Экстремальные задачи и вычислительные технологии анализа данных

Анна Владимировна Панасенко

КОНФЕРЕНЦИЯ

ЖЕНЩИНЫ В МАТЕМАТИКЕ

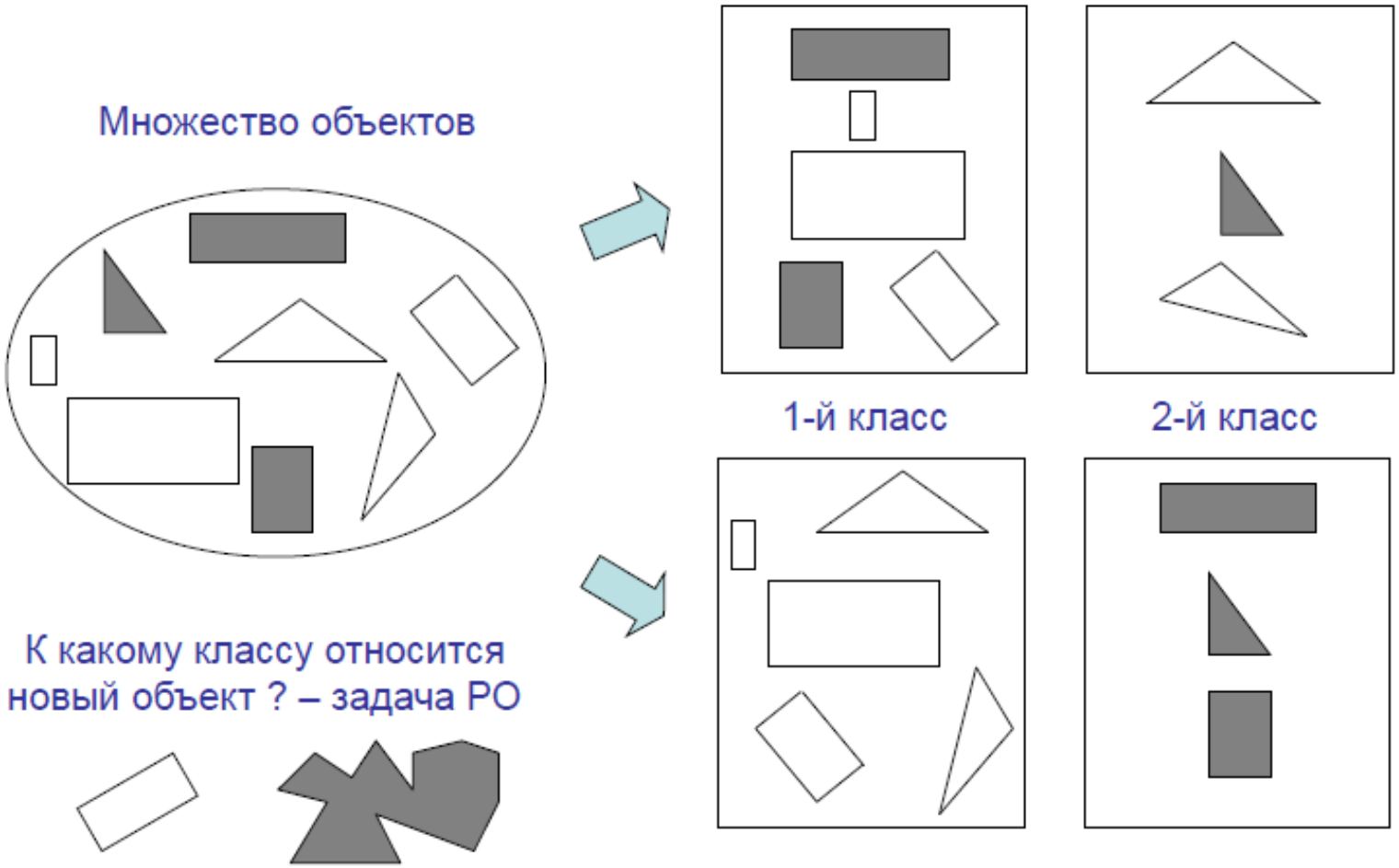
в Институте математики им. С.Л. Соболева

6 июня 2022

НОВОСИБИРСК, РОССИЯ

Пример задачи анализа данных и распознавания образов

Разбить множество объектов на классы похожих объектов – задача АД



Объект, предмет и цели научных исследований

Анализ данных (АД) и распознавание образов (РО) – научное направление, ориентированное на изучение **математических проблем**, возникающих при решении прикладных задач, типичных для широкого спектра естественно-научных, гуманитарных и технических сфер: физики, химии, биологии, медицины, геологии, лингвистики, экономики, социологии, криминалистики, промышленного производства, компьютерных технологий и др.

