Квадратичные евклидовы задачи 2-кластеризации: сложность и эффективные алгоритмы с оценками точности для их решения

Александр Кельманов

Институт математики им. С. Л. Соболева СО РАН,
Новосибирский государственный университет,
Новосибирск

17 всероссийская конференция
«Математические методы распознавания образов» (ММРО-17)

Светлогорск, 19–25 сентября 2015 г

Введение

План доклада

1. Введение. Предмет, цель и мотивация исследования

2. Несколько труднорешаемых квадратичных евклидовых задач 2-кластеризации:

   1. сложность и алгоритмы с оценками качества решения
   2. результативные методы, техники и приемы для этих задач
   3. открытые вопросы

3. Заключение. Актуальные проблемы

Введение

Предмет исследования —

задачи 2-разбиения конечного множества точек евклидова пространства на минимум квадратичных критериев.

Цель исследования —

1) анализ вычислительной сложности этих задач, построение эффективных алгоритмов с гарантированными оценками точности для их решения и обзор последних достижений по исследованию этих задач,
2) развитие результативных методов, техник и приемов построения эффективных приближенных алгоритмов для задач с квадратичным критерием.

Мотивация исследований —

наличие открытых 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 и геометрии.

2. Статистические проблемы совместного оценивания и проверки гипотез по неоднородным выборкам, которые содержат данные из нескольких (в частности — двух) распределений, причем информация о соответствии элементов выборки распределению отсутствует (недоступна).

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

4. 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. План доклада

Список рассматриваемых задач

1. Двухкластерная задача MSSC
(**Minimum Sum-of-Squares 2-Clustering**)

2. Квадратичная евклидова задача разбиения конечного множества точек на два кластера при заданном центре одного из кластеров
(**Minimum Sum-of-Squares 2-Clustering with given center of one cluster**)

3. Квадратичная евклидова задача разбиения конечной последовательности точек на два кластера при заданном центре одного из кластеров
(**Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster**)

Введение. План доклада

Список рассматриваемых задач

4. Квадратичная евклидова задача 2-кластеризации на минимум суммы по обоим кластерам внутрикластерных сумм квадратов попарных расстояний
(**Quadratic Euclidean Min-Sum All-Pairs 2-clustering**)

5. Квадратичная евклидова задача сбалансированной 2-кластеризации
(**Euclidean Balanced Variance-based 2-clustering**)

6. Квадратичная евклидова задача сбалансированной 2-кластеризации при заданном центре одного из кластеров
(**Euclidean Balanced Variance-based 2-clustering with given center**)

Основное внимание будет сосредоточено на задачах 2-кластеризации с заданным центром одного из кластеров.

Одной из самых известных (Fisher, 1958) NP-трудных (Aloise D., Deshpande A., Hansen P., Popat P., 2009) задач анализа данных является задача разбиения конечного множества точек на несколько кластеров. Двухкластерная задача имеет следующую формулировку.

Задача 2-MSSC (Minimum Sum-of-Squares 2-Clustering)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ — геометрические центры (центроиды) кластеров.

Источник задачи 2-MSSC —

статистическая проблема

Дано: неоднородная выборка $\mathcal{Y} = \{y_1, \ldots, y_N\}$ из двух q-мерных распределений. Соответствие элементов выборки распределению неизвестно.

Вопрос: верно ли, что элементы выборки принадлежат двум гауссовским распределениями с неизвестными средними и одной и той же известной диагональной ковариационной матрицей, у которой диагональные элементы идентичны.

Другой источник задачи 2-MSSC —

проблема аппроксимации

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$.
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ и точки $u \in \mathbb{R}^q$ и $v \in \mathbb{R}^q$ такие, что

$$\sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \to \min_{\mathcal{M}, u, v},$$

где

$$x_n = \begin{cases} u, & \text{если } n \in \mathcal{M}, \\ v, & \text{если } n \in \mathcal{N} \setminus \mathcal{M}. \end{cases}$$

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Близкой к 2-MSSC в постановочном плане является следующая задача (Кельманов, Кельманова, Хамидуллин, 2004)

Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

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

Задача 2-MSSC (Minimum Sum-of-Squares 2-Clustering)

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

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

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

Как и в задаче 2-MSSC, в этой задаче требуется разбить множество $\mathcal{Y}$ на 2 кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$. Однако центр одного из кластеров — $\mathcal{Y} \setminus \mathcal{C}$ — задан в начале координат, а геометрический центр $\overline{y}(\mathcal{C})$ кластера $\mathcal{C}$ неизвестен.

