$$S^x(\mathcal{M}) = \sum_{n\in\mathcal{M}} \|y_n - x\|^2 \to \min,$$

при ограничениях (2) на элементы набора $(n_1, \ldots, n_M)$.

Для описания алгоритма решения задачи 2 положим

$$g^x(n) = \|y_n - x\|^2, \quad n \in \mathcal{N}, \tag{21}$$

где x — фиксированная точка из $\mathbb{R}^q$. Тогда целевую функцию задачи 2 можно записать в эквивалентном виде

$$S^x(\mathcal{M}) = \sum_{n\in\mathcal{M}} g^x(n), \quad \mathcal{M} \subseteq \mathcal{N}.$$

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения $\mathcal{M}^x$ задачи 2. Схема опирается на результаты из [2, 14] и приводится здесь ради полноты изложения.

Лемма 7. *Для любого натурального $M > 1$ такого, что $(M-1)\,T_{\min} \le N - 1$, и для произвольной точки $x \in \mathbb{R}^q$ оптимальное значение $S^x_{\min} = \min_{\mathcal{M}} S^x(\mathcal{M})$ целевой функции задачи 2 находится по формуле*

$$S^x_{\min} = \min_{n\in\omega_M} S^x_M(n), \tag{22}$$

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

$$S^x_m(n) = \begin{cases} g^x(n), & \text{если } n \in \omega_1,\ m = 1; \\ g^x(n) + \min\limits_{j\in\gamma^-_{m-1}(n)} S^x_{m-1}(j), & \text{если } n \in \omega_m,\ m = 2,\dots,M, \end{cases} \tag{23}$$

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

$$\omega_m = \big\{n \mid 1 + (m-1)T_{\min} \le n \le N - (M-m)T_{\min}\big\}, \quad m = 1,\dots,M,$$
$$\gamma^-_{m-1}(n) = \big\{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \le j \le n - T_{\min}\big\},\ n \in \omega_m,\ m = 2,\dots,M.$$

Следствие. *Элементы* $n^x_1,\dots,n^x_M$ *оптимального набора* $\mathcal{M}^x$ *находятся по следующим рекуррентным формулам*:

$$n^x_M = \arg\min_{n\in\omega_M} S^x_M(n), \tag{24}$$

$$n^x_{m-1} = \arg\min_{n\in\gamma^-_m(n^x_m)} S^x_m(n), \quad m = M, M-1,\dots,2. \tag{25}$$

Запишем алгоритм, реализующий приведенную схему, в пошаговом виде.

Алгоритм $\mathcal{A}_1$.

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

Шаг 1. Вычислим значение $g^x(n)$ для каждого $n \in \mathcal{N}$ по формуле (21).

Шаг 2. Используя рекуррентные формулы (23), вычислим значение $S^x_m(n)$ для каждого $n \in \omega_m$ и $m = 1,\dots,M$.

Шаг 3. Найдем значение $S^x_{\min}$ по формуле (22) и набор $\mathcal{M}^x = (n^x_1,\dots,n^x_M)$ по формулам (24), (25).

Выход алгоритма: набор $\mathcal{M}^x = (n^x_1,\dots,n^x_M)$.

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

4. Приближенный алгоритм

Суть предлагаемого подхода к поиску приближенного решения задачи 1 состоит в следующем. Для каждой точки из входной последовательности строится куб с центром в этой точке. Размер куба вычисляется адаптивно (см. далее) так, чтобы хотя бы один из построенных кубов гарантированно включал неизвестный центроид оптимального решения. По заданной на входе относительной погрешности решения для каждого из кубов строится многомерная решетка (сетка), дискретизирующая куб с равномерным по всем координатам шагом. Шаг решетки, как и ее размер, вычисляется адаптивно для каждой из входных точек. Для каждого узла решетки методом динамического программирования с помощью алгоритма $\mathcal{A}_1$ решается вспомогательная задача 2. Найденное допустимое решение — набор номеров — включается в семейство претендентов на решение задачи 1.

