О НЕКОТОРЫХ ТРУДНОРЕШАЕМЫХ ЗАДАЧАХ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ

Кельманов А. В.

Институт математики им. С. Л. Соболева СО РАН
Новосибирск

Молодежная научная школа-семинар
«Дискретные модели и методы принятия решений»
21–23 июня 2013, Новосибирск

Введение

Предмет исследования —

дискретные экстремальные задачи кластерного анализа и поиска подмножеств векторов в конечном множестве векторов евклидова пространства.

Цель исследования —

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

Мотивация исследований —

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

Области приложений

1. Рассматриваемые задачи индуцируются, например, проблемами помехоустойчивого кластерного анализа данных, а также характерными (типовыми) проблемами «обучения компьютера» (Machine Learning) распознаванию образов.

2. С другой стороны, эти задачи можно трактовать как задачи компьютерной геометрии.

3. Кроме того, их можно интерпретировать как задачи приближения (аппроксимации).

4. Наконец, эти задачи напрямую связаны со статистическими проблемами совместного оценивания и проверки гипотез по выборкам, которые содержат данные из нескольких распределений, причем информация о принадлежности элементов выборки распределению отсутствует (недоступна).

Как возникают эти задачи?

Рассматриваемые экстремальные задачи возникают, в частности, при реализации подхода к решению проблем «обучения компьютера» распознаванию образов, состоящего в реализации следующей цепочки последовательно выполняемых шагов:

1) формулировка содержательной проблемы,

2) её формализация в виде оптимизационной модели,

3) выявление (аналитическими методами) дискретной экстремальной задачи, индуцированной этой моделью,

4) построение полиномиального алгоритма решения этой задачи с априорно гарантированной (доказуемой) оценкой точности,

5) компьютерная технология, ориентированная на решение прикладной проблемы.

Истоки задач

Комбинирование формализации содержательных проблем анализа данных и распознавания образов с классическими оптимизационными критериями оценивания и принятия решения (проверки гипотез) порождает необозримое многообразие неизученных дискретных экстремальных задач.

Статус сложности многих известных задач, моделирующих типовые содержательные проблемы анализа данных и распознавания образов, до настоящего времени не изучен, а алгоритмы с гарантированными оценками точности для их решения не построены.

Ниже будут рассмотрены несколько подобных задач, которые возникают при решении простейших содержательных проблем классификации данных.

Истоки задач

Общей чертой проблем, индуцирующих рассматриваемые ниже экстремальные задачи, является оптимизационная модель анализа данных, а именно:

эта модель формулируется как задача аппроксимации наблюдаемых данных по критерию **минимимума суммы квадратов расстояний** (уклонений) гипотетической априорно заданной (фиксированной) структурой (моделью порождения данных).

К **идентичным** формулировкам экстремальных задач приводит статистическая формализация содержательных проблем в случае, когда критерием принятия решения и оценивания является **максимум правдоподобия**, а моделью возмущающей помехи (ошибки) служат независимые одинаково распределенные гауссовские случайные величины.

Задачи поиска подмножеств

Одной из самых известных [1] (с 1965 г.) в общем случае NP-трудных [2] (Aloise D., Hansen P., 2007) задач анализа данных и распознавания образов, которая возникает в рамках описанного подхода, является

Задача MSSC (Minimum Sum-of-Squares Clustering)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $J > 1$.
Найти: разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

