УДК 514.14 + 519.25

Кратчайшие маршруты решетки кластерных разбиений конечного множества

Дронов С.В.
Алтайский государственный университет
dsv@math.asu.ru

Аннотация

В работе изучается метрическая структура семейства всех кластерных разбиений $\boldsymbol{\Xi}$ заданного конечного множества в кластерной метрике, введенной ранее автором работы. Оказывается, эта кластерная метрика вполне согласована со структурой решетки в частичном упорядочении $\boldsymbol{\Xi}$ по включению. Это означает, что кратчайший маршрут между двумя разбиениями в семействе $\boldsymbol{\Xi}$ может быть проложен по элементам, любые два соседних из которых сравнимы между собой. При этом маршрут между двумя разбиениями оказывается, вообще говоря, тем короче, чем более мелкими составляющими кластерами обладают лежащие на этом маршруте кластерные разбиения.

1. Основная задача. Определение кластерной метрики

Довольно часто при решении задач самых различных областей математики приходится иметь дело сразу с несколькими разбиениями одного и того же конечного множества на непустые попарно непересекающиеся части. Вероятно, чаще всего с такими разбиениями приходится иметь дело при применении аппарата кластерного анализа и решении разнообразных задач классификации. Поскольку автору ближе всего именно кластерные задачи, в работе используется терминология кластерного анализа, хотя, справедливости ради, следует отметить, что принципиально важная там концепция близости объектов одной из частей (кластеров) основного множества при относительной удаленности объектов из разных частей в некоторой метрике в настоящей работе напрямую нигде не используется. Правда, нет и никаких препятствий тому, чтобы ввести понятие близости объектов, используя уже имеющееся разбиение, считая близкими все объекты, относящиеся к одному и тому же кластеру.

Ниже под кластерным разбиением (иногда просто разбиением) основного множества U из n элементов будем понимать произвольный набор непустых его подмножеств $\mathbf{A} = \{A_1, ..., A_m\}$ таких, что

$$(\forall i, j)\ (i \neq j)\ \Rightarrow\ (A_i \cap A_j = \emptyset), \quad \bigcup_{j=1}^{m} A_j = U. \tag{1}$$

Элементы кластерного разбиения, допуская, как уже было сказано, некоторую вольность речи, будем называть кластерами.

Поскольку мы собираемся изучать степень различия разбиений множества U, полученных различными способами, то зададим метрику d на семействе $\boldsymbol{\Xi}$ всех возможных таких разбиений. После этого мы приходим к изучению специального метрического пространства $\langle \boldsymbol{\Xi}, d \rangle$. Заметим, что на том же семействе $\boldsymbol{\Xi}$ имеется естественным образом (по включению) определенный частичный порядок, в котором $\boldsymbol{\Xi}$ является решеткой. Теория

решеток изложена в [1,2], при этом [1] считается классической монографией, а в [2] можно найти некоторые новые результаты.

Основная задача работы состоит в том, чтобы выяснить связь метрической структуры семейства разбиений Ξ с упомянутым частичным порядком на этом семействе и попытаться увидеть в изучаемой метрической структуре аналогии и различия с обычной евклидовой метрической геометрией.

Будем далее выражаться точнее. Пусть $U = \{x_1, ..., x_n\}$ – основное множество, и на нем заданы два разбиения $\mathbf{A}$, $\mathbf{B}$. Обозначим для любого конечного множества D через $|D|$, количество его элементов. Тогда $|U| = n$. Из определения (1) следует, что для любого $x \in U$ в каждом из разбиений $\mathbf{A}$, $\mathbf{B}$ имеется ровно по одному множеству, элементом которых он является. Если использовать для этих множеств обозначения A_x, B_x соответственно, то, следуя [3], метрика d на заданной паре $\mathbf{A}, \mathbf{B} \in \Xi$ определяется следующей формулой:

$$d(\mathbf{A}, \mathbf{B}) = \sum_{x \in X} |A_x \Delta B_x|,$$

где $A_x \Delta B_x$ обозначена симметрическая разность множеств A_x, B_x. В той же работе приводится следующая формула, на практике позволяющая более легко вычислять расстояния между конкретными кластерными разбиениями: пусть $\mathbf{A} = \{A_1, ..., A_m\}$, $\mathbf{B} = \{B_1, ..., B_k\}$. Тогда

$$d(\mathbf{A}, \mathbf{B}) \ = \ \sum_{i=1}^{m} \sum_{j=1}^{k} |A_i \cap B_j| \cdot |A_i \Delta B_j|. \tag{2}$$

