Math-Net.Ru

Общероссийский математический портал

С. В. Козырев, Сети кластеров и билдинги Брюа–Титса,
ТМФ, 2014, том 180, номер 2, 234–244

DOI: <https://doi.org/10.4213/tmf8673>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.80.97.219

26 апреля 2023 г., 17:54:26

© 2014 г.

С. В. Козырев*

СЕТИ КЛАСТЕРОВ И БИЛДИНГИ БРЮА–ТИТСА

Рассматривается процедура кластеризации для случая, когда вместо фиксированной метрики используется семейство метрик. В этом случае вместо дерева классификации получается сеть классификации (ориентированный ациклический граф с неориентированными циклами). Обсуждается связь с билдингами Брюа–Титса. Вводится понятие размерности для общей сети кластеров.

Ключевые слова: кластеризация, сети кластеров, билдинги Брюа–Титса.

DOI: 10.4231/tmf8673

1. ВВЕДЕНИЕ

В настоящей работе обсуждается связь конструкции многомерных систем кластеров, описанной в работах [1], [2], и геометрии билдингов Брюа–Титса. Обсуждаются приложения к анализу данных. Процедура кластеризации задает алгоритм построения дерева кластеров с иерархией (частичным порядком), которая начинается с метрики на множестве точек. Эта процедура составляет основу важного метода анализа данных, в частности в приложениях к биоинформатике (построение таксономии).

В настоящей работе обсуждается следующий подход к кластеризации: пусть вместо одной метрики мы имеем семейство метрик, зависящих от набора параметров (такая ситуация является типичной для приложений). Тогда мы получим семейство кластеризаций. Обсудим вопрос, можно ли описать это семейство единым математическим объектом?

Мы будем пользоваться аналогией из p -адической геометрии. В p -адических пространствах имеются естественные иерархии (частично упорядоченные деревья) шаров. Такая иерархия может рассматриваться как семейство кластеров относительно метрики в p -адическом пространстве. В многомерной p -адической геометрии существует обобщение иерархии, описываемое аффинным билдингом Брюа–Титса, который связан с семействами шаров относительно нескольких ультраметрик.

Мы обсуждаем связь между билдингами Брюа–Титса и геометрией p -адических систем кластеров относительно семейства метрик. Мы также рассматриваем обобщение соответствующей структуры симплициального комплекса на общие системы

*Математический институт им. В.А. Стеклова РАН, Москва, Россия.
E-mail: kozyrev@mi.ras.ru

кластеров. В частности, мы обсуждаем определение размерности для общих систем кластеров. Другие приложения p -адических чисел к задачам анализа данных можно найти в работах [3], [4].

Структура настоящей работы следующая. В разделе 2 мы рассматриваем (известное) определение процедуры кластеризации и приводим некоторые примеры сетей кластеров для системы метрик. В разделе 3 мы напоминаем определение аффинных билдингов Брюа–Титса и обсуждаем их связь с сетями кластеров в p -адических пространствах. В разделе 4 мы рассматриваем структуру симплициального комплекса для общих сетей кластеров и определение размерности для общих сетей кластеров.

2. ПРИМЕРЫ СЕТЕЙ КЛАСТЕРОВ

Напомним стандартное определение кластеризации. Мы рассматриваем для простоты кластеризацию по ближайшему соседу (также можно рассматривать более общие методы кластеризации). Общее обсуждение понятия кластеризации можно найти в книге [5].

Пусть (M, ρ) есть некоторое метрическое пространство. Последовательность точек $a = x_0, x_1, \dots, x_{n-1}, x_n = b$ в пространстве (M, ρ) называется ε -цепью, соединяющей две точки a и b , если $\rho(x_k, x_{k+1}) \leq \varepsilon$ для всех $0 \leq k < n$. Если существует ε -цепь, соединяющая a и b , тогда a и b называются ε -связанными. Цепное расстояние между a и b определяется как $d(a, b) = \inf \varepsilon$, где точная нижняя грань вычисляется по всем ε -связанным a и b . Такое расстояние имеет все свойства ультраметрики, кроме невырожденности (т. е. несовпадающие точки могут иметь нулевое цепное расстояние). Таким образом, цепное расстояние определяет ультраметрику на множестве классов эквивалентности точек в M , где a, b лежат в одном классе эквивалентности, если $d(a, b) = 0$. В частности, если исходная метрика ρ является ультраметрикой, то соответствующее цепное расстояние d совпадает с ρ .

