Комбинаторика кластерных структур в компактных метрических пространствах*

Пушняков А. С.

141700, Московская область, г. Долгопрудный, Институтский пер., 9, МФТИ

pushnyakovalex@mail.ru

Метрический подход является основой многих методов анализа данных. Если метрика выбрана достаточно удачно, то выполнен принцип компактности: близкие объекты должны лежать скорее в одном классе, нежели в разных [1, 2]. В случае *хорошей* метрики можно полагать, что множество объектов распадается на несколько кластеров, отделенных друг от друга. В этом случае распределение попарных расстояний имеет характерные особенности. Например, можно выделить характерные *внутрикластерные* и *межкластерные* расстояния. Рассматривается следующий вопрос: какие ограничения следует наложить на распределение расстояний, чтобы гарантировать наличие *кластерной структуры* в метрическом пространстве.

Мы будем рассматривать компактное метрическое пространство (X, ρ) с ограниченной борелевской мерой μ (компактную метрическую тройку Громова [3, 4]). Под *r-кластером* мы будем понимать любое измеримое множество диаметра не более r.

Определение 1. *Семейство $2r$-кластеров $\mathcal{X} = \{X_1, \dots X_k\}$ будем называть r-кластерной структурой порядка k, если $\rho(X_i, X_j) \geqslant r$ при всех $1 \leqslant i < j \leqslant k$, где $\rho(A, B) = \inf\{\rho(x, y)\colon x \in A,\, y \in B\}$. Мерой $\mathcal{X}$ назовем величину $\mu(\mathcal{X}) \overset{def}{=} \sum\limits_{i=1}^{k} \mu(X_i)$.*

Если мера некоторой r-кластерной структурой порядка k близка к мере всего пространства, то можно считать, что пространство представляется объединением k кластеров. Наша задача состоит в получении нижней оценки на супремум мер r-кластерных структур порядка k (далее мы будем считать, что k и r фиксированы).

Перед тем, как мы приступим к описанию ограничений на распределение расстояний, отметим следующее

Утверждение 1. *Среди всех r-кластерных структур порядка k есть структура максимальной меры.*

*Работа поддержана грантом РФФИ 18-07-00741

Данное утверждение является прямым следствием компактности X. Идея доказательства состоит в применении теоремы Бляшке [5] к последовательности кластерных структур, меры которых стремятся к своему супремуму.

Пусть далее $\mathcal{X}^*$ — кластерная структура максимальной меры.

Предлагается следующая дискретизация попарных расстояний. Пару точек $(x, y) \in X^2$ будем называть *ребром*, длина ребра — это $\rho(x, y)$. Если $\rho(x, y) \leqslant r$, то будем называть ребро (x, y) *r-коротким*; если $\rho(x, y) > 3r$, то будем называть ребро (x, y) *r-длинным*; все остальные ребра — *r-средние*. Если понятно, о каком r идет речь, то приставка r будет опускаться. Данное разделение соответствует выделению характерных внутрикластерных и межкластерных расстояний. Пусть мера средних ребер *мала* в следующем смысле:

$$M(X) = \frac{1}{2}\mu\{(x, y) \in X^2\colon r < \rho(x, y) \leqslant 3r\} \leqslant \frac{1}{2}\delta\mu(X)^2, \tag{1}$$

где $\delta > 0$ —- параметр.

Следующее ограничение связано с тем, что мы ищем *ровно* k кластеров. Если метрическое пространство в точности является кластерной структурой порядка k, то среди любых $k{+}1$ точек хотя бы две будут из одного кластера. Набор точек $(x_1, \ldots, x_k)$ мы назовем *r-антикликой порядка k*, если $\rho(x_i, x_j) > r$ при всех $1 \leqslant i < j \leqslant k$. Поэтому будем полагать, что $k + 1$-антиклик также *мало* в следующем смысле:

$$T_{k+1}(X) = \frac{1}{(k+1)!}\mu\{(x_1, \ldots x_{k+1}) \in X^{k+1}\colon \rho(x_i, x_j) > r\} \leqslant \frac{\beta\mu(X)^{k+1}}{(k+1)!}, \tag{2}$$

где $\beta > 0$ — параметр.

В ограничениях (1) и (2) можно доказать оценку вида $\mu(\mathcal{X}^*) \geqslant \mu(X)(1 - \varphi(\beta, \delta))$, где $\varphi(\beta, \delta) = o(1)$ при $\beta + \delta \to 0$, однако, сходимость $\varphi(\beta, \delta) \to 0$ является достаточно медленной.