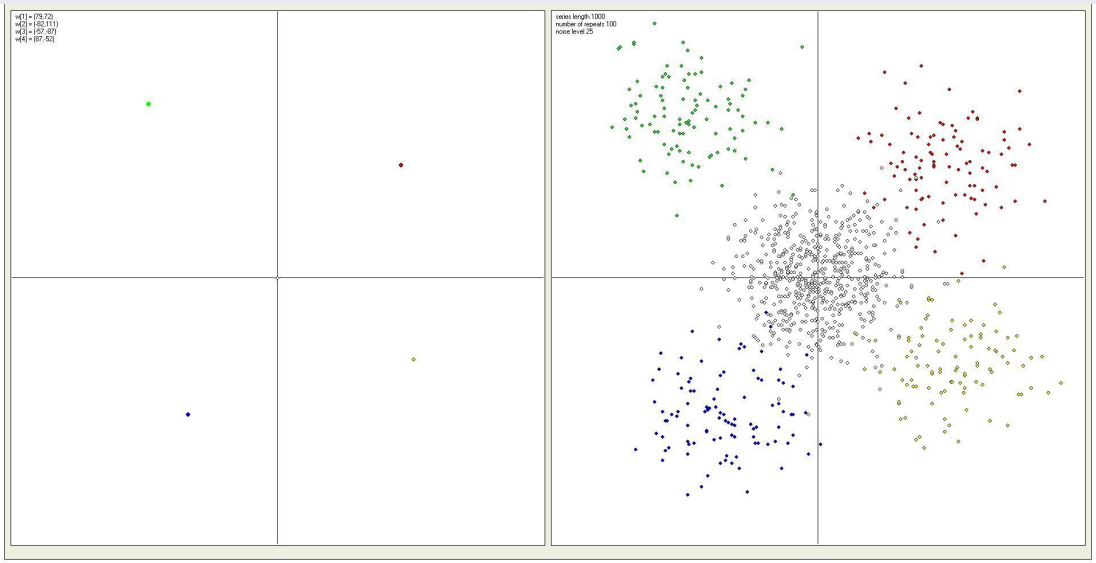
$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, — центр j-го кластера.

В некоторых публикациях эта задача фигурирует под названием k-Means, которое соответствует названию одного из первых эвристических алгоритмов для её решения.

Пример

Результаты измерений характеристик пяти объектов, изображенные на плоскости.



В качестве примера, приведем вывод целевой функции задачи MSSC из формализации содержательной проблемы в виде задачи аппроксимации.

Задача MSSC

Гипотетическая модель (идеальная структура) данных

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

$$x_n = \begin{cases} w_1, \ n \in \mathcal{M}_1, \\ \ldots \\ w_J, \ n \in \mathcal{M}_J, \end{cases} \tag{1}$$

где $\mathcal{M}_j \subset \mathcal{N}$, $\mathcal{M}_j \neq \emptyset$, $\mathcal{M}_j \cap \mathcal{M}_k = \emptyset$, $j \neq k$, $j, k = 1, \ldots, J$, $\cup_{j=1}^{J} \mathcal{M}_j = \mathcal{N}$.

Модель порождения наблюдаемых данных

Допустим, что для обработки доступна последовательность (таблица)

$$y_n = x_n + e_n, \ n \in \mathcal{N}, \tag{2}$$

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

Задача MSSC

Модель анализа данных

Учитывая зависимость элементов последовательности (1) от векторов и множеств, положим

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_J, w_1, \ldots, w_J) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \qquad (3)$$

Рассмотрим модель анализа данных в виде следующей оптимизационной задачи.

Задача 1 (аппроксимация последовательности)

Дано: последовательность y_n, $n \in \mathcal{N}$, векторов из $\mathbb{R}^q$ и натуральное число $J > 1$.

Найти: разбиение множества $\mathcal{N}$ на непустые подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_J$ и векторы $w_1, \ldots, w_J$, минимизирующие функционал $S(\cdot)$, при условии, что структура последовательности описывается формулами (1) и (2).

Экстремальная задача

Раскрыв сумму квадратов (3) с учетом формулы (1), получим

$$S(\cdot) = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - w_j\|^2. \tag{4}$$

Для любого разбиения множества $\mathcal{N}$ на непустые подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_J$ минимум функционала (4) достигается при $w_j = \frac{1}{|\mathcal{M}_j|} \sum_{n \in \mathcal{M}_j} y_n$, $j = 1, \ldots, J$, и равен

$$S_{min}(\mathcal{M}_1, \ldots, \mathcal{M}_J) = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - \frac{1}{|\mathcal{M}_j|} \sum_{n \in \mathcal{M}_j} y_n\|^2.$$

Положим $\mathcal{C}_j = \{y_n | n \in M_j\}$ и, переходя от суммирования по индексам к суммированию по элементам множеств, получим задачу дискретной оптимизации MSSC.

Статистический подход

Пусть в формуле (2) вектор $e_n \in \Phi_{0,\sigma^2 I}$. Тогда, учитывая, что $e_n = y_n - x_n$, функцию правдоподобия от выборки $\mathbf{e} = \{e_1, \ldots, e_N\}$ можно записать в виде:

$$\mathcal{F}(\mathbf{e}|0,\sigma^2) = \prod_{n=1}^{N} p_n$$

$$= \prod_{n=1}^{N} \frac{1}{(2\pi)^{q/2}(\sigma^2)^{q/2}} \exp\left\{-\frac{1}{2q\sigma^2}(y_n - x_n)^T(y_n - x_n)\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{qN/2}} \exp\left\{-\frac{1}{2q\sigma^2}\sum_{n=1}^{N}\|y_n - x_n\|^2\right\},$$

где p_n — плотность распределения вектора e_n.