Возможны варианты задачи, когда мощности кластеров заданы на входе и неизвестны.

В обобщениях этих задачи требуется разбить входное множество точек на несколько кластеров. При этом для части кластеров заданы желаемые центры, а геометрические центры остальных кластеров неизвестны.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Источник задачи —

содержательная проблема из анализа данных

Имеется таблица, содержащая 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 какому-либо состоянию объекта неизвестно.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Содержательная проблема из анализа данных

Требуется:

1. разбить таблицу (т.е. $\mathcal{Y}$), содержащую N q-мерных наборов, на подмножества ($\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$), соответствующие активному и пассивному состояниям объекта,

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

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Формализация содержательной проблемы с использованием критерия минимума суммы квадратов уклонений приводит к следующей задаче.

Один из истоков задачи — проблема аппроксимации

Дано: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$.
Найти: набор $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ и точку $v \in \mathbb{R}^q$ такие, что

$$\sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \to \min_{\mathcal{M}, v},$$

где

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

Эта задача приближения индуцирует сформулированную выше задачу 2-разбиения с заданным центром одного из кластеров.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Другой источник задачи — статистика

Задача проверки гипотез

Дано: неоднородная выборка $\mathcal{Y} = \{y_1, \ldots, y_N\}$ из двух q-мерных распределений. Соответствие элементов выборки распределению неизвестно.

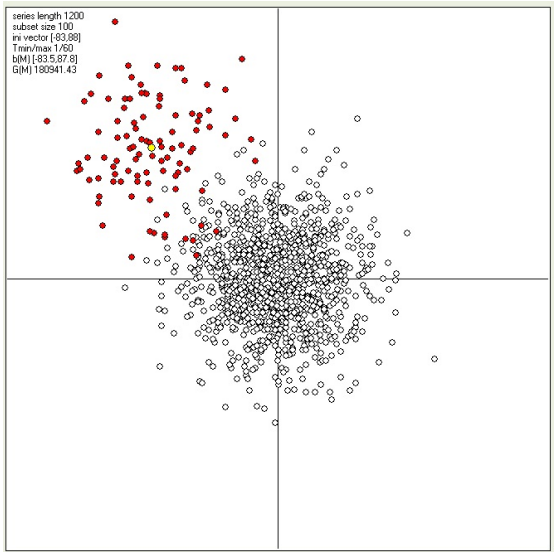
Вопрос: верно ли, что элементы выборки принадлежат двум гауссовским распределениями с одной и той же известной диагональной ковариационной матрицей, у которой диагональные элементы идентичны, причем среднее одного из распределений равно нулю, а среднее другого — неизвестно.

Эта задача также индуцирует сформулированную задачу 2-кластеризации с заданным в нуле центром.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Примеры:
(1) содержательная проблема из анализа данных,
(2) статистическая проблема

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



2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: сложность

1. Задача NP-трудна в сильном смысле.

K задаче с заданными размерами кластеров сведена задача Клика заданного размера (Бабурин, Гимади, Глебов, Кельманов, Кельманова, Пяткин, Хамидуллин, 2006, 2007)

K задаче с неизвестными размерами сведена задача 3-выполнимость (Кельманов, Пяткин, 2008).

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: алгоритмы

2. 2-приближённые алгоритмы, временная сложность — $\mathcal{O}(qN^2)$ (Долгушев, Кельманов, 2011, **заданные** размеры кластеров; Кельманов, Хандеев, 2013, **неизвестные** размеры кластеров).

Подход к построению алгоритмов — замена решения исходной задачи точным полиномиальным решением вспомогательной задачи и последующая оценка точности такой замены.

