

На правах рукописи

ДОЛГУШЕВ АЛЕКСЕЙ ВЛАДИМИРОВИЧ

**АЛГОРИТМЫ С ОЦЕНКАМИ
ДЛЯ РЕШЕНИЯ ЗАДАЧ АНАЛИЗА ДАННЫХ**

01.01.09 — дискретная математика и математическая кибернетика

А в т о р е ф е р а т
диссертации на соискание учёной степени
кандидата физико-математических наук

Новосибирск — 2012

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новосибирский государственный университет».

Научный руководитель: доктор физико-математических наук,
старший научный сотрудник,
Кельманов Александр Васильевич

Официальные оппоненты: доктор физико-математических наук,
профессор,
Гимади Эдуард Хайрутдинович

доктор физико-математических наук,
профессор,
Попков Владимир Константинович

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт математики и механики Уральского отделения Российской академии наук

Защита состоится "21" ноября 2012 г. в 16 часов на заседании диссертационного совета Д 003.015.01 при Федеральном государственном бюджетном учреждении науки Институте математики им. С. Л. Соболева Сибирского отделения РАН по адресу: 630090, г. Новосибирск, пр. Коптюга 4.

С диссертацией можно ознакомиться в библиотеке Института математики им. С. Л. Соболева Сибирского отделения РАН.

Автореферат разослан " __ " октября 2012 г.

Ученый секретарь
диссертационного совета,
доктор физико-математических наук,
старший научный сотрудник,

Шамардин Ю.В.

Общая характеристика работы

Актуальность работы¹. Объектом исследования настоящей работы являются проблемы оптимизации. Предмет исследования — дискретные экстремальные задачи, к которым сводятся актуальные проблемы помехоустойчивого анализа данных, распознавания образов и классификации. Цель исследования — анализ алгоритмической сложности этих задач и построение эффективных алгоритмов с гарантированными оценками точности для их решения.

Одной из наиболее известных экстремальных задач анализа данных и распознавания образов является задача MSSC (Minimum Sum-of-Squares Clustering) — кластеризации (разбиения) конечного множества векторов евклидова пространства по критерию минимума суммы квадратов расстояний. На протяжении нескольких десятилетий эта задача считалась NP-трудной. Однако совсем недавно² в доказательстве её труднорешаемости были обнаружены ошибки. В связи с этим снова стал актуальным анализ алгоритмической сложности этой задачи и её специальных случаев. Один из таких случаев проанализирован в настоящей работе.

К числу актуальных относится задача разбиения (по критерию минимума суммы квадратов расстояний) конечного множества векторов евклидова пространства на кластеры фиксированной мощности в случае, когда центр одного из кластеров определять не требуется (считается, что он фиксирован и равен нулю). Эта задача в постановочном плане близка к задаче MSSC, но не эквивалентна ей. Она важна для ряда приложений, связанных с помехоустойчивым анализом данных³. В настоящей работе изучается простейший в содержательном плане случай этой задачи, когда заданное множество требуется разбить на два кластера.

В приложениях, связанных с помехоустойчивой обработкой сигналов, актуальны задачи⁴ анализа и распознавания векторных последо-

¹Работа выполнена при поддержке грантов РФФИ (проекты № 09-01-00032, № 10-07-00195), а также ФЦП «Научные и научно-педагогические кадры инновационной России» (гос. контракт № 14.740.11.0362).

²Aloise D., Hansen P. On the Complexity of Minimum Sum-of-Squares Clustering // *Les Cahiers du GERAD*, G-2007-50. 2007. 12 p.

³Кельманов А.В. Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // *Тр. ИММ УрО РАН*. Екатеринбург. 2008. Т. 14. № 2. С. 81–88.

⁴Kel'manov A.V., Jeon B. A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // *IEEE Trans. on Signal Proc.*, 2004, Vol. 52, No. 3, pp. 645–656.

вательностей, структура которых в отсутствие помехи содержит ненулевые информационно значимые векторы, перемежающиеся с нулевыми векторами. В частных случаях эти задачи можно трактовать как задачи обнаружения разладки⁵ (изменения свойств) случайной последовательности. В тысячах публикаций рассматриваются разнообразные постановки подобных задач и последовательные (on-line) алгоритмы их решения, основанные на фундаментальной работе Вальда⁶. В то же время эффективные апостериорные (off-line) алгоритмы с гарантированными оценками точности для большинства из этих задач отсутствуют⁷. Несколько таких задач исследовано в настоящей работе.

Цель диссертационной работы — исследование задач кластеризации конечного множества векторов евклидова пространства, а также экстремальных задач связанных с помехоустойчивым анализом и распознаванием векторных последовательностей, структура которых характеризуется наличием повторяющихся информационно значимых векторов; построение эффективных алгоритмов с гарантированными оценками точности для решения этих задач.

