

Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 074 ГРУППЫ

«Комбинаторные оценки ε -разбиений метрических пространств»

Выполнил:

студент 4 курса 074 группы

Пушняков Алексей Сергеевич

Научный руководитель:

д.ф-м.н., чл.-корр. РАН

Рудаков Константин Владимирович

Содержание

1	Введение	3
2	Основные определения	4
3	Свойства максимальных ε -разбиений	5
4	Постановка задачи о максимальном кластере	7
5	Оценка мощности максимального кластера	8
6	Некоторые частные случаи	13
7	Заключение	16

Аннотация

Рассматривается метрическое пространство с конечным числом точек. Вводится понятие максимального ε -разбиения. Рассматриваются нижние оценки на мощность максимального множества диаметра не более ε' при ограничении сверху на число расстояний, превосходящих ε . Показано, что в случае $\varepsilon' < 2\varepsilon$ нельзя гарантировать линейную по мощности пространства оценку. В случае $\varepsilon' \geq 2\varepsilon$ получена наилучшая оценка.

Ключевые слова: метрическая конфигурация, максимальное ε -разбиение, лемма Холла.

1 Введение

Использование метрической информации является основой решения многих задач интеллектуального анализа данных. В задачах классификации и кластеризации метрика на множестве объектов обычно согласуется с *принципом компактности*: близкие объекты скорее должны лежать в одном классе, нежели в разных (см., например, [1, 2]). Таким образом, можно полагать, что в случае *хорошей* метрики, множество объектов распадется на подмножества *небольшого* диаметра.

В данной работе рассматривается произвольное метрическое пространство конечной мощности — метрическая конфигурация. Исследуется существование подмножества *малого* (по сравнению с диаметром всего пространства) диаметра ε' и *достаточно большой* мощности при ограничении на число расстояний превосходящих, значение другого *малого* параметра ε . Иными словами, ставится вопрос, всегда ли при известных и достаточно малых ε' и ε можно найти достаточно большой кластер, покрывающий *значительную* долю точек, например, близкую к единице. Отметим, что поиск такого рода кластеров или метрических сгущений важен для интеллектуального анализа данных [3], визуализации [4] и анализа самих задач классификации [5]. Соответствующим вычислительным методам посвящено значительное число работ (см., например, [6]). Особый интерес представляет связь ε' и ε . Далее будет показано, что в случае $\varepsilon' < 2\varepsilon$ нельзя гарантировать линейную по мощности пространства оценку. В случае $\varepsilon' = 2\varepsilon$ будет получена оценка, стремящаяся к единице при стремящейся к нулю доле расстояний, превосходящих ε . В случае $\varepsilon' > 2\varepsilon$ оценка не изменится, более того, будет показано, что она является неулучшаемой.

Однако введение дополнительных ограничений на метрику позволяет уменьшить ε' . В последнем разделе будут рассмотрены два частных случая: метрика является ультраметрикой и метрика вложима в некоторое евклидово пространство [7]. В случае ультраметрики получаемая неулучшаемая оценка для $\varepsilon' = \varepsilon$, а для евклидовой метрики получена оценка для $\varepsilon' = \sqrt{2}\varepsilon$.

Для получения свойств искомого кластера рассматриваются разбиения метрической конфигурации на подмножества диаметра не более ε' , среди которых *жадным* образом выделяется *максимальное*: среди всех подмножеств диаметра не более ε' вначале выбирается максимальное по мощности, затем из оставшихся точек снова

набирается максимальное по мощности, и так далее. Нижняя оценка на мощность первого подмножества в определяемом ниже *максимальном* ε' -разбиении и есть искомая оценка.

2 Основные определения

Рассмотрим метрическое пространство (X, ρ) с конечным числом точек. Напомним, что *диаметром* множества $Y \subset X$ называется величина $D(Y) = \sup_{x, y \in Y} \rho(x, y)$, причем, в нашем случае $|X| < \infty$, и супремум можно заменить на максимум.

Определение 1. *Набор подмножеств $\{X_0, \dots, X_s\}$ будем называть ε -разбиением метрической конфигурации (X, ρ) , если выполнены следующие условия:*

1. $\bigcup_{i=0}^s X_i = X$;
2. $X_i \cap X_j = \emptyset$ при всех $0 \leq i < j \leq s$;
3. $D(X_i) \leq \varepsilon$.

Зафиксируем произвольное число $\varepsilon > 0$. Из всех подмножеств X выберем такие, что их диаметр не превосходит ε :

$$W_\varepsilon = \{V \subseteq X : D(V) \leq \varepsilon\}. \quad (1)$$

Так как полученное семейство подмножеств конечно (в силу конечности множества X), то существует максимальное по мощности множество $X_0 \in W_\varepsilon$ (если таких несколько, то выберем любое из них). Далее выберем X_1 следующим образом: среди всех $V \in W_\varepsilon$ таких, что $V \cap X_0 = \emptyset$, берем любое из максимальных по мощности. Пусть уже построены множества X_0, \dots, X_k , тогда X_{k+1} строится аналогично: среди всех $V \in W_\varepsilon$ таких, что $V \cap X_i = \emptyset$ при всех $i = 0, \dots, k$, берем любое из максимальных по мощности. В итоге получаем разбиение X на подмножества $X_i, i = 0, \dots, s$, которое будем называть *максимальным ε -разбиением* метрической конфигурации X .

