Efficient approximation algorithms with performance guarantees for some discrete optimization problems in analysis and recognition of sequences

Alexander Kel'manov

Sobolev Institute of Mathematics
Siberian Branch of the Russian Academy of Sciences, Novosibirsk

10th International Conference
Intelligent Information Processing (IIP-10)
Crete, Greece,
October 6-10, 2014

Введение

За каждой содержательной проблемой анализа данных и распознавания образов скрывается задача дискретной оптимизации, её эффективное решение с теоретическими гарантиями по точности — ключ к решению проблемы

План доклада

Введение. Подходы к решению проблем.

Часть 1. Экстремальные задачи, индуцированные проблемами «обучения» распознаванию последовательностей (кластеризация, поиск похожих элементов).

Часть 2. Экстремальные задачи, моделирующие проблемы принятия решения о последовательности и поиска её структурных элементов (распознавание, обнаружение, сегментация и т.п.).

Заключение. Результативные методы, техники и приемы решения задач.

Введение

Предмет исследования —

задачи дискретной оптимизации, которые индуцируются, в частности, типовыми проблемами анализа, интерпретации и распознавания одномерных и многомерных последовательностей.

Цель исследования —

анализ вычислительной сложности этих задач, построение эффективных алгоритмов с гарантированными оценками точности для их решения и обзор последних достижений по исследованию этих задач.

Мотивация исследований —

наличие открытых математических проблем, слабая изученность задач и их актуальность для ряда математических и естественно-научных дисциплин, а также технических приложений.

Введение

Области приложений (истоки задач)

1. Проблемы аппроксимации.

2. Проблемы комбинаторной геометрии.

3. Статистические проблемы совместного оценивания и проверки гипотез по неоднородным выборкам, которые содержат данные из нескольких распределений, причем информация о принадлежности элементов выборки распределению отсутствует (недоступна).

4. Проблемы кластерного анализа, интерпретации данных (Data Mining), а также обучения компьютера (Machine Learning) распознаванию образов.

5. Прикладные проблемы технической и медицинской диагностики, мониторинга (геофизического, космического и др.), электронной разведки, биометрики и биоинформатики, эконометрики, криминалистики, обработки экспериментальных данных, анализа и распознавания сигналов и др.

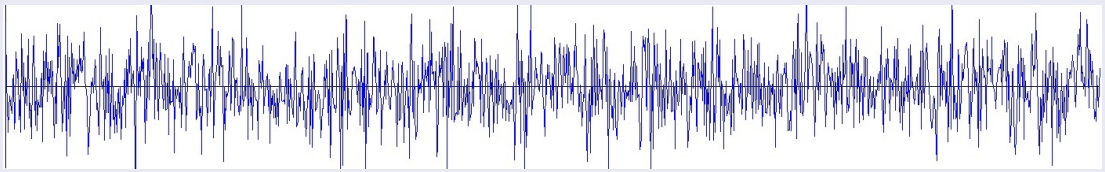
Введение

Типовые проблемы анализа данных и распознавания образов:

обнаружение, восстановление, оценивание, идентификация, разбиение, сегментация, кластеризация, обучение, принятие решения и т.п.

Пример типовой содержательной задачи

Дано: числовая последовательность (сигнал).



Найти: непериодически повторяющийся неизвестный фрагмент заданной размерности.

Задача типична для массы приложений.

Вопрос:

как решать эту простую в содержательном плане (но, как оказывается в действительности — NP-трудную) задачу?

Введение. Подходы к решению проблем

Типичный **поэтапный** подход

1. Фильтрация помех.
2. Обнаружение (принятие решения о «разладке», проверка гипотез).
3. Оценивание компонент вектор-фрагмента.

Суть этого подхода

1. Содержательная задача разбивается на последовательно решаемые подзадачи (этапы).
2. Каждая из этих подзадач решается подходящим известным (точным или приближенным, а зачастую эвристическим и без теоретических гарантий по точности) методом или алгоритмом.

Во многих случаях разбиение на подзадачи производится так, чтобы решение каждой подзадачи можно было найти либо одним из известных методов или алгоритмов, либо их незначительными модификациями.

Суть поэтапного подхода

Качество (точность и трудоемкость) поэтапного алгоритма решения задачи, очевидно, существенно зависит от варианта предлагаемого поэтапного разбиения задачи на подзадачи.

Поэтому подтверждение работоспособности поэтапного алгоритма, объединяющего решение всех подзадач и дающего решение проблемы в целом, получают:

1) в численных экспериментах;

2) на примерах решения конкретных прикладных (индивидуальных) задач (т.е. для узкого класса входных данных);

3) в соревновательных вычислительных экспериментах (конкурсах программ, реализующих предлагаемые алгоритмы).

Введение. Подходы к решению проблем

Суть поэтапного подхода

С использованием этого подхода во всем мире создано необозримое множество компьютерных технологий различного назначения.

Этот подход представлен в подавляющем числе докладов на международных и российских конференциях по анализу данных и распознаванию образов.

Суть поэтапного подхода

Однако, очевидно, что этот подход в общем случае **не гарантирует оптимальность** решения задачи в целом даже в случае оптимальности решений, получаемых на каждом из этапов или для каждой из подзадач.

Результат оптимизации, найденный по условным экстремумам, вычисленным на последовательно выполняемых этапах обработки данных, в общем случае может не совпадать с глобальным экстремумом.

Фундаментальный вопрос об априорно гарантированной точности поэтапного, по своей сути — приближенного, алгоритма, как правило, остается открытым.

Введение. Подходы к решению проблем

Альтернативный подход

Содержательная проблема не разбивается предварительно на последовательно решаемые подзадачи (этапы).

Суть подхода

1. Одновременное (совместное) решение всех подзадач в рамках адекватной оптимизационной модели содержательной проблемы анализа данных и распознавания образов.

2. Реализация этого подхода обуславливает решение специфических (как правило, труднорешаемых) дискретных экстремальных задач и построение полиномиальных алгоритмов с **теоретическими гарантиями** точности.

Введение. Подходы к решению проблем

Альтернативный подход

Совокупность существующих классических критериев решения задач в комбинации с многообразием реальных (практических) структур анализируемых данных порождает необозримое множество таких задач дискретной оптимизации.

Эти задачи имеют глубокую взаимосвязь с задачами теории приближения, геометрии, математической статистики, теории графов.

Пример, демонстрирующий эту взаимосвязь — классическая (Fisher, 1958) задача MSSC (Minimum Sum-of-Squares Clustering) или k-Means (Edwars, Cavalli-Sfarza, 1965) разбиения конечного множества векторов евклидова пространства.

Ниже представлены недавние результаты, полученные в ИМ СО РАН, по исследованию новых и известных, но слабо изученных задач, которые индуцируются оптимизационными моделями анализа данных и распознавания образов.

Задача 1-MSSC-S-NF (Minimum Sum-of-Squares Clustering, for the case of Sequence and NonFixed sizes of subsequences)

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

Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такой, что

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

где $\bar{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$, при ограничениях

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

на элементы искомого набора $\mathcal{M}$.

Имеется

таблица, содержащая упорядоченные по времени результаты многократных измерений набора числовых информационно значимых характеристик некоторого объекта, который может находиться в **пассивном** и **активном** состояниях.

Предполагается, что:

(1) в пассивном состоянии все числовые характеристики из набора равны нулю, а в активном — значение хотя бы одной характеристики не равно нулю;
(2) в каждом результате измерения, представленном в таблице, имеется ошибка;
(3) соответствие элементов таблицы состояниям объекта неизвестно.

(4) временной интервал между двумя последовательными активными состояниями объекта ограничен сверху и снизу некоторыми константами.

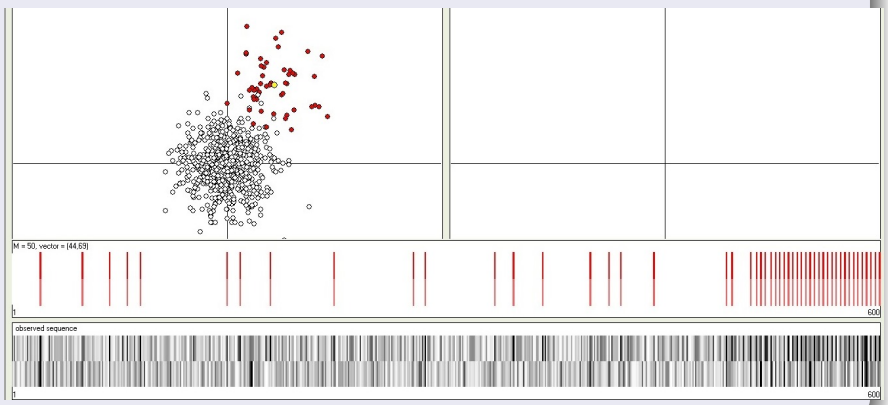
Требуется:

1. разбить последовательность (т.е. $\mathcal{Y}$), содержащую N q-мерных наборов, на подпоследовательности, соответствующие активному (с множеством номеров $\mathcal{M}$) и пассивному (с множеством номеров $\mathcal{N} \setminus \mathcal{M}$) состояниям объекта.

