О взаимосвязи мер кластеров и распределении расстояний в компактных метрических пространствах

Пушняков А. С.

Аннотация

Рассматривается компактное метрическое пространство с ограниченной борелевской мерой. Под r-кластером понимается любое измеримое множество диаметра не более r. Исследуется существование набора фиксированного числа $2r$-кластеров обладающего следующими свойствами: кластеры попарно отделены друг от друга на расстояние r, и мера набора — суммарная мера кластеров набора, — *близка* к мере всего пространства. Показано, что среди таких наборов существует набор максимальной меры. Для распределения расстояний вводится r-параметрическая дискретизация на *короткие*, *средние* и *длинные* расстояния. В терминах данной дискретизации получена нижняя оценка на меру набора максимальной меры.

Ключевые слова: кластеризация, компактное метрическое пространство, борелевская мера, метрика Хаусдорфа, теорема Бляшке, максимальное паросочетание.

1 Введение

Во многих задачах интеллектуального анализа данных для описания объектов используется метрическая информация [1, 2]. Для задач классификации и кластеризации [3, 4, 5] предполагается, что используемая метрика удовлетворяет так называемому *принципу компактности*: близкие объекты скорее должны лежать в одном классе, нежели в разных [6, 7]. В случае *хорошей* метрики можно полагать, что множество объектов распадается на несколько кластеров, отделенных друг от друга.

Для метрик, представимых в виде объединения кластеров, распределение расстояний имеет некоторые характерные особенности. Если у метрики есть некоторое характерное внутрикластерное расстояние r и межкластерное расстояние R, причем $R > r$, то доля расстояний в промежутке (r, R) *мала*. В данной статье исследуется следующий вопрос: что нужно потребовать от распределения расстояний, чтобы гарантировать наличие кластерной структуры в метрике?

Мы рассматриваем компактные пространства с ограниченной борелевской мерой (или, что тоже самое, компактные метрические тройки Громова [8, 9]). В данных

терминах удобно определить кластерную структуру как набор фиксированного числа отделенных друг от друга кластеров. Ниже будет показано, что среди таких структур найдется структура максимальной меры. Тогда факт *близости* меры данной структуры к мере всего пространства можно интерпретировать как представление метрики в виде объединения кластеров.

Следуя идее внутрикластерных и межкластерных расстояний, все расстояния классифицируются на *короткие*, *средние* и *длинные*, и мы требуем, чтобы доля *средних* расстояний была мала. Также мы потребуем некоторые дополнительные ограничения на распределение расстояний, обусловленные количеством кластеров в искомой кластерной структуре. В терминах параметрических ограничений на распределение расстояний мы получим нижнюю оценку на меру кластерной структуры максимальной меры. Вначале мы докажем искомую оценку для конечных полуметрических пространств с равномерной мерой, а затем, используя теорему Бляшке [10], обобщим оценку на случай компактного пространства.

2 Постановка задачи

Пусть дано компактное метрическое пространство (X, ρ) и ограниченная борелевская мера μ на X. Любое борелевское подмножество X диаметра не более r будем называть *r-кластером*.

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 $\mu(\mathcal{X}) \stackrel{def}{=} \sum\limits_{i=1}^{k} \mu(X_i)$.*

Верно следующее

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

Доказательство. Достаточно рассматривать структуры, содержащие только компактные множества. Для каждой структуры произвольно занумеруем входящие в неё множества. Пусть

$$\mu^* = \sup_{\mathcal{X}} \mu(\mathcal{X}) < \infty$$

Тогда найдется последовательность $\{\mathcal{X}_n\}_{n=1}^{\infty}$ такая, что $\lim\limits_{n\to\infty} \mu(\mathcal{X}_n) = \mu^*$. По теореме Бляшке метрическое пространство компактов из X по метрике Хаусдорфа ρ_H является компактным. Тогда без ограничения общности можно считать $X_{in} \stackrel{\rho_H}{\to} X_i^*$, где $\mathcal{X}_n = \{X_{in}\}_{i=1}^{k}$. Очевидно, что $\mathcal{X}^* = \{X_i^*\}_{i=1}^{k}$ является r-кластерной структурой порядка k. Пусть $\mu^* - \mu(\mathcal{X}) = \delta > 0$. Рассмотрим множества $X_i^{\varepsilon} = \{y\colon \rho(y, X_i^*) \leqslant \varepsilon\}$. При достаточно малых ε

$$\sum_{i=1}^{k} (\mu(X_i^{\varepsilon}) - \mu(X_i^*)) \leqslant \frac{\delta}{2},\ \mu^* - \sum_{i=1}^{k} \mu(X_i^{\varepsilon}) > \frac{\delta}{2}.$$

