УДК 514.172

Геометрия отрезка в семействе кластерных разбиений конечного множества

С.В. Дронов
Алтайский государственный университет (Барнаул, Россия)

The Geometry of a Segment in a Family of Cluster Partitions of a Finite Set

S.V. Dronov
Altai State University (Barnaul, Russia)

Рассматривается метрическое пространство семейства всех разбиений конечного множества на непустые дизъюнктные подмножества в кластерном расстоянии, предложенном автором в одной из предыдущих работ. Исследуется связь между структурой этого пространства и частичным порядком по включению на семействе разбиений. Оказывается, что при определении отрезка в этом пространстве в границах **A** и **B** как множества тех **C**, что сумма расстояний от него до **A** и до **B** равна расстоянию от **A** до **B**, он оказывается согласованным с частичным порядком. Это выражается в том, что расстояние между разбиениями соответствует наименьшей длине пути между ними по цепочкам в решетке соответствующего частичного порядка. Тем не менее определенный описанным образом отрезок обладает значительными отличиями от обычных отрезков в векторных пространствах, поэтому полной аналогии с теоремами обычной геометрии, к сожалению, не получается. Полученные результаты могут быть использованы при построении новых алгоритмов кластерного анализа, а также для нахождения точных вероятностных распределений расстояния между в некотором смысле правильным разбиением.

Ключевые слова: кластерная метрика, разбиения конечного множества, частичный порядок, отрезок в метрическом пространстве.

The paper considers the metric space of a family of all partitions of a finite set into non-empty disjoint subsets in the cluster distance proposed by the author in one of the previous papers. The relation between this space structure and the partial order generated by the inclusion on a family of partitions is investigated. It is found out that the segment it is coordinated with the partial order when the segment is determined in such space with the boundaries of **A** and **B** as the set of those **C** that the sum of the distances from it to **A** and **B** is equal to the distance from **A** to **B**. This is expressed by the fact that the distance between partitions corresponds to the smallest path length between them along the chains in the lattice of the corresponding partial order. Nevertheless, the segment defined this way has significant differences from ordinary segments in vector spaces. Therefore, it is not possible to completely carry out the analogy with theorems of usual geometry. The obtained results can be used in the construction of new algorithms for cluster analysis, as well as for finding the exact probability distributions of the distance between, in some sense, a correct partition and a partition constructed from the data of real observations.

Key words: cluster metric, partitions of a finite set, partial order, a segment in a metric space.

DOI 10.14258/izvasu(2018)1-13

1. Вводные замечания. Постановка задачи. При обработке результатов практически любых экспериментов очень часто встает проблема классификации объектов наблюдения по близости некоторых их признаков. Методики, согласно которым может производиться такая классификация, традиционно объединяются в группу алгоритмов, называемых кластерными, а та область анализа данных, которая занимается их созданием и изучением, называется кластерным анализом (например, [1,2]). При этом, в силу большого разнообразия методов, обычной является ситуация, когда в результате применения разных алгоритмов к одному и тому же множеству объектов получаются разные способы разбиения их на кластеры, что представляет на сегодня одну из важных проблем кластерного анализа (см. [3]). Поэтому актуальной является задача сравнения двух различных кластерных разбиений одного множества. Для такого сравнения на семействе всех возмож-

ных кластерных разбиений можно ввести метрику — расстояние между разбиениями. Это можно сделать разными способами. Один из них был реализован автором в [4]. Приведем здесь соответствующее определение.

Кластерное разбиение — это набор непустых дизъюнктных подмножеств основного множества U, объединение которых дает U. Кластерное разбиение далее условимся чаще называть просто разбиением, поскольку мы собираемся рассматривать все возможные такие наборы подмножеств U, не уточняя вид близости формирующих эти множества объектов. Разбиения будем обозначать полужирным шрифтом, множества, составляющие разбиение, — теми же прописными буквами.

Любой элемент x основного множества U входит в точно один из элементов каждого из кластерных разбиений **A, B**. Обозначим эти множества A_x, B_x соответственно. Если $|A|$ означает число элементов конечного множества A, то, по определению,

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

где $A\Delta B$ – симметрическая разность множеств.