Максимальное возможное свое значение $n(n-1)$ метрика d принимает только в случае, когда одно из разбиений совпадает с наиболее мелким из всех возможных разбиений основного множества $\underline{\mathbf{U}} = \{\{x_1\}, ..., \{x_n\}\}$ в то время, как второе совпадает с наиболее крупным из них $\overline{\mathbf{U}} = \{U\}$. Это также доказано в [3].

2. Сумма квадратов количеств элементов как характеристика разбиения

Пусть $\mathbf{A}$, $\mathbf{B}$ – два кластерных разбиения основного множества. Условимся писать $\mathbf{A} \subseteq \mathbf{B}$, и говорить о разбиении $\mathbf{A}$, как о более мелком, чем $\mathbf{B}$, если для произвольного $A \in \mathbf{A}$ найдется такое $B \in \mathbf{B}$, что $A \subseteq B$. Определенное сейчас отношение включения на семействе Ξ является отношением частичного порядка. При этом понятно, что Ξ относительно этого отношения образует решетку, причем $\underline{\mathbf{U}}, \overline{\mathbf{U}}$ являются в этой решетке наименьшим и наибольшим элементами соответственно.

Для кластерного разбиения $\mathbf{A} = \{A_1, ..., A_m\}$ основного множества определим

$$sq_{\mathbf{A}} = \sum_{j=1}^{m} |A_j|^2.$$

Понятно, что эта характеристика по сути представляет собой сумму квадратов m натуральных чисел, сумма которых равна n. Почти очевидно (а ниже это будет доказано строго), что $sq_{\mathbf{A}}$ принимает тем меньшие значения, чем мельче отдельные кластеры в $\mathbf{A}$.

Лемма 1. *Пусть* $\mathbf{A} = \{A_1, ..., A_m\}$, $\mathbf{B} = \{B_1, ..., B_k\}$, $\mathbf{A} \subseteq \mathbf{B}$. *Тогда* $d(\mathbf{A}, \mathbf{B}) = sq_{\mathbf{A}} - sq_{\mathbf{B}}$.

Доказательство. Из определения следует, что может быть построен такой набор непересекающихся множеств натуральных чисел $N(i)$, $i = 1, ..., m$, что $\bigcup\limits_{i=1}^{m} N(i) = \{1, ..., k\}$, и для произвольного i справедливо $B_i = \bigcup\limits_{j \in N(i)} A_j$.

При $j \in N(i)$ имеем

$$|B_i \Delta A_j| = |B_i| - |A_j| = \sum_{s \in N(i)} |A_s| - |A_j|.$$

Тогда, применяя (2), запишем

$$d(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{k} \sum_{j \in N(i)} |A_j| \left(\sum_{s \in N(i)} |A_s| - |A_j| \right) = \sum_{i=1}^{k} \sum_{j \in N(i)} |A_j| \cdot \sum_{s \in N(i)} |A_s| - sq_{\mathbf{A}}. \qquad (3)$$

Первая сумма в (3) может быть переписана в виде

$$\sum_{i=1}^{k} \sum_{j \in N(i)} |A_j| \cdot \left(\sum_{s \in N(i)} |A_s| \right) = \sum_{i=1}^{k} |B_i| \cdot \left(\sum_{s \in N(i)} |A_s| \right) = \sum_{i=1}^{k} |B_i|^2 = sq_{\mathbf{B}}.$$

Последние две формулы завершают доказательство леммы. □

Поскольку для произвольного разбиения $\mathbf{A}$ справедливо $\underline{\mathbf{U}} \subset \mathbf{A}$, то, согласно лемме 1,

$$d(\underline{\mathbf{U}}, \mathbf{A}) = sq_{\mathbf{A}} - n.$$

Отсюда вытекает любопытное соотношение: для разбиений $\mathbf{A}$ и $\mathbf{B}$, таких, что $\mathbf{A} \subseteq \mathbf{B}$,

$$d(\underline{\mathbf{U}}, \mathbf{B}) = d(\underline{\mathbf{U}}, \mathbf{A}) + d(\mathbf{A}, \mathbf{B}).$$