Окончательным решением объявляется такой набор из построенного семейства, для которого значение целевой функции задачи 1 минимально.

Для произвольной точки $z \in \mathbb{R}^q$ и положительных чисел h, H определим множество точек

$$\mathcal{D}(z, h, H) = \{d \mid d = z + h(j_1, \ldots, j_q),\ j_i \in \mathbb{Z},\ |\, hj_i| \le H,\ i = 1, \ldots, q\}, \tag{26}$$

т. е. решетку кубической формы размера $2H$ с расстоянием h между узлами и центром в точке z.

Замечание 2. Для любой точки $x \in \mathbb{R}^q$ такой, что $\|z - x\| \le H$, расстояние от этой точки до ближайшего узла решетки $\mathcal{D}(z, h, H + h/2)$, очевидно, не превосходит $\frac{h\sqrt{q}}{2}$.

Для числа узлов этой решетки справедлива оценка

$$|\mathcal{D}(z, h, H + h/2)| \le \left(2\left\lfloor\frac{H + h/2}{h}\right\rfloor + 1\right)^q \le \left(2\frac{H}{h} + 2\right)^q, \tag{27}$$

которая не зависит от z.

Лемма 5 (правая часть неравенства (13)) фактически определяет размер куба, который гарантированно содержит неизвестный центроид оптимального решения задачи 1 в том случае, когда центром решетки является точка $z = t$, т. е. точка, ближайшая к этому центроиду. Лемма 6 устанавливает условие на шаг решетки, при котором среди ее узлов найдется элемент, близкий (в смысле гарантированной погрешности ε) к центроиду оптимального решения.

Точка t неизвестна, но она в числе элементов последовательности $\mathcal{Y}$. Поэтому для адаптивного вычисления размера и шага решетки определим функции:

$$H(y) = \sqrt{\frac{1}{M}F(\mathcal{M}^y)}, \quad y \in \mathcal{Y}, \tag{28}$$

$$h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{qM}F(\mathcal{M}^y)}, \quad y \in \mathcal{Y},\ \varepsilon \in \mathbb{R}_+. \tag{29}$$

Замечание 3. Для любой точки $y \in \mathcal{Y}$ мощность $|\mathcal{D}(y, h, H + h/2)|$ решетки не превосходит значения

$$L = \left(\sqrt{\frac{2q}{\varepsilon}} + 2\right)^q \tag{30}$$

в силу (27)–(29).

Сформулируем следующий алгоритм решения задачи 1.

Алгоритм $\mathcal{A}$.

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

Для каждой точки $y \in \mathcal{Y}$ выполним шаги 1–6.

Шаг 1. С помощью алгоритма $\mathcal{A}_1$ найдем оптимальное решение $\mathcal{M}^y$ задачи 2 (при $x = y$); вычислим значение $F(\mathcal{M}^y)$ по формуле (1).

Шаг 2. Если $F(\mathcal{M}^y) = 0$, то положим $\mathcal{M}_{\mathcal{A}} = \mathcal{M}^y$; выход.

Шаг 3. Вычислим значения H и h по формулам (28) и (29) соответственно.

Шаг 4. Построим решетку $\mathcal{D}(y,h,H+h/2)$ по формуле (26).

Шаг 5. Для каждого узла d решетки $\mathcal{D}(y,h,H+h/2)$ с помощью алгоритма $\mathcal{A}_1$ найдем оптимальное решение $\mathcal{M}^d$ задачи 2 (при $x=d$) и вычислим значение $F(\mathcal{M}^d)$.

Шаг 6. Если $F(\mathcal{M}^d)=0$, то положим $\mathcal{M}_{\mathcal{A}}=\mathcal{M}^d$; выход.