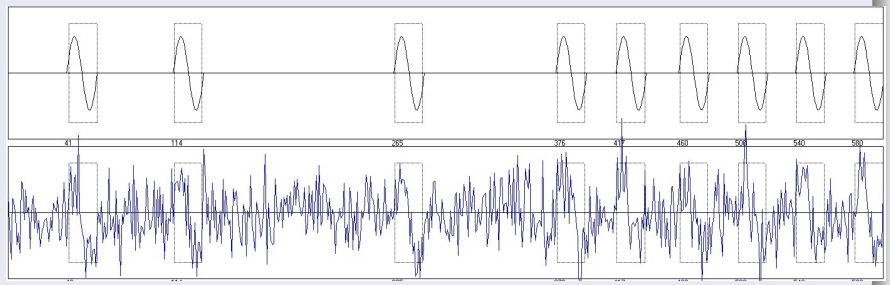
2. оценить набор (т.е. $\overline{y}(\mathcal{M})$) характеристик объекта в активном состоянии, учитывая, что данные содержат ошибку измерения.

Примеры результатов измерений характеристик объекта (в активном и пассивном состояниях).

Пример 1. Многомерный случай



Пример 2. Одномерный случай



Задача приближения

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ векторов из $\mathbb{R}^q$ (и, возможно, натуральные числа $T_{\min}$, $T_{\max}$).
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ и вектор $w \in \mathbb{R}^q$ такие, что

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

где

$$x_n = \begin{cases} w, & \text{если } n \in \mathcal{M}, \\ 0, & \text{если } n \in \mathcal{N} \setminus \mathcal{M}, \end{cases}$$

при ограничениях

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

на элементы искомого набора $\mathcal{M}$.

Известные результаты

1. Задача NP-трудна в сильном смысле. Поэтому для этой задачи не существует ни точного полиномиального, ни точного псевдополиномиального алгоритмов, если P≠NP (Кельманов, Пяткин, 2009).

2. Праметрический вариант — задача 1-MSSC-S-NF($T_{\min}$, $T_{\max}$) (Кельманов, Пяткин, 2013)

- NP-трудна в сильном смысле, когда $T_{\min} < T_{\max}$;
- разрешима за полиномиальное время при $T_{\min} = T_{\max}$.

3. Какие-либо полиномиальные алгоритмы с оценками точности до настоящего времени отсутствовали.

Новые результаты

1. Обоснован 2-приближенный полиномиальный алгоритм (Кельманов, Хамидуллин, 2014).

2. Установлено, что для этой задачи не существует полностью полиномиальной приближенной схемы (FPTAS), если P≠NP (Кельманов, 2014).

Суть подхода

1. Заменим решение исходной задачи 1-MSSC-S-NF решением более простой (вспомогательной) задачи.
2. Построим точный полиномиальный алгоритм её решения.
3. Оценим точность такой замены (приближения).

Подход опирается на записи целевой функции задачи 1-MSSC-S-NF в виде

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

$$= \sum_{j \in \mathcal{N}} \|y_j\|^2 - \sum_{i \in \mathcal{M}} \left(2(y_i, \overline{y}(\mathcal{M})) - \|\overline{y}(\mathcal{M})\|^2 \right).$$

Задача 1

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

Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов $\mathcal{Y}$ такой, что

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

при ограничениях

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

на элементы набора $\mathcal{M}$.

Алгоритм $\mathcal{A}$

Входами алгоритма являются $\mathcal{Y}$ (и, возможно, натуральные числа $T_{\min}$, $T_{\max}$).

Шаг 1. Для каждого вектора $b \in \mathcal{Y}$ найдем оптимальное решение $\widehat{\mathcal{M}}(b)$ и значение $G_{\max}(b)$ целевой функции задачи 1.

Шаг 2. Положим $b_A = \arg\max_{b \in \mathcal{Y}} G_{\max}(b)$, $\mathcal{M}_A = \mathcal{M}(b_A)$. Вычислим вектор $\overline{y}(\mathcal{M}_A) = \frac{1}{|\mathcal{M}_A|} \sum_{n \in \mathcal{M}_A} y_n$ и значение $F(\mathcal{M}_A)$ целевой функции; выход.

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

Замечание

Для реализации шага 1 предложена схема динамического программирования.

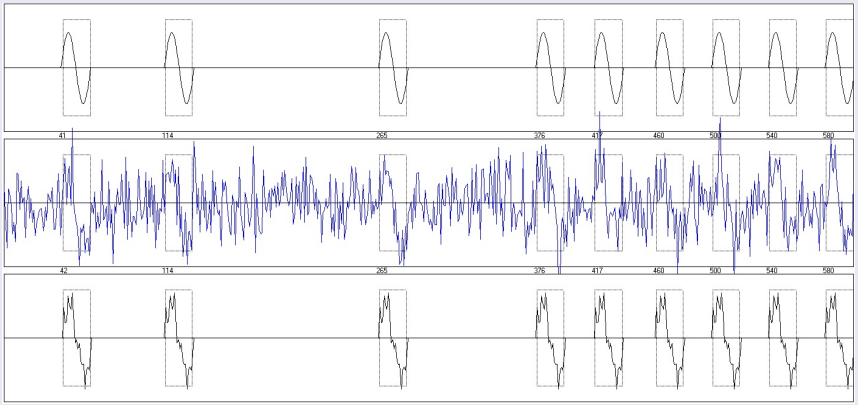
Теорема

Алгоритм $\mathcal{A}$ находит 2-приближенное решение задачи 1-MSSC-S-NF *за время* $\mathcal{O}(N^2((T_{\max} - T_{\min} + 1) + q))$.
Оценка 2 точности алгоритма асимптотически достижима.

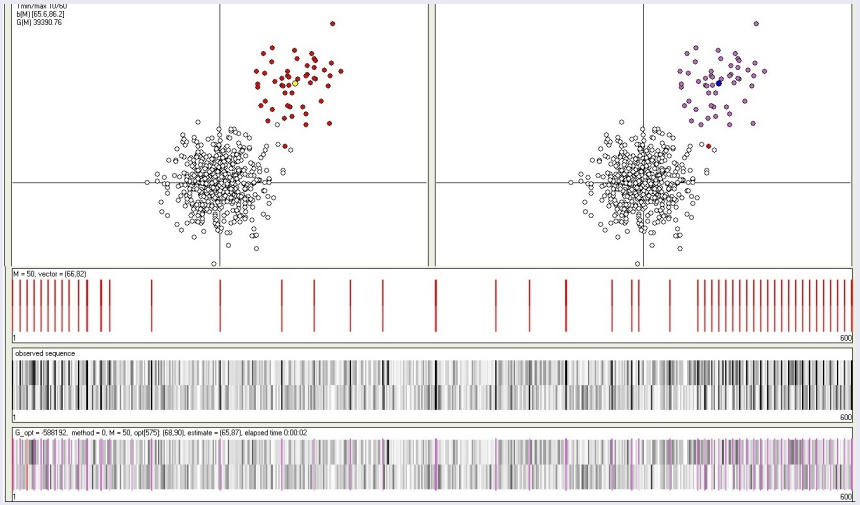
Замечание

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

Одномерный случай



Многомерный случай



Актуальные вопросы

1. Построение рандомизированного алгоритма.
2. Обоснование схемы PTAS.
3. Построение схемы FPTAS для специальных случаев задачи.
4. Обоснование точного псевдополиномиального алгоритма для специального случая задачи.
5. Разработка алгоритма для обобщения задачи на случай нескольких кластеров с оптимизируемыми центрами.

Задача 1-MSSC-F (Minimum Sum-of-Squares Clustering, Fixed cardinalities of sets)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.
Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ мощности M такое, что

$$Q(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min,$$

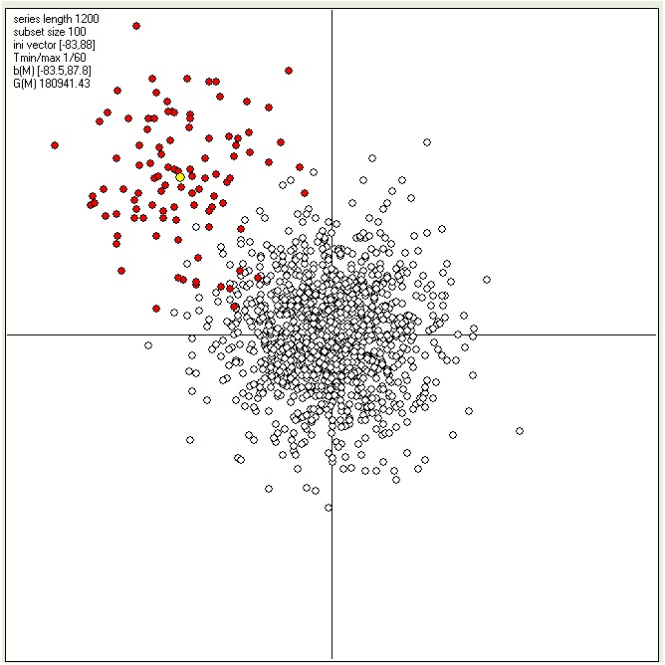
где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ — геометрический центр подмножества $\mathcal{C}$.

В этой задаче требуется разбить множество $\mathcal{Y}$ на 2 кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$. Центр кластера $\mathcal{Y} \setminus \mathcal{C}$ фиксирован в начале координат, а центр кластера $\mathcal{C}$ — оптимизируемая величина.

В обобщении — J-MSSC-F — этой задачи требуется разбить входное множество на $J + 1$ кластер. При этом геометрические центры J кластеров — оптимизируемые величины, центр одного из кластеров задан в начале координат, а мощности кластеров фиксированы (символ F).