В [4] также предложена другая формула для вычисления этой метрики, более простая в применении. Пусть $\mathbf{A} = \{A_1, ..., A_k\}, \mathbf{B} = \{B_1, ..., B_m\}$. Тогда

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

На семействе всех разбиений U, кроме образованной введением метрики структуры метрического пространства, имеется естественное отношение частичного порядка по включению. Будем писать $\mathbf{A} \subseteq \mathbf{B}$, если для произвольного $A \in \mathbf{A}$ найдется $B \in \mathbf{B}$, такое, что $A \subseteq B$ в обычном смысле. В этом случае каждое из множеств, составляющих $\mathbf{B}$, возможно составить из элементов $\mathbf{A}$ как из «кирпичиков»: для каждого значения i может быть указано такое множество натуральных чисел $A(i)$, что

$$\bigcup_{i=1}^{m} A(i) = \{1, ..., n\};\ (\forall i)\ B_i = \bigcup_{j\in A(i)} A_j, \qquad (2)$$

где $n = |U|$ — число элементов основного множества. Если $\mathbf{A} \subseteq \mathbf{B}$, но $\mathbf{A} \neq \mathbf{B}$, то будем писать $\mathbf{A} \subset \mathbf{B}$.

По отношению ко включению $\subseteq$, используемому как частичный порядок, семейство всех разбиений U образует решетку (необходимые определения и свойства решеток приводятся в классической монографии [5], современная точка зрения — в [6]). При этом элементом решетки, который меньше каждого из **A, B**, будет их пересечение, т.е. разбиение, образованное всеми возможными непустыми пересечениями множеств $A \cap B$, $A \in \mathbf{A}, B \in \mathbf{B}$. Это пересечение далее будем обозначать **AB**. Аналогичным образом может быть определено и объединение разбиений, которое больше каждого из них. В решетке разбиений множества U есть наибольший элемент — разбиение, состоящее из одного множества U, и наименьший — разбиение из n одноэлементных множеств.

Основной задачей работы является выяснение характера взаимодействия описанных структур метрического пространства и решетки на семействе разбиений. При этом главное внимание будет уделено понятию отрезка в семействе разбиений. Будем говорить, что разбиение $\mathbf{C}$ лежит на отрезке $[\mathbf{A};\ \mathbf{B}]$, если

$$d(\mathbf{A},\mathbf{B}) \;=\; d(\mathbf{A},\mathbf{C}) + d(\mathbf{C},\mathbf{B}). \qquad (3)$$

В руководствах по метрической геометрии [7,8] в аналогичной ситуации для $\mathbf{C}$, удовлетворяющего (3), используется термин «лежит между **A** и **B**». Далее мы увидим, что наше определение отрезка, действительно, не отвечает общепринятым интуитивным представлениям о свойствах прямолинейного отрезка в евклидовом пространстве. Это связано в первую очередь с тем, что в метрическом пространстве разбиений отношение порядка не является линейным.

2. Связь кластерной метрики с величинами отдельных кластеров. Оказывается, расстояние между двумя разбиениями существенно связано не с составом образующих эти разбиения множеств, а лишь с их размерами. Более точно, имеет место следующее утверждение. Для разбиения $\mathbf{A} = \{A_1, ..., A_k\}$ введем обозначение

$$sq_{\mathbf{A}} \;=\; \sum_{i=1}^{k} |A_i|^2.$$

Лемма 1. *Для произвольных двух разбиений*

$$d(\mathbf{A},\mathbf{B}) \;=\; sq_{\mathbf{A}} + sq_{\mathbf{B}} - 2sq_{\mathbf{AB}}.$$

Доказательство. Пусть

$$\mathbf{B} = \{B_1, ..., B_m\},\ \mathbf{C} = \mathbf{AB} = \{C_1, ..., C_f\}.$$

Для тех i, j, для которых соответствующее пересечение не пусто, найдется индекс $s(i,j)$ такой, что $A_i \cap B_j = C_{s(i,j)}$. Если формально ввести $C_0 = \emptyset$ и для остальных i, j положить $s(i,j) = 0$, то

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

и $s(i,j)$ принимает каждое ненулевое значение из возможных номеров множеств в пересечении кластерных разбиений ровно один раз. К тому же

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

даже для непересекающихся A_i, B_j. Учитывая (1), (4), запишем

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

и лемма доказана.

Поскольку непосредственно из определения пересечения разбиений следует, что при $\mathbf{A} \subseteq \mathbf{B}$ справедливо $\mathbf{AB} = \mathbf{A}$, и для произвольных двух разбиений $\mathbf{AB} \subseteq \mathbf{A}$, $\mathbf{AB} \subseteq \mathbf{B}$, то из леммы 1 немедленно вытекает

Лемма 2.