Статистический подход

Поэтому для логарифмической функции правдоподобия имеем:

$$\mathcal{L}(\mathbf{e}|0, \sigma^2) = -\frac{qN}{2} \ln 2\pi - \frac{qN}{2} \ln \sigma^2 - \frac{1}{2q\sigma^2} \sum_{i=1}^{N} \|y_n - x_n\|^2.$$

Отсюда видно, что при фиксированных N, σ^2 и q максимизация этой функции сводится к следующей задаче на минимум

$$\sum_{n=1}^{N} \|x_n - y_n\|^2 \to \min,$$

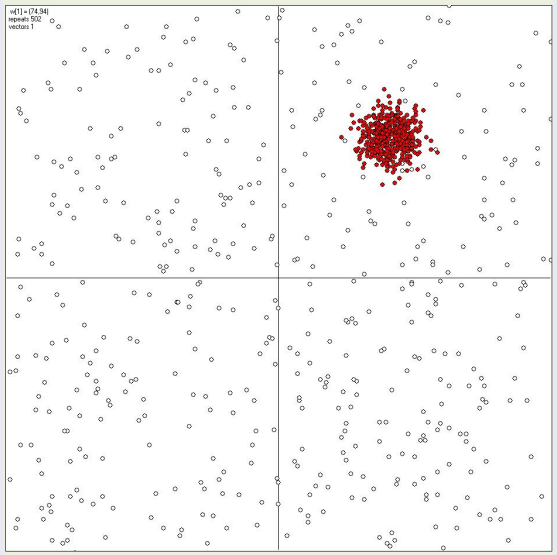
т.е. к рассмотренной выше задаче 1 аппроксимации.

Задачи поиска подмножеств

К числу менее известных и слабо изученных задач относятся 4 задачи, сформулированные ниже. Эти задачи близки к задаче MSSC в постановочном плане, но не эквивалентны ей. Они индуцируются одной и той же содержательной проблемой.

Пример

Результаты измерений характеристик множества объектов, изображенные на плоскости.



Задачи поиска подмножеств

Гипотетическая модель (структура) данных

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

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

где $\mathcal{M} \subseteq \mathcal{N}$, $\mathcal{M} \neq \emptyset$.

Модели порождения и анализа данных такие же, как и в задаче MSSC. Индуцированные этими моделями экстремальные задачи выводятся аналогичным образом.

Задачи поиска подмножеств

NP-трудность в сильном смысле этих задач установлена в работе [3] (Кельманов А.В., Пяткин А.В., 2010).

Задача VS-1 (Vector Subset 1)

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

$$\frac{1}{|\mathcal{C}|}\left\|\sum_{y\in\mathcal{C}} y\right\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y\|^2 \to \max,$$

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}} y$, при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Задача VS-2 (Vector Subset 2)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.

Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что

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

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$, при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Задача VS-3 (Vector Subset 3)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$, натуральное число $M > 1$.

Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что

$$\sum_{y \in \mathcal{C}} \sum_{z \in \mathcal{C}} \|y - z\|^2 \to \min,$$

при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Задача MSSC-Case (MSSC, special Case)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.

Найти: разбиение множества $\mathcal{Y}$ на $N - M + 1$ непустых кластеров $\mathcal{C}_1, \ldots, \mathcal{C}_{N-M+1}$ такое, что мощность одного из этих кластеров равна M и

$$\sum_{j=1}^{N-M+1} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$.