Но в силу сходимости $X_{in} \overset{\rho_H}{\to} X_i^*$ при фиксированном ε и достаточно больших n выполнено $X_{in} \subset X_i^\varepsilon$, и

$$\mu(\mathcal{X}_n) = \sum_{i=1}^{k} \mu(X_{in}) \leqslant \sum_{i=1}^{k} \mu(X_i^\varepsilon) < \mu^* - \frac{\delta}{2}.$$

Получили противоречие. □

Нашей дальнейшей задачей является определение условий, при которых можно гарантировать, что отношение $\dfrac{\mu(\mathcal{X}^*)}{\mu(X)}$ близко к единице, где $\mathcal{X}^*$ — r-кластерная структура порядка k максимальной меры. Везде далее мы считаем $k \geqslant 2$.

Рассмотрим модельный пример: пусть $X = X_1 \sqcup \ldots \sqcup X_k$, где все множества X_i являются r-кластерами, и $\rho(X_i, X_j) \geqslant R > r$ при всех $1 \leqslant i < j \leqslant k$. В данном случае мера $\frac{r}{2}$-кластерной структуры порядка k максимальной меры равна $\mu(X)$ и выполнены равенства:

$$\{(x_1, \ldots x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r,\ 1 \leqslant i < j \leqslant k+1\} = \varnothing$$

$$\{(x, y) \in X^2 \colon r < \rho(x, y) < R\} = \varnothing$$

Пару точек $(x, y) \in X^2$ будем называть *ребром*, длина ребра — это $\rho(x, y)$ (более наглядная аналогия будет видна в случае $|X| < \infty$). В выше описанном примере нет ребер длины которых лежат в интервале (r, R), а также среди любых $k+1$ точек есть ребро длины не больше r.

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*. Набор точек $(x_1, \ldots, x_k)$ назовем *r-антикликой порядка k*, если $\rho(x_i, x_j) > r$ при всех $1 \leqslant i < j \leqslant k$. Если понятно, о каком r идет речь, то приставка r будет опускаться.

В нижеследующих неравенствах (1) и (2) пары (x, y) и наборы $(x_1, \ldots, x_{k+1})$ мы считаем упорядоченными. Потребуем, чтобы в нашем метрическом пространстве (X, ρ) мера r-средних ребер была *мала* в следующем смысле:

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

а мера r-антиклик порядка $k+1$ была *мала* в следующем смысле:

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

где $\alpha, \beta > 0$ — параметры (мы будем их считать достаточно малыми).

Далее мы докажем, что при выполнении условии (1) и (2) верна оценка меры $\mathcal{X}^*$ вида

$$\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta)\mu(X), \tag{3}$$

где $\Psi(\alpha, \beta) \to 1$ при $\alpha \to 0$ и $\beta \to 0$.

Выбор верхней границы для интервала средних ребер объясняется техническими соображениями, и, гипотетически, может быть уменьшен. Нижняя же граница увеличена быть не может, если мы хотим получить оценку вида (3). Это следует из следующего утверждения.

Утверждение 2. *Для любых $\delta > 0$ и $\alpha > 0$ существует компактное метрическое пространство (X, ρ) такое, что*

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

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

Доказательство. Пусть $X = Y_1 \sqcup \ldots \sqcup Y_s$, и $|Y_i| = m$ при $1 \leqslant i \leqslant s$, а мера любого множества равна его мощности. Определим расстояние ρ следующим образом:

$$\rho(x, y) = \begin{cases} 2r', & x, y \in Y_i \\ r', & x \in Y_i, y \in Y_j,\ i \neq j \end{cases}$$

Тогда мера любого $2r$-кластера не превосходит s, а

$$\mu\{(x, y) \in X^2 \colon r < r' < \rho(x, y)\} = sm(m - 1)$$

Осталось взять $m > \frac{1}{\delta}$, $s > \frac{1}{\alpha}$. □

3 Жадная кластерная структура

Вначале мы получим оценку вида (3) в случае, когда (X, ρ) — конечное полуметрическое пространство. Отметим, что достаточно получить нижнюю оценку меры какой-то r-кластерной структуры порядка k. Рассмотрим следующую жадную процедуру. Пусть X_1 — множество максимальной мощности среди всех $2r$-кластеров (если таких множеств несколько, то выберем любое). Обозначим его окрестность r 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 $X \setminus \bigcup\limits_{i=1}^{m} Z_i$, а множество Z_{m+1} — r-окрестность X_{m+1} во множестве $X \setminus \bigcup\limits_{i=1}^{m} Z_i$, 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 оборвётся не некотором шаге.