Основные результаты:

1. Доказана NP-полнота задачи MSSC — кластеризации конечного множества векторов евклидова пространства по критерию минимума суммы квадратов расстояний — в случае, когда размерность пространства является, а число кластеров не является частью входа задачи.

2. Построен эффективный 2-приближённый алгоритм для решения NP-трудной задачи разбиения конечного множества векторов евклидова пространства на два кластера фиксированной мощности по критерию минимума суммы квадратов расстояний от элементов кластера до его центра в случае, когда центр одного из кластеров совпадает с началом координат.

3. Для этой же задачи обоснована приближённая полиномиальная схема (PTAS).

4. Разработаны полиномиальные алгоритмы, позволяющие находить оптимальное решение экстремальных задач, к которым сводится проблема помехоустойчивого off-line распознавания векторной последовательности как структуры, включающей повторяющийся вектор, при на-

⁵Ширяев А.Н. Статистический последовательный анализ // М.: Наука, 1976, 272 с.

⁶Wald A. Sequential analysis, New York: John Wiley, 1947. 224 p.

⁷Кельманов А.В. О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // 13-я Всероссийская конференция «Математические методы распознавания образов» (ММРО-13). Сборник докладов. М.: МАКС Пресс, 2007. – с. 261-264.

личии «посторонних» векторов-вставок: а) из произвольного множества ненулевых векторов; б) из заданного алфавита. Предполагается, что помеха является выборкой из нормального распределения с диагональной матрицей ковариаций, а критерием принятия решения является максимум функционала правдоподобия.

Научная новизна результатов состоит в следующем.

1. Факт NP-полноты указанного случая задачи MSSC установлен впервые.

2. Эффективные алгоритмы с константной оценкой точности, а также схема PTAS для решения задачи разбиения векторов евклидова пространства на два кластера фиксированной мощности в случае, когда не требуется определять центр одного из кластеров, до настоящего времени отсутствовали.

3. Полиномиальные off-line алгоритмы, гарантирующие оптимальное решение задач, к которым сводятся указанные выше проблемы анализа и распознавания последовательностей, предложены впервые.

Методы исследований. Основными инструментами исследований служили методы дискретной оптимизации, среднеквадратического приближения, математической статистики и полиномиальной сводимости. При обосновании приближенных алгоритмов применялась специальная техника, позволяющая находить гарантированную границу уклонения приближенного решения от оптимального. Эффективная разрешимость задач устанавливалась конструктивно (алгоритмически).

Теоретическая и практическая значимость. Работа носит теоретический характер. Её результаты могут быть использованы в математической теории распознавания образов и классификации, а также в теории дискретных экстремальных задач, например, в исследованиях, связанных с изучением алгоритмической сложности. Предложенные алгоритмы могут использоваться в компьютерных технологиях, ориентированных на решение прикладных задач.

На защиту выносятся совокупность результатов по исследованию алгоритмической сложности, а также эффективные алгоритмы с гарантированными оценками точности для задач дискретной оптимизации, к решению которых сводятся актуальные проблемы анализа данных, распознавания образов и классификации.

Личный вклад автора. Все выносимые на защиту результаты получены соискателем лично. Постановки проблем анализа данных и распознавания образов предложены научным руководителем. Подходы к анализу алгоритмической сложности и идеи алгоритмических решений редуцированных оптимизационных задач найдены совместно с научным

руководителем и соавтором. Доказательства утверждений выполнены соискателем. Конфликт интересов с соавторами отсутствует.

Апробация работы. Результаты диссертации докладывались на следующих международных и российских конференциях: XLV и XLVI международных студенческих конференциях «Студент и научно-технический прогресс» (г. Новосибирск, 2007, 2008), 15-й международной конференции «Проблемы теоретической кибернетики» (г. Казань, 2008), XIV Байкальской международной школе-семинаре «Методы оптимизации и их приложения» (г. Северобайкальск, 2008), международной конференции «Алгоритмический анализ неустойчивых задач» (г. Екатеринбург, 2008), международной конференции «Classification, Forecasting, Data Mining» (г. Варна, Болгария, 2009), IV Всероссийской конференции «Проблемы оптимизации и экономические приложения» (г. Омск, 2009), 14-й Всероссийской конференции «Математические методы распознавания образов» (г. Суздаль, 2009), Российской конференции «Дискретная оптимизация и исследование операций» (Алтай, 2010), 8-ой Международной конференции «Интеллектуализация обработки информации» (г. Пафос, Кипр, 2010), XIV Всероссийской конференции «Математическое программирование и приложения» (г. Екатеринбург, 2011), V Всероссийской конференции «Проблемы оптимизации и экономические приложения» (г. Омск, 2012), IX международной конференции «Интеллектуализация обработки информации» (г. Будва, Черногория, 2012). Кроме того, результаты работы обсуждались на семинарах кафедры Теоретической кибернетики Новосибирского государственного университета и на семинарах отдела Теоретической кибернетики Института математики им. С.Л. Соболева СО РАН.