Шаг 7. В семействе $\{\mathcal{M}^d \mid d\in\{\mathcal{D}(y,h,H+h/2),\ y\in\mathcal{Y}\}\}$ наборов, построенных на шагах 1–6, в качестве решения $\mathcal{M}_{\mathcal{A}}$ выберем тот набор $\mathcal{M}^d$, для которого значение $F(\mathcal{M}^d)$ минимально.

Выход алгоритма: набор $\mathcal{M}_{\mathcal{A}}$.

Теорема. *Для любого $\varepsilon>0$ алгоритм $\mathcal{A}$ находит $(1+\varepsilon)$-приближенное решение задачи 1 за время* $\mathcal{O}\Big(N^2(M(T_{\max}-T_{\min}+1)+q\Big)\Big(\sqrt{\frac{2q}{\varepsilon}}+2\Big)^q\Big)$.

Доказательство. Оценим точность алгоритма. По определению шага 7 имеем

$$F(\mathcal{M}_A)\le F(\mathcal{M}^d) \tag{31}$$

для любого узла $d\in\{\mathcal{D}(y,h,H+h/2),y\in\mathcal{Y}\}$.

Рассмотрим сначала случай, когда на шаге 2 алгоритма выполнилось условие $F(\mathcal{M}^y)=\ 0$ для некоторой точки $y\in\mathcal{Y}$. В этом случае подмножество $\mathcal{M}_A=\mathcal{M}^y\subseteq\mathcal{N}$ является оптимальным решением задачи 1, так как для любого подмножества $\mathcal{M}\subseteq\mathcal{N}$ справедливо неравенство $F(\mathcal{M})\ge 0$. Аналогичный оптимальный результат получаем на шаге 6.

Проанализируем случай, когда на шаге 2 условие $F(\mathcal{M}^y)=0$ не выполнилось. Рассмотрим точку t (см. лемму 5) из мультимножества $\{y_i \mid i\in\mathcal{M}^*\}$, ближайшую к центроиду этого множества. Очевидно, что в ходе работы алгоритма такая точка найдется. Для этой точки на шаге 4 будет построена решетка $\mathcal{D}(t,h,H+h/2)$.

Ясно, что среди узлов построенной решетки $\mathcal{D}(t,h,H+h/2)$ найдется узел

$$d^*(t)=\arg\min_{d\in\mathcal{D}(t,h,H+h/2)}\|d-\overline{y}(\mathcal{M}^*)\|,$$

ближайший к центроиду оптимального решения. Поскольку расстояние от $\overline{y}(\mathcal{M}^*)$ до ближайшего узла $d^*(t)$ решетки $\mathcal{D}(t,h,H+h/2)$ не превосходит $\frac{h\sqrt{q}}{2}$ (см. замечание 2), а шаг h решетки определяется формулой (29) (при $y=t$), имеем оценку

$$\|d^*(t)-\overline{y}(\mathcal{M}^*)\|^2\le\frac{h^2q}{4}=\frac{\varepsilon}{2M}F\big(\mathcal{M}^t\big).$$

Поэтому точка $d^*(t)$ удовлетворяет условиям леммы 6, и, следовательно,

$$F(\mathcal{M}^{d^*(t)})\le(1+\varepsilon)F(\mathcal{C}^*),$$

т. е. множество $\mathcal{M}^{d^*(t)}$ является $(1+\varepsilon)$-приближенным решением задачи 1. Поскольку $\mathcal{M}^{d^*(t)}\in\{\mathcal{M}^d \mid d\in\{\mathcal{D}(y,h,H+h/2),\ y\in\mathcal{Y}\}\}$, из (31) следует, что $\mathcal{M}_A$ — $(1+\varepsilon)$-приближенное решение задачи 1.

Оценим временну́ю сложность алгоритма. На шаге 1 решение задачи 2 выполняется за время $\mathcal{O}(N(M(T_{\max}-T_{\min}+1)+q))$ (см. замечание 1), а вычисление значения $F(\mathcal{M}^y)$ — за время $\mathcal{O}(qN)$. Шаги 2, 3 и 6 выполняются за $\mathcal{O}(1)$ операций. Для построения решетки на шаге 4 требуется $\mathcal{O}(qL)$ операций (по замечанию 3).