Большинство рассматриваемых в рамках этого направления проблем возникает в ситуациях, когда по совокупности результатов измерения *(или по результатам вычисления значений)* характеристик каких-либо объектов *(материальных или абстрактных)* **требуется решить одну из следующих содержательных задач:**

Объект, предмет и цели научных исследований

1. Задача разбиения объектов на кластеры

Разбить совокупность объектов на подмножества, содержащие элементы, похожие (или близкие) по некоторому заранее заданному критерию.

2. Задача принятия решения об объекте

По фиксированному критерию принять решение о сходстве некоторого объекта со всеми объектами одного из множеств (образов), входящих в заданное конечное семейство непересекающихся множеств, каждое из которых состоит из похожих (или близких) по этому же критерию объектов.

Объект, предмет и цели научных исследований

3. Задача группировки объектов в кластеры

В совокупности объектов найти семейство подмножеств такое, что каждое подмножество из этого семейства содержит объекты, похожие (или близкие) между собой по некоторому фиксированному критерию.

4. Задача селекции значимых характеристик

Из совокупности измеряемых (или вычисляемых) характеристик выбрать подмножество наименьшей мощности, достаточное для разбиения заданного множества объектов на подмножества похожих (или близких) по некоторому критерию объектов.

Направления исследований лаборатории анализа данных

1. Оптимизационные модели и дискретные экстремальные задачи в анализе данных и распознавании образов.

2. Логико-вероятностные модели и компьютерные технологии анализа данных, распознавания образов и прогнозирования.

3. Анализ символьных последовательностей.

4. Модели интеллектуального анализа данных, распознавания образов и прогнозирования.

Формулировка задачи и известные результаты

1. Cardinality-weighted variance-based 2-clustering with given center

Дано множество $\mathcal{Y} = (y_1, \ldots, y_N)$ точек в $\mathbb{R}^d$, положительное целое число $M < N$.

Найти разбиение $\mathcal{Y}$ на два непустых кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

$$F_1(\mathcal{C}) = M \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min ,$$

где $\overline{y}(\mathcal{C}) = \frac{1}{M} \sum_{y \in \mathcal{C}} y$ — центроид множества $\mathcal{C}$, причем $|\mathcal{C}| = M$.

Задача 1 *NP*-трудна в сильном смысле *[Кельманов, Пяткин, 2015]*.

Поэтому для Задачи 1 не существует точных полиномиального и псевдополиномиального алгоритмов, а также не существует полностью полиномиальной приближенной схемы, если $P \neq NP$ *[Кельманов, Пяткин, 2015]*.