- *Если* $\mathbf{A} \subseteq \mathbf{B}$, *то* $d(\mathbf{A},\mathbf{B}) = sq_{\mathbf{B}} - sq_{\mathbf{A}}$;
- $(\forall \mathbf{A},\mathbf{B})\ d(\mathbf{A},\mathbf{AB}) = sq_{\mathbf{A}} - sq_{\mathbf{AB}}$.

Сопоставление второго утверждения леммы 2 и леммы 1 дает очень важное для дальнейшего следствие.

Лемма 3. *Для двух произвольных разбиений* **A, B** основного множества

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

Теперь понятно, что величина $sq_{\mathbf{A}}$ несет в себе важную информацию о разбиении **A**. Фактически эта величина представляет собой сумму квадратов натуральных чисел, которые в сумме равны n — числу элементов основного множества. Заметим, что квадрат суммы натуральных чисел всегда строго больше суммы их квадратов, следовательно, при разбиении одного или нескольких элементов **A** на части $sq_{\mathbf{A}}$ строго уменьшается, а при объединении нескольких в одно множество — увеличивается. В частности, справедлива

Лемма 4.

- *Пусть* $\mathbf{C} \subseteq \mathbf{A}$. *Тогда* $sq_{\mathbf{C}} \le sq_{\mathbf{A}}$, *причем равенство достигается только в случае* $\mathbf{C}{=}\mathbf{A}$.
- *Если* $sq_{\mathbf{AB}} \ge sq_{\mathbf{A}}$, *то* $\mathbf{B} \supseteq \mathbf{A}$, *и* $sq_{\mathbf{AB}} = sq_{\mathbf{A}}$.

Лемма 5. *Для разбиений* $\mathbf{C} \subseteq \mathbf{A}$ *и произвольного разбиения* **B** *справедливо*

$$sq_{\mathbf{A}} - sq_{\mathbf{C}} \ge sq_{\mathbf{AB}} - sq_{\mathbf{BC}},$$

причем равенство достигается тогда и только тогда, когда $\mathbf{B} \supseteq \mathbf{A}$.

Доказательство. Пусть $\mathbf{A} = \{A_i,\ i = 1,...,k\}$, $\mathbf{C} = \{C_s,\ s = 1,...,f\}$. Введем $NA(i) = \{s:\ C_s \subseteq A_i\},\ i = 1,...,k$. Тогда каждое из $s : 1 \le s \le f$ попадает ровно в одно из введенных множеств. Пусть также c_s обозначает число элементов в C_s. Отсюда

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

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

$$sq_{\mathbf{A}} - sq_{\mathbf{C}} \ = \ 2\sum_{i=1}^{k}\sum_{*i} c_s c_t. \qquad (5)$$

Допустим, что $\mathbf{BC} = \{G_1,...,G_v\}$, $g_p = |G_p|$. Тогда все c_s есть суммы каких-то g_p. Введем $NAB(r)$ — множество номеров тех G_p, которые содержатся в r-м множестве **AB**. Тогда, полностью аналогично (5),

$$sq_{\mathbf{AB}} - sq_{\mathbf{BC}} = 2\sum_{i=1}^{k}\sum_{i|r}\sum_{**r} g_p g_q, \qquad (6)$$

где $\sum_{i|r}$ берется по тем r, что r-й элемент **AB** содержится в A_i, а $\sum_{**r}$ — двойная сумма по таким $p,q \in NAB(r)$, что $p < q$.

Зафиксируем индекс i. Заметим, что, если G_p, G_q оба лежат в r-м элементе **AB**, и r — один из индексов, входящих в $\sum_{i|r}$, то G_p, G_q являются подмножествами некоторых C_s и C_t, и $t,s \in NA(i)$. Следовательно, произведение таких g_p, g_q обязательно встретится после раскрытия скобок в (5). Поскольку это рассуждение можно провести для каждого $g_p g_q$ и для каждого i, то правая часть (5) не меньше правой части (6), и неравенство леммы доказано.

Наконец, если $\mathbf{B} \not\supseteq \mathbf{A}$, то при переходе к пересечению **AB** хотя бы один элемент **A** разобьется на части. Это означает, что найдутся G_p, G_q, содержащиеся в одном элементе **A**, но в разных элементах **AB**. Тем самым, $g_p g_q$ в (5) будет присутствовать, а в (6) нет, и неравенство леммы станет строгим.