Пример

1000 результатов измерений характеристик объекта, изображенные на плоскости.
100 раз были измерены характеристики объекта в активном состоянии и 900 раз — в пассивном.



Задача проверки гипотез

Дано: неоднородная выборка $\mathcal{Y} = \{y_1, \ldots, y_N\}$ из двух q-мерных гауссовских распределений (соответствие элементов выборки распределению неизвестно) с равными диагональными ковариационными матрицами и идентичными диагональными элементами.

Вопрос: верно ли, что M элементов выборки принадлежат к распределению с неизвестным средним, а $N - M$ элементов — к распределению с нулевым средним?

Решение этой задачи индуцирует задачу 1-MSSC-F.

Известные результаты

1. Задача NP-трудна в сильном смысле. Поэтому не существует ни точного полиномиального, ни точного псевдополиномиального алгоритмов, если P≠NP (Кельманов, Пяткин, 2008).

2. 2-приближённый алгоритм, временная сложность которого есть величина $\mathcal{O}(qN^2)$ (Долгушев, Кельманов, 2011).

3. Приближенная полиномиальная схема (PTAS), временная сложность которой $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — относительная погрешность (Долгушев, Кельманов, Шенмайер, 2012).

4. Рандомизированный алгоритм, позволяющий для установленного значения параметра при фиксированных ε и γ находить $(1+\varepsilon)$-приближённое решение с вероятностью 1-γ за время $\mathcal{O}(qN)$ (Кельманов, Хандеев, 2013).

5. Найдены условия, при которых алгоритм асимптотически точен и имеет временную сложность $\mathcal{O}(qN^2)$ (Кельманов, Хандеев, 2014).

Новые результаты

1. Установлено, что задача разрешима за время $\mathcal{O}(q^2 N^{2q})$, полиномиальное в случае, когда размерность q пространства фиксирована (Кельманов, Хандеев, 2014).

2. Точный псевдополиномиальный алгоритм для случая, когда компоненты векторов целочисленны, а размерность пространства фиксирована;
временная сложность алгоритма есть величина $\mathcal{O}(N(MD)^q)$; здесь D — максимальное абсолютное значение координат векторов входного множества;
алгоритм эффективнее известного полиномиального алгоритма при $MD < N^{2-\frac{1}{q}}$ (Кельманов, Хандеев, 2014).

3. Не существует полностью полиномиальной приближенной схемы (FPTAS), если P≠NP (Кельманов, 2014)

Суть подхода

Заменяем решение исходной труднорешаемой задачи 1-MSSC-F точным эффективным решением более простой вспомогательной задачи с последующей оценкой точности такой замены. Для этого:

1) аппроксимируем неизвестный центр одного из кластеров одним из узлов (векторов) специально построенной сетки (многомерной решетки) с рациональным значением шага. Шаг сетки подбираем так, чтобы искомый центр кластера совпал с одним из узлов сетки;

2) перебираем все векторы, соответствующие узлам построенной сетки, вычисляем значения целевой функции вспомогательной задачи во всех узлах и находим её оптимальное решение — узел сетки и подмножество, состоящее из M векторов, имеющих наибольшие проекции на направление, задаваемое этим узлом.

Пусть векторы из множества $\mathcal{Y}$ имееют целочисленные компоненты. Положим

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1,\ldots,q\}} |(y)^j|,$$

где $(y)^j$ — j-я координата вектора y.
Определим множество (совокупность узлов сетки)

$$\mathcal{D}_M = \{d \mid d \in \mathbb{R}^q, (d)^j = \frac{1}{M}(v)^j, (v)^j \in \mathbb{Z}, |(v)^j| \leq MD, j = 1, \ldots, q\}$$

векторов с рациональными координатами и целевую функцию

$$G(\mathcal{B}, b) = \sum_{y \in \mathcal{B}} \|y - b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2, \ \mathcal{B} \subseteq \mathcal{Y}, \ b \in \mathbb{R}^q$$

вспомогательной задачи.

Алгоритм $\mathcal{A}$

Вход алгоритма: множество $\mathcal{Y}$, натуральное число M.

Шаг 1. Для каждого вектора $b \in \mathcal{D}_M$ построим множество $\mathcal{B}(b)$, состоящее из M векторов множества $\mathcal{Y}$, имеющих наибольшие проекции на вектор b. Вычислим значение $G(\mathcal{B}(b), b)$.

Шаг 2. Найдём вектор $b_A = \arg\min\limits_{b \in \mathcal{D}} G(\mathcal{B}(b), b)$ и соответствующее ему подмножество $\mathcal{B}(b_A)$. В качестве решения задачи возьмём подмножество $\mathcal{C}_A = \mathcal{B}(b_A)$. Если решений несколько, то выберем любое из них.

Выход алгоритма: множество $\mathcal{C}_\mathcal{A}$.

Теорема

Пусть в условиях задачи 1-*MSSC-F векторы из множества* $\mathcal{Y}$ *имееют целочисленные компоненты из интервала* $[-D, D]$. *Тогда алгоритм* $\mathcal{A}$ *находит оптимальное решение задачи* 1-*MSSC-F за время* $\mathcal{O}(qN(2MD+1)^q)$.

Замечание

Если размерность q пространства фиксирована, то трудоёмкость алгоритма оценивается величиной $\mathcal{O}(N(MD)^q)$.
Время работы точного полиномиального алгоритма есть величина $\mathcal{O}(N^{2q})$. Поэтому предложенный псевдополиномиальный алгоритм более эффективен при $MD < N^{2-\frac{1}{q}}$.

Актуальные вопросы

1. Построение схемы FPTAS для специальных случаев задачи.

2. Обоснование алгоритма для обобщения задачи на случай нескольких кластеров с оптимизируемыми центрами.

3. Построение рандомизированного алгоритма для задачи 1-MSSC-NF, в которой мощности кластеров — оптимизируемые величины.

Квадратичная евклидова задача поиска подпоследовательности, содержащей заданное число "похожих"элементов (с одновременным **цензурированием** — удалением "мусора")

Задача VSS-2 (Vector Subsequence in a Sequence)

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ векторов из $\mathbb{R}^q$, натуральное число $M > 1$ (и, возможно, натуральные числа $T_{\min}$, $T_{\max}$).

Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов последовательности $\mathcal{Y}$ такой, что

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

где $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$, при ограничениях

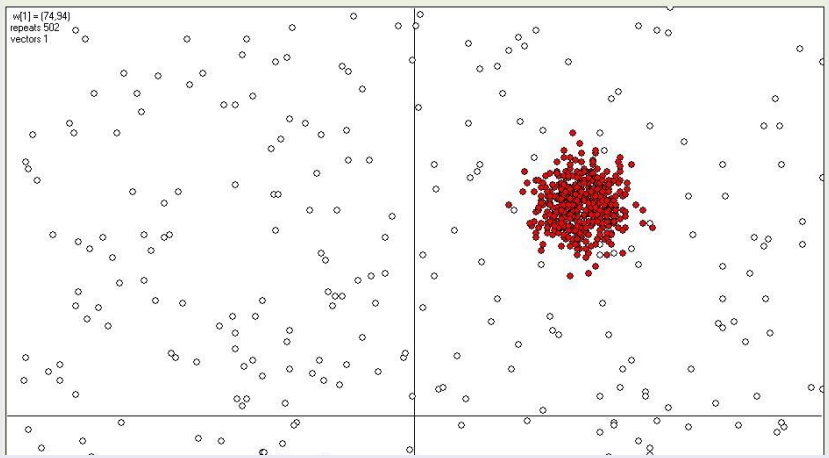
$$1 \leq T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \ \ m = 2, \ldots, M,$$

на элементы искомого набора $\mathcal{M}$.

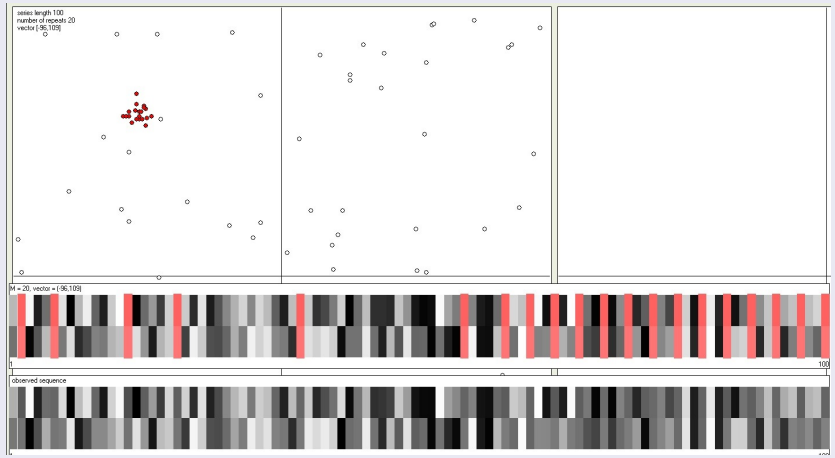
Часть I. 3. Задача VSS-2 поиска подпоследовательности.