Если интерпретировать значение $d(\mathbf{A}, \mathbf{B})$ как наименьшую длину маршрута между точками $\mathbf{A}$ и $\mathbf{B}$ в $\boldsymbol{\Xi}$, то это означает, что $\mathbf{A}$ лежит на кратчайшем маршруте от $\underline{\mathbf{U}}$ к содержащему его $\mathbf{B}$. Более того, в [2] случай, когда в метрическом пространстве с метрикой d для каких-то трех элементов выполнено

$$d(a, b) = d(a, c) + d(c, b)$$

обозначен термином “элемент c лежит между a и b”. Итак, в терминологии [2], кластерное разбиение $\mathbf{A}$, более мелкое, чем кластерное разбиение $\mathbf{B}$, обязательно лежит между $\mathbf{B}$ и наименьшим элементом $\underline{\mathbf{U}}$решетки кластерных разбиений.

Изучим подробнее введенную выше характеристику $sq_{\mathbf{A}}$ кластерного разбиения $\mathbf{A}$ основного множества. Это сумма квадратов натуральных чисел, которые, будучи сложенными сами по себе, дают число n элементов основного множества. Используем неравенство $(x + y)^2 > x^2 + y^2$, справедливое, если x, y – натуральные числа. Оно сразу же приводит к справедливости следующей леммы.

Лемма 2. *Если два любых кластера разбиения* $\mathbf{A}$ *объединить в один, то* $sq_{\mathbf{A}}$ *строго увеличится, а если какое-либо его кластеры разбить каждый на два или большее количество, то строго уменьшится.*

Назовем пересечением двух разбиений $\mathbf{A}$, $\mathbf{B}$ набор всех непустых попарных пересечений $A_i \cap B_j$, когда $A_i \in \mathbf{A}$, $B_j \in \mathbf{B}$. Ясно, что пересечение двух кластерных разбиений всегда является, в свою очередь, разбиением. Обозначим его $\mathbf{AB}$. Это кластерное разбиение представляет собой максимальный элемент решетки $\boldsymbol{\Xi}$, меньший обоих рассматриваемых разбиений в смысле введенного частичного порядка ($\min\{\mathbf{A}, \mathbf{B}\}$ в одном из вариантов обозначений [1]).

Имея в виду используемое там же определение максимума двух элементов решетки, можно ввести понятие объединения разбиений $\mathbf{A} \cup \mathbf{B}$. В этом разбиении два элемента

$x, y \in U$ относятся к одному кластеру тогда, когда они оба лежат либо в каком-то одном кластере в $\mathbf{A}$, либо оба относятся к одному и тому же кластеру в $\mathbf{B}$.

3. Кратчайшие маршруты по решетке

Теорема 1. *Для произвольных разбиений* $\mathbf{A} = \{A_1, ..., A_m\}$, $\mathbf{B} = \{B_1, ..., B_k\}$ *основного множества* U *справедливо* $d(\mathbf{A}, \mathbf{B}) = sq_{\mathbf{A}} + sq_{\mathbf{B}} - 2sq_{\mathbf{AB}}$, *или, иначе,*

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{A}, \mathbf{AB}) + d(\mathbf{AB}, \mathbf{B}).$$

Доказательство. Обозначим через f число множеств в разбиении $\mathbf{C} = \mathbf{AB}$. Введем дополнительно $C_0 = \emptyset$. Заметим, что для произвольных значений i, j найдется такое целое неотрицательное число $s(i, j) \leqslant f$, что $A_i \cap B_j = C_{s(i,j)}$. При этом $s(i, j)$ принимает каждое из значений $1, \ldots, f$ ровно один раз при изменении i, j. Значение 0 число $s(i, j)$ принимает каждый раз, когда соответствующее пересечение пусто. Имеют место соотношения

$$\bigcup_{j=1}^{k} C_{s(i,j)} = A_i, \quad \bigcup_{i=1}^{m} C_{s(i,j)} = B_j, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k.$$