Решение задачи 2 на шаге 5 и вычисление значения $F(\mathcal{M}^d)$ осуществляется, как и на шаге 1, за время $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$ и $\mathcal{O}(qN)$ для каждого узла решетки. Поэтому суммарное время вычислений для всех узлов решетки на этом шаге равно $\mathcal{O}(LN(M(T_{\max} - T_{\min} + 1) + q))$.

Поскольку шаги 1–6 выполняются N раз, трудоемкость выполнения этих шагов есть величина $\mathcal{O}(LN^2(M(T_{\max} - T_{\min} + 1) + q))$. Трудоемкость шага 7 оценивается величиной $\mathcal{O}(NL)$, а суммарные затраты на всех шагах равны $\mathcal{O}(LN^2(M(T_{\max} - T_{\min} + 1) + q))$. Поставив в эту формулу вместо L правую часть (30), получим итоговую оценку временно́й сложности алгоритма, приведенную в формулировке теоремы. □

Покажем, что при фиксированной размерности q пространства алгоритм $\mathcal{A}$ реализует схему FPTAS. Действительно, если $\varepsilon \in (0,\, 2q]$, то для одного из сомножителей в оценке вычислительной сложности имеем

$$\left(\sqrt{\frac{2q}{\varepsilon}} + 2\right)^q \le 2^q \left(\sqrt{\frac{2q}{\varepsilon}}\right)^q = 2^{3q/2} q^{q/2} (1/\varepsilon)^{q/2} = \mathcal{O}((1/\varepsilon)^{q/2}).$$

Кроме того, заметим, что величина $T_{\max} - T_{\min} + 1$ не превосходит N. Поэтому, если размерность q пространства фиксирована (т. е. ограничена сверху некоторой константой), то время алгоритма оценивается величиной $\mathcal{O}\big(MN^3 (1/\varepsilon)^{q/2}\big)$, которая ограничена полиномом как от размера входа задачи, так и от $1/\varepsilon$. А это значит, что предложенный алгоритм реализует схему FPTAS.

Заключение

В работе обоснован приближенный алгоритм для одной из NP-трудных в сильном смысле задач поиска в конечной последовательности точек евклидова пространства подпоследовательности, содержащей заданное число элементов. Показано, что предложенный алгоритм реализует схему FPTAS в случае фиксированной размерности пространства.

В алгоритмическом плане рассмотренная задача относится к числу слабоизученных проблем. Поэтому продолжение исследований алгоритмической аппроксимируемости задачи представляется важным делом ближайшей перспективы. В частности, значительный интерес представляют быстродействующие алгоритмы рандомизированного типа, которые позволяют находить решение с вероятностными гарантиями за линейное или сублинейное время. Интерес к подобным алгоритмам обусловлен необходимостью решения актуальной междисциплинарной проблемы большеразмерных данных.

Литература

1. **Кельманов А.В., Пяткин А.В.** О сложности некоторых задач выбора подпоследовательности векторов // Журн. вычисл. матем. и мат. физики. — 2012. — Т. 52, № 12. — С. 2284–2291.

2. **Кельманов А.В., Романченко С.М., Хамидуллин С.А.** Приближенные алгоритмы для некоторых труднорешаемых задач поиска подпоследовательности векторов // Дискретный анализ и исследование операций. — 2012. — Т. 19, № 3. — С. 27–38. — Перевод: Kel'manov A.V., Romanchenko S.M., and Khamidullin S.A. Approximation algorithms for some intractable problems of choosing a vector subsequence // J. Appl. Indust. Math. — 2012. — Vol. 6, № 4. — P. 443–450.