Примеры

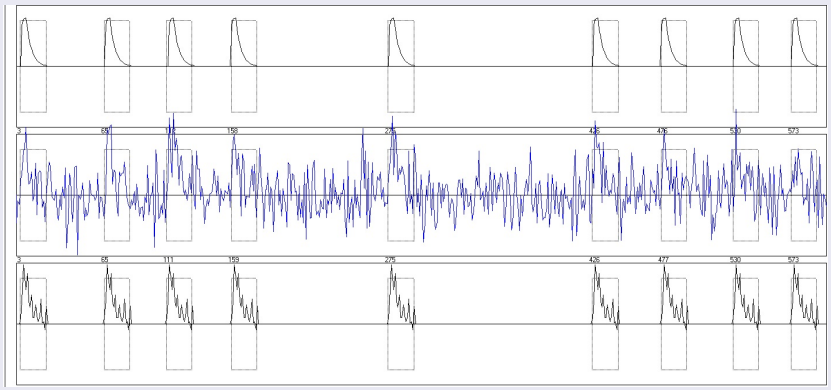
П1. Поиск подмножества



П2. Поиск подпоследовательности



П3. Одномерный случай



Задача проверки гипотез

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ q-мерных векторов.

Вопрос: верно ли, что $M > 1$ элементов этой последовательности принадлежат к одному и тому же гауссовскому распределению с неизвестным средним и диагональной ковариационной матрицей с идентичными диагональными элементами?

Решение этой задачи индуцирует задачу VSS-2.

Известные результаты

1. Задача NP-трудна в сильном смысле. (Кельманов, Пяткин, 2010).

2. Праметрический вариант — задача VSS-2($T_{\min}$, $T_{\max}$) (Кельманов, Пяткин, 2012)

- NP-трудна в сильном смысле, когда $T_{\min} < T_{\max}$;
- разрешима за полиномиальное время при $T_{\min} = T_{\max}$.

3. 2-приближённый алгоритм; временная сложность $\mathcal{O}(qN^2)$ (Кельманов, Романченко, 2011).

4. Схема (PTAS) для частного случая (поиска подмножества); временная сложность $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — относительная погрешность (Шенмайер, 2012).

5. Точный псевдополиномиальный алгоритм для случая, когда компоненты векторов целочисленны, а размерность пространства фиксирована; временная сложность алгоритма $\mathcal{O}(N(MD)^q)$; здесь D — максимальное абсолютное значение координат векторов входной последовательности; (Кельманов, Романченко, 2013).

Напомним формулировки классических (взвешенной и невзвешенной) задач Max Cut.

Взвешенная задача Max Cut (оптимизационная версия)

Вход: Полный неориентированный граф $G = (V, E)$, неотрицательные веса на ребрах $w(e) \geq 0$, $e \in E$.

Найти: Разрез максимального веса, то есть разбиение множества вершин V на два подмножества такое, что суммарный вес ребер между этими частями максимален.

Взвешенная задача Max Cut (верификационная версия)

Вход: Полный неориентированный граф $G = (V, E)$, неотрицательные веса на ребрах $w(e) \geq 0$, $e \in E$, число $K > 0$.

Вопрос: Содержит ли граф G разрез веса не меньше, чем K?

NP-полнота установлена в работе [Karp, 1972]. В классическом списке Карпа NP-полных задач числится под номером 21.

Невзвешенная задача соответствует случаю, когда $w(e) \in \{0, 1\}$, $e \in E$.

Невзвешенная задача Max Cut (оптимизационная версия)

Вход: Неориентированный граф $G = (V, E)$.

Найти: Разрез максимального размера, то есть разбиение множества вершин V на два подмножества такое, что сумммарное число ребер между этими частями максимально.

Невзвешенная задача Max Cut (верификационная версия)

Вход: Неориентированный граф $G = (V, E)$, положительное число K.

Вопрос: Содержит ли граф G разрез размера не меньше, чем K?

NP-полнота установлена в работе [Garey, Johnson, Stockmeyer, 1976].

Геометрические постановки (частные случаи) задач на графах

Задачи на полных графах со взвешенными ребрами имеют естественные геометрические постановки: вершины графа задаются точками некоторого пространства, веса (или длины) ребер расстояниями или функциями от расстояний между точками. Задачи с квадратами расстояний актуальны в помехоустойчивом анализе данных.

Метрический случай **Max Cut**

Известно, что задача **Max Cut** в случае произвольного метрического пространства NP-трудна в сильном смысле [de la Vega, Kenyon 2001].

Задача Euclidean Max Cut

Вход: множество $\mathcal{X} = \{x_1, \ldots, x_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{X}$ на два подмножества $\mathcal{Y}$ и $\mathcal{Z}$ таких, что

$$f(\mathcal{Y}, \mathcal{Z}) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \|y - z\| \longrightarrow \max.$$

Сложностной статус — открытый вопрос [Bern, Eppstein 1997].

Задача Quadratic Euclidean Max Cut

Дано: множество $\mathcal{X} = \{x_1, \ldots, x_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{X}$ на два подмножества $\mathcal{Y}$ и $\mathcal{Z}$ таких, что

$$g(\mathcal{Y}, \mathcal{Z}) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \|y - z\|^2 \longrightarrow \max.$$

Сложностной статус также не был установлен.

Часть I. 4. Известные результаты

В [Inaba, Katoh & Imai 1994] заявлено, что при фиксированной размерности q пространства **Quadratic Euclidean Max Cut** разрешима за время $O(N^q)$.

Line Max Cut и **Line Max Bisection** разрешимы за время $O(N^4)$ [Karpinski, Lingas, Sledneu 2012].

Max Cut полиномиально разрешима с точностью 0.878 [Goemans, Williamson 1995].

Max Cut APX-трудна [Papadimitriou & M. Yannakakis 1991]. Предел неаппроксимируемости для **Max Cut** $16/17 \approx 0.941176$ [Hastad 2001].

$(p, n-p)$ **Max Cut** полиномиально разрешима с точностью 0.5 [Ageev, Sviridenko 1998].

$(p, n-p)$ **Max-DiCut** полиномиально разрешима с той же точностью 0.5 [Ageev, Hassin, Sviridenko 2001].

Для **Metric Max Cut** существует рандомизированная полиномиальная приближенная схема [de la Vega & Kenyon 2001].

Часть I. 4. Евклидовы задачи о разрезе максимального веса. Новые результаты — решение открытых проблем

Теорема (Агеев, Кельманов, Пяткин, 2013-2014)

Задачи ***Euclidean Max Cut*** и ***Quadratic Euclidean Max Cut*** NP-трудны в сильном смысле.

При доказательстве построены полиномиальные сведения NP-трудной в сильном смысле [Bui et al. 1987] задачи *Minimum Bisection* в случае кубических графов к исследуемым задачам.

Задача Minimum Bisection

Вход: неориентированный граф G.
Найти: разбиение множества вершин графа G на две части равного размера (бисекцию) такое, что число ребер между этими частями минимально.

Утверждение (Агеев, Кельманов, Пяткин, 2013-2014)

Существование FPTAS как для ***Quadratic Euclidian Max Cut*** так и для ***Euclidean Max Cut*** влечёт P=NP.

Часть I. 4. Евклидовы задачи о разрезе максимального веса

Выводы

Из полученных результатов следует, что для рассмотренных геометрических случаев задачи **Max Cut** не существует точных полиномиальных и псевдополиномиальных алгоритмов, а также полностью полиномиальных приближенных схем в предположении P$\neq$NP.

Таким образом, евклидовы случаи задачи **Max Cut** столь же сложны с точки зрения аппроксимируемости, как и общий случай.

Открытые вопросы

— Сложностной статус евклидовых задач **Max Cut** при фиксированной размерности пространства.

— Построить детерминированную PTAS для **Metric Max Cut** или хотя бы для **Euclidean Max Cut**.

— Построить PTAS для **Quadratic Euclidean Max Cut**.

Формулировка общего случая задачи:

Задача m-Weighted Clique Problem (m-WCP)

Дано: полный неориентированый взвешенный граф $G = (V, E, a, c)$, где $a\colon V \to \mathbb{R}$, $c\colon E \to \mathbb{R}$, и натуральные числа $L_1, \ldots, L_m$ такие, что $\sum_{i=1}^{m} L_i \leqslant n$ (здесь $n = |V|$).

Найти: в графе G семейство $\mathcal{C} = \{C_1, \ldots, C_m\}$ дизъюнктных клик порядков $L_1, \ldots, L_m$ с минимальным суммарным весом вершин и ребер графа, входящих в эти клики.

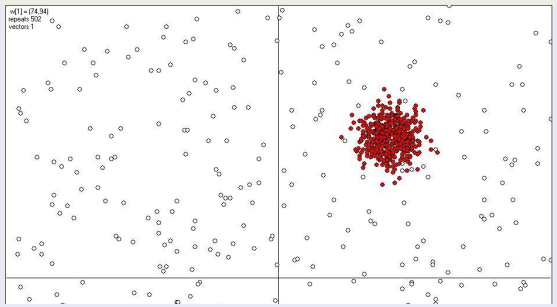
Особый интерес представляют специальные

геометрические случаи этой задачи:

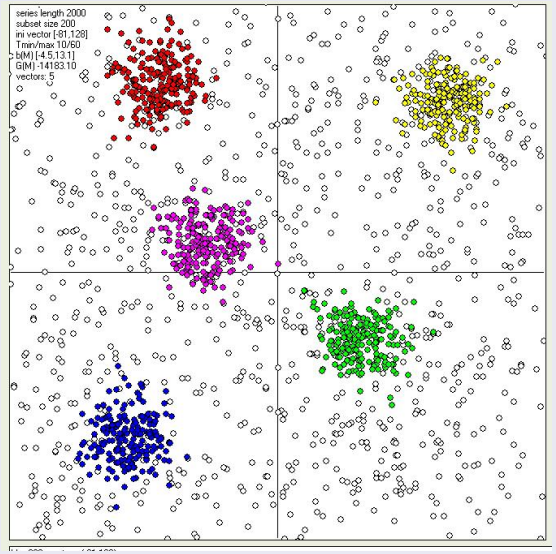
1. **Quadratic Euclidean m-WCP**,
когда веса ребер — квадраты расстояний между элементами некоторого семейства точек евклидова пространства;

2. **Metric m-WCP**,
когда веса ребер удовлетворяют неравенству треугольника.