(1) в задаче с **заданными размерами** кластеров:

для каждой точки входного множества строится $(M-1)$-элементное подмножество, состоящее из точек, имеющих наибольшие проекции на луч из начала координат в эту точку; в семействе построенных подмножеств выбирается наилучшее в смысле минимума целевой функции вспомогательной задачи.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: алгоритмы

(2) в задаче с **неизвестными размерами** кластеров:

для каждой точки входного множества строится гиперплоскость, перпендикулярная отрезку из начала координат в эту точку и делящая отрезок пополам. Подмножество точек, лежащих в полупространстве, не включающем начало координат, объявляется претендентом на решение задачи.
Из полученного семейства претендентов выбирается наилучший в смысле минимума целевой функции вспомогательной задачи.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: алгоритмы

3. Схема PTAS (для задачи с заданными размерами кластеров) с временной сложностью $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — относительная погрешность (Долгушев, Кельманов, Шенмайер, 2012).

Подход к построению схемы:
1) перебор линейных оболочек всех t-элементных подмножеств входного множества, т.к. одна из них включает точку, гарантирующую решение задачи с относительной погрешностью $1/t$;
2) поиск ограниченной области, включающей эту точку, с помощью многомерных решеток, построенных на этих линейных оболочках. Среди множества узлов решеток ищется наилучший в смысле минимума специальной оценочной функции. Число s^{t-1} узлов решетки (s и t — параметры алгоритма) определяет погрешность приближения к точке, аппроксимирующей геометрический центр с относительной погрешностью $1/t$.
Трудоемкость алгоритма записана для параметров $t = 2/\varepsilon$ и $s = 9\,t^{3/2}$.

Известные результаты: алгоритмы

4. Рандомизированный алгоритм для задачи с заданными размерами кластеров (Кельманов, Хандеев, 2014),

который в случае $M \geq \beta N$ для некоторого $\beta \in (0, 1)$ при заданных относительной погрешности ε и вероятности несрабатывания алгоритма γ находит $(1 + \varepsilon)$-приближённое решение задачи за время $\mathcal{O}(2^k q(k + N))$, где $k = \max\left(\left\lceil \frac{4}{\beta\gamma\varepsilon} \right\rceil, \left\lceil \frac{8}{\beta} \log \frac{2}{\gamma} \right\rceil\right)$, линейное по N и q при фиксированных значениях β, γ и ε.

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

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: алгоритмы

4. Рандомизированный алгоритм для задачи с заданными размерами кластеров (Кельманов, Хандеев, 2014).

Подход к построению алгоритма:

1) из входного множества формируется конечное мультимножество (случайная выборка с возвращением);

2) для каждого из непустых подмножеств мультимножества вычисляется центроид и формируется допустимое решение — набор из M точек входного множества, имеющих наибольшие проекции на направление из начала координат в этот центроид;

3) в семействе допустимых решений выбирается наилучшее (в смысле наименьшего значения целевой функции).

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Известные результаты: алгоритмы

5. Установлено, что задача с заданными размерами кластеров разрешима за время $\mathcal{O}(q^2 N^{2q})$, полиномиальное в случае, когда размерность q пространства фиксирована (Кельманов, Хандеев, 2014). Результат следует из разрешимости за такое же время полиномиально эквивалентной задачи поиска подмножества с максимальной нормой суммы элементов подмножества (Гимади, Пяткин, Рыков, 2008).
Из этого результата следует полиномиальная разрешимость задачи с оптимизируемыми размерами кластеров для этого же случая.

6. Точный псевдополиномиальный алгоритм для задачи с заданными размерами кластеров в случае, когда координаты точек целочисленны, а размерность пространства фиксирована; временная сложность алгоритма — $\mathcal{O}(NM(MD)^{q-1})$ (Гимади, Глазков, Рыков, 2008);
здесь D — максимальное абсолютное значение координат точек входного множества.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Новые результаты:

1. Точный псевдополиномиальный алгоритм для задачи с заданными размерами кластеров в случае, когда координаты точек целочисленны, а размерность пространства фиксирована; временная сложность алгоритма — $\mathcal{O}(N(MD)^q)$ (Кельманов, Хандеев, 2015);

здесь D — максимальное абсолютное значение координат точек входного множества.