Содержательная трактовка задачи

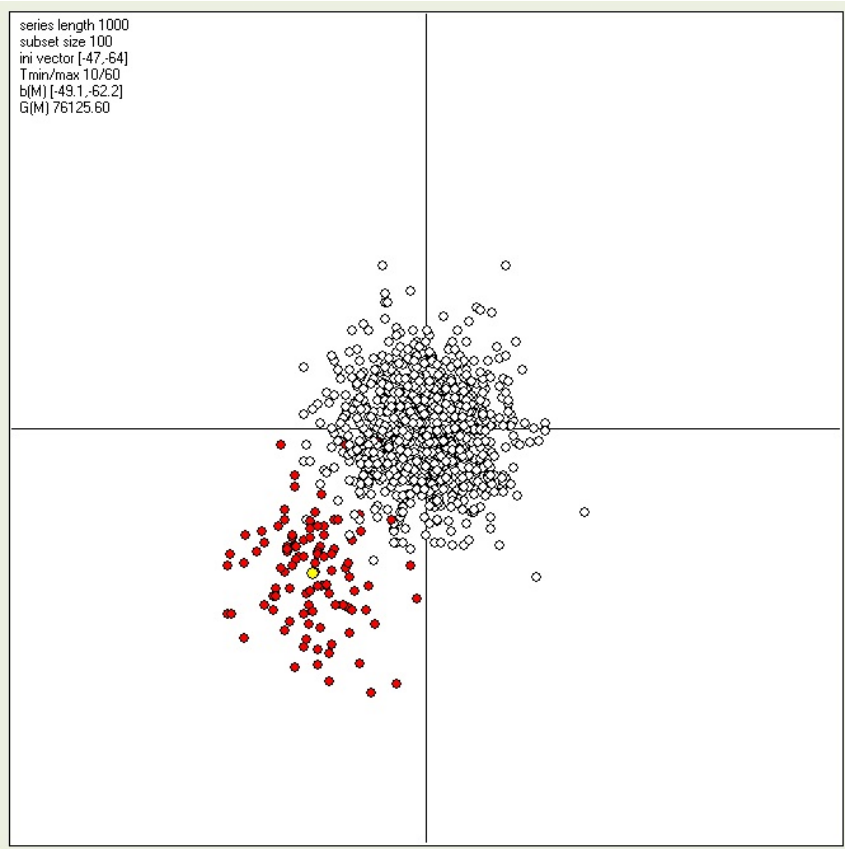


Рис.: Пример входного множества точек, когда центр одного из кластеров фиксирован в начале координат. Двумерность пространства соответствует ситуации, когда измеряются две характеристики объекта.

Введение: близкие в постановочном плане задачи

Задача Weighted variance-based 2-clustering problem

Дано множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^d$.
Найти разбиение множества $\mathcal{Y}$ на два непустых кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

где $\overline{y}(\mathcal{C})$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ — геометрические центры кластеров $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$.

Задача NP-трудна и эквивалентна следующей задаче.

Задача Min-sum all-pairs 2-clustering problem

Дано множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^d$.
Найти разбиение множества $\mathcal{Y}$ на два непустых кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

$$\sum_{x \in \mathcal{C}} \sum_{z \in \mathcal{C}} \|x - z\|^2 + \sum_{x \in \mathcal{Y} \setminus \mathcal{C}} \sum_{z \in \mathcal{Y} \setminus \mathcal{C}} \|x - z\|^2 \longrightarrow \min .$$

Введение: близкие в постановочном плане задачи

Наиболее значимые результаты для этих задач получены в работах:

1. Sahni S., GonzalezT.: P-Complete Approximation Problems. (1976)
2. Brucker P.: On the Complexity of Clustering Problems. (1978)
3. Hasegawa S., Imai H., Inaba M., Katoh N., Nakano J.: Efficient Algorithms for Variance-Based k-Clustering. (1993)
4. Inaba M., Katoh N., Imai H.: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering: (extended abstract). (1994)
5. de la Vega F., Kenyon C.: A Randomized Approximation Scheme for Metric Max-Cut. (2001)
6. de la Vega F., Karpinski M., Kenyon C., Rabani Y.: Polynomial Time Approximation Schemes for Metric Min-Sum Clustering. (2002)
7. Kel'manov A.V., Pyatkin A.V.: NP-Hardness of Some Quadratic Euclidean 2-clustering Problems. (2015)

Полученные результаты: Точный алгоритм для задачи на прямой

Идея алгоритма $\mathcal{A}_1$

1. Упорядочиваем множество точек по возрастанию.
2a. Если $M \geq N/2$, то формируем набор из $(N - M + 1)$ множества, каждое из которых состоит из M последовательных точек.
2b. Если $M < N/2$, то формируем набор из $(M + 1)$ множества, каждое из которых является дополнением до множества из $(N - M)$ последовательных точек.
3. В качестве решения алгоритм выбирает одно из построенных множеств, для которого значение целевой функции Задачи 1 минимально.

Теорема 1.1

Алгоритм $\mathcal{A}_1$ находит оптимальное решение одномерного случая Задачи 1 за время $\mathcal{O}(N \log N)$.

Полученные результаты: Точный псевдополиномиальный алгоритм для специального случая задачи

Идея алгоритма $\mathcal{A}_2$

1. Построим такую многомерную равномерную по каждой координате решетку (сетку) $\mathcal{D}$ с рациональным шагом, чтобы один из ее узлов гарантированно совпал с центроидом искомого подмножества.