$$|A_i \Delta B_j| = |A_i| + |B_j| - 2|C_{s(i,j)}|. \tag{4}$$

Учитывая (2), (4), запишем

$$d(A, B) = \sum_{i,j} |C_{s(i,j)}| \cdot \big(|A_i| + |B_j| - 2|C_{s(i,j)}|\big) =$$

$$= \sum_{i=1}^{m} |A_i| \sum_{j=1}^{k} |C_{s(i,j)}| + \sum_{j=1}^{k} |B_j| \sum_{i=1}^{m} |C_{s(i,j)}| - 2 \sum_{i,j} |C_{s(i,j)}|^2 =$$

$$= \sum_{i=1}^{m} |A_i|^2 + \sum_{j=1}^{k} |B_j|^2 - 2 \sum_{s=1}^{f} |C_s|^2 = sq_{\mathbf{A}} + sq_{\mathbf{B}} - 2sq_{\mathbf{AB}},$$

и доказательство завершено. □

Следовательно, вновь используя терминологию [2], мы доказали, что пересечение разбиений всегда расположено между точками $\boldsymbol{\Xi}$, соответствующими этим разбиениям. Если принять менее строгую терминологию и назвать совокупность всех тех разбиений $\mathbf{C}$, для которых

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{A}, \mathbf{C}) + d(\mathbf{C}, \mathbf{B}) \tag{5}$$

прямолинейным отрезком $[\mathbf{A}; \mathbf{B}] \subset \boldsymbol{\Xi}$, то возникает желание попробовать в некотором смысле использовать для получения дальнейших результатов все богатство классической евклидовой геометрии. К сожалению, аналогия с отрезками в евклидовом пространстве оказалась для определенных сейчас прямолинейных отрезков в $\boldsymbol{\Xi}$ далеко не полной, да и сам термин "прямолинейность" здесь, вообще говоря, корректно не применим. Некоторые причины этого будут приведены ниже в обсуждении результатов.

Пересечение двух разбиений представляет собой элемент, наиболее близкий к ним из сравнимых и с $\mathbf{A}$, и с $\mathbf{B}$, если рассматривать только более мелкие разбиения (лежащие по включению по направлению к $\underline{\mathbf{U}}$). Таким образом, кратчайший "нижний" маршрут по решетке из $\mathbf{A}$ в $\mathbf{B}$ должен проходить через $\mathbf{AB}$ и по длине представлять собой сумму длин маршрутов между $\mathbf{A}$ и $\mathbf{AB}$ и между $\mathbf{AB}$ и $\mathbf{B}$. Согласно теореме 1, $d(\mathbf{A}, \mathbf{B})$ как раз и представляет собой такую сумму.

Следовательно, d можно интерпретировать, как длину кратчайшего "нижнего" маршрута между разбиениями по решетке. Конечно же, если мы будем двигаться только по самой решетке (каждый раз переходя от какого-то разбиения к большему или меньшему его в рассматриваемом частичном порядке), то, кроме "нижнего" кратчайшего маршрута существует еще и кратчайший "верхний", проходящий через $\mathbf{A} \cup \mathbf{B}$ – ближайший элемент решетки, сравнимый с обоими рассматриваемыми разбиениями из больших каждого из них. Но этот маршрут оказывается, вообще говоря, длиннее, что будет продемонстрировано теоремой 2.

Но прежде дадим еще одно определение. Два кластерных разбиения $\mathbf{A}$ и $\mathbf{B}$ основного множества назовем совместимыми, если каждый кластер из $\mathbf{A}$ либо целиком содержится в некотором кластере $\mathbf{B}$, либо сам является объединением каких-либо из кластеров $\mathbf{B}$.

Отметим, что определение симметрично – то же, что было потребовано в нем для кластеров $\mathbf{A}$, автоматически окажется выполненным и для кластеров совместимого с ним кластерного разбиения $\mathbf{B}$. Понятно, что если $\mathbf{A} \subseteq \mathbf{B}$ или $\mathbf{B} \subseteq \mathbf{A}$, то разбиения $\mathbf{A}$, $\mathbf{B}$ совместимы. Но возможны и иные ситуации. Пусть, например, на множестве $U = \{1, 2, 3, 4, 5, 6\}$ заданы разбиения

$$\mathbf{A} = \{\{1, 2, 3\}; \{4, 5\}; \{6\}\}; \quad \mathbf{B} = \{\{1, 2\}; \{3\}; \{4, 5, 6\}\}.$$

Тогда эти два разбиения совместимы, хотя ни одно из них не является более мелким, чем другое.