Кластер $C(i, R)$ в метрическом пространстве (M, ρ) есть шар с центром i и радиусом R относительно цепного расстояния, т. е. множество $\{j \in M : d(i, j) \leq R\}$. Кластеризация метрического пространства M есть набор кластеров, имеющий следующие свойства:

- 1) каждый элемент из M лежит в некотором кластере;
- 2) для любой пары a, b элементов из M найдется минимальный кластер $\sup(a, b)$, содержащий оба элемента;
- 3) для произвольных вложенных кластеров $A \subset B$ каждая возрастающая последовательность вложенных кластеров $\{A_i\}$, $A \subset \dots \subset A_i \subset A_{i+1} \subset \dots \subset B$, конечна.

Процедура кластеризации генерирует частично упорядоченное дерево кластеров (дендрограмму) следующим образом:

- 1) вершины дерева суть кластеры;
- 2) частичный порядок задан вложением кластеров;
- 3) ребро соединяет два кластера, вложенных без промежуточных.

Многомерное обобщение процедуры кластеризации [1], [2] вводится как обобщение кластеризации на случай нескольких метрик. Пусть у нас есть семейство метрик (это стандартная ситуация в анализе данных), скажем, пусть это семейство параметризуется набором вещественных параметров. В этом случае вместо одного дерева

Рис. 1. Дерево кластеров $\mathcal{A}_1$. Рис. 2. Дерево кластеров $\mathcal{B}_1$. Рис. 3. Дерево кластеров $\mathcal{C}_1$.

кластеров мы получим семейство деревьев кластеров. Некоторые кластеры из различных деревьев могут совпадать как множества. Отождествляя такие вершины для разных деревьев, мы получим сеть (граф) кластеров. Далее мы обсудим связь между сетями кластеров такого вида и билдингами Брюа–Титса. Прежде чем мы перейдем к общим определениям, приведем примеры многомерной кластеризации.

ПРИМЕР 1. Пусть A, B, C – три точки на двумерной вещественной плоскости $\mathbb{R}^2$ со стандартной евклидовой метрикой. Параметры метрики суть координаты точек на плоскости. Дерево кластеров $\mathcal{A}_1$ изображено на рис. 1. Набор кластеров содержит A, B, C, AB, ABC (вершины дерева кластеров), ребра соединяют вершины в соответствии с ростом кластеров, т. е. дерево кластеров содержит ребра

$$(A, AB), \quad (B, AB), \quad (AB, ABC), \quad (C, ABC).$$

Здесь мы обозначаем через ABC кластер, содержащий точки A, B и C .

Рис. 4. Дерево кластеров $\mathcal{A}_2$.

Дерево кластеров $\mathcal{B}_1$ изображено на рис. 2. Рассмотрим деформацию метрики (движение точек на плоскости $\mathbb{R}^2$), которое заменяет описанный выше набор кластеров набором кластеров A, B, C, BC, ABC с соответствующими ребрами

$$(B, BC), (C, BC), (BC, ABC), (A, ABC).$$

На рис. 3 показана сеть кластеров $\mathcal{C}_1$. Такая сеть есть объединение деревьев кластеров $\mathcal{A}_1$ и $\mathcal{B}_1$ (где мы отождествляем вершины, которые отвечают кластерам, совпадающим как множества). Набор вершин сети $\mathcal{C}_1$ содержит кластеры

$$A, B, C, AB, BC, ABC,$$

набор ребер сети $\mathcal{C}_1$ содержит ребра

$$(A, AB), (B, AB), (AB, ABC), (C, ABC), \\ (B, BC), (C, BC), (BC, ABC), (A, ABC).$$

Частичный порядок вершин задан включением кластеров. Циклы в графе описывают различные истории роста кластеров (рост кластеров относительно различных метрик).