2. Для каждого узла $y \in \mathcal{D}$ точным образом решим вспомогательную задачу и найдем подмножество $\mathcal{B}^y$ из M элементов множества $\mathcal{Y}$.

3. В семействе $\{\mathcal{B}^y \mid y \in \mathcal{D}\}$ найденных подмножеств выберем в качестве решения то подмножество, для которого значение целевой функции вспомогательной задачи минимально *(в которой центроид заменен на конкретную точку y)*.

Теорема 1.2

Пусть элементы множества $\mathcal{Y}$ имеют целочисленные компоненты из интервала $[-D, D]$. Тогда Алгоритм $\mathcal{A}_2$ находит оптимальное решение Задачи 1 за время $\mathcal{O}(dN(2MD+1)^d)$.

Полученные результаты: 2-приближенный алгоритм

Идея алгоритма $\mathcal{A}_3$

1. Для каждой точки $y \in \mathcal{Y}$ точным образом решим вспомогательную задачу и найдем подмножество $\mathcal{B}^y$ из M элементов множества $\mathcal{Y}$.

2. В семействе найденных подмножеств выберем в качестве решения то подмножество, для которого значение целевой функции Задачи 1 минимально.

Теорема 1.3

Алгоритм $\mathcal{A}_3$ находит 2-приближенное решение Задачи 1 за время $\mathcal{O}(dN^2)$.

Полученные результаты: FPTAS для специального случая задачи

Идея алгоритма $\mathcal{A}_4$

1. Для каждой точки входного множества построим многомерную решетку (сетку) с адаптивными размером и шагом так, чтобы центроид искомого подмножества гарантированно принадлежал области одной из построенных адаптивных решеток.
2. Для каждого узла y построенной решетки точным образом решим вспомогательную задачу и найдем подмножество $\mathcal{B}^y$ из M элементов множества $\mathcal{Y}$.
3. В семействе найденных подмножеств выберем в качестве решения то подмножество, для которого значение целевой функции Задачи 1 минимально.

Теорема 1.4

Для любого $\varepsilon \in (0,1)$ *алгоритм* $\mathcal{A}_4$ *находит* $(1+\varepsilon)$*-приближенное решение Задачи 1 за время* $\mathcal{O}(dN^2(\sqrt{\frac{2d}{\varepsilon}}+2)^d)$.

Полученные результаты: PTAS

Идея алгоритма $\mathcal{A}_5$

1. Для каждой точки входного множества формируется подмножество из q точек, не включающее рассматриваемую в данный момент точку входного множества.
2. Для каждого сформированного подмножества рассматривается такая область пространства, что центр искомого подмножества обязательно попадает в одну из этих областей.
3. Для заданных параметров решения и рассматриваемого на данном этапе подмножества строится решетка, которая дискретизирует область с постоянным шагом по всем направлениям.
4. Для каждого узла сетки формируется подмножество искомой величины, минимизирующее целевую функцию вспомогательной задачи. Это подмножество объявляется кандидатом для решение исходной задачи.
5. В качестве итогового решения выбирается то множество из множества кандидатов, на котором значение целевой функции исходной задачи является наименьшим.

Полученные результаты: PTAS

Теорема 1.6

Для любых фиксированных $s, t > 0$ Алгоритм $\mathcal{A}_5$ находит $(1/t + 8\zeta(t,s))$-приближенное решение Задачи 1, где

$$\zeta(t,s) = \sqrt{t-1}/s + (t-1)/s^2,$$

за время $\mathcal{O}\left(dN^2\binom{N-1}{q}(L_R + q)\right)$, *где* $q = \min\{M, d+1, t\} - 1$.

Свойство 1.1

В случае $t = 2/\varepsilon$, где $\varepsilon > 0$, и $s = 9t^{3/2}$, алгоритм решает задачу за время $O(dN^{2/\varepsilon+1}((9/\varepsilon)^{3/\varepsilon} + 2/\varepsilon - 1))$ с относительной ошибкой ε.