Определение 2. *Набор подмножеств $\{X_0, \dots, X_s\}$ будем называть *максимальным ε -разбиением метрической конфигурации (X, ρ) , если выполнены следующие условия:**

1. набор является ε -разбиением (X, ρ) ;

2. при всех $0 \leq i \leq s$ для любого $Y \in W_\varepsilon$ такого, что $Y \cap \bigcup_{j=0}^{i-1} X_j = \emptyset$, выполнено $|Y| \leq |X_i|$.

Отметим, что максимальное разбиение может быть неединственным. Это иллюстрирует следующий простой пример. Пусть $X = \{x, y, z\}$, $\rho(x, y) = \rho(x, z) = \varepsilon$, $\rho(y, z) = 2\varepsilon$. Тогда ε -разбиения $\{x, y\}, \{z\}$ и $\{x, z\}, \{y\}$ являются максимальными.

3 Свойства максимальных ε -разбиений

Пусть дано некоторое максимальное ε -разбиение $\{X_0, \dots, X_s\}$. Вначале докажем почти очевидное утверждение.

Утверждение 1. Для всех целых неотрицательных i и любой точки $x \in \bigcup_{k=i+1}^s X_k$ существует $y \in X_i$ такое, что $\rho(x, y) > \varepsilon$.

Доказательство. Предположим противное: пусть нашлась такая точка $x \in \bigcup_{k=i+1}^s X_k$, что $\rho(x, y) \leq \varepsilon$, для любого $y \in X_i$. Так как $x \notin \bigcup_{k=0}^i X_k$, то при построении множества X_i к нему можно добавить точку x , и диаметр $X_i \cup \{x\}$ не будет превосходить ε , что противоречит максимальнойности X_i согласно построению. \square

Теперь докажем более сложное утверждение, использующее свойство максимальнойности X_i и лемму Холла о паросочетаниях [8]. Напомним, что *паросочетанием* в графе $G = (E, V)$ называется подмножество ребер $M \subset E$ такое, что никакие два ребра из M не инцидентны одной и той же вершине.

Лемма 1 (Холл). В двудольном графе $G = (E, V)$ с долями V_0 и V_1 существует паросочетание, покрывающее все вершины V_1 , тогда и только тогда, когда для любого подмножества $U \subset V_1$ мощность U не превосходит мощности множества вершин, смежных с вершинами U .

Утверждение 2. Для всех целых неотрицательных i и любого подмножества $Z \subseteq \bigcup_{k=i+1}^s X_k$ такого, что $D(Z) \leq \varepsilon$, существует инъективное отображение $\varphi : Z \rightarrow X_i$, удовлетворяющее условию $\rho(x, \varphi(x)) > \varepsilon$ для любого $x \in Z$.

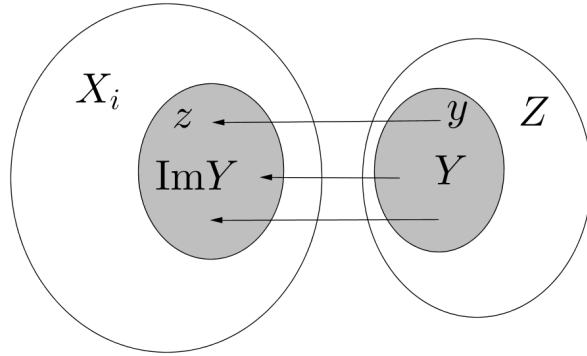


Рис. 1:

Доказательство. Представим множества X_i и Z как доли двудольного графа G , в котором вершины $u \in X_i$ и $v \in Z$ соединены ребром тогда и только тогда, когда $\rho(u, v) > \varepsilon$. Рассмотрим произвольное подмножество $Y \subseteq Z$, $|Y| = k$ и соответствующее ему подмножество $\text{Im}Y \subseteq Z$, определяемое следующим образом:

$$\text{Im}Y = \{z \in X_i : \exists y \in Y, \rho(z, y) > \varepsilon\}. \quad (2)$$

Иными словами, $\text{Im}Y$ — множество вершин графа, смежных с вершинами Y . Покажем, что $|Y| \leq |\text{Im}Y|$. Будем следовать идее доказательства утверждения 1. Предположим противное. Тогда рассмотрим множество $X'_i = (X_i \setminus \text{Im}Y) \cup Y$; в силу предположения имеем $|X'_i| > |X_i|$. Так как для любых $y \in Y$, $x \in X'_i$ верно $\rho(x, y) \leq \varepsilon$, то получаем, что $D(X'_i) \leq \varepsilon$. А это в свою очередь противоречит выбору X_i как максимального по мощности.