ПРИМЕР 2. Рассмотрим множество из четырех точек A, B, C, D , расположенных на плоскости $\mathbb{R}^2$ в вершинах некоторого четырехугольника. Мы получим следующий набор кластеров. Дерево кластеров $\mathcal{A}_2$ изображено на рис. 4. Кластеризация по метрике плоскости дает кластеры

$$A, B, C, D, AB, CD, ABCD.$$

Набор ребер имеет вид

$$(A, AB), (B, AB), (C, CD), (D, CD), (AB, ABCD), (CD, ABCD).$$

На рис. 5 приведено дерево кластеров $\mathcal{B}_2$. Деформация описанного выше четырехугольника дает набор кластеров

$$A, B, C, D, AC, BD, ABCD$$

Рис. 5. Дерево кластеров $\mathcal{B}_2$. Рис. 6. Дерево кластеров $\mathcal{C}_2$.

и набор ребер

$$(A, AC), \quad (C, AC), \quad (B, BD), \quad (D, BD), \quad (AC, ABCD), \quad (BD, ABCD).$$

Сеть кластеров $\mathcal{C}_2$ (см. рис. 6) есть объединение деревьев кластеров $\mathcal{A}_2$ и $\mathcal{B}_2$. Такая сеть содержит объединение наборов вершин и ребер из сетей $\mathcal{A}_2$ и $\mathcal{B}_2$.

3. АФФИННЫЕ БИЛДИНГИ БРЮА–ТИТСА И СЕТИ КЛАСТЕРОВ

В данном разделе мы покажем, что для сети шаров в пространстве $\mathbb{Q}_p^d$ существует естественная структура симплициального комплекса, связанная с аффинным билдингом Брюа–Титса. Теорию билдингов можно найти, например, в монографии [6]. О p -адической геометрии (в частности, решеток) см. книгу [7].

3.1. Аффинный билдинг Брюа–Титса. Вершины билдинга суть классы эквивалентности решеток. Решетка в $\mathbb{Q}_p^d$ есть открытый компактный $\mathbb{Z}_p$ -модуль в $\mathbb{Q}_p^d$. Любая решетка может быть задана в виде

$$\bigoplus_{i=1}^d \mathbb{Z}_p e_i,$$

где $\{e_i\}$ есть базис в $\mathbb{Q}_p^d$. Две решетки являются эквивалентными, если одна получается из другой умножением на скаляр. Две решетки L_1 и L_2 называются смежными

(соединены ребром), если некоторые представители из классов эквивалентности L_1 и L_2 удовлетворяют условию

$$pL_1 \subset L_2 \subset L_1.$$

$(k-1)$ -Симплексы определяются как классы эквивалентности k соединенных ребрами решеток, т. е. как последовательности

$$pL_k \subset L_1 \subset L_2 \subset \dots \subset L_k.$$

Здесь $1 \leq k \leq d$.

Апартамент в аффинном билдинге есть симплицальный подкомплекс, отвечающий фиксированному базису $\{e_i\}$ в пространстве $\mathbb{Q}_p^d$, который содержит классы эквивалентности решеток $\bigoplus_{i=1}^d \mathbb{Z}_p p^{a_i} e_i$, $a_i \in \mathbb{Z}$.

3.2. Многомерная p -адическая метрика. Рассмотрим метрику $s_{q_1, \dots, q_d}(x, y)$ в $\mathbb{Q}_p^d$, определенную нормой $N_{q_1, \dots, q_d}(z)$:

$$s_{q_1, \dots, q_d}(x, y) = N_{q_1, \dots, q_d}(x - y), \quad (1)$$

$$N_{q_1, \dots, q_d}(z) = \max_{i=1, \dots, d} (q_i |z_i|_p), \quad q_i > 0. \quad (2)$$

Растяжения $p^k \mathbb{Z}_p^d$, $k \in \mathbb{Z}$, являются шарами относительно всех норм $N_{q_1, \dots, q_d}$, если $p^{-1} < q_i \leq 1$. Здесь мы используем следующее определение нормы для $\mathbb{Q}_p^d$: норма есть функция $N(\cdot)$ на $\mathbb{Q}_p^d$, принимающая значения в $[0, \infty)$ и удовлетворяющая следующим условиям:

- 1) невырожденность: $N(x) = 0 \Leftrightarrow x = 0$;
- 2) линейность: $N(ax) = |a|_p N(x)$ для всех $x \in \mathbb{Q}_p^d$, $a \in \mathbb{Q}_p$;
- 3) сильное неравенство треугольника: $N(x + y) \leq \max[N(x), N(y)]$ для всех $x, y, z \in \mathbb{Q}_p^d$.