Подход к построению алгоритма:

1. В области пространства, определяемой максимальным абсолютным значением координат входных точек, строится многомерная равномерная по каждой координате сетка (решётка) с рациональным шагом. Шаг сетки выбирается так, чтобы один из её узлов совпал с геометрическим центром одного из искомых кластеров.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Новые результаты:

Подход к построению алгоритма:

2. Для каждого узла построенной сетки решается задача максимизации вспомогательной целевой функции — M-элементной суммы проекций точек входного множества на луч из начала координат в этот узел.
В результате решения находится подмножество точек, доставляющее максимум этой функции (подмножество точек с наибольшими проекциями).
Найденный набор объявляется претендентом на решение.

3. В качестве окончательного решения выбирается тот набор, для которого значение целевой функции исходной задачи минимально.

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Новые результаты:

2. Для задачи с заданными размерами кластеров **не существует** схемы **FPTAS**, если P≠NP, и такая схема построена для случая фиксированной размерности пространства.

Временная сложность схемы — $\mathcal{O}(N^2\varepsilon^{-q/2})$, где ε — относительная погрешность (Кельманов, Хандеев, 2015).

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Новые результаты:

Подход к построению схемы **FPTAS**:
1) для каждой точки входного множества строится область (куб), гарантированно включающая центроид искомого подмножества;

2) в кубе строится решетка с шагом, определяющим относительную погрешность приближенного решения;

3) для каждого узла решетки строится подмножество (допустимое решение) мощности M, состоящее из точек, имеющих наибольшие проекции на направление из начала координат в этот узел;

4) в семействе допустимых решений выбирается наилучшее (в смысле наименьшего значения целевой функции).

2. Minimum Sum-of-Squares 2-Clustering with given center of one cluster

Актуальные вопросы

1. Построение схемы FPTAS для задачи с неизвестными размерами кластеров, в случае, когда размерность пространства фиксирована, без перебора по множеству пар допустимых размеров искомых кластеров.

2. Обоснование рандомизированного алгоритма для общего случая задачи с неизвестными размерами кластеров, без перебора по множеству пар допустимых размеров искомых кластеров.

3. Построение эффективных алгоритмов с оценками точности для обобщения задачи на случай нескольких кластеров с неизвестными геометрическими центрами.

4. Обоснование приближенных алгоритмов, обеспечивающих решение задачи за линейное время.

3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Задача впервые сформулирована в 2004 г. (Кельманов, Кельманова, Хамидуллин).

Формулировка задачи

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

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

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

где $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum\limits_{i \in \mathcal{M}} y_i$, 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 $(n_1, \ldots, n_M)$.

3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Известные результаты: сложность