Теперь осталось заметить, что построение инъективного отображения $\varphi : Z \rightarrow X_i$, удовлетворяющего условию, эквивалентно нахождению паросочетания в графе G , покрывающего Z . Наличие последнего следует из леммы Холла, условие применения которой не что иное, как только что доказанное неравенство $|Y| \leq |\text{Im}Y|$. \square

До сих пор мы пользовались только свойством симметричности метрики. Следующее утверждение уже использует неравенство треугольника. Пусть $D(V) \leq \varepsilon$, а $Y \cap V = \emptyset$. Аналогично (2) в более общем случае определим:

$$\text{Im}_\varepsilon(Y, V) = \{v \in V : \exists y \in Y, \rho(y, v) > \varepsilon\}. \quad (3)$$

Утверждение 3. Пусть выполнено $D(V) \leq \varepsilon$, $D(Y) \leq 2\varepsilon$, $Y \cap V = \emptyset$ и $\text{Im}_\varepsilon(Y, V) \neq V$, тогда верно $D(Y \cup V) \leq 2\varepsilon$.

Доказательство. Так множество $V \setminus \text{Im}_\varepsilon(Y, V)$ непусто, то возьмём любой его элемент z_0 . Тогда для любого $v \in V$ верно $\rho(v, z_0) \leq \varepsilon$, и для любого $y \in Y$ также выполнено $\rho(y, z_0) \leq \varepsilon$. По неравенству треугольника получаем, что $\rho(v, y) \leq \rho(v, z_0) + \rho(z_0, y) \leq 2\varepsilon$ для любых $v \in V, y \in Y$. А так как $D(V) \leq \varepsilon, D(Y) \leq 2\varepsilon$, то $D(Y \cup V) \leq 2\varepsilon$. \square

4 Постановка задачи о максимальном кластере

Теперь применим вышеописанную конструкцию для решения следующей задачи. Пусть выбран некоторый порог ε , и известно, что из всех $\binom{|X|}{2}$ попарных расстояний число таких, которые больше ε , *мало* в следующем смысле:

$$|\Lambda_\varepsilon| = |\{(x, y) : \rho(x, y) > \varepsilon\}| \leq \frac{\delta |X|^2}{2}, \quad (4)$$

где $\delta > 0$ — некоторое число. В задаче требуется определить некоторый новый порог ε' такой, чтобы *значительная* часть точек пространства попала в некоторое множество диаметра не более ε' . Иначе говоря, требуется найти гарантированную оценку снизу на мощность максимального множества диаметра не более ε' , которое и будем называть *максимальным кластером*.

В начале приведем отрицательный результат, который определит дальнейший выбор ε' .

Утверждение 4. Пусть $\varepsilon \leq \varepsilon' < 2\varepsilon$, тогда для любых $\alpha > 0, \delta > 0$ существует метрическое пространство (X, ρ) такое, что $|\Lambda_\varepsilon| \leq \delta |X|^2$, и $\max\{|V| : D(V) \leq \varepsilon'\} < \alpha |X|$.

Доказательство. Положим $X = \bigcup_{i=1}^s Y_i$, где $Y_i = \{x_{i1}, \dots, x_{im}\}$, $|X| = ms$. Определим метрику следующим образом:

$$\rho(x_{ij}, x_{kl}) = \begin{cases} 2\varepsilon, & i \neq k, j = l, \\ 0, & i = k, j = l, \\ \varepsilon, & \text{иначе.} \end{cases} \quad (5)$$

Очевидно, что данная функция является метрикой. Тогда имеем $D(Y_i) \leq \varepsilon \leq \varepsilon'$, то есть $\max\{|V| : D(V) \leq \varepsilon'\} \geq m$, но для любого $(m+1)$ -элементного подмножества $V \subseteq X$, в нем найдутся x_{ij} и x_{kl} такие, что $i \neq k, j = l$, и тогда $D(V) > \varepsilon'$.

Значит, $\max \{|V| : D(V) \leq \varepsilon'\} = m$. Тогда $|\Lambda_\varepsilon| = m \binom{s}{2}$, и

$$\frac{|\Lambda_\varepsilon|}{|X|^2} = \frac{ms(s-1)}{2s^2m^2} \leq \frac{1}{2m},$$

$$\frac{\max \{|V| : D(V) \leq \varepsilon'\}}{|X|} = \frac{m}{ms} = \frac{1}{s}.$$

Осталось только положить $m = \lfloor \frac{2}{\delta} \rfloor + 1$ и $s = \lfloor \frac{1}{\alpha} \rfloor + 1$. □

Таким образом, чтобы получить достаточно большое значение $\max \{|V| : D(V) \leq \varepsilon'\}$, необходимо брать $\varepsilon' \geq 2\varepsilon$. Мы остановимся на случае $\varepsilon' = 2\varepsilon$. Очевидно, что в случае $\varepsilon' > 2\varepsilon$ оценка как функция от δ не ухудшится.

5 Оценка мощности максимального кластера

Построим, как было описано выше, *максимальное 2ε -разбиение* множества X на множества X_i , $i = 0, \dots, s$, диаметра не более 2ε .