Задачи VS-1, VS-2, VS-3, MSSC-Case

Задача MSSC-Case эквивалентна задаче VS-2.
Задачи VS-1, VS-2 и VS-3 полиномиально эквивалентны, так как

$$\sum_{y\in\mathcal{Y}}\|y\|^2-\left(\frac{1}{|\mathcal{C}|}\left\|\sum_{y\in\mathcal{C}}y\right\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2\right)$$

$$=\sum_{y\in\mathcal{C}}\|y-\overline{y}(\mathcal{C})\|^2=\frac{1}{2|\mathcal{C}|}\sum_{x\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|x-z\|^2.$$

Задачи поиска подмножеств. Известные результаты

Задачи VS-1, VS-2, VS-3, MSSC-Case

1. В [5] (Кельманов А.В., Романченко., 2010) для этих задач обоснованы точные псевдополиномиальные алгоритмы для случая, когда размерность пространства фиксирована, а компоненты векторов имеют целочисленные значения. Временная сложность этих алгоритмов есть величина $\mathcal{O}(qN(2MB)^q)$, где B — максимальное абсолютное значение координаты векторов входного множества.

2. Для общего случая задачи VS-1 какие-либо полиномиальные алгоритмы с гарантированными оценками точности на сегодняшний день неизвестны.

Задачи поиска подмножеств. Известные результаты

Задачи VS-1, VS-2, VS-3, MSSC-Case

3. 2-приближенные полиномиальные алгоритмы для задач VS-2 и VS-3 и MSSC-Case были предложены в [4] (Кельманов А.В., Романченко., 2011). Алгоритмы находят решение задач за время $\mathcal{O}(qN^2)$.

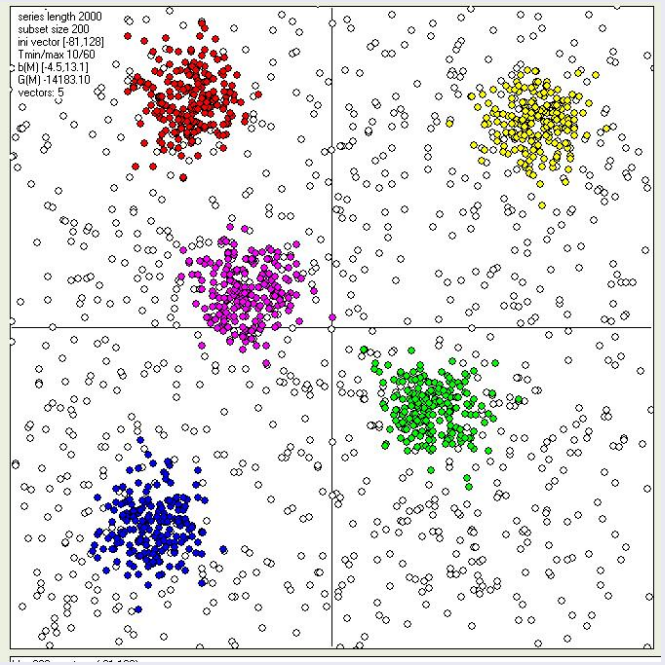
4. В [6] (Шенмайер В.В., 2012) для задачи VS-2 построена приближенная полиномиальная схема (PTAS), обеспечивающая решение задачи с произвольной относительной погрешностью ε за время $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$. Эта же схема может использоваться и для решения задач VS-3 и MSSC-Case.

5. Наконец, для случая фиксированной размерности q пространства в работе [7] (Кельманов А.В., Романченко., 2013) обоснована полностью полиномиальная приближённая схема (FPTAS), гарантирующая $(1+\varepsilon)$-приближённое решение задачи за время $\mathcal{O}(qN^2(M/\varepsilon)^q)$.

Задачи поиска подмножеств

Следующая слабо изученная NP-трудная в сильном смысле задача является обобщением задачи VS-2 на случай поиска нескольких подмножеств.

Пример



В отличие от задачи MSSC в этой задаче требуется найти не разбиение множества, а непересекающиеся подмножества фиксированной мощности.

Гипотетическая модель (структура) данных

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

$$x_n = \begin{cases} w_1, \ n \in \mathcal{M}_1, \\ \ldots \\ w_J, \ n \in \mathcal{M}_J, \\ v_n, \ n \in \mathcal{N} \backslash (\cup_{j=1}^{J} \mathcal{M}_j), \end{cases}$$

где $\mathcal{M}_j \subset \mathcal{N}$, $\mathcal{M}_j \neq \emptyset$, $\mathcal{M}_j \cap \mathcal{M}_k = \emptyset$, $j \neq k$ $j, k = 1, \ldots, J$.

Индуцированная экстремальная задача выводится также, как и ранее.

Задача FDVS-($\mathbb{R}^q$)SD (Family of Disjoint Vector Subsets, Euclidean case for Squared Distances)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$.
Найти: непересекающиеся подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$, множества $\mathcal{Y}$, такие, что

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, на мощности искомых подмножеств.