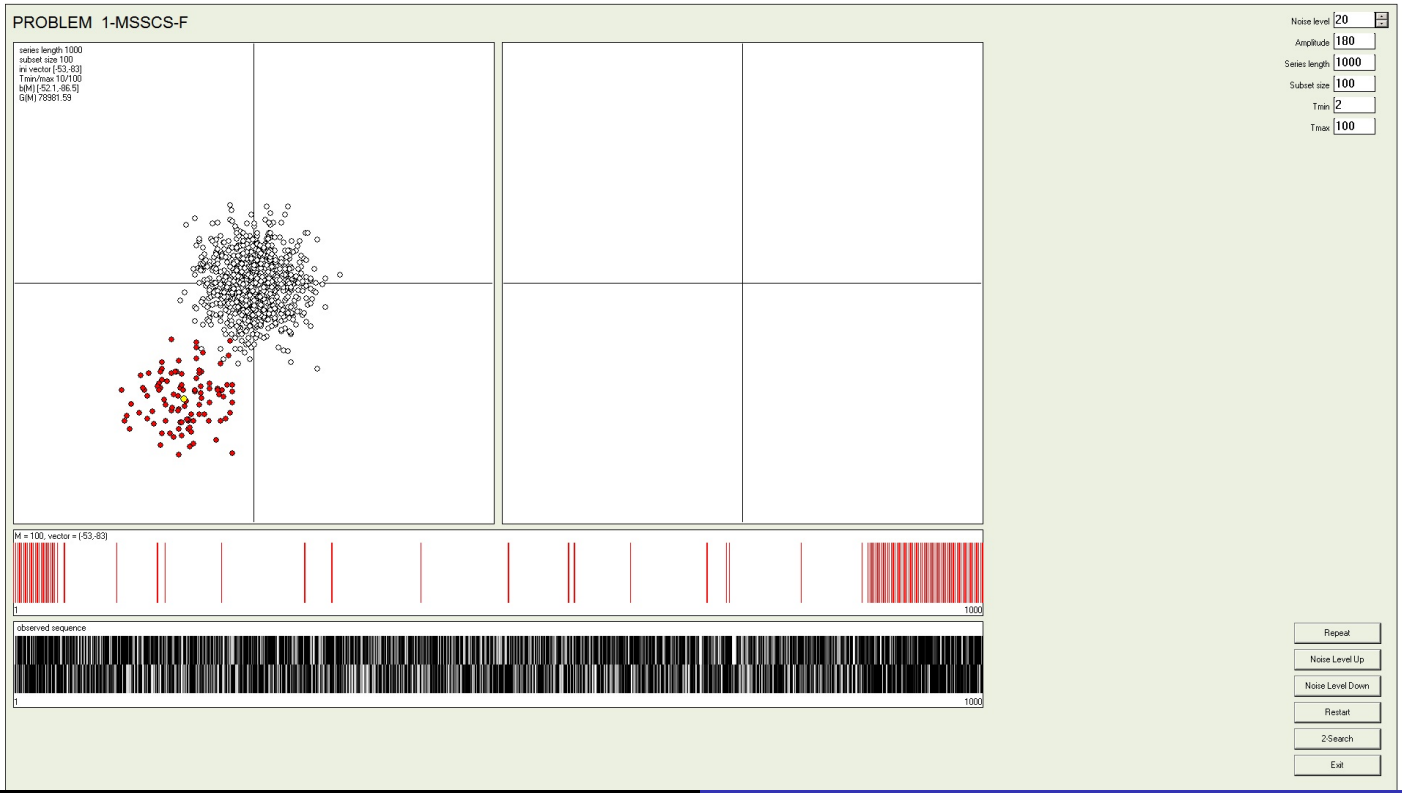
Частный случай этой задачи **кластеризации последовательности**, когда $T_{\min} = 1$ и $T_{\max} = N$, эквивалентен рассмотренной задаче Minimum Sum-of-Squares 2-Clustering with given center of one cluster **кластеризации множества**.
Поэтому эта задача NP-трудна в сильном смысле, если $T_{\min}$ и $T_{\max}$ являются частью входа (как в случае **заданного** размера искомого набора, так и в случае, когда размерность набора неизвестна).

Установлено (Кельманов, Пяткин, 2013), что если $T_{\min}$ и $T_{\max}$ не являются частью входа (параметрический вариант задачи), то задача кластеризации последовательности 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.

3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Содержательная трактовка

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



3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Известные результаты: алгоритмы

(1) вариант задачи, в котором **мощность набора $\mathcal{M}$ задана** на входе:

1. Полиномиальный 2-приближенный алгоритм, временная сложность которого есть величина $\mathcal{O}(N^2(MN+q))$ (Кельманов, Хамидуллин, 2013).

2. Точный псевдополиномиальный алгоритм для случая, когда координаты входных точек целочисленны, а размерность пространства фиксирована. Трудоемкость алгоритма оценивается величиной $\mathcal{O}(N^3(MD)^q)$ (Кельманов, Хамидуллин, Хандеев, 2014).

(2) вариант задачи, в котором **мощность набора $\mathcal{M}$ неизвестна**:

3. Полиномиальный 2-приближенный алгоритм, временная сложность которого есть величина $\mathcal{O}(N^2(N+q))$ (Кельманов, Хамидуллин, 2014).

3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Новый результат:

Установлено, что для этой задачи не существует схемы FPTAS, если P≠NP, и такая схема построена для варианта задачи с **заданной мощностью** искомого набора в случае, когда размерность пространства фиксирована.

Временная сложность схемы — $\mathcal{O}(N^4(1/\varepsilon)^{q/2})$, где ε — относительная погрешность (Кельманов, Хамидуллин, Хандеев).