Публикации. По результатам исследований опубликовано 16 работ. В их числе 4 статьи в журналах, рекомендованных ВАК РФ, 4 статьи в рецензируемых трудах конференций и 8 тезисов докладов.

Структура и объем диссертации. Диссертация изложена на 112 страницах и включает введение, две главы, заключение и список цитируемой литературы из 172 наименования.

Содержание работы

Во Введении обоснована актуальность темы диссертационной работы, приведено краткое изложение её содержания, сформулированы основные результаты, выносимые на защиту.

В Первой главе представлен обзор задач дискретной оптимизации

ции, возникающих в рамках оптимизационных моделей кластеризации и поиска подмножеств, а также подходов и методов их решения.

Анализируется известная задача MSSC и задача разбиения множества векторов евклидова пространства на два кластера при условии, что центр одного из кластеров совпадает с началом координат. Для одного из случаев первой задачи установлена NP-полнота. Для второй задачи предложен эффективный приближённый алгоритм решения с константной оценкой точности и приближённая полиномиальная схема решения.

В форме верификации свойств первая из этих задач формулируется следующим образом.

Задача MSSC. ДАНО: множество $\mathcal{V} = \{v_1, \dots, v_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$ и положительное число A . ВОПРОС: существуют ли такое разбиение множества $\mathcal{V}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \dots, \mathcal{C}_J$, что имеет место неравенство

$$\sum_{j=1}^J \sum_{v \in \mathcal{C}_j} \|v - \bar{v}(\mathcal{C}_j)\|^2 \leq A,$$

где $\bar{v}(\mathcal{C}_j) = \sum_{v \in \mathcal{C}_j} v / |\mathcal{C}_j|$, $j = 1, \dots, J$, — центр j -го кластера?

Одномерный вариант задачи разрешим за полиномиальное время⁸. Четыре возможных случая многомерного варианта этой задачи индуцируются комбинированием размерности пространства и числа кластеров как элементов, которые либо являются, либо не являются частью входа задачи. Ранее была установлена полиномиальная разрешимость одного⁹ и доказана¹⁰ NP-трудность¹¹ ещё двух из четырёх возможных случаев задачи. Сложность оставшегося случая устанавливает

Теорема 1.1. *Задача MSSC NP-полна в случае, когда размерность пространства является, а число кластеров не является частью входа задачи соответственно.*

Доказательство основано на полиномиальном сведении задачи J -MSSC к частному случаю задачи $(J + 1)$ -MSSC и NP-полноте задачи

⁸Rao M. Cluster Analysis and Mathematical Programming // *J. Amer. Statist. Assoc.* 1971. Vol. 66. P. 622–626.

⁹Inaba M., Katch N., Imai H. Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based Clustering // *Proc. Annual Symp. on Comput. Geom.* 1994. P. 332–339.

¹⁰Aloise D., Deshpande A., Hansen P., Popat P. NP-Hardness of Euclidean Sum-of-Squares Clustering // *Les Cahiers du GERAD*, G-2008-33. 2008. 4 p.

¹¹Mahajan M., Nimbhorkar P., Varadarajan K. The Planar k-means Problem is NP-Hard // *Lecture Notes in Computer Science*. 2009. Vol. 5431. P. 284–285.

2-MSSC¹⁰.

В этой же главе рассматривается
Задача VS¹² (Vector Subset). ДАНО: множество $\mathcal{Y} = \{y_1, \dots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$. НАЙТИ: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ мощности M такое, что целевая функция

$$S(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2, \quad (1)$$

где $\bar{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$, минимальна.

Задача VS NP-трудна в сильном смысле, так как целевая функция $S(\mathcal{C})$ минимальна тогда и только тогда, когда максимальна целевая функция NP-трудной¹³ в сильном смысле¹⁴ задачи MLSVS (Maximum Length of the Sum of Vectors in a Subset) поиска подмножества векторов с максимальной нормой суммы.

Для решения задачи MLSVS ранее были предложены: полиномиальный приближённый алгоритм¹³ (без априорной оценки точности), FPTAS¹⁴ для случая фиксированной размерности пространства, полиномиальный алгоритм¹⁵, гарантирующий нахождение оптимального решения для случая фиксированной размерности пространства, а также псевдополиномиальные алгоритмы¹⁶, гарантирующие оптимальность решения задачи в случае, когда компоненты векторов имеют целочисленные значения. Анализ этих результатов указывает на актуальность поиска таких приближенных алгоритмов решения задачи VS и MLSVS, сложность которых не зависит экспоненциально от размерности пространства. Для решения задачи VS в работе обоснован

¹²Кельманов А.В., Пяткин А.В. О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // *Журн. выч. мат. и мат. физ.* 2009, Т.49, №11. С.2059-2067.

¹³Гимади 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(25). С.55–74.