Определение 2. *Построенное разбиение* $X = \bigsqcup_{i=1}^{n} Z_i$ *мы назовем жадным кластерным разбиением, а семейство* $2r$*-кластеров* $\{X_1, \dots X_k\}$ *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 $|Z_i|$. Следующим шагом мы покажем, что в условиях (1) и (2) и при достаточно малых α и β первые k по мощности Z_i покрывают *почти все* множество X, т.е. верно неравенство

$$\sum_{i=1}^{k} W_i \geqslant \Phi(\alpha, \beta)|X|, \tag{4}$$

где $\Phi(\alpha, \beta) \to 1$ при $\alpha \to 0$ и $\beta \to 0$.

4 Нижняя оценка числа антиклик

Пусть $T_s(i_1, \dots, i_s)$ — число r-антиклик порядка s, таких, что ровно по одной вершине содержится в каждом из множеств Z_{i_j}. Понятно, что при $s = 1$ выполнено $T_s(i_j) = |Z_{i_j}|$. Нам понадобится следующее рекуррентное соотношение на $T_s(i_1, \dots, i_s)$.

Утверждение 3. *Пусть* $i_1 < \dots < i_s$*, то при* $s \geqslant 2$

$$T_s(i_1, \dots, i_s) \geqslant \frac{|Z_{i_1}|}{s} T_{s-1}(i_2, \dots, i_s), \tag{5}$$

Доказательство. Пусть $x_2, \dots, x_s$ — вершины некоторой антиклики, $x_j \in Z_{i_j}$. Для каждой из вершин x_j рассмотрим множества

$$S(x_j) = \{y \in Z_{i_1} \colon \rho(x_j, y) \leqslant r\},$$

Так как диаметр $S(x_j)$ не более $2r$, то $|S(x_j)| \leqslant |X_{i_1}|$. Пусть $Y = Z_{i_1} \setminus \bigcup_{j=2}^{s} S(x_j)$, тогда $|Y| \geqslant \dfrac{|Z_{i_1}|}{s}$.

Для любой точки $y \in Y$ вершины $y, x_2, \ldots, x_s$ образуют r-антиклику порядка s. Тогда имеем

$$T_s(i_1, \ldots, i_s) \geqslant \sum_{(x_2,\ldots,x_s)} \frac{|Z_{i_1}|}{s} = \frac{|Z_{i_1}|}{s} T_{s-1}(i_2, \ldots, i_s)$$

□

Из утверждения (3) сразу же получаем

$$T_s(i_1, \ldots, i_s) \geqslant \frac{1}{s!} \prod_{j=1}^{s} |Z_{i_j}| \tag{6}$$

Теперь мы получим нижнюю оценку на $T_{k+1}(X)$ — число r-антиклик порядка $k+1$. Нам осталось только просуммировать неравенство (6) по всем наборам из $k+1$ множеств Z_i. Введем обозначение для симметрического многочлена от n переменных

$$\sigma_s(y_1, \ldots y_n) \stackrel{\text{def}}{=} \sum_{1 \leqslant i_1 < \ldots < i_s \leqslant n} \prod_{j=1}^{s} y_j, \tag{7}$$

тогда, используя (2) и (6), получим

$$\frac{1}{(k+1)!} \sigma_{k+1}(z_1, \ldots, z_n) \leqslant T_{k+1}(X) \leqslant \frac{1}{(k+1)!} \beta |X|^{k+1} \tag{8}$$

Разделим каждое z_i на $|X|$ и упорядочим по убыванию: получим набор $w_1 \geqslant \ldots \geqslant w_n$, и тогда

$$\sigma_{k+1}(w_1, \ldots, w_n) \leqslant \beta \tag{9}$$

5 Нижняя оценка для $\sum_{j=1}^{k} W_j$

По сути мы получили следующую задачу оптимизации

$$\begin{cases} f(\mathbf{w}) = \sum\limits_{j=1}^{k} w_j \to \min\limits_{\mathbf{w}} \\ w_i \geq 0 \\ w_i \geq w_j, \ i \leq j \\ \sum\limits_{i=1}^{n} w_i = 1 \\ \sigma_{k+1}(w_1, \ldots, w_n) \leqslant c \end{cases} \tag{10}$$

Очевидно, что задача (10) имеет решение. Мы будем далее считать, что $n > k$, иначе решение задачи (10) очевидно. Нам понадобятся следующие простые утверждения.