Сформулируем сначала почти очевидное неравенство.

Утверждение 5. *При выполнении неравенства (4) верно неравенство $|X_0| \geq |X|(1 - \delta)$.*

Доказательство. Общее количество расстояний не превосходящих ε не меньше, чем $\frac{|X|(|X| - 1)}{2} - \frac{|X|^2\delta}{2}$. Представим X как граф, в котором вершины u и v соединены ребром тогда и только тогда, когда $\rho(u, v) \leq \varepsilon$. Тогда суммарная степень вершин не менее $|X|(|X| - 1 - \delta|X|)$, а значит, есть вершина z_0 степени не менее $|X| - 1 - \delta|X|$. Тогда в замкнутом шаре $B_\varepsilon(z_0) = \{x : \rho(z_0, x) \leq \varepsilon\}$ есть хотя бы $|X| - \delta|X|$ точек. Осталось заметить, что $D(B_\varepsilon(z_0)) \leq 2\varepsilon$. □

Теперь перейдем к получению более точной оценки. Мы будем оценивать снизу число расстояний превосходящих ε , то есть мощность множества Λ_ε . Как и раньше, удобно рассматривать (X, ρ) как взвешенный полный граф. Будем называть ребро (u, v) *коротким*, если $\rho(u, v) \leq \varepsilon$, все остальные ребра — *длинными*.

В этом графе рассмотрим двудольный подграф $G_1 = (X, E)$ с долями X_0 и $X \setminus X_0$, причем $(u, v) \in E$ тогда и только тогда, когда u и v лежат в разных долях, и $\rho(u, v) > 2\varepsilon$. Согласно утверждению 2 для любого $Y \subseteq X \setminus X_0$ такого, что $D(Y) \leq$

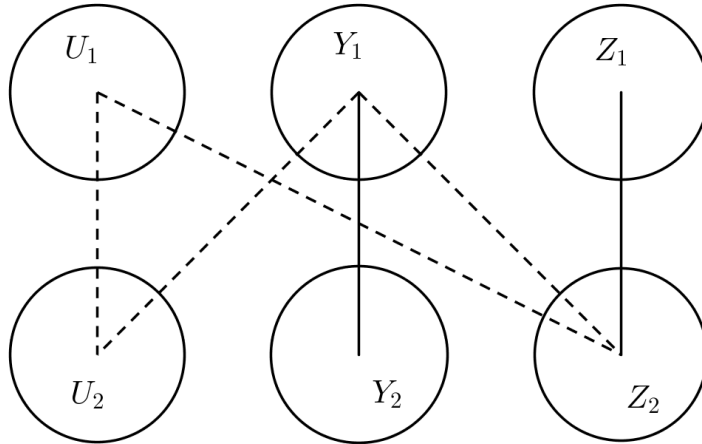


Рис. 2:

2ε , существует инъекция по ребрам G_1 в X_0 . Далее мы покажем, что существует подмножество $X'_0 \subseteq X_0$ мощности не более $|X \setminus X_0|$ такое, что для любого $Y \subseteq X \setminus X_0$ такого, что $D(Y) \leq 2\varepsilon$, существует инъекция по ребрам G_1 в X'_0 .

Рассмотрим следующую конструкцию. Пусть $M \subseteq E$ — максимальное паросочетание в $G_1 = (X, E)$, определенном выше. Путь в G_1 назовем *чередующимся*, если в нем ребра из M и из $E \setminus M$ чередуются. Пусть U_1 — множество вершин в X_0 , не покрытое паросочетанием M , аналогично определяется U_2 . В силу максимальной M , между U_1 и U_2 рёбер нет. Далее, обозначим за Y_1 вершины из X_0 , покрытые M , и до которых существует *чередующийся* путь из U_1 ; остальные вершины X_0 образуют Z_1 . Аналогично определим Y_2 как покрытые M вершины из $X \setminus X_0$, до которых существует *чередующийся* путь из U_1 ; остальные вершины $X \setminus X_0$ образуют Z_2 .

На рис. 2 схематично изображена структура графа G_1 . Сплошными линиями изображены рёбра паросочетания M . Пунктирными линиями обозначено отсутствие рёбер между соответствующими компонентами. Следующее утверждение объясняет, почему граф устроен именно так.

Утверждение 6. *Для вышеописанной конструкции выполнено:*

1. между парами компонент (U_1, U_2) , (U_1, Z_2) , (Y_1, U_2) и (Y_1, Z_2) нет ребер;
2. ребра паросочетания соединяют вершины Y_1 с вершинами Y_2 и вершины Z_1 с вершинами Z_2 .