3. Задача Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster

Подходы к построению алгоритмов

основаны на **комбинировании** результативных алгоритмических техник для задач кластеризации множества и индивидуальных (для каждой из задач) схем динамического программирования, обеспечивающих корректный учет ограничений на номера элементов последовательности, включаемых в тот или иной кластер.

Актуальные вопросы:

1. Построение рандомизированных алгоритмов.

2. Обоснование более быстрых алгоритмов (точных псевдополиномиальных и схемы FPTAS) для варианта задачи с неизвестными мощностями кластеров.

3. Построение схем PTAS.

4. Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering

Квадратичная евклидова задача 2-кластеризации на минимум суммы по обоим кластерам внутрикластерных сумм квадратов попарных расстояний

Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на подмножества $\mathcal{X}$ и $\mathcal{Z}$ такие, что

$$h(\mathcal{X}, \mathcal{Z}) = \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{X}} \|x - z\|^2 + \sum_{x \in \mathcal{Z}} \sum_{z \in \mathcal{Z}} \|x - z\|^2 \longrightarrow \min .$$

4. Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering

Известные результаты

1. Неевклидова задача (на графах) Min-Sum 2-clustering впервые сформулирована в 1976 г. (Sahni, Gonzalez); установлена её труднорешаемость.

2. Метрический случай — задача Min-Sum All-Pairs 2-clustering NP-труден (de la Vega, Kennyon, 2001 г.).

3. Сложностной **статус** рассматриваемого квадратичного евклидова случая — задачи Quadratic Euclidean Min-Sum All-Pairs 2-clustering до настоящего времени **не был установлен**, хотя интуитивно и гипотетически он считался труднорешаемым.

4. В 1994 г. высказано предположение (Inaba, Katoh, Imai), что квадратичная евклидова задача разрешима за время $O(N^q)$, полиномиальное при фиксированной размерности q пространства.

4. Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering

Известные результаты

5. В 2000 г. установлено (Schulman), что квадратичная евклидова задача разрешима за время $\mathcal{O}(N^{(q+2)})$, полиномиальное в случае фиксированной размерности q пространства.

6. Предложена схема PTAS, которая для любого k (число кластеров) и для любого фиксированного $\varepsilon > 0$ находит приближенное решение за время $\mathcal{O}(N^{(k/\varepsilon)})$.
(de la Vega, Karpinski, Kenion, Rabani, 2002, 2003).

Однако, как отмечено выше, сложностной статус задачи многие годы оставался невыясненным.

4. Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering

Новый результат

Задача Quadratic Euclidean Min-Sum All-Pairs 2-clustering NP-трудна в сильном смысле и для не существует схемы FPTAS, если P≠NP. (Кельманов, Пяткин, 2015).

Легко проверить справедливость следующего равенства

$$h(\mathcal{X},\mathcal{Z}) = \sum_{x\in\mathcal{Y}}\sum_{z\in\mathcal{Y}} \|x-z\|^2 - \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} \|x-z\|^2.$$

В правой части этого равенства первая двойная сумма — константа, а вторая — целевая функция NP-трудной в сильном смысле (Агеев, Кельманов, Пяткин, 2013–2014) задачи **Quadratic Euclidean Max-Cut**. Отсюда следует NP-трудность в сильном смысле сформулированной задачи.

До настоящего времени открыт вопрос о разрешимости задачи на числовой прямой (как и задачи Quadratic Euclidean Max-Cut).

Квадратичная евклидова задача сбалансированной 2-кластеризации. Размеры кластеров — балансировочные множители (Inaba, Katoh, Imai, 1994)

Задача Euclidean balanced variance-based 2-clustering

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на подмножества $\mathcal{X}$ и $\mathcal{Z}$ такие, что

$$g(\mathcal{X}, \mathcal{Z}) = |\mathcal{X}| \sum_{x \in \mathcal{X}} \|x - \overline{x}(\mathcal{X})\|^2 + |\mathcal{Z}| \sum_{z \in \mathcal{Z}} \|z - \overline{z}(\mathcal{Z})\|^2 \longrightarrow \min,$$

где $\overline{x}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum\limits_{x \in \mathcal{X}} x$ и $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum\limits_{z \in \mathcal{Z}} z$ — центроиды множеств $\mathcal{X}$ и $\mathcal{Z}$ соответственно.