3. **Кельманов А.В., Романченко С.М., Хамидуллин С.А.** Точные псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подпоследовательности векторов // Журн. вычисл. матем. и мат. физики. — 2013. — Т. 53, № 1. — С. 143–153.
4. **Aggarwal C.C.** Data Mining: The Textbook. — Springer International Publishing, 2015.
5. **Tak-chung Fu.** A review on time series data mining // Engineering Applications of Artificial Intelligence. — 2011. — Vol. 24, iss. 11. — P. 164–181.
6. **Kuenzer C., Dech S., and Wagner W.** Remote Sensing Time Series. Remote Sensing and Digital Image Processing. Vol. 22. — Switzerland: Springer International Publishing, 2015.
7. **Warren Liao T.** Clustering of time series data — a survey // Pattern Recognition. — 2005. — Vol. 38, № 11. — P. 1857–1874.
8. **Кельманов А.В., Пяткин А.В.** NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. — 2010. — Т. 17, № 5. — С. 37–45. — Перевод: Kel'manov A.V., Pyatkin A.V. NP-completeness of some problems of choosing a vector subset // J. Appl. Indust. Math. — 2011. — Vol. 5, № 3. — P. 352–357.
9. **Aggarwal A., Imai H., Katoh N., and Suri S.** Finding k points with minimum diameter and related problems // J. Algorithms. — 1991. — Vol. 12. — P. 38–56.
10. **Кельманов А.В., Романченко С.М.** Приближенный алгоритм решения одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. — 2011. — Т. 18, № 1. — С. 61–69. — Перевод: Kel'manov A.V., Romanchenko S.M. An approximation algorithm for solving a problem of search for a vector subset // J. Appl. Indust. Math. — 2012. — Vol. 6, № 1. — P. 90–96.
11. **Шенмайер В.В.** Аппроксимационная схема для одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. — 2012. — Т. 19, № 2. — С. 92–100. — Перевод: Shenmaier V.V. An approximation scheme for a problem of search for a vector subset // J. Appl. Indust. Math. — 2012. — Vol. 6, № 3. — P. 381–386.
12. **Кельманов А.В., Романченко С. М.** Псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. — 2012. — № 2. — С. 156–162. — Перевод: Kel'manov A.V., Romanchenko S.M. Pseudopolynomial algorithms for certain computationally hard vector subset and cluster analysis problems // Automation and Remote Control. — 2012. — Vol. 73, № 2. — P. 349–354.
13. **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. — Перевод: Kel'manov A.V., Romanchenko S.M. An FPTAS for a vector subset search problem // J. Appl. Indust. Math. — 2014. — Vol. 8, № 3. — P. 329–336.
14. **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. — Перевод: Kel'manov A.V., Khamidullin S.A. Posterior detection of a given number of identical subsequences in a quasi-periodic sequence // Comput. Mathem. and Math. Phys. — 2001. — Vol. 41, № 5. — P. 762–774.

Поступила в редакцию 1 сентября 2016 г.,
в окончательном варианте 8 января 2017 г.

Литература в транслитерации

1. **Kel'manov A.V., Pyatkin A.V.** O slozhnosti nekotoryh zadach vybora podposledovatel'nosti vektorov // Zhurn. vychisl. matem. i mat. fiziki. — 2012. — T. 52, № 12. — S. 2284–2291.