Пример 1



Пример 2



В анализе данных и распознавании образов нередки ситуации, когда входом проблемы является матрица **попарных сравнений** объектов. При этом требуется найти семейство, состоящее из непересекающихся подмножеств "похожих"объектов, отбросив результаты сравнений с произвольными (случайными или "ошибочными") объектами, которые трактуются как "мусор". Обобщение этой проблемы на случай, когда на входе дополнительно заданы веса объектов, моделирует задача m-Weighted Clique Problem.

В простейших случаях этих задач — **Quadratic Euclidean WCP** и **Metric WCP** — требуется найти **одно** L-элементное подмножество — клику с минимальной суммой весов ребер.

Известные результаты для Quadratic Euclidean WCP и Metric WCP

1. Оба геометрических случая задачи NP-трудны в сильном смысле (Еремин, Гимади, Кельманов, Пяткин, Хачай, 2013).

2. Для этих случаев задач обоснован 2-приближенный алгоритм, временная сложность которого есть величина $\mathcal{O}(n^2)$ (Еремин, Гимади, Кельманов, Пяткин, Хачай, 2013).

Новые результаты для обобщений этих задач — Quadratic Euclidean m-WCP и Metric m-WCP

1. Обе задачи NP-трудны в сильном смысле как в случае, когда m — часть входа задачи (что очевидно), так и в случае, когда m фиксировано (что доказано).

2. Доказано, что исходные задачи **Quadratic Euclidean m-WCP** и **Metric m-WCP** можно записать в эквивалентном виде для полных реберно взвешенных графов с нулевыми весами вершин и модифицированными весами ребер.

3. Построен приближенный алгоритм решения этих задач, временная сложность которого есть величина $\mathcal{O}(n^{m+2}\log n)$. При фиксированном m алгоритм полиномиален и имеет установленные (см. далее) гарантированные достижимые оценки точности.

(авторы результатов — Гимади, Кельманов, Пяткин, Хачай, 2014)

Теорема

Задачи **Quadratic Euclidean** m-**WCP** и **Metric** m-**WCP** NP-трудны в сильном смысле.

При доказательстве показано, что к этим задачам полиномиально сводятся NP-трудные в сильном смысле задачи Quadratic Euclidean WCP и Metric WCP.

Подход к построению алгоритма

1. Преобразуем исходную матрицу весов ребер вместе с весами вершин в матрицу весов ребер графа (с модифицированными весами ребер и нулевыми весами вершин).

2. Заменяем решение исходных труднорешаемых задач точным эффективным решением более простой вспомогательной задачи с последующей оценкой точности такой замены.

Вспомогательная задача

.

Дано: полный n-вершинный неориентированный взвешенный граф G и натуральные числа $L_1, \ldots, L_m$ такие, что $\sum_{i=1}^{m} L_i \leqslant n$.

Найти в графе G семейство $\mathcal{B} = \{B_1, \ldots, B_m\}$ из m дизъюнктных звезд такое, что $|B_k| = L_k$, $k = 1, \ldots, m$, и

$$\mathcal{S}(\mathcal{B}) = \sum_{k=1}^{m} L_k S(B_k) \to \min,$$

где $S(B_k)$ — вес k-й звезды.

Алгоритм $\mathcal{A}$

Входами алгоритма являются:
1) набор весов объектов,
2) матрица попарных сравнений объектов, либо матрица квадратов попарных расстояний между объектами,
3) размеры m искомых клик.

Шаг 1. Преобразуем входной набор данных в матрицу, состоящую из весов ребер модифицированного полного графа.

Шаг 2. Для каждой (из $\mathcal{O}(n^m)$) допустимой комбинации центров звезд находим точное решение вспомогательной задачи — отыскания дизъюнктных звезд с помощью известного алгоритма (Кляйншмидт, Шаннатх, 1995) транспортного типа за время $\mathcal{O}(n^2 \log n)$. В семействе найденных решений выбираем наилучшее.

Выходом алгоритма (решением задачи) объявляем совокупности вершин, образующих найденные звезды.

Теорема

Алгоритм $\mathcal{A}$ находит решение задач за время $\mathcal{O}(n^{m+2}\log n)$ и имеет:
1) для задачи **Metric** m-**WCP** — достижимую оценку точности

$$2\left(1-\frac{\sum_{k=1}^{m} S(B_k^*)}{\sum_{k=1}^{m} L_k S(B_k^*)}\right),$$

где $S(B_k^*)$ — суммарный вес вершин и ребер в k-й звезде вспомогательной задачи, $k=1,\ldots,m$;
2) для задачи **Quadratic Euclidean** m-**WCP** — достижимую оценку точности, равную 2.

(авторы результатов — Гимади, Кельманов, Пяткин, Хачай, 2014)

Замечание

При фиксированном m алгоритм полиномиален.

Часть I. 5. Геометрические задачи поиска нескольких клик

Открытые вопросы

— Сложностной статус евклидовой задачи Euclidean m-WCP.

— Алгоритмы с гарантированными оценками точности в для общего случая геометрических задач, а также для других специальных случаев этих задач.

Часть II. Проблемы принятия решения о последовательности и поиска её структурных элементов: распознавание, обнаружение, сегментация... (в научно-популярной форме)

В этой части доклада речь пойдет о нескольких десятках экстремальных задач, моделирующих указанные проблемы.

Основная трудность в отыскании решения этих задач — **мощность** множества допустимых решений **растет экспоненциально** с увеличением размера их входа. Этот факт пугает прикладников, предпочитающих решать задачи поэтапно.

Тем не менее, для всех рассмотренных далее задач нами **конструктивно показана их полиномиальная разрешимость**, т.е. построены эффективные алгоритмы, гарантирующие отыскание оптимального решения.

Эти алгоритмы лежат в основе созданных компьютерных технологий и успешно применяются для решения прикладных задач (как в России, так и за рубежом).

Часть II. Проблемы принятия решения о последовательности и поиска её структурных элементов

Основные шаги решения проблем

1. Каждая из содержательных проблем формулируется в виде задачи приближения входной последовательности гипотетической (модельной) последовательностью (по критерию минимума суммы квадратов уклонений или максимума правдоподобия).

2. Задача приближения решается аналитически.
В результате выявляется задача комбинаторной оптимизации (целевая функция), индуцированная содержательной проблемой.

Здесь же из модели выводятся аналитически (а не вводятся искусственно) «меры похожести» или «меры различия» между распознаваемыми объектами. Эти меры разнообразны (расстояние, скалярное произведение, функции над ними) и существенно зависят от решаемой проблемы.

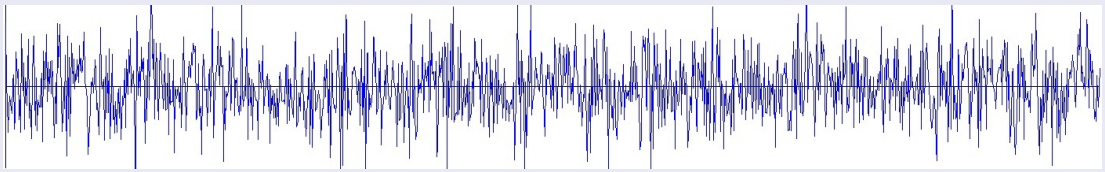
3. Для выявленной экстремальной задачи строится полиномиальный алгоритм решения и доказываются, что он доставляет оптимум соответствующей целевой функции.

Часть II. 1. Обнаружение элементов (векторов или фрагментов) последовательности, похожих на заданный шаблон

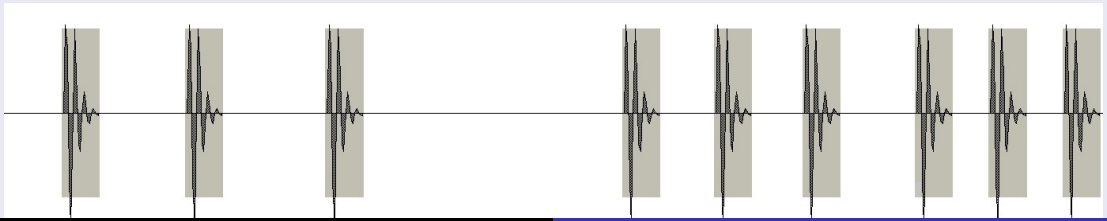
Дано: зашумленная числовая или векторная последовательность и шаблон (вектор).
Найти: все места расположения этого шаблона в последовательности.

В этой задаче мощность допустимого множества — $\mathcal{O}(N^M)$.