Доказательство. Так как любом *чередующемся* пути из U_1 в Y_1 первое ребро не лежит в M , то тогда все ребра этого пути, идущие в X_0 , будут лежать в M . Тогда последнее ребро будет принадлежать M . Следовательно, предпоследняя вершина на этом пути лежит в Y_2 (так как до Z_2 нет чередующихся путей из U_1). Более того, среди ребер M нет таких, которые соединяют Y_2 и Z_1 , иначе бы нашелся *чередующийся* путь из U_1 в Z_1 . Следовательно, паросочетание M содержит только ребра между Y_1 и Y_2 и между Z_1 и Z_2 . Значит, $|Y_1| = |Y_2|$ и $|Z_1| = |Z_2|$. Далее, если бы между Z_2 и Y_1 были ребра, то они лежали бы в $E \setminus M$, и мы бы нашли *чередующийся* путь из U_1 в Z_2 . Значит между Z_2 и Y_1 нет ребер.

Теперь покажем, что ребер из U_2 в Y_1 нет. Предположим противное. Пусть $a_0 \in U_2$, $a_1 \in Y_1$, $(a_0, a_1) \in E \setminus M$. Пусть $a_1 a_2 \dots a_{2k+1}$ — *чередующийся* путь из Y_1 в U_1 , тогда $a_0 a_1 \dots a_{2k+1}$ — *чередующийся* путь из U_1 в U_1 . Причем ребра (a_{2i-1}, a_{2i}) лежат в паросочетании, а остальные — нет. Тогда заменим в паросочетании M все ребра (a_{2i-1}, a_{2i}) на (a_{2i}, a_{2i+1}) , и мощность M увеличится на единицу, что противоречит максимальнойности M . Противоречие. □

Теперь используем вышеописанную конструкцию для доказательства следующего утверждения.

Утверждение 7. *Для любого $V \subseteq X \setminus X_0$ такого, что $D(V) \leq 2\varepsilon$, существует инъекция φ по ребрам G_1 в $X'_0 \stackrel{\text{def}}{=} Y_1 \cup Z_1$.*

Доказательство. Обозначим $V_u = V \cap U_2$, $V_y = V \cap Y_2$ и $V_z = V \cap Z_2$, причем $V = V_x \sqcup V_y \sqcup V_z$. Для всех $u \in V_y$ в качестве $\varphi(u)$ возьмем вершину, соединенную с u по ребру из M , то есть $(u, \varphi(u)) \in M$ и $\varphi(u) \in Y_1$. Отметим, что из $U_2 \cup Z_2$ нет ребер в $U_1 \cup Y_1$. Тогда, применяя утверждение 2 для $V_u \cup V_z$, получаем, что для всех $u \in V_u \cup V_z$ верно включение $\varphi(u) \in Z_1$. □

Отметим также, что $|X_1| \leq |X'_0| \leq |X \setminus X_0|$. Теперь перейдем непосредственно к оценке числа *длинных* ребер.

Утверждение 8. *Для любого $u \in X \setminus X'_0$ существует не менее $|X \setminus X_0|$ длинных ребер во множество вершин $X'_0 \cup (X \setminus X_0)$.*

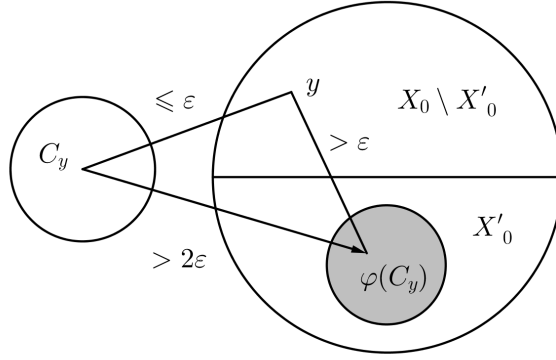


Рис. 3:

Доказательство. Рассмотрим множество $C_y = \{x \in X \setminus X_0 : \rho(x, y) \leq \varepsilon\}$. По неравенству треугольника для любых $x_1, x_2 \in C_y$ имеем $\rho(x_1, x_2) \leq \rho(x_1, y) + \rho(x_2, y)$, а значит, $D(C_y) \leq 2\varepsilon$ и $|C_y| \leq |X_1|$. Тогда по утверждению 7 существует инъекция $\varphi : C_y \rightarrow X'_0$. И для каждого $x \in C_y$ имеем $\rho(y, \varphi(x)) \geq \rho(x, \varphi(x)) - \rho(x, y) > 2\varepsilon - \varepsilon = \varepsilon$. Мы получили $|C_y|$ длинных ребер, идущих в X'_0 , и $|X \setminus X_0| - |C_y|$, идущих в $X \setminus X_0$. \square

Следствие 1. Число длинных ребер, исходящих из $X_0 \setminus X'_0$, не менее $|X \setminus X_0| \cdot |X_0 \setminus X'_0|$.

Утверждение 9. Число длинных ребер между вершинами множества $X'_0 \cup (X \setminus X_0)$ не менее $|X \setminus X_0| \cdot |X'_0|$.

Доказательство. Для каждого $x \in Y_2 \cup Z_2$ обозначим за $\varphi(x)$ такую вершину, что ребро $(x, \varphi(x))$ лежит в максимальном паросочетании M ; по построению $\rho(x, \varphi(x)) > 2\varepsilon$. Также имеем $\bigcup_{x \in Y_2 \cup Z_2} \varphi(x) = Y_1 \cup Z_1 = X'_0$.