Задачи поиска подмножеств

Эта задача имеет эквивалентную формулировку в виде задачи кластеризации (задача MSSC-Case).

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$.
Найти: разбиение множества $\mathcal{Y}$ на $N - (M_1 + \ldots + M_J) + J$ непустых кластеров $\mathcal{C}_1, \ldots, \mathcal{C}_{N-(M_1+\ldots+M_J)+J}$ такое, что

$$\sum_{j=1}^{N-(M_1+\ldots+M_J)+J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$.

Задачи поиска подмножеств. Известные результаты

Задача FDVS-$(\mathbb{R}^q)$SD

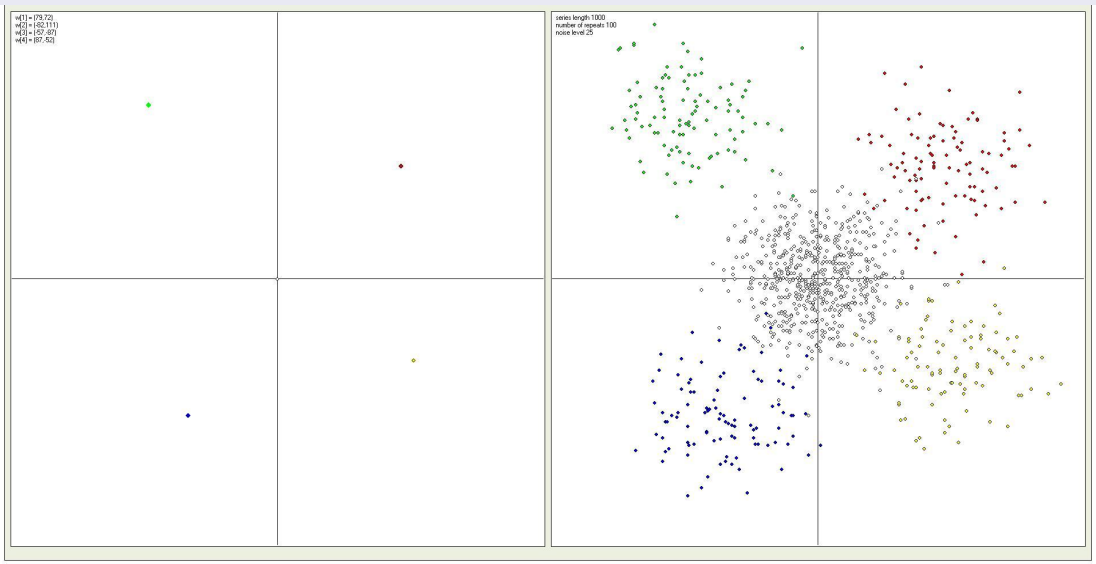
Приближенный алгоритм с оценкой точности 2 для этой задачи предложен в [9] (Галашов А.Е., Кельманов А.В., 2013). Алгоритм полиномиален в случае, когда число J искомых подмножеств фиксировано, и находит решение за время $\mathcal{O}(N^2(N^{J+1}+q))$.

Задачи поиска подмножеств

В двух следующих задачах, в отличие от задачи MSSC, центр одного из кластеров определять не требуется. Считается, что центр этого кластера совпадает с началом координат.

Пример

Результаты измерений характеристик пяти объектов, изображенные на плоскости. Один из объектов находился в пассивном состоянии, когда значения измеряемых характеристик равны нулю.



Задачи поиска подмножеств

Гипотетическая модель (структура) данных

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

$$x_n = \begin{cases} w_1, \ n \in \mathcal{M}_1, \\ \ldots \\ w_J, \ n \in \mathcal{M}_J, \\ 0, \ n \in \mathcal{N} \backslash (\cup_{j=1}^{J} \mathcal{M}_j), \end{cases}$$

где $\mathcal{M}_j \subset \mathcal{N}$, $\mathcal{M}_j \cap \mathcal{M}_k = \emptyset$, $j \neq k$ $j, k = 1, \ldots, J$, $\mathcal{M}_j \neq \emptyset$, $\mathcal{N} \backslash (\cup_{j=1}^{J} \mathcal{M}_j)$.

Индуцированные этой моделью экстремальные задачи выводятся аналогично.