Утверждение 4. *Пусть* $\mathbf{w}$ *— решение задачи (10), тогда либо* $w_i = w_j$ *при всех* $1 \leqslant i \leqslant j \leqslant n$, *либо* $\sigma_{k+1}(w_1, \ldots, w_n) = c$.

Доказательство. Предположим противное. Пусть $w_l > w_{l+1}$. Рассмотрим вектор

$$\mathbf{w}^\varepsilon = (w_1^\varepsilon, \ldots w_n^\varepsilon) = (w_1 - (n-l)\varepsilon, \ldots, w_l - (n-l)\varepsilon, w_{l+1} + l\varepsilon, \ldots, w_n + l\varepsilon)$$

Так как $\sigma_{k+1}(w_1, \ldots, w_n) < c$, то при достаточно малых $\varepsilon > 0$ вектор $\mathbf{w}^\varepsilon$ будет допустимым для задачи (10). Но $f(\mathbf{w}^\varepsilon) < f(\mathbf{w})$, получили противоречие. □

Утверждение 5. *Пусть* $\mathbf{w}$ *— решение задачи (10) и* $w_k = \lambda > 0$. *Тогда* $\mathbf{w} = (w_1, \underbrace{\lambda, \ldots, \lambda}_{s}, \mu, 0, \ldots, 0)$, *где* $s \geqslant k-1$ *и* $\mu < \lambda$.

Доказательство. Если все w_i попарно равны λ, то утверждение верно. Для любых $1 \leqslant i < j \leqslant n$ имеем

$$\sigma_{k+1}(w_1, \ldots, w_n) = w_i w_j \sigma_{k-1}(w_1, \ldots, \hat{w_i}, \ldots, \hat{w_j}, \ldots, w_n) +$$

$$+(w_j + w_i)\sigma_k(w_1, \ldots, \hat{w_i}, \ldots, \hat{w_j}, \ldots, w_n) + \sigma_{k+1}(w_1, \ldots, \hat{w_i}, \ldots, \hat{w_j}, \ldots, w_n)$$

Пусть нашлось $2 \leqslant i \leqslant k$ такое, что $w_i > \lambda = w_{i+1} = \ldots = w_k$. Тогда рассмотрим вектор

$$\mathbf{w}' = (w_1 + w_i - \lambda, w_2, \ldots, w_{i-1}, \lambda, \ldots, \lambda)$$

$$\sigma_{k+1}(\mathbf{w}) - \sigma_{k+1}(\mathbf{w}') = (w_1 w_i - (w_1 + w_i - \lambda)\lambda)\sigma_{k-1}(w_1, \ldots, \hat{w_i}, \ldots, \hat{w_j}, \ldots, w_n) > 0$$

Пусть w_l — последняя ненулевая компонента $\mathbf{w}$. Пусть j — первая компонента $\mathbf{w}$ такая, что $w_j < \lambda$, и $j < l$. Тогда рассмотрим вектор

$$\mathbf{w}' = (w_1, \ldots w_{j-1}, \min\{\lambda, w_l + w_j\}, w_{j+1}, \ldots, w_{l-1}, \max\{w_l + w_j - \lambda, 0\}, \ldots, 0)$$

$$\sigma_{k+1}(\mathbf{w}) - \sigma_{k+1}(\mathbf{w}') = (w_j w_l - \min\{\lambda, w_l + w_j\} \max\{w_l + w_j - \lambda, 0\})\sigma_{k-1}(w_1, \ldots, \hat{w_i}, \ldots, \hat{w_j}, \ldots, w_n) > 0$$

Тогда по утверждению (4) получаем, что w — не решение задачи (10). □

Итак, рассмотрим вектор $\mathbf{w} = (w_1, \underbrace{\lambda, \ldots, \lambda}_{s}, \mu, 0, \ldots, 0)$.