¹⁴Бабурин А.Е., Гимади Э.Х., Глебов Н.И., Пяткин А.В. Задача отыскания подмножества векторов с максимальным суммарным весом // *Дискретный анализ и исследование операций*. Сер. 2. 2007. Т.14, №1. С.32–42.

¹⁵Гимади Э.Х., Пяткин А.В., Рыков И.А. О полиномиальной разрешимости некоторых задач выбора подмножества векторов в евклидовом пространстве фиксированной размерности // *Дискретный анализ и исследование операций*. 2008. Т.15, №6. С.11–19.

¹⁶Гимади Э.Х., Глазков Ю.В., Рыков И.А. О двух задачах выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммой разности // *Дискретный анализ и исследование операций*. 2008. Т.15, №5. С.30–43.

Алгоритм $\mathcal{A}_1$.

Шаг 1. Для каждого $y \in \mathcal{Y}$, найдем подмножество $\mathcal{B}(y) \subseteq \mathcal{Y}$, состоящее из вектора y и $M - 1$ векторов, имеющих максимальные проекции на направление вектора y .

Шаг 2. Выберем тот вектор $y^* \in \mathcal{Y}$ и подмножество $\mathcal{B}(y^*) \subseteq \mathcal{Y}$, для которых целевая функция

$$D(\mathcal{B}(y), y) = \sum_{x \in \mathcal{B}(y)} \|x - y\|^2 + \sum_{x \in \mathcal{Y} \setminus \mathcal{B}(y)} \|x\|^2 \quad (2)$$

минимальна. Подмножество $\mathcal{B}(y^*)$ объявляем результатом работы алгоритма.

Теорема 1.2. *Алгоритм $\mathcal{A}_1$ находит 2-приближённое решение задачи VS за время $O(qN^2)$. Оценка 2 точности алгоритма достижима.*

В этой же главе для решения задачи VS предложена схема PTAS. Алгоритм, реализующий эту схему, заключается в следующем. Пусть t и s — натуральные параметры алгоритма.

Алгоритм $\mathcal{A}_2$.

Шаг 1. Найдём 2-приближённое решение $\mathcal{B}^*$ и значение целевой функции $S(\mathcal{B}^*)$ с помощью алгоритма $\mathcal{A}_1$. По значению целевой этого решения и параметру алгоритма s вычислим шаг $h = \frac{4}{s} \sqrt{S(\mathcal{B}^*)/M}$ дискретной сетки.

Шаг 2. Для каждого подмножества фиксированной мощности $t \in [2, N]$ множества $\mathcal{Y}$ построим линейную оболочку, которую дискретизируем с шагом h .

Шаг 3. Для каждого элемента z дискретной сетки, лежащего в шаре радиуса $2\sqrt{S(\mathcal{B}^*)/M}$, построим подмножество $\mathcal{F}(z) \subseteq \mathcal{Y}$, состоящее из M векторов, имеющих наибольшие проекции на направление вектора z .

Шаг 4. По формуле (2) вычислим значение функции $D(\mathcal{F}(z), z)$.

Шаг 5. Среди найденных на шагах 2, 3 подмножеств выберем подмножество $\mathcal{F}^*$, для которого значение функции S минимально. Подмножество $\mathcal{F}^*$ объявляем решением задачи VS.

Теорема 1.3. *Алгоритм $\mathcal{A}_2$ находит приближённое решение задачи VS с относительной погрешностью $\epsilon \leq 1/t + 8\zeta(t, s)$, где $\zeta(t, s) = \sqrt{t-1}/s + (t-1)/s^2$, за время $O(qN^{t+1}s^{t-1})$.*

Во Второй главе рассматриваются проблемы помехоустойчивого off-line анализа и распознавания векторных последовательностей по критерию минимума суммы квадратов уклонений и максимума правдоподобия, анализируются дискретные экстремальные задачи, к которым

сводится решение этих проблем, обосновываются эффективные алгоритмы нахождения оптимального решения редуцированных задач. Эти задачи входят в большое семейство¹⁷ дискретных экстремальных задач, к которым сводятся актуальные проблемы¹⁸ помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы.

Одна из возможных содержательных трактовок рассматриваемых проблем состоит в следующем. Сканирующий источник сообщений через канал связи с помехой передает данные об активном и пассивном состояниях основного и посторонних объектов в виде последовательности наборов измеряемых информационно важных характеристик. В пассивном состоянии значения всех компонент набора равны нулю. В активном состоянии значение хотя бы одной компоненты не равно нулю. Имеется конечная совокупность основных объектов. Каждому объекту соответствует уникальный набор измеряемых характеристик. На приёмную сторону поступает зашумлённая последовательность квазипериодически перемежающихся наборов, в которой кроме набора, соответствующего активному состоянию основного объекта, имеются ненулевые наборы-вставки, соответствующие активным состояниям посторонних объектов. Термин «квазипериодически» означает, что интервал времени между двумя последовательными ненулевыми наборами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Необходимо определить (распознать), от какого из основных объектов были получены данные, а также найти в последовательности повторяющийся набор, соответствующий его активному состоянию. Анализируется два случая: а) наборы-вставки являются произвольными ненулевыми наборами; б) наборы-вставки принадлежат заданному конечному алфавиту.