Рассмотрим вначале любые две вершины $x, y \in Y_2 \cup Z_2$. Покажем, что среди ребер (x, y) , $(\varphi(x), y)$, $(x, \varphi(y))$, $(\varphi(x), \varphi(y))$ есть хотя бы два длинных. Действительно, если $\rho(x, \varphi(y))$ и $\rho(x, \varphi(y))$ длинные, то утверждение верно. Если же, например, $\rho(x, \varphi(y)) \leq \varepsilon$, то по неравенству треугольника получаем, что $\rho(x, y) \geq \rho(y, \varphi(y)) - \rho(x, \varphi(y)) > 2\varepsilon - \varepsilon = \varepsilon$ и $\rho(\varphi(x), \varphi(y)) \geq \rho(x, \varphi(x)) - \rho(x, \varphi(y)) > 2\varepsilon - \varepsilon = \varepsilon$, то есть ребра (x, y) и $(\varphi(x), \varphi(y))$ длинные. Также имеем, что все ребра $(x, \varphi(x))$ длинные. Тогда получаем следующую оценку на λ_1 — число длинных ребер между вершинами множества $X'_0 \cup Y_2 \cup Z_2$:

$$\lambda_1 \geq |X'_0| + 2 \frac{|X'_0|(|X'_0| - 1)}{2} = |X'_0|^2.$$

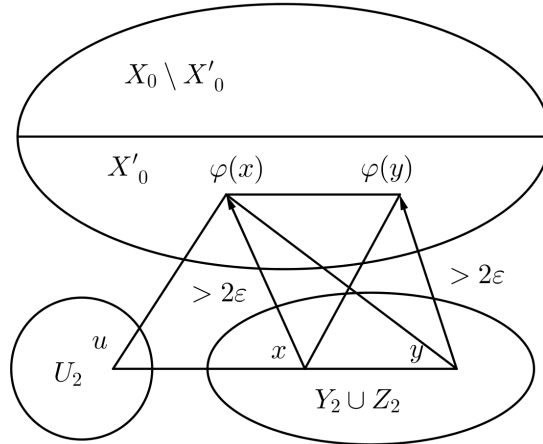


Рис. 4:

Теперь рассмотрим произвольную вершину $u \in U_2$ и пару $(x, \varphi(x)) \in (Y_2 \cup Z_2) \times X'_0$. Опять же используя неравенство треугольника для вершин $x, \varphi(x)$ и u получаем, что одно из ребер (u, x) и $(u, \varphi(x))$ является *длинным*. Тогда мы получаем следующую оценку на число ребер из U_2 в $X'_0 \cup Y_2 \cup Z_2$:

$$\lambda_2 \geq |U_2| \cdot |X'_0|.$$

Складывая полученные неравенства и вспоминая, что $|U_2| + |X'_0| = |X \setminus X_0|$, получаем требуемое неравенство. \square

Из следствия 1 и утверждения 9 сразу получаем центральное утверждение.

Следствие 2. Число длинных ребер в X не меньше, чем $|X_0| \cdot |X \setminus X_0|$.

Теперь воспользуемся неравенством (4) и получим:

$$|X| \cdot |X_0| - |X_0|^2 = |X_0| \cdot |X \setminus X_0| \leq |\Lambda_\varepsilon| \leq \frac{\delta |X|^2}{2}. \quad (6)$$

Будем считать, что $0 < \delta \leq \frac{1}{2}$. Согласно утверждению 5, имеем $X_0 \geq \frac{|X|}{2}$. Тогда, решая неравенство (6) и выбирая решения, подходящие под условие $X_0 \geq \frac{|X|}{2}$, получаем:

$$|X_0| \geq \frac{|X|(1 + \sqrt{1 - 2\delta})}{2}. \quad (7)$$

Следующий простой пример показывает, что полученная оценка на число *длинных* ребер неулучшаема в общем случае. Пусть $X = X_0 \sqcup X_1$ и для любых $x \in X_0, y \in X_1$

$\rho(x, y) = 3\varepsilon$, а все остальные расстояния равны ε . Тогда число *длинных* ребер в точности равно $|X_0| \cdot |X_1|$. Этот пример также подходит для случая $\varepsilon' > 2\varepsilon$ при замене 3ε на $\varepsilon' + 1$.

Наконец, проверим, что мы получили оценку более сильную, чем в утверждении 5. В силу $0 < \delta \leq \frac{1}{2}$ имеем $\sqrt{1 - 2\delta} \geq 1 - 2\delta$ и тогда получаем:

$$\frac{|X|(1 + \sqrt{1 - 2\delta})}{2} \geq |X|(1 - \delta),$$

причем равенство достигается только при $\delta = \frac{1}{2}$. Также исследуем асимптотическое поведение оценки (7) при $\delta \rightarrow +0$:

$$\frac{|X|(1 + \sqrt{1 - 2\delta})}{2} = \frac{|X|(1 + 1 - \delta + o(\delta))}{2} \approx |X| \left(1 - \frac{\delta}{2}\right).$$

6 Некоторые частные случаи

Выше, в утверждении 4, было показано, что в случае $\varepsilon' < 2\varepsilon$ линейную по $|X|$ оценку на $|X_0|$ гарантировать нельзя. Однако при некоторых дополнительных условиях на метрику значение ε' может быть уменьшено. Далее будут рассмотрены два частных случая: метрика является ультраметрикой и метрика вложима в евклидово пространство.