$$\sigma_{k+1}(\mathbf{w}) = w_1 \mu \sigma_{k-1}(\underbrace{\lambda, \ldots, \lambda}_{s}) + (w_1 + \mu)\sigma_k(\underbrace{\lambda, \ldots, \lambda}_{s}) + \sigma_{k+1}(\underbrace{\lambda, \ldots, \lambda}_{s}) =$$

$$= w_1 \mu \binom{s}{k-1} \lambda^{k-1} + (w_1 + \mu)\binom{s}{k}\lambda^k + \binom{s}{k+1}\lambda^{k+1} \leqslant c \qquad (11)$$

Мы рассмотрим несколько случаев.

1. $s = k - 1$. Неравенство (11) переходит в

$$\frac{1}{k+1}\mu\lambda^{k-1} \leqslant w_1\mu\lambda^{k-1} \leqslant c$$

$$f(\mathbf{w}) = 1 - \mu \geqslant 1 - \min\left\{\lambda, \frac{c(k+1)}{\lambda^{k-1}}\right\} \geqslant 1 - (c(k+1))^{\frac{1}{k}}$$

2. $s = k$.

$$\frac{1}{k+1}\lambda^k \leqslant w_1\mu k\lambda^{k-1} + (w_1+\mu)\lambda^k \leqslant c, \;\; \lambda \leqslant (c(k+1))^{\frac{1}{k}}$$

$$f(\mathbf{w}) \geqslant 1 - 2\lambda \geqslant 1 - 2(c(k+1))^{\frac{1}{k}}$$

3. $s \geqslant k+1$

$$\left(\frac{s\lambda}{k+1}\right)^{k+1} \leqslant \binom{s}{k+1}\lambda^{k+1} \leqslant c, \;\; s\lambda \leqslant (k+1)c^{\frac{1}{k+1}}$$

$$f(\mathbf{w}) \geqslant 1 - s\lambda \geqslant 1 - (k+1)c^{\frac{1}{k+1}}$$

Так как при $k \geqslant 2$ и $c \leqslant \frac{1}{2}$ выполнено $(k+1)c^{\frac{1}{k+1}} \geqslant 2(c(k+1))^{\frac{1}{k}}$, то верно следующее

Утверждение 6. *Пусть* $\mathbf{w}$ *— решение задачи (10) и* $(k+1)c^{\frac{1}{k+1}} \leqslant 1$. *Тогда*

$$f(\mathbf{w}) \geqslant 1 - (k+1)c^{\frac{1}{k+1}}. \tag{12}$$

Используя соотношения (9) и (12) получаем

$$\sum_{i=1}^{k} W_i \geqslant |X|\left(1 - (k+1)\beta^{\frac{1}{k+1}}\right) \stackrel{\text{def}}{=} \Phi(\alpha,\beta)|X|, \tag{13}$$

6 Оценка меры жадной r-кластерной структуры

Далее мы будем рассматривать только внутреннюю структуру множеств Z_i. Поэтому без ограничения общности можно полагать, что $|Z_i| = W_i$.

Нам осталось доказать, что

$$\sum_{i=1}^{k}|Z_i| - \sum_{i=1}^{k}|X_i| = o(1)|X|, \;\; \alpha+\beta \to 0$$

Рассмотрим множества Z_i и $X_i \subset Z_i$. Для любых $x \in X_i$ и $z \in Z_i$ выполнено $\rho(x,z) \leqslant 3r$, поэтому концы всех длинных ребер лежат в $Z_i \setminus X_i$. Рассмотрим во множестве $Z_i \setminus X_i$ максимальное паросочетание из длинных ребер, которое покрывает множество W_i. Пусть $Y_i = Z_i \setminus (X_i \cup W_i)$, тогда $X_i \cap Y_i$ является $3r$-кластером. Докажем простое утверждение, связывающее мощность X_i и число средних ребер в $X_i \cap Y$.