Символы F и NF в кратких названиях сформулированных ниже задач образованы от английских слов Fixed и NonFixed, и соответствуют наличию (Fixed) или отсутствию (NonFixed) ограничений на мощности искомых кластеров.

Задача J-MSSC-F

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$.
Найти: разбиение множества $\mathcal{Y}$ на непустые подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и $\mathcal{Y} \setminus (\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_J)$ такое, что

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_J)} \|y\|^2 \to \min, \qquad (5)$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, при ограничениях $|\mathcal{C}_j| = M_j, j = 1, \ldots, J$, на мощности искомых подмножеств.

Задача J-MSSC-NF

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$.
Найти: разбиение множества $\mathcal{Y}$ на непустые подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и $\mathcal{Y} \setminus (\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_J)$ такое, что имеет место (5).

Задачи поиска подмножеств. Известные результаты

NP-трудность в сильном смысле этих задач установлена в работе [11] (Кельманов, Пяткин, 2009).
Очевидно, что обобщения этих задач на случай, когда центр одного из кластеров фиксируется не в начале координат, а в произвольной точке евклидова пространства, также относится к числу труднорешаемых в сильном смысле задач.

Так как для любого непустого подмножества $\mathcal{C}$ конечного множества $\mathcal{Y}$ векторов из $\mathbb{R}^q$ справедливо равенство

$$\sum_{y\in\mathcal{C}}\|y-\overline{y}(\mathcal{C})\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2=\sum_{y\in\mathcal{Y}}\|y\|^2-\frac{1}{|\mathcal{C}|}\|\sum_{y\in\mathcal{C}}y\|^2, \qquad (6)$$

частному случаю задачи J-MSSC-F — задаче 1-MSSC-F полиномиально эквивалентна NP-трудная в сильном смысле [14, 15] (Бабурин, Гимади, Глебов, Кельманов, Пяткин, Хамидуллин, 2006-2008)

Задачи поиска подмножеств. Известные результаты

Задача LVS (subset with the Longest Vectors Sum)

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

$$\|\sum_{y \in \mathcal{C}} y\| \to \max .$$

В этой задаче требуется найти подмножество векторов, которое доставляет максимум длине суммы векторов.

В силу формулы (6) частному случаю задачи J-MSSC-NF — задаче 1-MSSC-NF полиномиально эквивалентна NP-трудная в сильном смысле [16, 17] (Кельманов, Пяткин, 2008)

Задачи поиска подмножеств. Известные результаты

Задача MNLVS (subset with the Maximum Normalized Length of Vectors Sum)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$.
Найти: непустое подмножество $\mathcal{C} \subseteq \mathcal{Y}$ такое, что

$$\frac{1}{|\mathcal{C}|} \| \sum_{y \in \mathcal{C}} y \|^2 \to \max,$$

В этой задаче требуется найти подмножество векторов, которое доставляет максимум нормированному на мощность значению квадрата длины суммы векторов из искомого подмножества.

В алгоритмическом плане задачи J-MSSC-NF и J-MSSC-NF слабо изучены: лишь для случая $J = 1$ построены алгоритмы с гарантированными оценками точности.

Задачи поиска подмножеств. Известные результаты

Задачи 1-MSSC-F, 1-MSSC-NF и LVS, MNLVS

1. Прежде всего заметим, что при фиксированной размерности q пространства обе задачи — 1-MSSC-F и 1-MSSC-NF — разрешимы за полиномиальное время $\mathcal{O}(q^2 N^{2q})$. Этот результат следует из формулы (6) и полиномиальной разрешимости за это же время задач LVS и MNLVS, которая была установлена в [12] (Гимади, Пяткин, Рыков, 2008).

2. Кроме того, для задач LVS и MNLVS обоснованы:

1) точные псевдополиномиальные алгоритмы для случая, когда размерность пространства фиксирована, а компоненты векторов целочисленны [13] (Гимади, Глазков, Рыков, 2008); трудоемкость этих алгоритмов есть величина $\mathcal{O}(qMN(2MB)^{q-1})$ и $\mathcal{O}(qN^{q+1}(2B)^{q-1})$ соответственно, где B — максимальное абсолютное значение координаты векторов входного множества,

Задачи поиска подмножеств. Известные результаты

Задачи 1-MSSC-F, 1-MSSC-NF и LVS, MNLVS

2) схемы FPTAS для случая, когда размерность пространства фиксирована [14] (Бабурин, Гимади, Глебов, Пяткин, 2008), [16, 17] (Кельманов, Пяткин, 2008); алгоритмы находят решение задач за время $\mathcal{O}(Nq^2(2l)^{q-1})$ с относительной погрешностью $\varepsilon \leq (q-1)/(8l^2)$ и $\varepsilon \leq (q-1)/(4l^2)$ соответственно, где l — параметр алгоритмов.

