An exact pseudopolynomial algorithm for a bi-partitioning problem

Alexander Kel'manov, Vladimir Khandeev

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

10th International Conference
Intelligent Information Processing (IIP-10)
Crete, Greece
October 4-11, 2014

Введение

Предмет исследования —

квадратичная евклидова NP-трудная в сильном смысле задача разбиения конечного множества векторов.

Цель исследования —

обоснование точного псевдополиномиального алгоритма для специального случая этой задачи (когда размерность пространства фиксирована).

Области приложений:

анализ данных и распознавание образов, комбинаторная геометрия, математическая статистика, теория приближения.

Введение

Одной из самых известных задач анализа данных и распознавания образов является

Задача MSSC (Minimum Sum-of-Squares Clustering) [Fisher, 1958]

Дано: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$.
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 $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, — центр j-го кластера.

Эта задача также известна под названием k-Means [Edwars, Cavalli-Sfarza, 1965].

К числу слабоизученных относится задача

Задача J-MSSC-F
(Minimum Sum-of-Squares Clustering for the case of Fixed cardinalities and one cluster with known center)

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

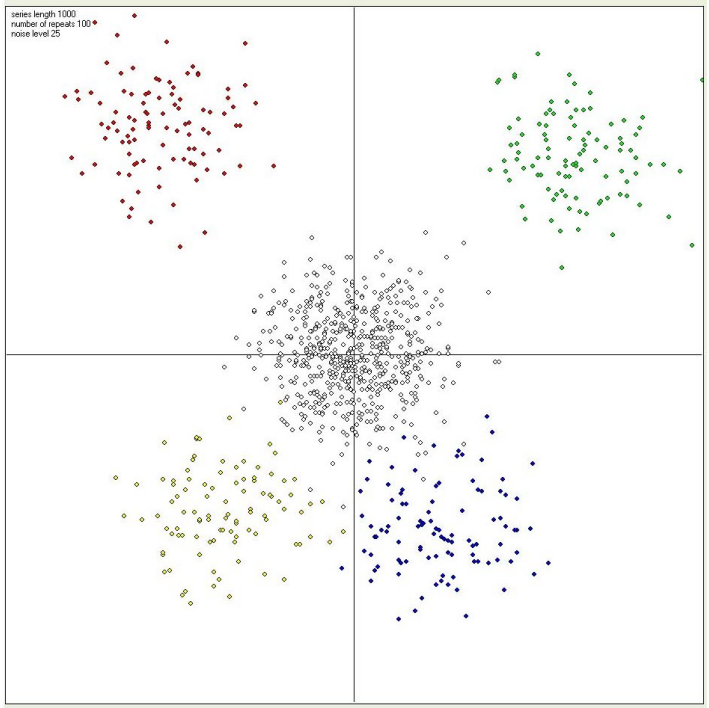
$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_J)} \|y\|^2 \to \min,$$

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

Введение

Пример 1

1000 результатов измерений характеристик 4-х объектов, изображённые на плоскости.
Каждый объект может находиться в активном и пассивном состоянии.



Формулировка задачи

В настоящей работе рассматривается частный случай задачи J-MSSC-F при $J = 1$.

Задача 1-MSSC-F

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

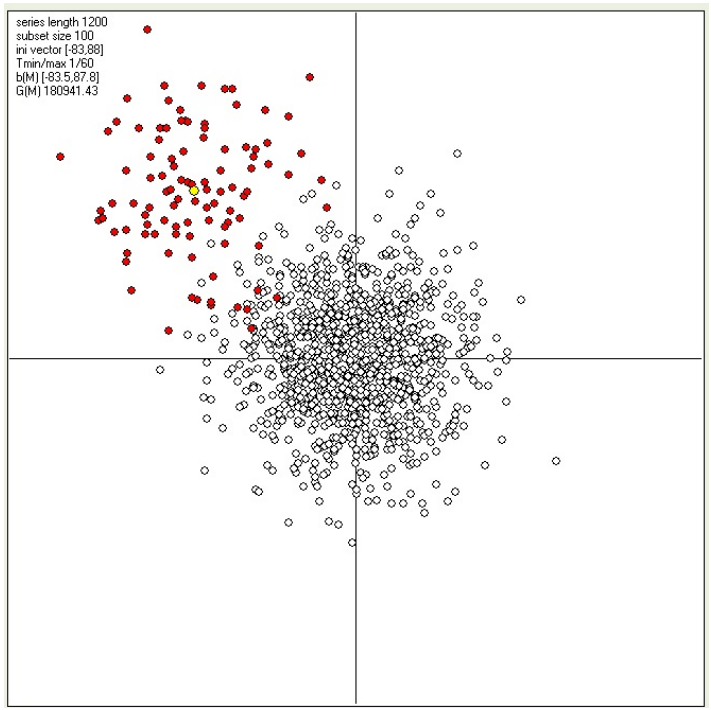
$$S(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min,$$

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

Содержательная трактовка

Пример 2

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



Известные результаты

- Задача NP-трудна в сильном смысле (Кельманов А. В., Пяткин А. В. 2008).
- Предложены:
  - 2-приближённый алгоритм с временной сложностью $\mathcal{O}(qN^2)$ (Долгушев А. В., Кельманов А. В. 2011),
  - схема PTAS, временная сложность которой $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, где ε — относительная погрешность (Долгушев А. В., Кельманов А. В., Шенмайер В. В. 2012),
  - 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 А. В., Хандеев В. И. 2013).