Утверждение 7. *Пусть (A, ρ) — конечное полуметрическое пространство диаметра не более $3r$, и множество B является $2r$-кластером максимальной мощности. Тогда число средних ребер не менее $M(A) \geqslant \frac{1}{2}|A||A \setminus B|$.*

Доказательство. Пусть x_0 — точка, из которой выходит максимально число коротких ребер, а S — замкнутый шар радиуса r с центром в x_0. Тогда

$$M(A) \geqslant \frac{1}{2}|A|(|A| - |S|) \geqslant \frac{1}{2}|A|(|A| - |B|)$$

□

Также для любого ребра (u_1, u_2) из паросочетания, покрывающего W_i, и точки $x \in X_i$ хотя бы одно из ребер (x, u_j) является средним. В купе с утверждением (7) получаем следующее неравенство:

$$M(Z_i) \geqslant \frac{1}{2}(|X_i| + |Y_i|)|Y_i| + \frac{1}{2}|W_i||X_i| \tag{14}$$

Сейчас мы применим технику аналогичную той, что использовалась при оценке числа антиклик.

Утверждение 8. *Пусть $T_s(Z_i)$ — число r-антиклик порядка s во множестве Z_i. Тогда при $s \geqslant 3$*

$$T_s(Z_i) \geqslant \frac{1}{s}(|Z_i| - (s-1)|X_i|)_+ T_{s-1}(Z_i)$$

Доказательство. Доказательство почти дословно совпадает с доказательством утверждения (3). Пусть $x_1, \dots x_{s-1}$ образуют некоторую атиклику. Для каждой из вершин x_j рассмотрим множества

$$S(x_j) = \{y \in Z_i \colon \rho(x_j, y) \leqslant r\},$$

Так как диаметр $S(x_j)$ не более $2r$, то $|S(x_j)| \leqslant |X_{i_1}|$. Пусть $Y = Z_i \setminus \bigcup_{j=2}^{s} S(x_j)$, тогда

$$|Y| \geqslant (|Z_i| - (s-1)|X_i|)_+.$$

Для любой точки $y \in Y$ вершины $y, x_1, \dots, x_{s-1}$ образуют антиклику порядка s. Осталось заметить что каждую антиклику порядка s мы посчитали не более s раз, тогда имеем

$$T_s(Z_i) \geqslant \sum_{(x_1,\dots,x_{s-1})} \frac{1}{s}(|Z_i| - (s-1)|X_i|)_+ = \frac{1}{s}(|Z_i| - (s-1)|X_i|)_+ T_{s-1}(Z_i).$$

□

Из утверждения (6) и равенства $T_1(Z_i) = |Z_i|$ сразу следует неравенство

$$T_{k+1}(Z_i) \geqslant \frac{1}{(k+1)!}\prod_{j=1}^{k+1}(|Z_i| - (j-1)|X_i|)_+ \tag{15}$$

Если $|X_i|(k+1) \leqslant |Z_i|$, то

$$T_{k+1}(Z_i) \geqslant \left(\frac{|Z_i|}{k+1}\right)^{k+1}$$

Пусть I_1 — множество всех индексов от $1 \leqslant i \leqslant k$ таких, что $|X_i|(k+1) \leqslant |Z_i|$, тогда

$$\left(\frac{\sum_{i\in I_1}|Z_i|}{k(k+1)}\right)^{k+1} \leqslant \frac{1}{(k+1)^{k+1}}\sum_{i\in I_1}|Z_i|^{k+1} \leqslant \sum_{i\in I_1}T_{k+1}(Z_i) \leqslant T_{k+1}(X) \leqslant \frac{\beta|X|^{k+1}}{(k+1)!}$$

$$\sum_{i\in I_1}|Z_i| \leqslant ek\beta^{\frac{1}{k+1}}|X|$$

Если же $|X_i|(k+1) > |Z_i|$, то из неравенства (14) получаем

$$M(Z_i) \geqslant \frac{1}{2(k+1)}|Z_i|(|Y_i| + |W_i|) = \frac{1}{2(k+1)}|Z_i|(|Z_i| - |X_i|)$$

$$|Z_i| - |X_i| \leqslant \frac{2(k+1)M(Z_i)}{|Z_i|}$$

Рассмотрим I_2 — множество таких индексов от $1 \leqslant i \leqslant k$, что $i \notin I_1$ и $|Z_i| \geqslant \sqrt{\alpha}|X|$. Тогда суммируя предыдущее неравенство по множеству I_2:

$$\sum_{i\in I_2}(|Z_i| - |X_i|) \leqslant \frac{2(k+1)M(X)}{\sqrt{\alpha}|X|} \leqslant \sqrt{\alpha}(k+1)|X|$$