Этой содержательной проблеме соответствует модель анализируемых данных, в которой векторная последовательность $x_n \in \mathbb{R}^q$, $n \in \mathcal{N} = \{1, 2, \dots, N\}$, обладает свойством:

$$x_n = \begin{cases} u, & u \in \mathcal{U}, n \in \mathcal{K}; \\ v_n, & v_n \in \mathcal{V}, n \in \mathcal{L}; \\ 0, & n \in \mathcal{N} \setminus (\mathcal{K} \cup \mathcal{L}), \end{cases} \quad (3)$$

¹⁷<http://math.nsc.ru/~serge/qpsl>

¹⁸Kel'manov A.V., Mikhailova L.V., Khamidullin S.A. QPSLab System for Analysis and Recognition of Signals With a Quasiperiodic Structure // 9-th Intern. Conf. «Pattern Recognition and Image Analysis: New Information Technologies»: Conference Proceedings. Nizhni Novgorod, 2008. Vol. 1. p. 412-418.

где $\mathcal{K} \cup \mathcal{L} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$, $\mathcal{K} \cap \mathcal{L} = \emptyset$, $|\mathcal{K}| = K > 0$, $|\mathcal{L}| = L \geq 0$; $\mathcal{U}, \mathcal{V} \subset \mathbb{R}^q$, $\mathcal{U} \cap \mathcal{V} = \emptyset$. Предполагается, что $\mathcal{K} = \{n_{\mu_1}, \dots, n_{\mu_K}\}$, $\mathcal{L} = \{n_{\nu_1}, \dots, n_{\nu_L}\}$, где $\{\mu_1, \dots, \mu_K\}$ и $\{\nu_1, \dots, \nu_L\}$ — подмножества порядковых номеров из множества $\{1, \dots, M\}$, а элементы набора $\{n_1, \dots, n_M\}$ удовлетворяют ограничениям

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

где натуральные константы $T_{\min}$ и $T_{\max}$ задают допустимый интервал между ближайшими номерами двух ненулевых векторов последовательности (3). Доступной для обработки считается последовательность

$$y_n = x_n + e_n, \quad n \in \mathcal{N}, \quad (5)$$

где e_n — вектор помехи (ошибки измерения), независимый от x_n .

В этой модели вектор u соответствует набору характеристик активного состояния основного объекта в содержательной проблеме; v_n — наборам характеристик активного состояния «посторонних» объектов, набор $\{n_1, \dots, n_M\}$ — порядковым номерам ненулевых наборов в принятой последовательности, $\{\mu_1, \dots, \mu_K\}$ и $\{\nu_1, \dots, \nu_L\}$ — порядковым номерам ненулевых наборов от основного и «посторонних» объектов, соответственно. Константы $T_{\min}$ и $T_{\max}$ задают минимальный и максимальный интервалы между двумя последовательными наборами в принятой последовательности, соответствующими активным состояниям объектов (интервал квазипериодичности).

Оптимизационные модели содержательных проблем распознавания формулируются в виде задач минимизации функционала

$$S(u, \{v_n, n \in \mathcal{L}\}, n_1, \dots, n_M, \mu_1, \dots, \mu_K) = \sum_{n=1}^N \|y_n - x_n\|^2, \quad (6)$$

где структура последовательностей x_n и y_n , $n \in \mathcal{N}$, задаётся формулами (3), (4) и (5). В работе установлено, что к задачам минимизации функционала (6) сводятся статистические постановки проблем в случае, когда вектор e_n есть выборка из q -мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а критерием решения является максимум функционала правдоподобия.

Из модели (3)-(5) следует, что минимизация функционала (6) сводится к решению следующих оптимизационных задач.

РАСПОЗНАВАНИЕ(А). ДАНО: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, конечное множество $\mathcal{U} \subset \mathbb{R}^q$, неотрицательное целое число L и натуральные K , $T_{\min}$ и $T_{\max}$. НАЙТИ: вектор $u \in \mathcal{U}$, а также наборы

$\{n_1, \dots, n_M\} \subseteq \mathcal{N}$ и $\{\mu_1, \dots, \mu_K\} \subseteq \{1, \dots, M\}$, такие что

$$F_A(u, n_1, \dots, n_M, \mu_1, \dots, \mu_K) = \sum_{i=1}^K k(n_{\mu_i}, u) + \sum_{j=1}^L r(n_{\nu_j}) \rightarrow \max, \quad (7)$$

где

$$k(n, u) = 2(y_n, u) - \|u\|^2, \quad n \in \mathcal{N}, \quad u \in \mathcal{U}, \quad (8)$$

$$r(n) = \|y_n\|^2, \quad n \in \mathcal{N}, \quad (9)$$

при условии, что $M = K + L$, а элементы множества $\{n_1, \dots, n_M\}$ удовлетворяют ограничениям (4).