Напомним, что метрика ρ на X называется *ультраметрикой*, если для любых $x, y, z \in X$ выполнено $\rho(x, y) \leq \max\{\rho(y, z), \rho(x, z)\}$. Для ультраметрик известно следующее свойство.

Утверждение 10. *Если ρ — ультраметрика на X , то любой треугольник является равнобедренным. Если треугольник не равносторонний, то основание меньше боковых сторон.*

Покажем, что в случае ультраметрики для получения наилучшей оценки (7) достаточно взять $\varepsilon' = \varepsilon$. Построим максимальное ε -разбиение множества X на множества X_i , $i = 0, \dots, s$.

Утверждение 11. *Если ρ — ультраметрика на X , и выполнено неравенство (4) при $\delta \leq \frac{1}{2}$, тогда верно (7).*

Доказательство. Рассмотрим любую точку $x \in X \setminus X_0$. Согласно утверждению 1 найдем $\varphi(x) \in X_0$ такую, что $\rho(x, \varphi(x)) > \varepsilon$. Рассмотрим произвольную точку $y \in X_0$ такую, что $y \neq \varphi(x)$. Тогда, применяя утверждение 10 к треугольнику $x\varphi(x)y$, получаем $\rho(x, y) > \varepsilon$. Значит, число *длинных* ребер из точки x во множество X_0 равно $|X_0|$. Таким образом, получаем, что число *длинных* ребер не меньше $|X_0| \cdot |X \setminus X_0|$. Отсюда с учётом $\delta \leq \frac{1}{2}$ сразу получается (7). \square

Перейдем к случаю евклидовых метрик. Напомним, что метрическое пространство (X, ρ) называется *вложимым* в евклидово пространство, если существует изометрия из (X, ρ) в некоторое евклидово пространство E . Далее будем для удобства считать, что $X \subset E$.

Рассмотрим произвольную точку $x_0 \in X$ и векторы $v_i = x_i - x_0$, $i = 1, \dots, |X| - 1$. Выразим их скалярные произведения через попарные расстояния

$$\Gamma_{ij} = (v_i, v_j) = \frac{1}{2}(\rho^2(x_i, x_0) + \rho^2(x_j, x_0) - \rho^2(x_i, x_j)), \quad i, j = 1, 2, \dots, |X| - 1. \quad (8)$$

Справедливо следующее утверждение.

Утверждение 12. *Метрическое пространство (X, ρ) , $|X| < \infty$ вложимо в некоторое евклидово пространство тогда и только тогда, когда матрица $\Gamma = \|\Gamma_{ij}\|_{i,j=1}^{|X|-1}$ является неотрицательно определенной.*

Доказательство. Пусть метрическое пространство (X, ρ) вложимо в некоторое евклидово пространство. Матрица Γ есть матрица Грама векторов v_i , откуда сразу следует, что Γ является неотрицательно определенной.

Пусть теперь матрица Γ является неотрицательно определенной. Тогда существует такая ортогональная матрица S и диагональная матрица $D = \text{diag}(\lambda_1, \dots, \lambda_{|X|-1})$, $\lambda_i \geq 0$, что $\Gamma = S^T D S$. Рассмотрим матрицу $U = \sqrt{D} S = \|u_1, \dots, u_{|X|-1}\|$, тогда $\Gamma = U^T U$. Искомое вложение в $\mathbb{R}^{|X|-1}$ строим следующим образом: точка x_0 переходит в начало координат, а любая другая x_i переходит в точку с координатами u_i .

Проверим условие изометрии. Евклидово расстояние в $\mathbb{R}^{|X|-1}$ обозначим за ρ_e . Для любых $i \geq 1$ имеем:

$$\rho_e^2(x_i, x_0) = (u_i, u_i) = \Gamma_{ii} = \rho^2(x_i, x_0).$$

Для любых $i, j \geq 1$, применяя предыдущее равенство, имеем:

$$\rho_e^2(x_i, x_j) = (u_i, u_i) + (u_j, u_j) - 2(u_i, u_j) = \rho_e^2(x_i, x_0) + \rho_e^2(x_j, x_0) - 2\Gamma_{ij} = \rho^2(x_i, x_j).$$

□

Получим удобное следствие из предыдущего утверждения.

Утверждение 13. Пусть $a, b, c, d \in X$, тогда $\rho^2(a, b) + \rho^2(b, c) + \rho^2(c, d) + \rho^2(d, a) \geq \rho^2(a, c) + \rho^2(b, d)$.