Свойство 1.2

В случае $t = 2/\varepsilon$, где $\varepsilon > 0$, $s = 9t^{3/2}$, и фиксированной размерности d пространства (или ограниченной сверху константой), алгоритм решает задачу за время $\mathcal{O}(N^{d+2}\varepsilon^{-3d/2})$ с относительной ошибкой ε.

Полученные результаты: Рандомизированный алгоритм

Идея алгоритма $\mathcal{A}_6$

1. Сформируем мультимножество $\mathcal{T}$ из k независимых случайных выборок по одному элементу из $\mathcal{Y}$, где k — натуральный параметр.

2. Для каждого непустого $\mathcal{H} \subseteq \mathcal{T}$ вычислим центроид $\overline{y}(\mathcal{H})$ и сформируем подмножество $\mathcal{C}$ из M элементов множества $\mathcal{Y}$ (решим вспомогательную задачу).

3. В семействе найденных подмножеств выберем в качестве решения то подмножество, для которого значение целевой функции Задачи 1 минимально.

Полученные результаты: Рандомизированный алгоритм

Теорема 1.7

Для произвольного вещественного $\delta \in (0,1)$ и натуральных $t \leq k$ алгоритм $\mathcal{A}_6$ находит $(1+\frac{1}{\delta t})$-приближенные решения задачи 1 за время $\mathcal{O}(2^k d(k+N))$ вероятностью не менее $1-(\delta+\alpha)$, где

$$\alpha = \sum_{i=0}^{t-1} \binom{k}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{k-i}.$$

Следствие 1.1

Пусть $M \geq \beta N$, где $\beta \in (0,1)$ — константа, $\varepsilon > 0$ и $\gamma \in (0,1)$. Тогда при фиксированном параметре $k = \max\left(\left\lceil \frac{2}{\beta}\left\lceil \frac{2}{\gamma\varepsilon}\right\rceil\right\rceil, \left\lceil \frac{8}{\beta}\ln\frac{2}{\gamma}\right\rceil\right)$ алгоритм $\mathcal{A}_6$ находит $(1+\varepsilon)$-приближенное решение задач 1 за время $\mathcal{O}(dN)$ с вероятностью не менее $1-\gamma$.

Теорема 1.8

Пусть $k = \lceil \log_2 N \rceil$ и $M \geq \beta N$, где $\beta \in (0,1)$ — константа. Тогда алгоритм $\mathcal{A}_6$ находит $(1+\varepsilon_N)$-приближенные решения Задачи 1 с вероятностью не менее $1-\gamma_N$ за время $\mathcal{O}(dN^2)$, где $\varepsilon_N \xrightarrow[N\to\infty]{} 0$, $\gamma_N \xrightarrow[N\to\infty]{} 0$.

Задача 2 (на последовательности точек)

Дано последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек в $\mathbb{R}^d$, положительные целые числа $T_{\min}$, $T_{\max}$, $M > 1$.

Найти подмножество $\mathcal{M} = \{n_1, n_2, \ldots\} \subset \mathcal{N} = \{1, \ldots, N\}$ индексов $\mathcal{Y}$ таких, что

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

где $\overline{y}(\{y_n | n \in \mathcal{M}\}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$ — центроид множества $\{y_n | n \in \mathcal{M}\}$, с ограничениями

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

и $|\mathcal{M}| = M$.

Задача 3 (с произвольными весами)

Дано множество $\mathcal{Y} = (y_1, \ldots, y_N)$ точек в $\mathbb{R}^d$, положительное целое $M > 1$ и вещественные числа $w_1 > 0$ and $w_2 \geq 0$.

Найти подмножество $\mathcal{C} \subset \mathcal{Y}$ такое, что

$$F(\mathcal{C}) = w_1 \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + w_2 \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min ,$$

где $\overline{y}(\mathcal{C}) = \frac{1}{M} \sum_{y \in \mathcal{C}} y$ — центроид множества C, причем $|\mathcal{C}| = M$.

Близкая в постановочном смысле задача

Задача 4 (о нахождении подмножества наибольшей мощности)

Дано N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и число $\alpha \in (0, 1)$.
Найти подмножество $\mathcal{C} \subset \mathcal{Y}$ наибольшей мощности такое, что

$$F_1(\mathcal{C}) = M \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \leq \alpha N \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2 .$$

Спасибо за внимание!