Наконец, получаем

$$\sum_{i=1}^{k}(|Z_i| - |X_i|) \leqslant \sum_{i\in I_1}|Z_i| + \sum_{i\in I_2}(|Z_i| - |X_i|) + \sum_{i\notin I_1\cup I_2}|Z_i| \leqslant$$

$$\leqslant ek\beta^{\frac{1}{k+1}}|X| + \sqrt{\alpha}(k+1)|X| + \sqrt{\alpha}k|X| = (\sqrt{\alpha}(2k+1) + ke\beta^{\frac{1}{k+1}})|X|$$

Итак, мы доказали следующую теорему

Теорема 1. *Пусть (X, ρ) конечное полуметрическое пространство с равномерной мерой μ, а $\mathcal{X}^*$ — r-кластерная структура максимальной меры. Тогда, если выполнены неравенства (1) и (2), то*

$$\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta)|X|, \tag{16}$$

где

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

7 Обобщение на случай произвольного компактного пространства

Мы будем использовать технику, аналогичную той, что была использована при доказательстве утверждения (1).

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 (16).*

Доказательство. Фиксируем произвольное $0 < \varepsilon < 1$. В X существует конечная ε-сеть, а значит и разбиение X на конечное число N_ε ε-кластеров $\{A_i\}_{i=1}^{N_\varepsilon}$. Выберем N_ε положительных рациональных чисел $q_1, \dots q_{N_\varepsilon}$ так, что $\mu(A_i) \geqslant q_i$ при $1 \leqslant i \leqslant N_\varepsilon$ и $q_i \geqslant \mu(A_i)(1-\varepsilon)$.

Рассмотрим полуметрическое пространство конечной мощности $X_\varepsilon = B_1 \sqcup \ldots \sqcup B_s$, где $\dfrac{|B_i|}{|B_j|} = \dfrac{q_i}{q_j}$, а функция расстояния ρ_ε определяется следующим образом:

$$\rho_\varepsilon(x,y) = \begin{cases} 0, & x, y \in B_i \\ \rho(A_i, A_j), & x \in B_i, y \in B_j,\ i \neq j \end{cases}$$

Отметим, что

$$|B_i| = \frac{q_i |X_\varepsilon|}{\sum\limits_{j=1}^{N_\varepsilon} q_j} \leqslant \frac{\mu(A_i)|X_\varepsilon|}{(1-\varepsilon)\sum\limits_{j=1}^{N_\varepsilon} \mu(A_j)} = \frac{\mu(A_i)|X_\varepsilon|}{(1-\varepsilon)\mu(X)}$$

Если $x \in B_i$, $y \in B_j$ и $\rho_\varepsilon(x,y) \in (r, 3r]$, то для всех $v \in A_i$, $u \in A_j$ верно $\rho(v,u) \in (r, 3r + 2\varepsilon]$. Отсюда получаем оценку на число r-средних ребер в X_ε:

$$M(X_\varepsilon) = \sum_{1 \leqslant i < j \leqslant N_\varepsilon} [\rho(A_i, A_j) > r] |B_i||B_j| \leqslant$$

$$\leqslant \frac{|X_\varepsilon|^2}{(1-\varepsilon)^2 \mu(X)^2} \sum_{1 \leqslant i < j \leqslant N_\varepsilon} [\rho(A_i, A_j) > r] \mu(A_i)\mu(A_i) \leqslant$$

$$\leqslant \frac{|X_\varepsilon|^2}{2(1-\varepsilon)^2 \mu(X)^2} \mu\{(x,y) \in X^2 \colon r < \rho(x,y) \leqslant 3r + 2\varepsilon\} \overset{(1)}{\leqslant}$$

$$\leqslant \frac{|X_\varepsilon|^2}{2(1-\varepsilon)^2} \left(\alpha + \frac{1}{\mu(X)^2} \mu\{(x,y) \in X^2 \colon 3r < \rho(x,y) \leqslant 3r + 2\varepsilon\} \right) \overset{\text{def}}{=} \frac{1}{2} \alpha_\varepsilon |X_\varepsilon|^2$$