Более общая норма (A -вращение нормы $N_{q_1, \dots, q_d}$) определяется как

$$N_{q_1, \dots, q_d}^A(z) = N_{q_1, \dots, q_d}(Az), \quad (3)$$

где A есть матрица из $Gl_d(\mathbb{Q}_p)$. Метрика $s_{q_1, \dots, q_d}^A$ определена нормой $N_{q_1, \dots, q_d}^A$, как указано выше (см. формулу (1)). В частности, для нормы $N_{q_1, \dots, q_d}$ вида (2), в которой

$$p^{-1} < q_1 < \dots < q_d \leq 1, \quad (4)$$

набор промежуточных $N_{q_1, \dots, q_d}$ -шаров (т. е. шаров относительно нормы $N_{q_1, \dots, q_d}$) между $p\mathbb{Z}_p^d$ и $\mathbb{Z}_p^d$ содержит шары

$$B_j = \mathbb{Z}_p \times \dots \times \mathbb{Z}_p \times p\mathbb{Z}_p \times \dots \times p\mathbb{Z}_p$$

с j компонентами $\mathbb{Z}_p$ и $d - j$ компонентами $p\mathbb{Z}_p$, $j = 0, \dots, d$.

3.3. Симплицальный комплекс шаров. Определим структуру симплицального комплекса на сети $\mathcal{C}$ шаров относительно определенного выше семейства метрик $s_{q_1, \dots, q_d}^A$, $A \in Gl_d(\mathbb{Q}_p)$. Пусть s – метрика из описанного семейства и I – s -шар (т. е. шар относительно метрики s), содержащий ноль, где ноль есть вектор

в $\mathbb{Q}_p^d$ с нулевыми координатами. Тогда растяжение pI также есть s -шар. Содержащие ноль s -шары I и J соединены ребром, если $pI \subset J \subset I$. Далее, $(k-1)$ -симплексы определяются как семейства k смежных s -шаров вида

$$pI_k \subset I_1 \subset I_2 \subset \dots \subset I_k.$$

Рассмотрим максимальную последовательность вложенных s -шаров между pI и I . Если параметры q_i нормы имеют общее положение (любые два параметра не могут быть сделаны равными друг другу умножением на степени p , например параметры удовлетворяют соотношению (4)), то такая последовательность содержит $d+1$ шар и определяет $(d-1)$ -симплекс. Общие симплексы в симплициальном комплексе шаров относительно метрики $s_{q_1, \dots, q_d}^A$ определяются как сдвиги описанных симплексов (сдвиги понимаются как сдвиги наборов множеств в $\mathbb{Q}_p^d$). Симплициальный комплекс $\mathcal{C}$ шаров относительно семейства $\{s_{q_1, \dots, q_d}^A\}$ метрик есть объединение комплексов шаров относительно различных метрик $s_{q_1, \dots, q_d}^A$. В данном случае мы отождествляем s -шар и s' -шар, совпадающие как множества (а также отождествляем s - и s' -симплексы, которые совпадают как наборы шаров).

3.4. Связь между нормами в $\mathbb{Q}_p^d$ и симплексами в аффинном билдинге. Шар относительно нормы в $\mathbb{Q}_p^d$, содержащий ноль, есть решетка. Это следует из сильного неравенства треугольника. Рассмотрим для определенного выше симплициального комплекса $\mathcal{C}$ шаров подкомплекс $\mathcal{C}_0$ шаров, содержащих ноль. Для любого шара I , содержащего ноль, растяжение $p^k I$, $k \in \mathbb{Z}$, также является шаром (относительно той же нормы). Это же имеет место для симплексов. Следовательно, фактор $\mathcal{C}_0/\Gamma$ по группе растяжений на степени p^k , $k \in \mathbb{Z}$, также есть симплициальный комплекс.

Существует естественное симплициальное отображение из симплициального комплекса $\mathcal{C}_0/\Gamma$ в аффинный билдинг Брюа–Титса, ставящее в соответствие шару отвечающую ему решетку. Это отображение представляет собой вложение комплекса $\mathcal{C}_0/\Gamma$ в аффинный билдинг. Покажем, что это отображение сюръективно (т.е. является изоморфизмом симплициальных комплексов).

Мы будем говорить, что две нормы эквивалентны, если они генерируют одно и то же семейство шаров. Покажем, что любому максимальному симплексу в аффинном билдинге Брюа–Титса можно поставить в соответствие класс эквивалентности норм в пространстве $\mathbb{Q}_p^d$.