Доказательство. Пусть $x = b - a$, $y = c - a$, $z = d - a$. Рассмотрим матрицу Грама $\Gamma(x, y, z)$ векторов x, y, z ; как следствие утверждения 12, она является неотрицательно определенной, и в частности, при $\xi = (1, -1, 1)^T$ выполнено $\xi^T \Gamma(x, y, z) \xi \geq 0$. Подставляя (8), получаем:

$$\begin{aligned} & \rho^2(a, b) + \rho^2(a, c) + \rho^2(d, a) + (\rho^2(a, b) + \rho^2(d, a) - \rho^2(b, d)) - \\ & - (\rho^2(a, c) + \rho^2(d, a) - \rho^2(c, d)) - (\rho^2(a, c) + \rho^2(a, b) - \rho^2(c, b)) \geq 0, \\ & \rho^2(a, b) + \rho^2(a, c) + \rho^2(d, a) + \rho^2(c, b) + \rho^2(c, d) \geq \rho^2(a, c) + \rho^2(b, d). \end{aligned}$$

□

Следствие 3. Пусть $a, b, c, d \in X$ и $\rho(a, c) > \sqrt{2}\varepsilon$ и $\rho(b, d) > \sqrt{2}\varepsilon$, тогда

$$\max\{\rho(a, b), \rho(a, d), \rho(b, c), \rho(c, d)\} > \varepsilon.$$

Теперь положим $\varepsilon' = \sqrt{2}$ и рассмотрим максимальное ε' -разбиение множества X на множества X_i , $i = 0, \dots, s$. Покажем, что в этом случае можно гарантировать линейную по $|X|$ оценку на $|X_0|$.

Утверждение 14. Пусть ρ вложима в евклидово пространство, и выполнено неравенство (4) при $\delta \leq \frac{1}{4}$, тогда верна следующая оценка:

$$|X_0| \geq |X|(1 - 2\sqrt{\delta}).$$

Доказательство. Для каждого нечетного по номеру подмножества X_{2k+1} в максимальном $\sqrt{2}\varepsilon$ -разбиении построим согласно утверждению 2 инъекцию φ по ребрам длины больше $\sqrt{2}\varepsilon$ во множество X_{2k} и получим некоторое паросочетания M , мощность которого:

$$|M| = \sum_{1 \leq 2k+1 \leq s} X_{2k+1} \geq \frac{1}{2} \sum_{k=1}^s X_k = \frac{1}{2}(|X| - |X_0|).$$

Для каждой пары ребер $x\varphi(x)$ и $y\varphi(y)$ из M рассмотрим четырёхугольник $xy\varphi(x)\varphi(y)$, и по следствию 3 среди ребер xy , $x\varphi(y)$, $y\varphi(x)$, $\varphi(x)\varphi(y)$ найдется хотя бы одно *длинное* (больше, чем ε), причем, каждое из этих ребер однозначно определяется парой ребер паросочетания M . Тогда получим, что общее число *длинных* ребер не меньше, чем

$$|M| + \binom{|M|}{2} \geq \frac{1}{2}|M|^2 \geq \frac{1}{8}(|X| - |X_0|)^2$$

Тогда в силу (4) получаем:

$$\frac{1}{8}(|X| - |X_0|)^2 \leq \frac{\delta|X|^2}{2}, \quad |X_0| \geq |X|(1 - 2\sqrt{\delta}).$$

□

7 Заключение

При ограничениях на число расстояний, превосходящих ε , получена неуплучшаемая нижняя оценка на мощность максимального по включению множества диаметра не более 2ε . Показано, что при $\varepsilon' < 2\varepsilon$ в общем случае нельзя гарантировать существования линейной по мощности метрической конфигурации нижней оценки на мощность множества диаметра не более ε' . Рассмотрены частные случаи ультраметрик и евклидовых метрик, для которых значение ε' может быть уменьшено.

Список литературы

- [1] *Аркадьев А., Браверман Э.* Обучение машины классификации объектов. — Наука, 1971.
- [2] *Загоруйко Н.* Гипотезы компактности и λ -компактности в методах анализа данных // *Сибирский журнал индустриальной математики.* — 1998. — Vol. 1, no. 1. — Pp. 114–126.
- [3] *Никитов Г., Рудаков К.* О структуре метрических технологий data mining // *Искусственный интеллект.* — 2002. — Vol. 2.
- [4] Выявление и визуализация метрических структур на множествах пользователей и ресурсов интернет / К. Воронцов, К. Рудаков, В. Лексин, А. Ефимов // *Искусственный Интеллект.* — 2006. — Vol. 2. — Pp. 285–288.
- [5] *Рудаков К., Черепнин А., Чехович Ю.* О метрических свойствах пространств задач классификации // *Доклады Академии наук.* — 2007. — Vol. 416. — Pp. 457–460.
- [6] *Стрижов В., Кузнецов М., Рудаков К.* Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // *Математическая биология и биоинформатика.* — 2012. — Vol. 7. — P. 345–359.
- [7] *Деза М., Лоран М.* Геометрия разрезов и метрик. — М.: МЦНМО, 2001. — P. 736.
- [8] *Hall P.* On representatives of subsets // *J. London Math. Soc.* — 1935. — Vol. 10. — P. 26–30.