Ограничение (2) отвечает за то, что число кластеров не больше k. Можно ввести аналогическое ограничение, отвечающее за то, что число *достаточно больших* кластеров не менее k: среди наборов из k точек значительная доля должна содержать лишь точки из различных кластеров. Это ограничение формулируется в виде следующей нижней оценки на меру k-антиклик:

$$T_k(X) = \frac{1}{k!}\mu\{(x_1, \ldots x_k) \in X^k\colon \rho(x_i, x_j) > r,\ 1 \leqslant i < j \leqslant k\} \geqslant \frac{\alpha\mu(X)^k}{k!}, \tag{3}$$

где $\alpha > 0$ — параметр. Добавление данного ограничения позволяет значительно улучшить оценку $\mu(\mathcal{X}^*)$.

Мы получим искомую оценку $\mu(\mathcal{X}^*)$ в три шага. Первым шагом является сведение задачи к случаю конечного полуметрического пространства с равномерной мерой. Пусть мы доказали оценку вида $\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta, \delta)\mu(X)$ для конечных пространств, тогда, используя теорему Бляшке, можно получить аналогичную оценку в случае произвольного компактного пространства при условии, что функция $\Psi(\alpha, \beta, \delta)$ и функция распределения $\rho(x, y)$ непрерывны. Далее мы будем считать, что X конечно, а

$\mu(A) = |A|$. В случае конечного пространства мы конструктивно построим разбиение $X = \bigsqcup_{i=1}^{n} Z_i$, обладающее следующим свойством: если в каждом Z_i выбрать максимальный по мощности $2r$-кластер X_i, то набор множеств X_i будет образовывать кластерную структуру в X. Вторым шагом мы покажем, что суммарная мощность k максимальных по мощности Z_i близка к $|X|$. Последним шагом мы оценим разности $|Z_i| - |X_i|$ и покажем, что величина $\sum(|Z_i| - |X_i|)$, где суммирование ведется по выбранным на предыдущем шаге Z_i, мала.

Перейдем к описанию предлагаемой конструкции. Рассмотрим следующую жадную процедуру. Пусть X_1 — множество максимальной мощности среди всех $2r$-кластеров (если таких множеств несколько, то выберем любое). Обозначим его r-окрестность за Z_1, т.е.

$$Z_1 = \{x \in X \colon \rho(x, X_1) < r\}$$

Пусть у нас есть попарно непересекающиеся множества $Z_1, \ldots, Z_m$. Тогда X_{m+1} — множество максимальной мощности среди всех $2r$-кластеров в $X \setminus \bigcup_{i=1}^{m} Z_i$, а множество Z_{m+1} — r-окрестность X_{m+1} во множестве $X \setminus \bigcup_{i=1}^{m} Z_i$, т.е.

$$Z_{m+1} = \left\{ x \in X \setminus \bigcup_{i=1}^{m} Z_i \colon \rho(x, X_{m+1}) < r \right\}$$

Так как мощность X конечна, то процедура оборвется не некотором шаге.

Определение 2. *Построенное разбиение $X = \bigsqcup_{i=1}^{n} Z_i$ мы назовем жадным кластерным разбиением, а семейство $2r$-кластеров $\{X_1, \ldots X_k\}$ назовем жадной r-кластерной структурой порядка k.*

Сделаем несколько замечаний относительно последнего определения. Во-первых, последовательности Z_i и X_i определяются неоднозначно — далее считается, что фиксирована некоторая пара последовательностей (X_i, Z_i). Во-вторых, из построения очевидно, что жадная r-кластерная структура порядка k является r-кластерной структурой порядка k по определению 1. Отметим, что последовательность $\{|X|_i\}_{i=1}^n$ монотонно убывает, однако, для последовательности $\{|Z|_i\}_{i=1}^n$ свойство монотонности в общем случае не выполняется.

Пусть $\{W_i\}_{i=1}^n$ упорядоченные по убыванию мощности множеств Z_i. Оценивая число антиклик в терминах симметрических многочленов от W_i, можно получить следующую оценку:

$$\sum_{i=1}^{k} W_i \geqslant |X| \left(1 - (k+1)\beta^{\frac{1}{k+1}}\right) \tag{4}$$

Уже на данном этапе мы получили коэффициент $\beta^{\frac{1}{k+1}}$. Из следующего простого утверждения следует, что без добавления ограничения 3 оценка (4) неулучшаема в асимптотическом смысле.