Пусть L – решетка в $\mathbb{Q}_p^d$ и

$$pL = L_0 \subset L_1 \subset \dots \subset L_d = L \quad (5)$$

есть максимальная последовательность (различных) вложенных решеток (или, что эквивалентно, максимальный симплекс в аффинном билдинге). Для каждой пары $L_j \supset L_{j-1}$, $j = 1, \dots, d$, последовательных решеток в такой последовательности выберем элемент $f_j \in L_j$ такой, что $f_j \notin L_{j-1}$. Это дает набор $\{f_1, \dots, f_d\}$ векторов в $\mathbb{Q}_p^d$. Имеет место следующая лемма.

ЛЕММА 1. *Определенный выше набор $\{f_1, \dots, f_d\}$ есть базис в $\mathbb{Q}_p^d$; решетки L_j имеют вид*

$$L_j = \left(\bigoplus_{i=1}^j \mathbb{Z}_p f_i \right) \oplus \left(\bigoplus_{i=j+1}^d p \mathbb{Z}_p f_i \right). \quad (6)$$

Введем норму в $\mathbb{Q}_p^d$ следующим образом. Поставим в соответствие решеткам L_j из (5) некоторые положительные числа q_j такие, что $p^{-1} < q_1 < \dots < q_d \leq 1$. Определим функцию $N(x)$ на $\mathbb{Q}_p^d$ следующим образом. Для $x \in L_j \setminus L_{j-1}$, $j = 1, \dots, d$, положим $N(x) = q_j$. Для $x = 0$ положим $N(x) = 0$. Определим функцию $N(\cdot)$ на всем пространстве $\mathbb{Q}_p^d$, используя условие $N(p^k x) = p^{-k} N(x)$, $k \in \mathbb{Z}$.

ЛЕММА 2. *Функция $N(\cdot)$ является нормой в $\mathbb{Q}_p^d$, удовлетворяющей сильному неравенству треугольника. Последовательность (5) решеток есть максимальная последовательность шаров относительно нормы $N(\cdot)$, лежащих между шарами pL и L .*

Введенная норма N принадлежит семейству (3) с параметрами, удовлетворяющими условию (4). В частности, матрица A может быть выбрана как матрица, отображающая базис $\{f_1, \dots, f_d\}$ из леммы 1 на координатный базис в $\mathbb{Q}_p^d$, при этом $N(\cdot) = N_{q_1, \dots, q_d}^A(\cdot)$. Любые две нормы, определенные таким образом, будут эквивалентны (будут порождать одно и то же семейство шаров).

Итак, мы построили норму вида (3) с условием (4), начав с максимального симплекса из аффинного билдинга. Аналогично рассмотрим для нормы $N_{q_1, \dots, q_d}^A(\cdot)$, определенной формулами (3), (4), набор решеток (6), где базис $\{f_1, \dots, f_d\}$ задается матрицей A описанным выше образом. Данный набор определяет симплекс в аффинном билдинге. Таким образом, мы показали, что существует взаимно однозначное соответствие между классами эквивалентности норм, удовлетворяющих соотношениям (3), (4), и максимальными симплексами в аффинном билдинге Брюа–Титса.

Требование, что мы рассматриваем нормы с параметрами общего положения, является существенным. Возьмем общую норму N вида (2), (3) и выберем N -шар L . Возможно, что (например, в случае, когда какие-то параметры q_i в (2) равны друг другу) множество промежуточных шаров между L и pL содержит менее чем $d + 1$ шаров, следовательно, не может определять $(d - 1)$ -симплекс в аффинном билдинге.

4. ОБЩИЕ СЕТИ КЛАСТЕРОВ

В настоящем разделе мы обсудим обобщение на общие системы кластеров описанной выше сети кластеров $\mathcal{C}$, связанной с семейством метрик. Мы рассматриваем структуру симплицеального комплекса на такой сети и определение размерности для системы кластеров.

Пусть X – локально компактное ультраметрическое пространство с определенным на нем некоторым конечным семейством ультраметрик $\mathbf{s}$. Более того, пусть для каждой пары метрик $s, r \in \mathbf{s}$ любой s -шар будет конечным объединением r -шаров. Семейство $\mathbf{s}$ ультраметрик на X совместно, если для любых двух шаров (s -шара I и r -шара J , $s, r \in \mathbf{s}$) пересечение $I \cap J$ является шаром относительно некоторой ультраметрики $t \in \mathbf{s}$.