РАСПОЗНАВАНИЕ(Б). ДАНО: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, конечные множества $\mathcal{U} \subset \mathbb{R}^q$ и $\mathcal{V} \subset \mathbb{R}^q \setminus \mathcal{U}$, неотрицательное целое число L и натуральные K , $T_{\min}$ и $T_{\max}$. НАЙТИ: вектор $u \in \mathcal{U}$, а также наборы $\{n_1, \dots, n_M\} \subseteq \mathcal{N}$ и $\{\mu_1, \dots, \mu_K\} \subseteq \{1, \dots, M\}$, такие что

$$F_B(u, n_1, \dots, n_M, \mu_1, \dots, \mu_K) = \sum_{i=1}^K k(n_{\mu_i}, u) + \sum_{j=1}^L c(n_{\nu_j}) \rightarrow \max, \quad (10)$$

где $k(n, u)$ определяется формулой (8),

$$c(n) = \max_{v \in \mathcal{V}} (2(y_n, v) - \|v\|^2), \quad n \in \mathcal{N}, \quad (11)$$

при условии, что $M = K + L$, а элементы набора $\{n_1, \dots, n_M\}$ удовлетворяют ограничениям (4).

Установлено, что максимумы функций F_A и F_B находятся по правилу

$$\begin{aligned} & \max_{u, \{n_1, \dots, n_M\}, \{\mu_1, \dots, \mu_M\}} F_*(u, n_1, \dots, n_M, \mu_1, \dots, \mu_M) = \\ & \max_u \max_{\{n_1, \dots, n_M\}, \{\mu_1, \dots, \mu_M\}} \left(\sum_{i=1}^K k(n_{\mu_i}, u) + \sum_{j=1}^L l_*(n_{\nu_j}) \right) \end{aligned} \quad (12)$$

где $F_* = F_A$, $l_*(n) = r(n)$, $n \in \mathcal{N}$, для задачи РАСПОЗНАВАНИЕ(А), $F_* = F_B$, $l_*(n) = c(n)$, $n \in \mathcal{N}$, для задачи РАСПОЗНАВАНИЕ(Б). Для построения алгоритмов решения задач РАСПОЗНАВАНИЕ(А) и (Б) рассматривается следующая базовая экстремальная задача.

Задача В. ДАНО: числовые последовательности $g_1(n)$ и $g_2(n)$, $n \in \mathcal{N}$, неотрицательное целое число L и натуральные K , $T_{\min}$ и $T_{\max}$.

НАЙТИ: наборы $\{n_1, \dots, n_M\} \subseteq \mathcal{N}$ и $\{\mu_1, \dots, \mu_K\} \subseteq \{1, \dots, M\}$ такие, что

$$G(n_1, \dots, n_M, \mu_1, \dots, \mu_K) = \sum_{i=1}^K g_1(n_{\mu_i}) + \sum_{j=1}^L g_2(n_{\nu_j}) \rightarrow \max, \quad (13)$$

где $\{\nu_1, \dots, \nu_L\} = \{1, \dots, M\} \setminus \{\mu_1, \dots, \mu_K\}$, при условии, что $M = K + L$, а элементы набора $\{n_1, \dots, n_M\}$ удовлетворяют ограничениям (4).

Доказательство полиномиальной разрешимости этой задачи носит конструктивный характер и основывается на реализации схемы динамического программирования, обоснованной в следующем утверждении.

Лемма 2.1. Максимум целевой функции задачи В определяется по формуле

$$G_{\max} = \max_{n \in \omega_M(M)} \max_{t \in \{K, K+1, \dots, M\}} G(n, K, t, M), \quad (14)$$

а значения функции $G(n, K, t, M)$, $n \in \omega_M(M)$, $t \in \{K, K+1, \dots, M\}$, вычисляются по следующим рекуррентным формулам:

$$G(n, k, t, m) = \begin{cases} g_1(n), & \text{если } k = 1, t = 1, m = 1, n \in \omega_1(M), \\ g_1(n) + \max_{j \in \gamma_{m-1}^-(n)} H(j, m-1), & \text{если} \\ & k = 1, t = 2, \dots, M, m = t, n \in \omega_m(M), \\ g_1(n) + \max_{j \in \gamma_{m-1}^-(n)} \max_{s \in \{k-1, \dots, m-1\}} G(j, k-1, s, m-1), & \text{если} \\ & k = 2, \dots, K, t = k, \dots, M, m = t, n \in \omega_m(M), \\ g_2(n) + \max_{j \in \gamma_{m-1}^-(n)} G(j, k, t, m-1), & \text{если} \\ & k = 1, \dots, K, t = k, \dots, M, m = t+1, \dots, M, n \in \omega_m(M), \end{cases} \quad (15)$$