Теорема 2. *Для произвольных разбиений* $\mathbf{A} = \{A_1, ..., A_m\}$, $\mathbf{B} = \{B_1, ..., B_k\}$ *основного множества* U *справедливо*

$$d(\mathbf{A}, \mathbf{B}) \le d(\mathbf{A}, \mathbf{A} \cup \mathbf{B}) + d(\mathbf{A} \cup \mathbf{B}, \mathbf{B}),$$

причем равенство в этом неравенстве достигается тогда и только тогда, когда разбиения $\mathbf{A}$ *и* $\mathbf{B}$ *совместимы.*

Доказательство. С учетом леммы 1 и теоремы 1 преобразуем доказываемое неравенство к виду

$$sq_{\mathbf{A}\cup\mathbf{B}} - sq_{\mathbf{A}} + sq_{\mathbf{A}\cup\mathbf{B}} - sq_{\mathbf{B}} \ge sq_{\mathbf{A}} + sq_{\mathbf{B}} - 2sq_{\mathbf{AB}},$$

или

$$sq_{\mathbf{A}\cup\mathbf{B}} + sq_{\mathbf{AB}} \ge sq_{\mathbf{A}} + sq_{\mathbf{B}}. \tag{6}$$

Если $\mathbf{AB} = \{C_1, ..., C_f\}$, то кластеры остальных трех разбиений из (6) составлены из C_j, как из "кирпичиков". Пусть c_s обозначает число элементов в C_s, $s = 1, ..., f$. Ясно, что любое из чисел c_j^2, $j = 1, ..., f$ встречается и в левой, и в правой части (6) ровно по два раза (соответствующий "кирпичик" входит по разу в каждое из четырех кластерных разбиений). После сокращения этих квадратичных слагаемых неравенство (6) превратится в соотношение между попарными произведениями вида $2c_ic_j$, $i \ne j$.

Если неверно, что $\mathbf{A} \subset \mathbf{B}$ или $\mathbf{B} \subset \mathbf{A}$, то найдутся такие $A \in \mathbf{A}$, $B \in \mathbf{B}$, что $C_j = A \cap B \ne \emptyset$, и имеются два кластера C_s, C_t, $s \ne t$ из $\mathbf{AB}$, для которых

$$A \supseteq C_j \cup C_s, \quad B \supseteq C_j \cup C_t. \tag{7}$$

При этом из того, что пересечение A, B не пусто, следует, что в $\mathbf{A}\cup\mathbf{B}$ найдется кластер D, содержащий их объединение, и, следовательно, такой, что

$$|D| \ge c_j + c_s + c_t.$$

Таким образом, в левую часть (6) входят слагаемые

$$S_{j,s} + S_{j,t} + S_{s,t} = 2c_jc_s + 2c_jc_t + 2c_sc_t. \tag{8}$$

При этом $S_{j,s}$, $S_{j,t}$ содержатся в $sq_{\mathbf{A}}$, $sq_{\mathbf{B}}$ соответственно (см. (7)). Но последнего слагаемого из (8) в правой части (6) нет, т.к. C_s, C_t являются подмножествами ("кирпичиками") разных кластеров в $\mathbf{A}$ и в $\mathbf{B}$. Это значит, что неравенство (6), доказано, а вместе с ним и эквивалентное ему неравенство теоремы.

Далее, если, $\mathbf{A}$ и $\mathbf{B}$ совместимы, то для любой пары пересекающихся кластеров $A \in \mathbf{A}$, $B \in \mathbf{B}$ в соотношении типа (7) один из C_s, C_t обязательно заменится на $\emptyset$, и неравенство теоремы превратится в итоге в равенство.