3. Предложенные в указанных работах приближенные и асимптотически точные алгоритмы можно использовать для получения аналогичных решений задач 1-MSSC-F и 1-MSSC-NF.

Задачи поиска подмножеств. Известные результаты

Задачи 1-MSSC-F, 1-MSSC-NF и LVS, MNLVS

4. В работах [18] (Долгушев, Кельманов, 2011), [19] (Кельманов, Хандеев, 2011) для задач 1-MSSC-F и 1-MSSC-NF обоснованы 2-приближенные алгоритмы, полиномиальные относительно N и q. Временная сложность этих алгоритмов есть величина $\mathcal{O}(qN^2)$.

5. В [20] (Долгушев, Кельманов, Шенмайер, 2012) для задачи 1-MSSC-F предложена приближенная полиномиальная схема (PTAS), позволяющая решать задачу с произвольной относительной погрешностью ε за время $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$.

6. Для этой же задачи в [21] (Кельманов, Хандеев, 2013) обоснован приближенный рандомизированный алгоритм и установлены условия его асимптотической точности. Подход к построению алгоритма можно использовать для обоснования рандомизированного алгоритма решения задачи 1-MSSC-NF.

Задачи поиска подмножеств

Следующая задача на протяжении нескольких десятилетий интуитивно и гипотетически, но бездоказательно считалась NP-трудной.

Задача MSSC-PD (MSSC, the case for Pairwise Distances)

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$.
Найти: разбиение этого множества на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

$$\sum_{j=1}^{J} \sum_{x \in \mathcal{C}_j} \sum_{z \in \mathcal{C}_j} \|x - z\|^2 \to \min .$$

В случае разбиения на 2 кластера справедливо равенство

$$\sum_{x \in \mathcal{C}_1} \sum_{z \in \mathcal{C}_1} \|x - z\|^2 + \sum_{x \in \mathcal{C}_2} \sum_{z \in \mathcal{C}_2} \|x - z\|^2 + \sum_{x \in \mathcal{C}_1} \sum_{z \in \mathcal{C}_2} \|x - z\|^2$$
$$= \sum_{x \in \mathcal{Y}} \sum_{z \in \mathcal{Y}} \|x - z\|^2 = const$$

Поэтому частному случаю задачи MSSC-PD (когда $J = 2$) полиномиально эквивалентна следующая

Задача Max-Cut-$(\mathbb{R}^q)$SD (Max-Cut in Euclidean space, the case for Squared Distances).

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

$$\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \|x - z\|^2 \to \max.$$

Оказалось, что сложностной статус этой задачи ранее не был установлен. В работе [10] (Кельманов, Пяткин, 2013) доказано, что эта задача NP-трудна в сильном смысле. Из этого результата и упомянутой эквивалентности следует NP-трудность в сильном смысле задачи MSSC-PD.

Эквивалентная формулировка задачи MSSC-PD

Для целевой функции задачи MSSC-PD справедливо равенство

$$\sum_{j=1}^{J} \sum_{x \in \mathcal{C}_j} \sum_{z \in \mathcal{C}_j} \|x - z\|^2 = 2 \sum_{j=1}^{J} |\mathcal{C}_j| \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2.$$

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$. Поэтому задача кластеризации по критерию минимума правой части последнего равенства NP-трудна в сильном смысле.

Задачи поиска подмножеств. Известные результаты

Для решения задачи Max-Cut-$(\mathbb{R}^q)$SD, очевидно, применимы все существующие приближенные алгоритмы с оценками точности, ориентированные на общий случай задачи.

Однако, интерес представляют и менее трудоёмкие алгоритмы, учитывающие специфику (геометрические свойства) входных данных.

Задачи поиска подмножеств

В анализе данных и распознавании образов нередки ситуации, когда данные задаются в виде матрицы **попарных** сравнений объектов.

При этом требуется найти подмножество из наиболее «похожих» объектов.