где

$$H(n, m) = \begin{cases} g_2(n), & \text{если } m = 1, n \in \omega_1(M), \\ g_2(n) + \max_{j \in \gamma_{m-1}^-(n)} H(j, m-1), & \text{если} \\ & m = 2, \dots, M, n \in \omega_m(M); \end{cases} \quad (16)$$

$$\omega_m(M) = [1 + (m-1)T_{\min}, N - (M-m)T_{\min}], \quad m = 1, \dots, M,$$

— интервал допустимых значений m -ой компоненты набора $\{n_1, \dots, n_M\}$,

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

– интервал допустимых значений $(t - 1)$ -ой компоненты набора $\{n_1, \dots, n_M\}$ при условии, что t -ая компонента равна n .

В следствии к этой лемме установлены формулы вычисления оптимальных наборов $\{\hat{n}_1, \dots, \hat{n}_M\}$ и $\{\hat{\mu}_1, \dots, \hat{\mu}_K\}$, которые опускаются здесь в виду громоздкости.

Алгоритм $\mathcal{A}_B$ решения базовой экстремальной задачи заключается в вычислении $G_{\max}$ по формулам из леммы 2.1 (прямой ход) и нахождении оптимальных наборов по формулам следствия (обратный ход).

Теорема 2.1. *Алгоритм $\mathcal{A}_B$ находит точное решение задачи B за время $O(NKM^2(T_{\max} - T_{\min} + 1))$.*

В работе построены алгоритмы $\mathcal{A}_{R1}$ и $\mathcal{A}_{R2}$ решения задач РАСПОЗНАВАНИЯ (А) и (Б). Эти алгоритмы включают в себя три шага: 1) вычисление функций $k(n, u) = g_1(n|u)$, $l_*(n) = g_2(n)$, $n \in \mathcal{N}$, ($l_*(n) = r(n)$, $n \in \mathcal{N}$, для задачи РАСПОЗНАВАНИЕ(А), $l_*(n) = c(n)$, $n \in \mathcal{N}$, для задачи РАСПОЗНАВАНИЕ(Б)) для каждого фиксированного $u \in \mathcal{U}$; 2) вычисления $G_{\max}(u)$, $u \in \mathcal{U}$, с помощью алгоритма $\mathcal{A}_B$; 3) вычисления $G_{\max} = \max_{u \in \mathcal{U}} G_{\max}(u)$ и соответствующего этому значению целевой функции вектора $\hat{u}$ и наборов $\{\hat{n}_1, \dots, \hat{n}_M\}$, $\{\hat{\mu}_1, \dots, \hat{\mu}_K\}$.

Установлена истинность следующих утверждений.

Теорема 2.2. *Алгоритм $\mathcal{A}_{R1}$ находит оптимальное решение задачи РАСПОЗНАВАНИЕ(А) за время $O(|\mathcal{U}|N(K(K+L)^2(T_{\max} - T_{\min} + 1) + q))$.*

Теорема 2.3. *Алгоритм $\mathcal{A}_{R2}$ находит оптимальное решение задачи РАСПОЗНАВАНИЕ(Б) за время $O(N(|\mathcal{U}|K(K+L)^2(T_{\max} - T_{\min} + 1) + (|\mathcal{U}| + |\mathcal{V}|)q))$.*

Значения $(T_{\max} - T_{\min} + 1)$, K и L не превышают N . Поэтому алгоритмы $\mathcal{A}_B$, $\mathcal{A}_{R1}$, $\mathcal{A}_{R2}$ полиномиальны.

В частном случае, когда $|\mathcal{U}| = 1$, алгоритмы $\mathcal{A}_{R1}$ и $\mathcal{A}_{R2}$ гарантируют оптимальность решения актуальных задач помехоустойчивого обнаружения заданного вектора в последовательности при наличии «посторонних» векторов-вставок: а) из произвольного множества ненулевых векторов; б) из заданного алфавита.

В Заключение перечислены основные результаты диссертационной работы, очерчен круг возможных теоретических и практических применений полученных результатов, обозначены перспективы для дальнейших исследований.

Автор выражает искреннюю благодарность и глубокую признательность своему научному руководителю Александру Васильевичу Кельманову, а также всем сотрудникам кафедры Теоретической кибернетики НГУ и лабораторий Анализа данных и Дискретных задач в исследова-

нии операций ИМ СО РАН за постоянное внимание к работе, ценные замечания, плодотворные дискуссии и поддержку.