3. Строение отрезка в семействе разбиений. Определение отрезка $[\mathbf{A};\mathbf{B}]$ в семействе разбиений множества U было дано выше (см. (3)). Докажем сейчас некоторые его свойства. Из леммы 3, например, сразу следует, что $\mathbf{AB} \in [\mathbf{A};\mathbf{B}]$. Получим условие попадания **C** в $[\mathbf{A};\mathbf{B}]$ в терминах сумм квадратов sq.

Лемма 6 (основная). *Пусть* **A, B, C** — *произвольные разбиения.* $\mathbf{C} \in [\mathbf{A};\mathbf{B}]$ *в том и только том случае, если*

$$sq_{\mathbf{C}} + sq_{\mathbf{AB}} \ = \ sq_{\mathbf{AC}} + sq_{\mathbf{BC}}.$$

Доказательство. Определение (3) и лемма 1 дают необходимое и достаточное условие принадлежности **C** нужному отрезку:

$$sq_{\mathbf{A}} + sq_{\mathbf{B}} - 2sq_{\mathbf{AB}} = \\ = sq_{\mathbf{A}} + sq_{\mathbf{C}} - 2sq_{\mathbf{AC}} + sq_{\mathbf{C}} + sq_{\mathbf{B}} - 2sq_{BC},$$

что после приведения подобных и сокращения на 2 совпадает с требуемым равенством.

Лемма 7. *Пусть* $\mathbf{A} \subseteq \mathbf{C} \subseteq \mathbf{B}$. *Тогда* **C** *лежит на отрезке* $[\mathbf{A};\mathbf{B}]$.

Это утверждение немедленно следует из основной леммы, если заметить, что в рассматриваемом случае $\mathbf{AB} = \mathbf{A},\ \mathbf{AC} = \mathbf{A},\ \mathbf{BC} = \mathbf{C}$.

Лемма 8. *Пусть* $\mathbf{C} \subset \mathbf{AB}$. *Тогда* $\mathbf{C}$ *не лежит на отрезке* $[\mathbf{A}; \mathbf{B}]$.

Доказательство. В этом случае из леммы 4 получаем $sq_{\mathbf{AB}} > sq_{\mathbf{C}}$. К тому же $\mathbf{C} \subset \mathbf{A},\ \mathbf{C} \subset \mathbf{B}$, откуда $sq_{\mathbf{AC}} = sq_{\mathbf{BC}} = sq_{\mathbf{C}}$. Но тогда

$$sq_{\mathbf{AC}} + sq_{\mathbf{BC}} = 2sq_{\mathbf{C}} < sq_{\mathbf{C}} + sq_{\mathbf{AB}}.$$

Осталось применить основную лемму 6.

Лемма 9. *Если* $\mathbf{C} \supset \mathbf{A}$ *и* $\mathbf{C} \not\subseteq \mathbf{B}$ *или наоборот, то* $\mathbf{C}$ *не лежит на отрезке* $[\mathbf{A}; \mathbf{B}]$.

Доказательство. В рассматриваемой ситуации $sq_{\mathbf{AC}} = sq_{\mathbf{A}}$, и условие принадлежности $\mathbf{C}$ из основной леммы может быть переписано в виде

$$sq_{\mathbf{C}} - sq_{\mathbf{A}} = sq_{\mathbf{BC}} - sq_{\mathbf{AB}}.$$

Но в соответствии с утверждением леммы 5 это равенство выполнено тогда и только тогда, когда $\mathbf{B} \supseteq \mathbf{C}$. Лемма доказана.

Лемма 8 обеспечивает отсутствие элементов $[\mathbf{A}; \mathbf{B}]$ «внутри» пересечения $\mathbf{AB}$. Но оказывается, что их нет и среди не сравнимых с этим пересечением разбиений.

Лемма 10. *Если* $\mathbf{C} \not\supseteq \mathbf{AB}$, *то* $\mathbf{C}$ *не лежит на* $[\mathbf{A}; \mathbf{B}]$.

Доказательство. Если бы $\mathbf{C}$ лежало на отрезке, то из леммы 6 и леммы 5 мы бы получили

$$sq_{\mathbf{AC}} - sq_{\mathbf{AB}} = sq_{\mathbf{C}} - sq_{\mathbf{BC}} \geq sq_{\mathbf{AC}} - sq_{\mathbf{ABC}},$$

откуда $sq_{\mathbf{ABC}} \geq sq_{\mathbf{AB}}$, что, как следует из второго утверждения леммы 4, возможно лишь при условии $\mathbf{C} \supseteq \mathbf{AB}$.