Полученные результаты

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

2. Построен 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 $MD < N^{2-\frac{1}{q}}$.

1. Полиномиальная разрешимость задачи при фиксированной размерности пространства

Утверждение

Задача 1-MSSC-F разрешима за время $\mathcal{O}(q^2 N^{2q})$, полиномиальное при фиксированной размерности q пространства.

Справедливость следует из равенства

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

и результатов Гимади Э. Х., Пяткина А. В., Рыкова И. А. (2008), устанавливающих полиномиальную разрешимость при фиксированной размерности пространства задачи максимизации второго члена в правой части (1).

2. Точный псевдополиномиальный алгоритм для специального случая задачи

Положим

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

где $\mathcal{B} \subseteq \mathcal{Y}$, $|\mathcal{B}| = M$, $b \in \mathbb{R}^q$.

Лемма

- *При любом фиксированном подмножестве $\mathcal{B} \subseteq \mathcal{Y}$ минимум функционала (2) доставляется вектором $\overline{y}(\mathcal{B}) = \frac{1}{M} \sum_{y \in \mathcal{B}} y$ и равен $S(\mathcal{B})$.*
- *При любом фиксированном векторе $b \in \mathbb{R}^q$ минимум функционала (2) достигается на множестве, состоящем из M векторов множества $\mathcal{Y}$, имеющих наибольшие проекции на вектор b.*

2. Точный псевдополиномиальный алгоритм для специального случая задачи

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

Положим

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

и определим множество

$$\mathcal{D} = \{d \mid d \in \mathbb{R}^q, (d)^j = \frac{1}{M}(v)^j, \\ (v)^j \in \mathbb{Z}, |(v)^j| \leq MD, j = 1, \ldots, q\}. \quad (3)$$

2. Точный псевдополиномиальный алгоритм для специального случая задачи

Замечание 1.

Из определения (3) следует, что центр $\overline{y}(\mathcal{C}) = \frac{1}{M} \sum\limits_{y \in \mathcal{C}} y$ любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности M лежит во множестве $\mathcal{D}$. Следовательно, и центр оптимального подмножества лежит в этом же множестве.

Замечание 2.

$$|\mathcal{D}| = (2MD + 1)^q.$$

2. Точный псевдополиномиальный алгоритм для специального случая задачи

Алгоритм $\mathcal{A}$

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

Шаг 1. Для каждого вектора $b \in \mathcal{D}$ 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 $\mathcal{C}_A$.

2. Точный псевдополиномиальный алгоритм для специального случая задачи

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* $\mathcal{O}(qN(2MD+1)^q)$.

Доказательство Теоремы

По определению шага 2 алгоритма имеем

$$G(\mathcal{C}_A, b_A) \leq G(\mathcal{B}(y^*), y^*), \tag{4}$$

где $y^* = \frac{1}{M} \sum\limits_{y \in \mathcal{C}^*} y$ — центр оптимального подмножества $\mathcal{C}^*$.

Из первого утверждения леммы следует оценка

$$S(\mathcal{C}_A) \leq G(\mathcal{C}_A, b_A), \tag{5}$$

а из второго — равенство

$$G(\mathcal{B}(y^*), y^*) = S(\mathcal{C}^*). \tag{6}$$

Объединяя (4) — (6), получаем оценку

$$S(\mathcal{C}_A) \leq S(\mathcal{C}^*).$$

Доказательство Теоремы

С другой стороны, так как множество $\mathcal{C}_A$ является допустимым решением задачи 1-MSSC-F, то справедливо неравенство

$$S(\mathcal{C}^*) \leq S(\mathcal{C}_A),$$

что устанавливает равенство значений $S(\mathcal{C}^*)$ и $S(\mathcal{C}_A)$.

2. Точный псевдополиномиальный алгоритм для специального случая задачи

Доказательство Теоремы

Оценим временную сложность алгоритма.

- Шаг 1. Для каждого из $(2MD+1)^q$ векторов множества $\mathcal{D}$:
  - вычисление проекций на этот вектор — $\mathcal{O}(qN)$ операций;
  - выбор M векторов, имеющих наибольшие проекции — $\mathcal{O}(N)$ операций;
  - вычисление значения функции $G(\mathcal{B}(b), b)$ — $\mathcal{O}(qN)$ операций.
- Шаг 2. Поиск наименьшего элемента — $\mathcal{O}((2MD+1)^q)$ операций.

Итоговая временная сложность — $\mathcal{O}(qN(2MD+1)^q)$.

2. Точный псевдополиномиальный алгоритм для специального случая задачи

Замечание

Если размерность q пространства фиксирована, то трудоёмкость $\mathcal{O}(qN(2MD+1)^q)$ 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 $MD < N^{2-\frac{1}{q}}$.

Заключение

Основные результаты

1. Установлена полиномиальная разрешимость задачи 1-MSSC-F в случае, когда размерность пространства фиксирована.

2. Обоснован псевдополиномиальный алгоритм, гарантирующий отыскание оптимального решения задачи в случае, когда компоненты векторов целочисленны и размерность пространства фиксирована. Установлены условия, при которых этот алгоритм быстрее полиномиального алгоритма.

Заключение

Актуальные вопросы

1. Построение FPTAS для случая фиксированной размерности пространства.

2. Рассмотрение обобщения задачи на случай нескольких кластеров — задачи J-MSSC-F.

Спасибо за внимание!