Наконец, допустим, что левая часть (6) равна правой. Тогда, вновь прослеживая приведенное выше рассуждение, видим, что, если хотя бы для одной пары пересекающихся кластеров различных разбиений в (7) получится, что и $C_s \neq \emptyset$, и $C_t \neq \emptyset$, то это автоматически приведет к строгому неравенству в (6). Получающееся противоречие предположению о равенстве в (6) показывает, что хотя бы один в каждой из пар имеющих непустое пересечение кластеров обязан целиком содержаться в другом. Теорема доказана. □

Таким образом, из теоремы 2, например, вытекает, что для двух разбиений, пример которых приведен перед ее формулировкой, длины маршрутов от $\mathbf{A}$ к $\mathbf{B}$ через их пересечение и через их объединение должны быть одинаковыми. Действительно,

$$\mathbf{AB} = \{\{1,2\};\{3\};\{4,5\};\{6\}\}; \quad \mathbf{A} \cup \mathbf{B} = \{\{1,2,3\};\{4,5,6\}\}.$$

Вычисления показывают $sq_{\mathbf{A}} = sq_{\mathbf{B}} = 14$, $sq_{\mathbf{AB}} = 10$, $sq_{\mathbf{A}\cup\mathbf{B}} = 18$. Привлекая утверждения теоремы 1 и леммы 1, получаем

$$d(\mathbf{A},\mathbf{B}) = 14 + 14 - 20 = 8; \quad d(\mathbf{A},\mathbf{A}\cup\mathbf{B}) = 18 - 14 = 4 = d(\mathbf{A}\cup\mathbf{B},\mathbf{B}),$$

$$d(\mathbf{A},\mathbf{AB}) = d(\mathbf{B},\mathbf{AB}) = 14 - 10 - 4.$$

Отсюда уже видно, что

$$d(\mathbf{A},\mathbf{B}) = d(\mathbf{A},\mathbf{AB}) + d(\mathbf{AB},\mathbf{B}) = d(\mathbf{A},\mathbf{A}\cup\mathbf{B}) = d(\mathbf{A}\cup\mathbf{B},\mathbf{B}),$$

т.е. оба маршрута действительно одинаковы по длине.

Лемма 3. *Если для кластерных разбиений* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *справедливо* $\mathbf{A} \subseteq \mathbf{C} \subseteq \mathbf{B}$, *то*

$$d(\mathbf{A},\mathbf{B}) = d(\mathbf{A},\mathbf{C}) + d(\mathbf{C},\mathbf{B}).$$

Эта лемма является простым следствием леммы 1. Она показывает, что можно считать кратчайшие маршруты между разбиениями $\mathbf{A}$ и $\mathbf{B}$ в случае, когда $\mathbf{A} \subseteq \mathbf{B}$, проходящими по ребрам решетки. Это значит, что маршрут в качестве промежуточных своих пунктов содержит все разбиения, лежащие в рассматриваемом частичном порядке между его началом и концом. В частности, кратчайший "нижний" маршрут проходит через все разбиения $\mathbf{C}$, $\mathbf{D}$, для которых $\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{A}$, $\mathbf{AB} \subseteq \mathbf{D} \subseteq \mathbf{B}$.

Из теорем 1, 2 вытекает, что кратчайший в смысле метрики d маршрут между двумя разбиениями из двух ближайших сравнимых с ними обоими разбиений содержит их пересечение, а не объединение, т.е. расположен на решетке ближе к ее наименьшему элементу $\underline{\mathbf{U}}$. Учитывая то, что переход от некоторого кластерного разбиения к его пересечению с любым другим уменьшает его кластеры, в некотором смысле приближая итоговое разбиение к $\underline{\mathbf{U}}$, можно сказать, что заключительный результат настоящей работы интуитивно очевиден – "более низко" расположенные маршруты короче.

Теорема 3. *Пусть* $\mathbf{A}$, $\mathbf{C}$ *– кластерные разбиения,* $\mathbf{C} \subseteq \mathbf{A}$. *Тогда для произвольного разбиения* $\mathbf{B}$ *основного множества справедливо* $d(\mathbf{A},\mathbf{C}) \geq d(\mathbf{AB},\mathbf{CB})$, *причем равенство в этом неравенстве достигается только тогда, когда* $\mathbf{B} \supseteq \mathbf{A}$, *т.е.* $\mathbf{AB} = \mathbf{A}$, $\mathbf{CB} = \mathbf{C}$.