Утверждение 2. *Пусть фиксированы $r > 0$ и $0 < \beta < 1$. Существует конечное метрическое пространство (X, ρ) с равномерной мерой такое, что число средних ребер равно нулю, число $k+1$-антиклик удовлетворяет неравенству $T_{k+1}(X) \leqslant \dfrac{1}{(k+1)!}\beta|X|^{k+1}$ и $|X| - \mu(\mathcal{X}^*) \geqslant \dfrac{1}{k+1}\beta^{\frac{1}{k}}$.*

При добавлении ограничения (3) можно получить следующую оценку:

$$\sum_{i=1}^{k} W_i \geqslant \left(1 - \frac{(k+1)!\beta}{\alpha'}\right)|X|, \tag{5}$$

где $\alpha' = \alpha - \dfrac{1}{2}\lambda k^3$, $\lambda = \dfrac{k+1}{2}\delta + \dfrac{(k+1)^2\beta^2}{2\alpha^2}$. Видно, что при фиксированном α оценка (5) по параметру β асимптотически лучше (4).

Завершающий шаг фактически сводится к задаче об оценке мощности максимального кластера в метрическом пространстве при ограничениях вида (1). Более подробно данная задача рассматривается в [6]. Основным результатом является следующая

Теорема 1. *Пусть (A, ρ) — конечное полуметрическое пространство, и множество B является $2r$-кластером максимальной мощности. Тогда суммарное число средних и длинных ребер не менее $\frac{1}{2}\max\{|A|, 2|B|\}|A \setminus B|$.*

Отметим, что множества Z_i являются $4r$-кластерами по построению и описанная выше оценка в общем случае учитывает длинные ребра, число которых никак не ограничено. Однако можно показать, что суммарная мощность Z_i, содержащих *много* длинных ребер, *мала*.

В [6] также показано, что описанная в теореме 1 оценка точна, а граница $2r$ существенна. Последнее напрямую следует из следующего утверждения.

Утверждение 3. *Для любых $\delta > 0$, $\alpha > 0$ и $r' > r$ существует компактное метрическое пространство (X, ρ) такое, что*

$$\mu\{(x, y) \in X^2 \colon r' < \rho(x, y)\} \leqslant \alpha\mu(X)^2,$$

и мера любого $2r$ кластера не более $\delta\mu(X)$.

Данное утверждение очевидным образом переносится на кластерные структуры, причем всегда можно построить пространство так, чтобы $T_{k+1}(X) = 0$. Именно поэтому r-кластерная структура определяется как набор $2r$-кластеров.

Наконец, сформулируем итоговые оценки в зависимости от наличия ограничения (3).

Теорема 2. *Пусть (X, ρ) компактное метрическое пространство с ограниченной борелевской мерой μ, а $\mathcal{X}^*$ — r-кластерная структура максимальной меры. Тогда, если выполнены неравенства (1) и (2), то выполнено неравенство*

$$\mu(\mathcal{X}^*) \geqslant \Psi_1(\delta, \beta)\mu(X), \tag{6}$$

где

$$\Psi_1(\delta, \beta) = 1 - \sqrt{\delta}(2k+1) - (k(e+1)+1)\beta^{\frac{1}{k+1}}$$

Теорема 3. *Пусть (X, ρ) компактное метрическое пространство с ограниченной борелевской мерой μ, $\mathcal{X}^*$ — r-кластерная структура максимальной меры, и функция распределения величины $\rho(x, y)$ непрерывна. Тогда, если выполнены неравенства (1), (2), (3) и $\delta + \dfrac{(k+1)\beta^2}{\alpha^2} \leqslant \dfrac{2}{(k+1)^3}$, то выполнено неравенство*

$$\mu(\mathcal{X}^*) \geqslant \Psi_2(\alpha, \beta, \delta)\mu(X), \tag{7}$$

где

$$\Psi_2(\alpha, \beta, \delta) = 1 - \sqrt{\delta}(2k+1) - \frac{k!(k+2)\beta}{\alpha - \frac{1}{2}k^3\lambda}$$

$$\lambda = \frac{k+1}{2}\delta + \frac{(k+1)^2\beta^2}{2\alpha^2}$$

Список литературы

[1] Загоруйко Н. Г. Гипотезы компактности и λ-компактности в методах анализа данных // Сибирский журнал индустриальной математики. — 1998. — Т. 1, № 1. — С. 114–126.

[2] Браверман Э. М. Опыты по обучению машины распознаванию зрительных образов // Автоматика и телемеханика. — 1962. — Т. 23, № 3. — С. 349–365.

[3] Gromov Mikhail. Metric structures for Riemannian and non-Riemannian spaces. — Springer Science & Business Media, 2007.

[4] Вершик А. М. Универсальное пространство Урысона, метрические тройки Громова и случайные метрики на натуральном ряде // Успехи математических наук. — 1998. — Т. 53, № 5 (323). — С. 57–64.

[5] Половинкин Е. С., Балашов М. В. Элементы выпуклого и сильно выпуклого анализа. — М.: Физматлит, 2004.

[6] Пушняков А. С. О комбинаторных оценках максимальных ε-разбиений метрических конфигураций // Машинное обучение и анализ данных. — 2014. — Т. 7, № 1. — С. 854–862.