Список литературы

- [1] Долгушев А.В., Кельманов А.В. К вопросу об алгоритмической сложности одной задачи кластерного анализа // «Дискретный анализ и исследование операций», 2010, том 17, № 2, с. 39-45.
- [2] A.V. Dolgushev, A.V. Kel'manov. On the Algorithmic Complexity of a Problem in Cluster Analysis // «Journal of Applied and Industrial Mathematics». 2011. Vol. 5, No.2, pp. 191-194.
- [3] Долгушев А.В., Кельманов А.В. Приближённый алгоритм для решения одной задачи кластерного анализа // «Дискретный анализ и исследование операций», 2011, том 18, № 2, с. 29-40.
- [4] A.V. Dolgushev, A.V. Kel'manov. An Approximation Algorithm for Solving a Problem of Cluster Analysis // «Journal of Applied and Industrial Mathematics», 2011, Vol. 5, No. 4, pp. 551-558.
- [5] Долгушев А.В., Кельманов А.В. Об одном варианте задачи обнаружения в числовой последовательности образца фрагмента среди квазипериодически перемежающихся фрагментов // Труды XIV Байкальской международной школы-семинара «Методы оптимизации и их приложения». Иркутск, Байкал, 2-8 июля 2008 г. Т. 1 (Математическое программирование): Иркутск, ИСЭМ СО РАН, 2008, с. 356-362.
- [6] Алексей Долгушев, Александр Кельманов. Об одной задаче распознавания последовательности, включающей повторяющийся вектор // Proceedings of the Intern. Conference «Classification, Forecasting, Data Mining» (CFDM 2009), June 22-July 2, 2009, Varna, Bugaria // Intern. Book Series «Information Science and Computing», No. 8 / Suppl. to the Intern. Journal, «Information Technologies and Knowledge», Vol. 3, 2009, pp. 91-96.
- [7] Долгушев А.В., Кельманов А.В. Алгоритм помехоустойчивого распознавания последовательности, включающей повторяющийся вектор, при наличии посторонних векторов-вставок из алфавита //

Математические методы распознавания образов: 14-я Всероссийская конференция (ММРО-14). Владимирская обл., г. Суздаль, 21-26 сентября 2009 г. : Сборник докладов. — М.: МАКС Пресс, 2009, с. 225-228.

- [8] Долгушев А.В. Обнаружение и идентификация заданного числа наборов эталонных фрагментов в квазипериодической последовательности // Тез. докл. XLV Международной научной студенческой конференции «Студент и научно-технический прогресс». Новосибирск, 10-12 апреля 2007, с. 216.
- [9] Долгушев А.В. Задача обнаружения и идентификации двух перемежающихся фрагментов в числовой последовательности // Тез. докл. XLVI Международной научной студенческой конференции «Студент и научно-технический прогресс». Новосибирск, 26-30 апреля 2008, с. 185.
- [10] Долгушев А.В., Кельманов А.В. Задача обнаружения и идентификации двух квазипериодически перемежающихся фрагментов в числовой последовательности // Тез. докл. 15-й международной конф. «Проблемы теоретической кибернетики» (Казань, 2-7 июня 2008). Под ред. Ю.И. Журавлева. — Казань: Отечество, 2008, с. 28.
- [11] Долгушев А.В. Об одном варианте задачи обнаружения образца фрагмента среди неизвестных перемежающихся фрагментов // Тез. докл. междунар. конф. «Алгоритмический анализ неустойчивых задач», посвященной 100-летию со дня рождения В.К. Иванова, 1-6 сентября 2008 г. — Екатеринбург: изд-во Уральского университета, 2008, с. 266-267.
- [12] Долгушев А.В., Кельманов А.В. Об одной задаче поиска вектора в векторном алфавите // Материалы IV Всероссийской конференции «Проблемы оптимизации и экономические приложения». Омск, 29 июня - 4 июля 2009, с. 123.
- [13] Долгушев А.В., Кельманов А.В. К вопросу о сложности задачи MSSC // Материалы Российской конференции «Дискретная оптимизация и исследование операций», DOOR-2010, Алтай, 27 июня - 3 июля 2010. — Новосибирск: Изд-во Института математики СО РАН, 2010, с. 186. <http://math.nsc.ru/conference/door2010/book.html>.

- [14] Долгушев А.В., Кельманов А.В. Приближённый алгоритм решения одной задачи кластерного анализа // Тез. докл. Всероссийской конференции «Математическое программирование и приложения». Екатеринбург, 2011, с. 84.
- [15] Долгушев А.В., Кельманов А.В., Шенмайер В.В. Аппроксимационная схема для одной задачи кластерного анализа // Материалы V Всероссийской конференции «Проблемы оптимизации и экономические приложения». Омск, 2012, с. 120.
- [16] Долгушев А.В., Кельманов А.В., Шенмайер В.В. Приближённая полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: 9-ая международная конференция (ИОИ-9). Черногория, г. Будва, 16-22 сентября, 2012 г.: Сборник докладов — М: МАКС Пресс, 2012 — с. 242–244.