Аналогично имеем оценку для r-антиклик порядка $k+1$:

$$T_{k+1}(X_\varepsilon) = \sum_{1 \leqslant i_1 < \cdots < i_{k+1} \leqslant N_\varepsilon} \prod_{1 \leqslant j < l \leqslant k+1} [\rho(A_{i_j}, A_{i_l}) > r] \prod_{j=1}^{k+1} |B_{i_j}| \leqslant$$

$$\leqslant \frac{|X_\varepsilon|^{k+1}}{(1-\varepsilon)^{k+1}\mu(X)^{k+1}} \sum_{1\leqslant i_1<\dots<i_{k+1}\leqslant N_\varepsilon} \prod_{1\leqslant j<l\leqslant k+1} [\rho(A_{i_j}, A_{i_l}) > r] \prod_{j=1}^{k+1} \mu(A_{i_j}) \overset{(2)}{\leqslant}$$

$$\leqslant \frac{|X_\varepsilon|^{k+1}\beta}{(k+1)!(1-\varepsilon)^{k+1}} \overset{\text{def}}{=} \frac{1}{(k+1)!}\beta_\varepsilon |X_\varepsilon|^{k+1}$$

Заметим, что при $\varepsilon \to 0$ $\alpha_\varepsilon \to \alpha$ и $\beta_\varepsilon \to \beta$. В силу теоремы 1 получаем, что в X_ε существует r-кластерная структура порядка k $\mathcal{C}_\varepsilon = \{C_{1\varepsilon}, \dots C_{k\varepsilon}\}$ меры не менее $\Psi(\alpha_\varepsilon, \beta_\varepsilon)|X|_\varepsilon$.

Понятно, что каждое B_i либо полностью содержится в каком-то множестве семейства $\mathcal{C}_\varepsilon$, либо никакой элемент B_i не входит ни в какое множество семейства $\mathcal{C}_\varepsilon$. Для каждого $C_{i\varepsilon} = B_{j_1} \sqcup \dots \sqcup B_{j_l}$ рассмотрим множество $X_{i\varepsilon} = cl(A_{j_1} \sqcup \dots \sqcup A_{j_l})$ в X. Заметим, что множество $X_{i\varepsilon}$ является $(r + 2\varepsilon)$-кластером, и для любых $1 \leqslant i < j \leqslant k$ выполнено $\rho(X_{i\varepsilon}, X_{j\varepsilon}) \geqslant r$.

Настало время снова применить теорему Бляшке. Пусть $\varepsilon \to 0$. Рассматривая последовательность наборов $(X_{1\varepsilon}, \dots X_{k\varepsilon})$, без ограничения общности можно считать, что $X_{i\varepsilon} \overset{\rho_H}{\to} X_i^*$. Очевидно, что множества X_i^* образуют r-кластерную структуру порядка k, назовем её $\mathcal{X}^*$ Более того

$$\sum_{i=1}^{k} \mu(X_{i\varepsilon}) \geqslant \frac{(1-\varepsilon)\mu(X)}{|X_\varepsilon|} \sum_{i=1}^{k} |C_{i\varepsilon}| \geqslant (1-\varepsilon)\Psi(\alpha_\varepsilon, \beta_\varepsilon)\mu(X) \overset{\varepsilon\to 0}{\to} \Psi(\alpha, \beta)\mu(X).$$

Почти дословно повторяя доказательство утверждения (1), получаем $\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta)\mu(X)$.

□

Список литературы

[1] *Журавлев Ю. И., Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // *Кибернетика.* — 1971. — no. 3. — Pp. 1–11.

[2] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — Наука,, 1970.

[3] *Celebi M. E., Kingravi H. A., Vela P. A.* A comparative study of efficient initialization methods for the k-means clustering algorithm // *Expert Systems with Applications.* — 2013. — Vol. 40, no. 1. — Pp. 200–210.

[4] *De Amorim R. C., Mirkin B.* Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering // *Pattern Recognition.* — 2012. — Vol. 45, no. 3. — Pp. 1061–1075.

[5] *Aggarwal C. C., Reddy C. K.* Data clustering: algorithms and applications. — CRC Press, 2013.

[6] *Загоруйко Н. Г.* Гипотезы компактности и λ-компактности в методах анализа данных // *Сибирский журнал индустриальной математики.* — 1998. — Vol. 1, no. 1. — Pp. 114–126.

[7] *Браверман Э. М.* Опыты по обучению машины распознаванию зрительных образов // *Автоматика и телемеханика.* — 1962. — Vol. 23, no. 3. — Pp. 349–365.

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

[9] *Вершик А. М.* Универсальное пространство урысона, метрические тройки громова и случайные метрики на натуральном ряде // *Успехи математических наук.* — 1998. — Vol. 53, no. 5 (323). — Pp. 57–64.

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