Доказательство. Пусть $\mathbf{A} = \{A_i,\ i = 1, ..., m\}$, $\mathbf{C} = \{C_s,\ s = 1, ..., f\}$. Введем в рассмотрение $NA(i) = \{s : C_s \subseteq A_i\},\ i = 1, ..., m$. Ясно, что каждое из $s : 1 \le s \le f$ попадает ровно в одно из множеств $NA(i)$. Через c_s, как и ранее, обозначим число элементов в C_s. Тогда

$$sq_{\mathbf{A}} = \sum_{i=1}^{m} \left(\sum_{s \in NA(i)} c_s \right)^2 = sq_{\mathbf{C}} + 2 \sum_{i=1}^{m} \sum_{*i} c_s c_t,$$

где двойная сумма $\sum\limits_{*i}$ берется по всем тем $s, t \in NA(i)$, что $s > t$. Следовательно,

$$d(\mathbf{A}, \mathbf{C}) = sq_{\mathbf{A}} - sq_{\mathbf{C}} = 2 \sum_{i=1}^{m} \sum_{*i} c_s c_t. \tag{9}$$

Пусть теперь $\mathbf{B} = \{B_t,\ t = 1, ..., k\}$, $h_{s,t} = |C_s \cap B_t|$, $s = 1, .., f$, $t = 1, ..., k$ (некоторые из этих чисел равны 0). Отметим, что

$$(\forall i, j) |A_i \cap B_j| = \sum_{s \in N(i)} h_{s,j} \quad \Rightarrow \quad sq_{\mathbf{AB}} = \sum_{i=1}^{m} \sum_{j=1}^{k} \left(\sum_{s \in N(i)} h_{s,j} \right)^2.$$

С другой стороны,

$$sq_{\mathbf{BC}} = \sum_{j=1}^{k} \sum_{s=1}^{f} h_{s,j}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{s \in N(i)} h_{s,j}^2.$$

Тогда, полностью аналогично (9), выводим

$$d(\mathbf{AB}, \mathbf{CB}) = 2 \sum_{i=1}^{m} \sum_{u=1}^{k} \sum_{*i} h_{s,u} h_{t,u}.$$

Фиксируем i. При $s \in NA(i)$ справедливо $\sum\limits_{u=1}^{k} h_{s,u} = c_s$, и, следовательно,

$$h_{s,u} \le c_s,\ s \in NA(i),\ u = 1, ..., k. \tag{10}$$

Из (10) и (9) получаем

$$d(\mathbf{AB}, \mathbf{CB}) \le 2 \sum_{i=1}^{m} \sum_{*i} \left(c_s \cdot \sum_{u=1}^{k} h_{t,u} \right) = d(\mathbf{A}, \mathbf{C}).$$

Ясно, что для замены неравенства теоремы равенством необходимо и достаточно, чтобы равенство одновременно достигалось бы во всех неравенствах (10). Но это возможно тогда и только тогда, когда при каждом сочетании s, i найдется такое u, что $C_s \subseteq B_u$. Заметим, что, в силу включения $\mathbf{AB} \supseteq \mathbf{BC}$ при всех $s \in NA(i)$ это u обязано быть одним и тем же. Но тогда $(\forall i)(\exists u)\, A_i \subseteq B_u$, что, по определению, означает $\mathbf{A} \subseteq \mathbf{B}$. Теорема доказана. □

4. Обсуждение и выводы

После анализа полученных в работе результатов становится ясно, что кластерная метрика d, ранее введенная автором в [3], вполне согласована со структурой решетки по

включению, естественным образом заданной на семействе $\boldsymbol{\Xi}$ всевозможных разбиений основного множества U на непустые дизъюнктные части. Согласованность эта может быть описана следующим образом. Естественно считать величину $d(\mathbf{A},\mathbf{B})$ длиной кратчайшего маршрута от разбиения $\mathbf{A}$ до разбиения $\mathbf{B}$ в рассматриваемом метрическом пространстве. Тогда оказывается, что, даже если рассматриваемые разбиения не сравнимы между собой в частичном порядке по включению, всегда можно выбрать маршрут от $\mathbf{A}$ к $\mathbf{B}$ так, чтобы соседние промежуточные точки этого маршрута были сравнимы между собой, а длина его в точности равнялась $d(\mathbf{A},\mathbf{B})$. Возможность сравнения между собой промежуточных точек маршрута можно интерпретировать как пролегание его по ребрам решетки.