Поставим в соответствие метрике $s \in \mathbf{s}$ дерево $\mathcal{T}(X, s)$ s -шаров в X . Вершины этого дерева суть s -шары, две вершины соединены ребром, если соответствующие шары вложены без промежуточных шаров. Граф $\mathcal{C}(X, \mathbf{s})$ (сеть кластеров в X относительно семейства метрик $\mathbf{s}$) есть объединение деревьев $\mathcal{T}(X, s)$ s -шаров, $s \in \mathbf{s}$. Множество вершин $\mathcal{C}(X, \mathbf{s})$ есть объединение множеств s -шаров, $s \in \mathbf{s}$, ребра соединяют s -шары (с одинаковым s), вложенные без промежуточных. Частичный порядок в $\mathcal{C}(X, \mathbf{s})$ определен вложением подмножеств в X . Если некоторый s -шар совпадает с некоторым r -шаром как множество, то они определяют одну и ту же вершину в множестве $\mathcal{C}(X, \mathbf{s})$.

Определим симплексы в сети $\mathcal{C}(X, \mathbf{s})$. Пусть $\mathbf{r} \subset \mathbf{s}$ есть некоторое подсемейство метрик в X . Выберем некоторый $\mathbf{r}$ -шар I (т.е. I есть s -шар для всех $s \in \mathbf{r}$). Пусть J – наименьший $\mathbf{r}$ -шар, строго больший I . Определим симплекс как подмножество (мощностью не менее двух) множества s -шаров, лежащих между I и J (для метрик $s \in \mathbf{r}$); $(k - 1)$ -симплекс будет содержать k s -шаров, в частности ребро в $\mathcal{T}(X, s)$ (пара шар – максимальный подшар) будет 1-симплексом.

Итак, мы определили структуру симплициального комплекса на дереве $\mathcal{T}(X, s)$. Объединение по метрикам $s \in \mathbf{s}$ таких симплициальных комплексов задает структуру симплициального комплекса на сети $\mathcal{C}(X, \mathbf{s})$ (где мы, как обычно, отождествляем вершины, отвечающие шарам, которые совпадают как множества, и отождествляем симплексы, которые совпадают как наборы вершин). Идея такого определения симплексов взята из p -адического случая, обсуждавшегося в предыдущем разделе. В этом случае симплексы в аффинном билдинге Брюа–Титса заданы наборами шаров, лежащих между шарами I и pI . Если I – шар относительно нескольких метрик $s_{q_1, \dots, q_d}^A$ (скажем, заданными различными вращениями A или различным порядком индексов q_j), то pI также будет шаром относительно тех же метрик. Следовательно, приведенное выше определение симплициального комплекса $\mathcal{C}(X, \mathbf{s})$ обобщает определение симплициального комплекса $\mathcal{C}$ шаров в $\mathbb{Q}_p^d$ из предыдущего раздела (вместо рассмотрения симплексов как подмножеств последовательности вложенных шаров, лежащих между I и растяжением pI , мы рассматриваем циклы в сети кластеров $\mathcal{C}(X, \mathbf{s})$, при этом симплекс есть подмножество пути из минимальной вершины цикла в максимальную).

4.1. Размерность сети кластеров. Обобщим определение размерности (равной числу p -адических параметров) на общие системы кластеров.

Пусть $\mathbf{r} \subset \mathbf{s}$ есть подсемейство (семейства $\mathbf{s}$) метрик в X . Пусть I, J – пара $\mathbf{r}$ -шаров (s -шаров относительно всех метрик $s \in \mathbf{r}$), где $I \subset J$ и J есть минимальный $\mathbf{r}$ -шар, содержащий I . Рассмотрим максимальный s -симплекс, $s \in \mathbf{r}$, отвечающий шарам I и J (т.е. максимальная последовательность вложенных шаров между I и J). Ранг такого симплекса¹⁾ мы назовем $\mathbf{r}$ -размерностью для пары I, J .