Поскольку для любого непустого конечного множества $\mathcal{Y}$ точек евклидова пространства справедливо равенство

$$\sum_{x\in\mathcal{Y}}\sum_{z\in\mathcal{Y}}\|x-z\|^2 = 2|\mathcal{Y}|\sum_{y\in\mathcal{Y}}\|y-\overline{y}(\mathcal{Y})\|^2,$$

где $\overline{y}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} y$ — геометрический центр множества $\mathcal{Y}$, целевые функции задач Quadratic Euclidean Min-Sum All-Pairs 2-clustering и Euclidean balanced variance-based 2-clustering связаны формулой

$$h(\mathcal{X},\mathcal{Z}) = 2g(\mathcal{X},\mathcal{Z}).$$

Поэтому труднорешаемость задачи сбалансированной 2-кластеризации следует из труднорешаемости (Кельманов, Пяткин, 2015) задачи Quadratic Euclidean Min-Sum All-Pairs 2-clustering. Отсюда же следует несуществование схемы FPTAS.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Квадратичная евклидова задача сбалансированной 2-кластеризации при заданном (желаемом) центре одного из кластеров

Euclidean Balanced Variance-based 2-clustering with given center

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на подмножества $\mathcal{X}$ и $\mathcal{Z}$ такие, что

$$f(\mathcal{X}, \mathcal{Z}) = |\mathcal{X}| \sum_{x \in \mathcal{X}} \|x - \overline{x}(\mathcal{X})\|^2 + |\mathcal{Z}| \sum_{z \in \mathcal{Z}} \|z\|^2 \longrightarrow \min,$$

где $\overline{x}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum\limits_{x \in \mathcal{X}} x$ — геометрический центр подмножества $\mathcal{X}$.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Известные результаты

1. Вопрос о статусе сложности задачи до настоящего времени был открытым.

2. Какие-либо эффективные алгоритмы с оценками точности до настоящего времени отсутствовали.

Новые результаты

1. Задача NP-трудна в сильном смысле.
2. Не существует схемы FPTAS, если P$\neq$NP.
(Кельманов, Пяткин, 2015).

3. Обоснован точный псевдополиномиальный алгоритм для случая задачи, в котором размерность пространства фиксирована, а координаты точек целочисленны. Временная сложность алгоритма — $O(N(MD)^q)$, где D — максимальное абсолютное значение координат точек входного множества
(Кельманов, Моткова, 2015).

6. Задача Euclidean balanced variance-based 2-clustering with given center

Суть подхода к доказательству

Полиномиальное сведение к рассматриваемой задаче известной NP-трудной в сильном смысле задачи *Minimum Bisection* для *кубических* графов.

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.

Известно [Bui et al. 1987], что задача *Minimum Bisection* NP-трудна в случае *кубических* графов.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Схема доказательства

1. По примеру задачи Minimum Bisection для кубических графов строим пример (вход) рассматриваемой задачи.

2. Допускаем, что рассматриваемая задача разрешима за полиномиальное время, т.е. существует точный полиномиальный алгоритм её решения.

3. Показываем, что алгоритмическому решению в построенном примере рассматриваемой задачи однозначно соответствует оптимальное решение задачи Minimum Bisection; получаем **противоречие**, т.к. задача Minimum Bisection NP-трудна в сильном смысле.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Схема доказательства

Для доказательства противоречия устанавливаются два факта:

1. В оптимальном 2-разбиении входных данных из примера рассматриваемой задачи мощности искомых подмножеств равны.

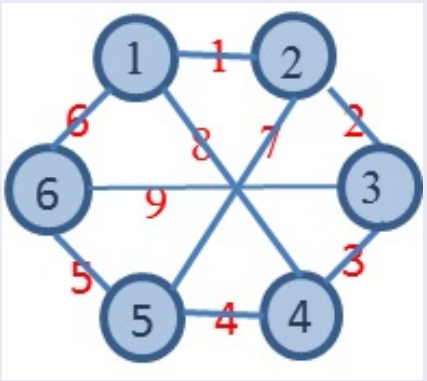
2. Мощность бисекции графа G, однозначно соответствующая оптимальному 2-разбиению в рассматриваемом примере, минимальна среди всех возможных бисекций графа G.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Пример задачи Minimum Bisection для кубического графа

Кубический граф $G = (V, E)$ с числом вершин $|V| = n = 2k$ (число вершин чётно) и числом ребер $|E| = m$.