Входная (наблюдаемая) последовательность



Одна из множества допустимых аппроксимирующих (ненаблюдаемых) последовательностей (решений)



Часть II. 1. Поиск фрагментов, похожих на заданный шаблон. Число фрагментов **задано**

Задача поиска фрагментов по образцу

Дано: числовая последовательность $\mathcal{Y} = (y_0, \ldots, y_{N-1})$, вектор $U \in \mathbb{R}^q$, натуральное число $M > 1$ (и, возможно, $T_{\min}$, $T_{\max}$).
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\}$ номеров элементов последовательности $\mathcal{Y}$ такой, что

$$\sum_{j \in \mathcal{M}} (Y_j, U) \to \max,$$

где

$$Y_j = (y_j, \ldots, y_{j+q-1}), \;\; j = 0, \ldots, N-q,$$

при ограничениях

$$0 \le n_1 \le N - q, \quad 0 \le n_M \le N - q,$$

$$0 < q \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N - q, \;\; m = 2, \ldots, M,$$

на элементы набора $\mathcal{M}$.

Часть II. 1. Поиск фрагментов, похожих на заданный шаблон. Число фрагментов **не задано**

Задача поиска фрагментов по образцу

Дано: числовая последовательность $\mathcal{Y} = (y_0, \ldots, y_{N-1})$, вектор $U \in \mathbb{R}^q$ (и, возможно, натуральные числа $T_{\min}$, $T_{\max}$).
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\}$ номеров элементов последовательности $\mathcal{Y}$ такой, что

$$\sum_{j \in \mathcal{M}} \{2(Y_j, U) - \|U\|^2\} \to \max,$$

где

$$Y_n = (y_n, \ldots, y_{n+q-1}), \quad n = 0, \ldots, N-q,$$

при ограничениях

$$0 \le n_1 \le N - q, \quad 0 \le n_M \le N - q,$$

$$0 < q \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N - q, \ m = 2, \ldots, M,$$

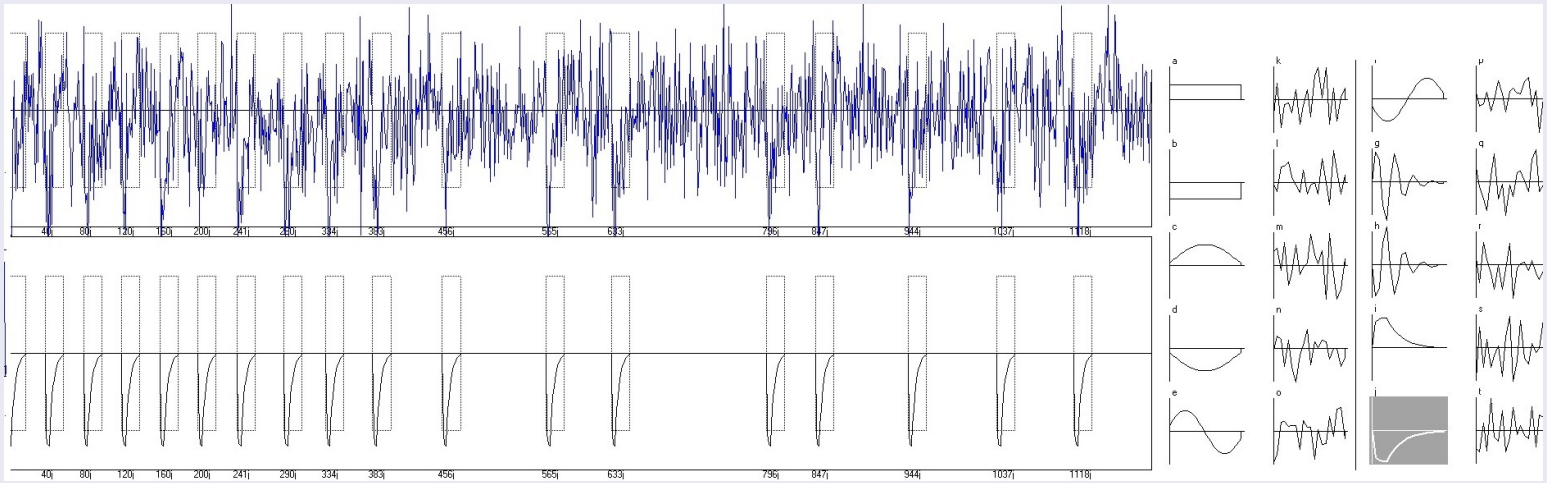
на элементы набора $\mathcal{M}$.

Часть II. 2. Распознавание последовательности как структуры, порожденной непериодически повторяющимся вектор-шаблоном из заданного алфавита

Дано: зашумленная числовая или векторная последовательность и алфавит шаблонов (векторов).
Найти: шаблон в алфавите, который непериодически повторяется в этой последовательности.

В этой задаче мощность допустимого множества — $\mathcal{O}(KN^M)$.

Поиск шаблона в алфавите

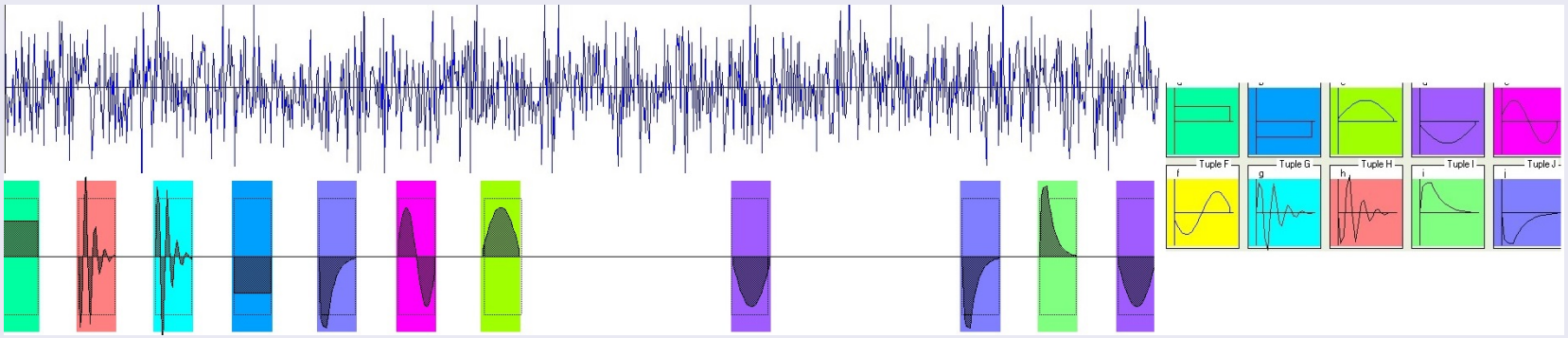


Часть II. 3. Совместное обнаружение и идентификация вектор-фрагментов в последовательности как элементов из заданного алфавита вектор-шаблонов

Дано: зашумленная числовая или векторная последовательность и алфавит шаблонов (векторов).
Найти: шаблоны из алфавита, которые непериодически встречаются в этой последовательности и их расположение.

В этой задаче мощность допустимого множества — $\mathcal{O}(N^M \times K^M)$.

Поиск шаблона в алфавите

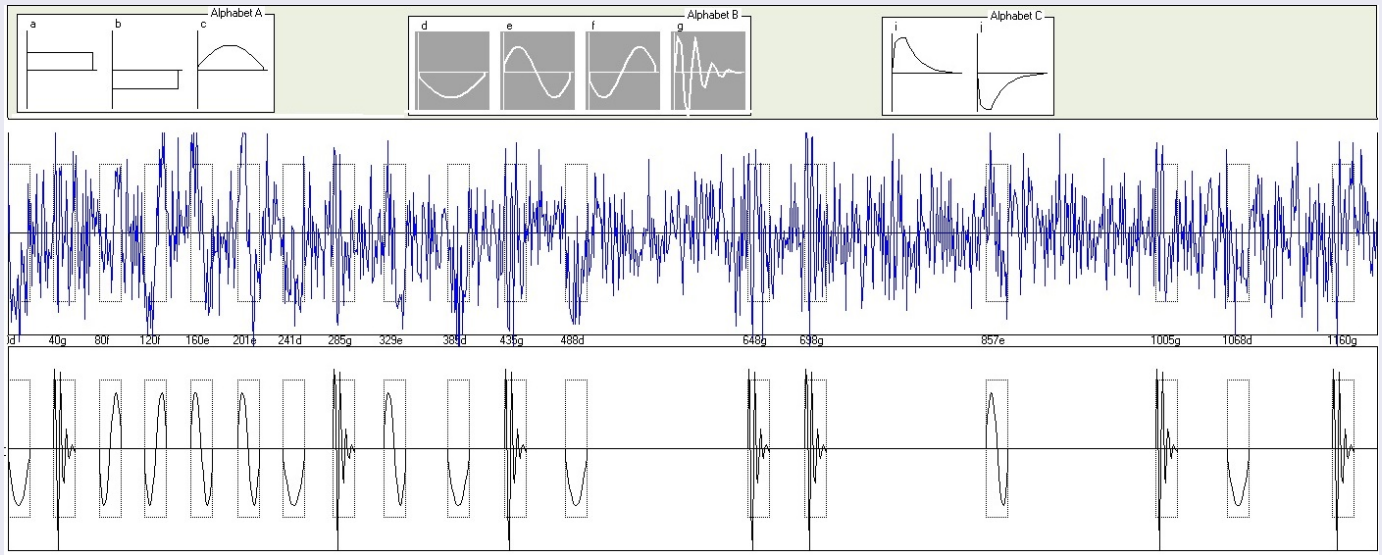


Часть II. 4. Распознавание последовательности как структуры, порожденной алфавитом вектор-шаблонов

Дано: зашумленная числовая или векторная последовательность и несколько алфавитов шаблонов (векторов).
Найти: алфавит, элементы которого непериодически встречаются в этой последовательности и их расположение.