Эту ситуацию моделирует следующая задача поиска подмножества строк симметрической матрицы.

Задача RSSM-($\mathbb{R}^q$)PSD (Row's Subset of Symmetric Matrix, Euclidean case for Pairwise Squared Distances)

Дано: матрица $(w_{ij}), 1 \le i, j \le N$, квадратов расстояний между точками евклидова пространства натуральное число $M > 1$.
Найти: подмножество $\mathcal{M}$ мощности M строк этой матрицы такое, что

$$\sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}} w_{ij} \to \min .$$

NP-трудность этой задачи в сильном смысле следует из очевидного полиномиального сведения к ней задачи VS-3.

2-приближенный алгоритм решения этой задачи за время $\mathcal{O}(N^2)$ предложен в [4] (Кельманов, Романченко, 2011) и в [8] (Еремин, Гимади, Кельманов, Пяткин, Хачай, 2013).

Литература

1. Anil K. Jain K. Data Clustering: 50 Years Beyond k-Means // Pattern Recognition Lett. 2010. Vol. 31. P. 651–666.

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

3. Кельманов А.В., Пяткин А.В. NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. 2010. Т.17, № 5. С. 37–45.

4. Кельманов А.В., Романченко С.М. Приближённый алгоритм для решения одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. 2011. Т.18, № 1. С. 61-69.

5. Кельманов А.В., Романченко. Псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. 2012. № 2, С. 156-162.

6. Шенмайер В.В. Аппроксимационная схема для одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. 2012. Т.19, № 2. С. 92–100.

Литература

7. Кельманов А.В.,Романченко С.М. FPTAS для одной NP-трудной задачи поиска подмножества векторов // Труды 8-й международной конференции Дискретная оптимизация и исследование операций. Новосибирск 2013.

8. Еремин И.И., Гимади Э.Х., Кельманов А.В., Пяткин А.В., Хачай М.Ю. 2-прибли-женный алгоритм поиска клики с минимальным весом вершин и рёбер // Труды Института математики и механики УрО РАН. 2013. Т. 19, № 2 (принята в печать).

9. Галашов А.Е.,Кельманов А.В. 2-Приближенный алгоритм для одной задачи поиска семейства непересекающихся подмножеств // Труды 8-й международной конференции Дискретная оптимизация и исследование операций. Новосибирск 2013.

10. Кельманов А.В., Пяткин А.В. О сложности одной задачи о разрезе максимального веса // Труды 8-й международной конференции Дискретная оптимизация и исследование операций. Новосибирск 2013.

Литература

11. Кельманов А. В., Пяткин А. В. О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. физики. 2009. Т.49, № 11. С. 2059–2067.

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

13. Гимади Э.Х., Глазков Ю.В., Рыков И.А. Задача выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы // Дискретный анализ и исследование операций. 2008. Т. 15, № 4. С. 31–43.

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

Литература

15. Гимади Э.Х., Кельманов А.В., Кельманова М.А., Хамидуллин С.А. Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. журн. индустр. математики. 2006. Т. 9, № 1(25). С. 55–74.

16. Кельманов А.В., Пяткин А.В. О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Доклады РАН. 2008. Т. 421, № 5. С. 590–592.

17. Кельманов А.В., Пяткин А.В. Об одном варианте задачи выбора подмножества векторов // Дискретный анализ и исследование операций. 2008. Т. 15, № 5. С. 25–40.

18. Долгушев А.В., Кельманов А.В. Приближенный алгоритм решения одной задачи кластерного анализа // Дискретный анализ и исследование операций. 2011. Т.18, № 2. С. 29–40.

Литература

19. А.В. Кельманов, В.И. Хандеев. Полиномиальный алгоритм с оценкой точности 2 для решения одной задачи кластерного анализа // Дискретный анализ и исследование операций 2013 (принята в печать).

20. А.В. Долгушев, А.В. Кельманов, В.В. Шенмайер. Приближенная полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: 9-я международная конференция. Республика Черногория, г. Будва, 16–22 сентября 2012 г.: Сборник докладов. М.: Торус Пресс, 2012. С. 242-244.

21. А.В. Кельманов, В.И. Хандеев. Рандомизированный алгорритм для одной задачи кластерного анализа // Дискретный анализ и исследование операций 2013. // Труды 8-й международной конференции Дискретная оптимизация и исследование операций. Новосибирск 2013.

Спасибо за внимание!