2. **Kel'manov A.V., Romanchenko S.M., Hamidullin S.A.** Priblizhennye algoritmy dlya nekotoryh trudnoreshaemyh zadach poiska podposledovatel'nosti vektorov // Diskretnyy analiz i issledovanie operatsiy. — 2012. — T. 19, № 3. — S. 27–38. — Perevod: Kel'manov A.V., Romanchenko S.M., and Khamidullin S.A. Approximation algorithms for some intractable problems of choosing a vector subsequence // J. Appl. Indust. Math. — 2012. — Vol. 6, № 4. — P. 443–450.
3. **Kel'manov A.V., Romanchenko S.M., Hamidullin S.A.** Tochnye psevdopolinomial'nye algoritmy dlya nekotoryh trudnoreshaemyh zadach poiska podposledovatel'nosti vektorov // Zhurn. vychisl. matem. i mat. fiziki. — 2013. — T. 53, № 1. — S. 143–153.
4. **Aggarwal C.C.** Data Mining: The Textbook. — Springer International Publishing, 2015.
5. **Tak-chung Fu.** A review on time series data mining // Engineering Applications of Artificial Intelligence. — 2011. — Vol. 24, iss. 11. — P. 164–181.
6. **Kuenzer C., Dech S., and Wagner W.** Remote Sensing Time Series. Remote Sensing and Digital Image Processing. Vol. 22. — Switzerland: Springer International Publishing, 2015.
7. **Warren Liao T.** Clustering of time series data — a survey // Pattern Recognition. — 2005. — Vol. 38, № 11. — P. 1857–1874.
8. **Kel'manov A.V., Pyatkin A.V.** NP-polnota nekotoryh zadach vybora podmnozhestva vektorov // Diskretnyy analiz i issledovanie operatsiy. — 2010. — T. 17, № 5. — S. 37–45. — Perevod: Kel'manov A.V., Pyatkin A.V. NP-completeness of some problems of choosing a vector subset // J. Appl. Indust. Math. — 2011. — Vol. 5, № 3. — P. 352–357.
9. **Aggarwal A., Imai H., Katoh N., and Suri S.** Finding k points with minimum diameter and related problems // J. Algorithms. — 1991. — Vol. 12. — P. 38–56.
10. **Kel'manov A.V., Romanchenko S.M.** Priblizhennyy algoritm resheniya odnoy zadachi poiska podmnozhestva vektorov // Diskretnyy analiz i issledovanie operatsiy. — 2011. — T. 18, № 1. — S. 61–69. — Perevod: Kel'manov A.V., Romanchenko S.M. An approximation algorithm for solving a problem of search for a vector subset // J. Appl. Indust. Math. — 2012. — Vol. 6, № 1. — P. 90–96.
11. **Shenmayer V.V.** Approksimatsionnaya skhema dlya odnoy zadachi poiska podmnozhestva vektorov // Diskretnyy analiz i issledovanie operatsiy. — 2012. — T. 19, № 2. — S. 92–100. — Perevod: Shenmaier V.V. An approximation scheme for a problem of search for a vector subset // J. Appl. Indust. Math. — 2012. — Vol. 6, № 3. — P. 381–386.
12. **Kel'manov A.V., Romanchenko S. M.** Psevdopolinomial'nye algoritmy dlya nekotoryh trudnoreshaemyh zadach poiska podmnozhestva vektorov i klasternogo analiza // Avtomatika i telemekhanika. — 2012. — № 2. — S. 156–162. — Perevod: Kel'manov A.V., Romanchenko S.M. Pseudopolynomial algorithms for certain computationally hard vector subset and cluster analysis problems // Automation and Remote Control. — 2012. — Vol. 73, № 2. — P. 349–354.
13. **Kel'manov A.V., Romanchenko S.M.** FPTAS dlya odnoy zadachi poiska podmnozhestva vektorov // Diskretnyy analiz i issledovanie operatsiy. — 2014. — T. 21, № 3. — S. 41–52. — Perevod: Kel'manov A.V., Romanchenko S.M. An FPTAS for a vector subset search problem // J. Appl. Indust. Math. — 2014. — Vol. 8, № 3. — P. 329–336.
14. **Kel'manov A. V., Hamidullin S. A.** Aposteriornoe obnaruzhenie zadannogo chisla odinakovyh podposledovatel'nostey v kvaziperiodicheskoy posledovatel'nosti // Zhurn. vychisl. matematiki i mat. fiziki. — 2001. — T. 41, № 5. — S. 807–820. — Perevod: Kel'manov A.V., Khamidullin S.A. Posterior detection of a given number of identical subsequences in a quasi-periodic sequence // Comput. Mathem. and Math. Phys. — 2001. — Vol. 41, № 5. — P. 762–774.