В этой задаче мощность допустимого множества — $\mathcal{O}(J(N \times K)^M)$ (J — число алфавитов, K — мощность наибольшего по размеру алфавита).

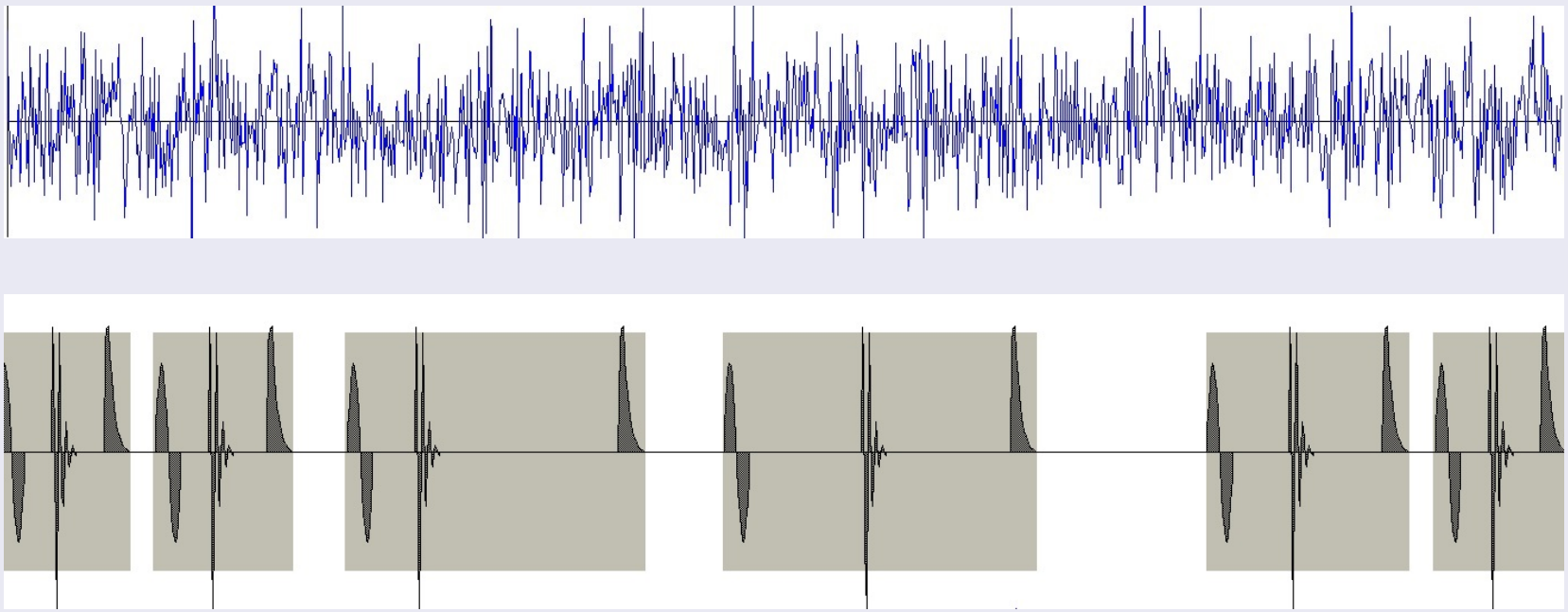
Распознавание алфавита



Часть II. 5. Обнаружение повторяющегося паттерна — набора вектор-шаблонов

Дано: зашумленная числовая или векторная последовательность и комбинация вектор-шаблонов.
Найти: повторы шаблона и расположение всех его элементов.

Поиск паттерна и его элементов



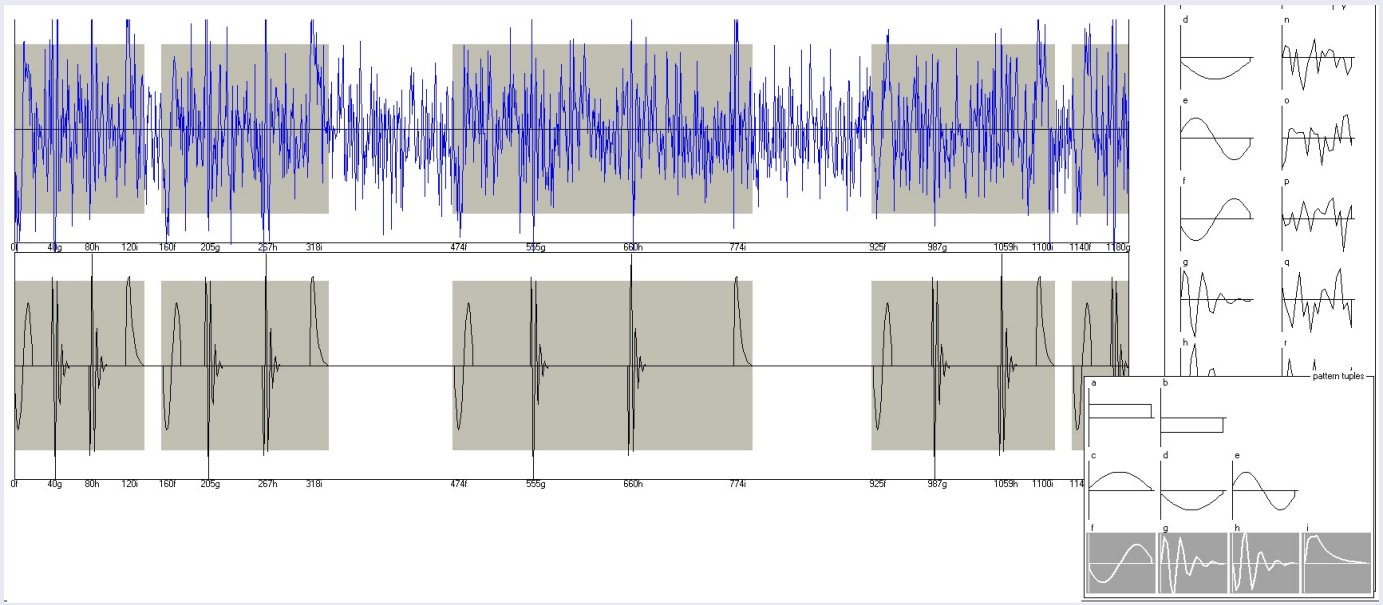
Обобщение — задача обнаружения паттерна, когда допустимы всевозможные перестановки его элементов, также разрешима за полиномиальное время.

Часть II. 6. Распознавание последовательности, включающей повторяющийся паттерн (слово) из заданного семейства (словаря)

Дано: зашумленная числовая или векторная последовательность и словарь паттернов.

Найти: паттерн (слово), который повторяется в последовательности, и расположение элементов (букв) этого паттерна в последовательности.

Поиск паттерна из словаря

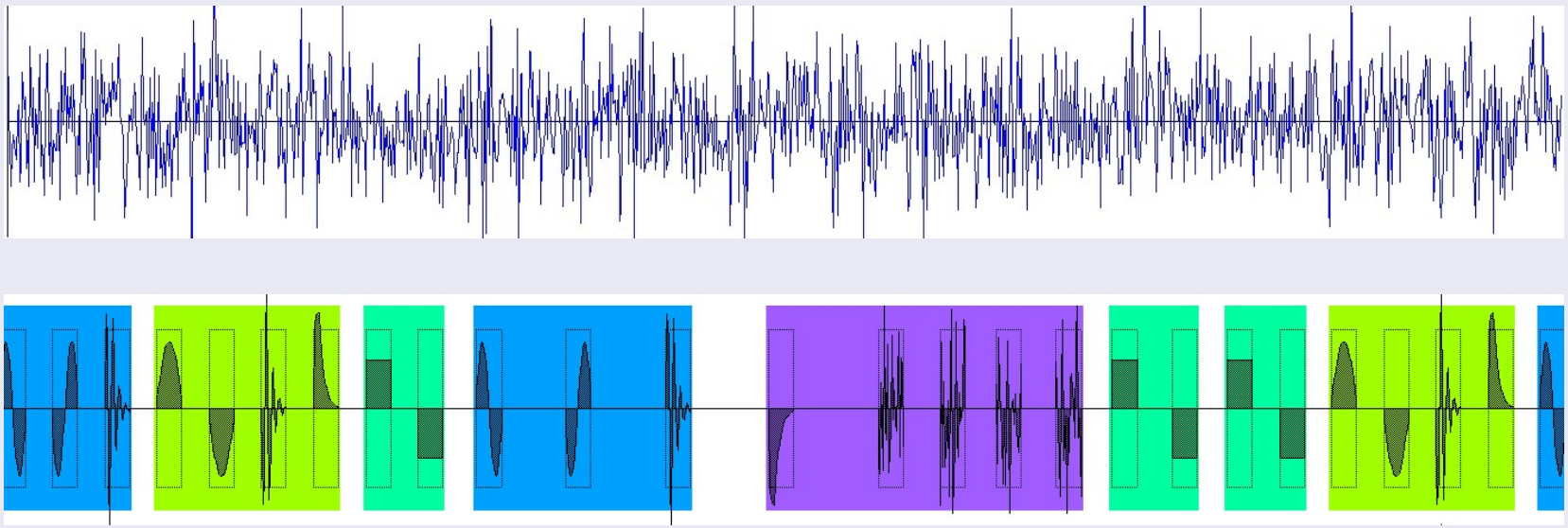


Часть II. 7. Обнаружение и идентификация паттернов (слов) из заданного семейства (словаря)

Дано: зашумленная числовая или векторная последовательность и словарь паттернов.