Одним из таких маршрутов всегда является тот, который в работе назван "нижним" – он проходит через пересечение разбиений (теорема 1). При некоторых дополнительных условиях на соотношение разбиений $\mathbf{A}$, $\mathbf{B}$ кратчайших маршрутов может оказаться несколько. Например, если $\mathbf{A}$, $\mathbf{B}$ – совместимые разбиения (их пересечение состоит из кластеров, лежавших либо в $\mathbf{A}$, либо в $\mathbf{B}$, никаких новых кластеров при переходе к пересечению не возникает), то таким же кратчайшим маршрутом является "верхний", проходящий через $\mathbf{A}\cup\mathbf{B}$, что следует из теоремы 2. Этот маршрут не совпадает с "нижним" в случае, когда ни одно из совместимых разбиений не является более мелким, чем другое (они не сравнимы в имеющемся частичном порядке).

Если рассматривать наименьший элемент решетки $\underline{\mathbf{U}}$ как начало отсчета (ноль в некоторой условной системе координат), то, согласно лемме 1, расстояние между любым кластерным разбиением $\mathbf{A}$ и этим началом отсчета (измеряемое по "прямолинейному" по отношению к включению отрезку) совпадает с $\langle|\mathbf{A}|\rangle = sq_{\mathbf{A}} - n$.

Аналогия с векторными пространствами наводит на мысль о том, что эта характеристика может служить некоторым вариантом нормы кластерного разбиения. Правда, для получения полной аналогии нужно было бы ввести на семействе кластерных разбиений $\boldsymbol{\Xi}$ операции сложения и умножения на число. И, если в роли суммы двух разбиений, вероятно, может выступить их объединение, то как умножить разбиение даже на натуральное число, совершенно непонятно.

Подойдем к рассматриваемой аналогии с другой стороны. Метрика и норма в векторном пространстве обычно связаны соотношением $d(x,y) = ||x-y||$, что позволяет надеяться на определение разности разбиений $\mathbf{A},\mathbf{B}$ как такого разбиения $\mathbf{D}$, для которого справедливо

$$d(\mathbf{A},\mathbf{B}) = \langle|\mathbf{D}|\rangle,$$

а далее через это понятие подойти уже к "правильному" определению суммы разбиений и умножения разбиения на скаляр. Но эти идеи еще только предстоит проработать.

К сожалению, полная аналогия с евклидовой геометрией здесь все равно невозможна. Вероятно, главным источником этой невозможности служит отсутствие линейного порядка на $\boldsymbol{\Xi}$ и, как следствие, невозможность корректного определения прямых линий так, чтобы они обладали привычными нам свойствами. Например, рассмотренные перед теоремой 2 разбиения множества из шести элементов

$$\mathbf{A} = \{\{1,2,3\};\{4,5\};\{6\}\};\quad \mathbf{B} = \{\{1,2\};\{3\};\{4,5,6\}\}$$

оба лежат на отрезке $[\mathbf{AB};\ \mathbf{A}\cup\mathbf{B}]$, но, тем не менее, несравнимы между собой. Это указывает на принципиально иную структуру отрезков, определяемых соотношением (5), в $\boldsymbol{\Xi}$, чем в евклидовых (или векторных) пространствах. Подробно структура отрезка в семействе кластерных разбиений будет изучена в одной из следующих работ автора.

Список литературы

1. Биргхоф Г. Теория решеток. — М. : Главная редакция физ-мат литературы, 1984.

2. Gratzer G. Lattice Theory: Foundations. — Springer Science & Business Media, 2011. — DOI 10.1007/978-3-0348-0018-1.

3. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. — 2011. — № 1/2 (69). — С. 32–35.

4. Бураго Д.Ю., Бураго Ю.Д., Иванов С.В. Курс метрической геометрии. — Москва–Ижевск : Институт компьютерных исследований, 2004.