Пример. Кубический граф $K_{3,3}$



6. Задача Euclidean balanced variance-based 2-clustering with given center

Пример входа задачи Euclidean balanced variance-based 2-clustering with given center

A — матрица с числом строк n и числом столбцов $m + n$, состоящая из двух частей: слева матрица инциденций кубического графа G, справа диагональная матрица размера $n \times n$ с диагональными элементами $\gamma = \lceil \sqrt{k^2 + 3k - 5} \rceil$.

Пример. Матрица A

	Ребра графа														
	1	2	3	4	5	6	7	8	9						
y_1	1	0	0	0	0	1	0	1	0	γ	0	0	0	0	0
y_2	1	1	0	0	0	0	1	0	0	0	γ	0	0	0	0
y_3	0	1	1	0	0	0	0	0	1	0	0	γ	0	0	0
y_4	0	0	1	1	0	0	0	1	0	0	0	0	γ	0	0
y_5	0	0	0	1	1	0	1	0	0	0	0	0	0	γ	0
y_6	0	0	0	0	1	1	0	0	1	0	0	0	0	0	γ

$n = 6,\ m = 9,\ q = n + m = 15$, строки матрицы — вход рассматриваемой задачи

6. Задача Euclidean balanced variance-based 2-clustering with given center

Несуществование FPTAS

Существование FPTAS для рассматриваемой задачи влечёт P=NP.

Доказательство этого факта опирается на следующие утверждения (см., например, М.Гэри, Д.Джонсон. Вычислительные машины и труднорешаемые задачи М.: Мир, 1982.)

Теорема 6.8

Пусть П — NP-трудная задача на минимум такая, что её целевая функция целочисленна и ограничена полиномом от максимального числового значения на входе задачи. Тогда из существования схемы FPTAS для решения задачи П следует существование псевдополиномиального алгоритма решения задачи П.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Следствие из этой теоремы

Пусть П — оптимизационная задача, удовлетворяющая предположениям теоремы 6.8. Тогда если П является NP-трудной в сильном смысле задачей, то (при P≠NP) задача П не может быть решена схемой FPTAS.

Для удвоенной целевой функции рассматриваемой задачи имеем равенство

$$2f(\mathcal{X},\mathcal{Z}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{X}} \|x-y\|^2 + 2|\mathcal{Z}|\sum_{z\in\mathcal{Z}} \|z\|^2,$$

в котором при целочисленности координат входных векторов очевидна целочисленность правой части и её ограниченность полиномом от максимального абсолютного значения координаты входных векторов. Отсюда следует факт несуществования FPTAS.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Точный псевдополиномиальный алгоритм. Суть подхода

1. В области пространства, определяемой максимальным абсолютным значением координат входных точек, строится многомерная равномерная по каждой координате сетка (решетка) с рациональным шагом.
Шаг решётки выбирается так, чтобы один из её узлов совпал с геометрическим центром одного из искомых кластеров.

2. Для каждого узла построенной решетки решается задача минимизации вспомогательной целевой функции. В результате решения находится подмножество, доставляющее минимум этой функции.

3. Найденное подмножество включается в семейство претендентов на решение исходной задачи. В качестве окончательного решения выбирается то подмножество из построенного семейства, для которого значение целевой функции исходной задачи минимально.

6. Задача Euclidean balanced variance-based 2-clustering with given center

Актуальные вопросы

Задача новая и на сегодняшний день какие-либо эффективные алгоритмы с оценками точности для ее решения ещё не построены (за исключением обоснованного точного псевдополиномиального алгоритма для случая задачи, в котором размерность пространства фиксирована, а координаты точек целочисленны).
Построение таких алгоритмов — предмет предстоящих исследований.

Спасибо за внимание!