Найти: паттерны (слова) и расположение всех элементов (букв) этих паттернов в последовательности.

Поиск и идентификация паттернов

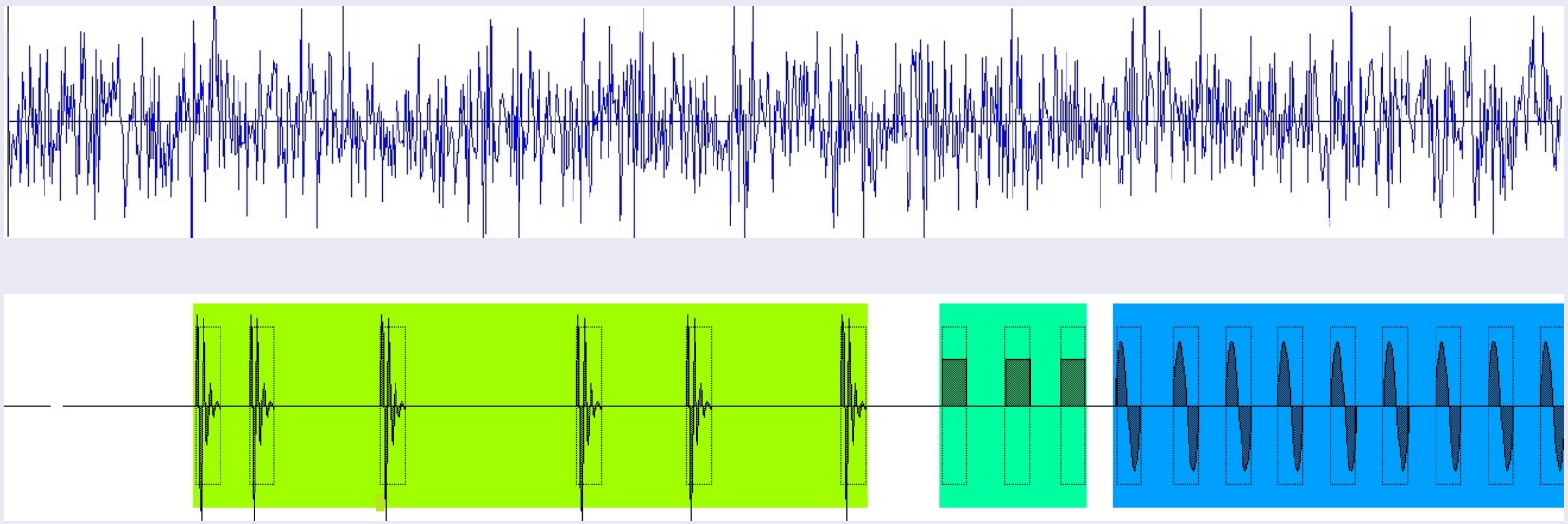


Часть II. 8. Разбиение последовательности по заданному шаблону на участки, включающие серии повторов элементов шаблона

Дано: зашумленная числовая или векторная последовательность и векторный паттерн (шаблон).

Найти: участки последовательности, включающие серии повторов элементов паттерна, и элемены паттерна.

Поиск серийных участков по образцу

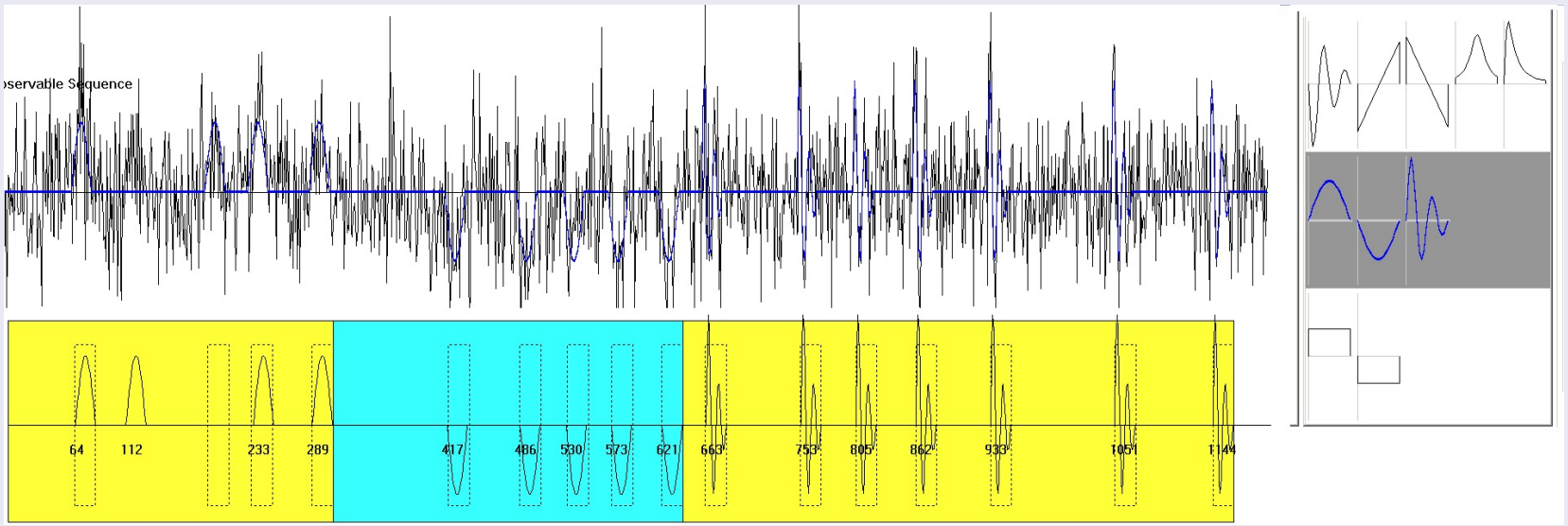


Часть II. 9. Распознавание последовательности как серийной структуры, порожденной шаблоном из заданного словаря

Дано: зашумленная числовая или векторная последовательность и словарь векторных паттернов (слов).

Найти: Слово в словаре, элементы (буквы) которого соответствуют серийным участкам последовательности.

Распознавание слов, порождающих серийные участки

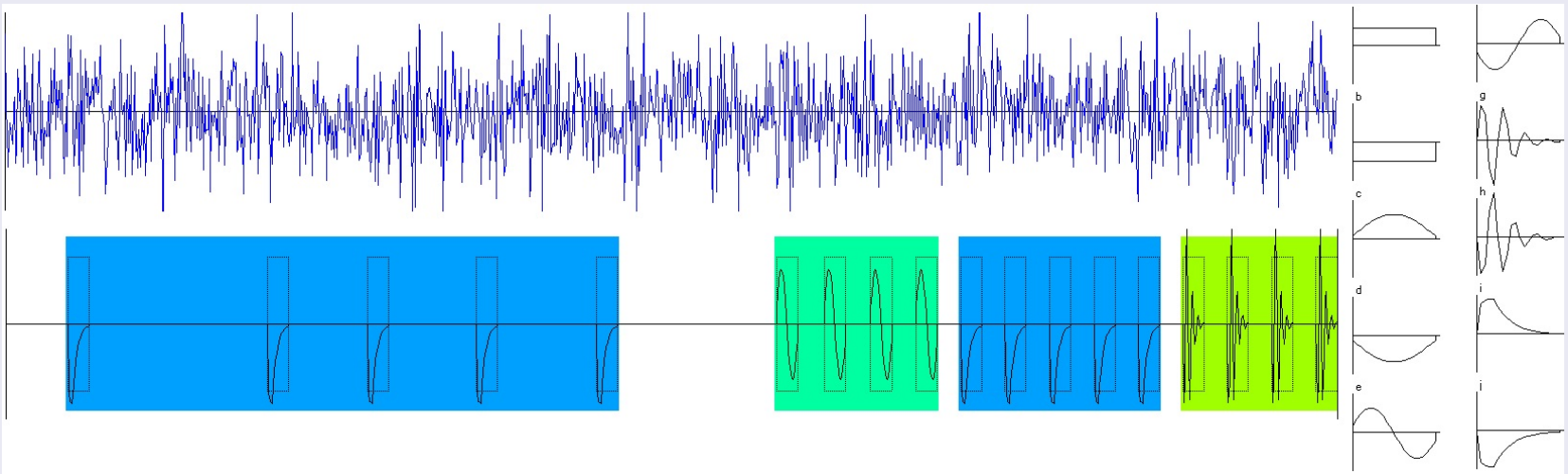


Часть II. 10. Разбиение последовательности на серийные участки, похожие на элементы из заданного алфавита векторных шаблонов

Дано: зашумленная числовая или векторная последовательность и алфавит векторов.

Найти: Найти все векторы из алфавита, которые соответствуют серийным участкам последовательности, и указать их расположение в последовательности.

Распознавание слов, порождающих серийные участки



Часть II. Проблемы принятия решения о последовательности и поиска её элементов

Выше приведен далеко не полный список изученных содержательных проблем анализа и распознавания последовательностей.

За каждой из перечисленных содержательных проблем стоят 4 экстремальные задачи с различными целевыми функциями и предложенные для их решения алгоритмы, обеспечивающие отыскание точного решения за полиномиальное время.

Полученные результаты нашли свое применение при решении прикладных задач в области космического мониторинга, геофизики, атомной энергетики, нефтеразведки, диагностики машиностроительного оборудования.

Спасибо за внимание!