В p -адическом случае семейство $\mathbf{r}$ задано набором метрик $s_{q_1, \dots, q_d}^A$ с фиксированной матрицей A и всевозможными перестановками параметров

$$p^{-1} < q_1 < \dots < q_d \leq 1.$$

¹⁾В принципе такой ранг может отличаться для различных $s \in \mathbf{r}$, тогда мы выбираем максимальный ранг.

Заданная выше размерность будет равна числу d p -адических параметров. Введенная таким образом размерность не равна VC-размерности (размерности Вапника–Червоненкиса, или комбинаторной размерности) [8].

4.2. Приложения к анализу данных. Кластеризация представляет собой один из методов анализа данных и имеет различные приложения, в частности к биоинформатике. Множество X данных может генерироваться сложным образом, в них могут присутствовать независимо генерируемые вклады. Подобные независимости должны как-то описываться на уровне сетей кластеров. Классификационные деревья (деревья кластеров) описывают разнообразие данных, многомерное обобщение должно отвечать ситуации, когда имеются независимые источники разнообразия. При этом размерность сети кластеров будет определять число источников разнообразия.

Применение к таксономии в задачах сетевой эволюции. Обсудим применение описанных выше сетей кластеров в биоинформатике. Одной из задач биоинформатики является построение филогенетических классификационных деревьев путем сравнения генетических маркеров (некоторых подпоследовательностей в геномах). Такие деревья строятся при помощи кластеризации с использованием некоторых метрик на генетических маркерах.

Рассмотрим случай, когда имеется несколько генетических маркеров. В этом случае метрика определена не единственным образом и имеет вид суммы вкладов от различных генетических маркеров:

$$d(X, Y) = \sum_{j=1}^N w_j d_j(X_j, Y_j),$$

где $w_j \geq 0$ суть веса, X и Y – геномы, X_j и Y_j представляют собой генетические маркеры, а d_j есть метрика для j -го генетического маркера. Различные наборы весов генерируют (вообще говоря) различные классификационные деревья. В частности, для случая, когда только один вес d_j отличен от нуля, мы получим классификационное дерево для j -го генетического маркера. Объединение таких деревьев иногда называют “лес жизни” [9].

Для некоторых генетических маркеров соответствующие классификационные деревья будут различными, поскольку эти маркеры могут иметь различную эволюционную историю. В этом случае эволюция является сетевой – некоторые части генома могут иметь разное происхождение благодаря гибридизации или горизонтальному переносу генов. Тогда вместо филогенетических деревьев рассматривают филогенетические сети. Обсуждение математических методов анализа филогенетических сетей можно найти в работах [10], [11]. Рассмотренные в настоящей работе классификационные сети могут определить общие рамки для построения филогенетических сетей. При таком подходе вместо воспроизведения детальной генетической истории популяций (что в общем случае не представляется возможным) можно использовать классификационные сети для грубого описания эволюции ансамблей генов.

Благодарности. Работа была частично поддержана программой “Современные проблемы теоретической математики” Отделения математики РАН.

Список литературы

- [1] S. Albeverio, S. V. Kozyrev, *p-Adic Numbers Ultrametric Anal. Appl.*, **4**:3 (2012), 167–178, arXiv: 1204.5952.
- [2] С. В. Козырев, *ТМФ*, **164**:3 (2010), 394–400.
- [3] J. Benois-Pineau, A. Khrennikov, *Computer J.*, **53**:4 (2010), 417–431.
- [4] Ж. Бенуа-Пиньо, А. Ю. Хренников, Н. Котович, *Докл. РАН*, **381**:5 (2001), 604–609.
- [5] F. Murtagh, *Multidimensional Clustering Algorithms*, Physica-Verlag, Heidelberg, 1985.
- [6] P. B. Garrett, *Buildings and Classical Groups*, Chapman and Hall, London, 1997.
- [7] А. Вейль, *Основы теории чисел*, Мир, М., 1972.
- [8] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [9] Е. В. Кунин, *Логика случая. О природе и происхождении биологической эволюции*, Центрполиграф, М., 2014.
- [10] D. H. Huson, R. Rupp, C. Scornavacca, *Phylogenetic Networks*, Cambridge Univ. Press, Cambridge, 2010.
- [11] A. Dress, K. T. Huber, J. Koolen, V. Moulton, A. Spillner, *Basic Phylogenetic Combinatorics*, Cambridge Univ. Press, Cambridge, 2012.

Поступила в редакцию 10.03.2014