Лемма 11. *Пусть* $\mathbf{C} \supset \mathbf{AB}$, *и* $\mathbf{C}$ *не сравнимо ни с* $\mathbf{A}$, *ни с* $\mathbf{B}$. *Тогда* $\mathbf{C}$ *не может лежать на отрезке* $[\mathbf{A}; \mathbf{B}]$.

Доказательство. Предположим, что $\mathbf{AB}$ состоит из множеств $D_j, j = 1, ..., q, x_j = |D_j|$. В рассматриваемом случае $\mathbf{ABC}$=$\mathbf{AB}$, а значит, все элементы разбиений $\mathbf{C}$, $\mathbf{AC}$ и $\mathbf{BC}$ можно составить из D_j, как из «кирпичиков», т. е. количества элементов любого из множеств указанных разбиений будут равны суммам каких-то из x_j, причем в этих суммах каждого из разбиений каждое x_j будет участвовать ровно один раз. Отметим, что

$$sq_{\mathbf{AB}} = \sum_{j=1}^{q} x_j^2,$$

а из условия $\mathbf{C} \neq \mathbf{AB}$ следует, что в аналогичной формуле для $sq_{\mathbf{AC}}$ хотя бы в одном случае в квадрат будет возводиться сумма не менее чем двух разных x_j.

В необходимом и достаточном условии для попадания $\mathbf{C}$ на отрезок (лемма 6) каждое слагаемое вида $x_j^2, j = 1, ..., q$ и в левой, и в правой части встретится после раскрытия скобок ровно по два раза. Значит, его выполнение будет определяться только удвоенными парными произведениями $x_i x_j,\ i \neq j$. Поскольку пересечение разбиений $\mathbf{AB}$ и $\mathbf{BC}$ совпадает с $\mathbf{ABC} = \mathbf{AB}$, то в правой части условия общее x_j у двух разных квадратов может быть максимум одно, а следовательно, каждое удвоенное произведение нужного нам вида встретится там не более одного раза.

К тому же если $x_i x_j$ образуется, например, при раскрытии скобок в $sq_{\mathbf{AC}}$, то оно образуется и после преобразования $sq_{\mathbf{C}}$, потому что тогда D_i, D_j лежат в одном множестве $\mathbf{AC}$, а значит, и в одном множестве $\mathbf{C} \supseteq \mathbf{AC}$. Это же можно сказать и о множествах, составляющих $\mathbf{BC}$.

Наконец, поскольку $\mathbf{C}$, $\mathbf{A}$ не сравнимы, то при переходе к их пересечению хотя бы один элемент $\mathbf{C}$ разобьется на части: D_i, D_j, содержавшиеся в каком-то C_s, окажутся в разных множествах $\mathbf{AC}$. Тогда получится, что левая часть условия содержит $2x_i x_j$, а правая — нет. Отсюда

$$sq_{\mathbf{C}} + sq_{\mathbf{AB}} > sq_{\mathbf{AC}} + sq_{\mathbf{BC}},$$

условие попадания $\mathbf{C}$ в отрезок нарушено, и лемма доказана.

Мы подошли к основному результату работы.

Теорема. *Разбиение* $\mathbf{C}$ *лежит на отрезке* $[\mathbf{A}; \mathbf{B}]$ *тогда и только тогда, когда* $\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{A}$ *или* $\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{B}$.

Доказательство. Необходимость. Из лемм 8 и 10 обязательно $\mathbf{C} \supseteq \mathbf{AB}$. Согласно лемме 11, $\mathbf{C}$ сравнимо либо с $\mathbf{A}$, либо с $\mathbf{B}$. Если $\mathbf{A}$, $\mathbf{B}$ не сравнимы, то по лемме 9 возможны только случаи $\mathbf{C} \subseteq \mathbf{A}$ или $\mathbf{C} \subseteq \mathbf{B}$. Случай сравнимых $\mathbf{A}$, $\mathbf{B}$ обосновывается перебором возможных ситуаций. Проверим достаточность. Если, например, $\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{B}$, то $\mathbf{BC} = \mathbf{C}$ и $\mathbf{AC} = \mathbf{AB}$, следовательно, то, что $\mathbf{C}$ лежит на $[\mathbf{A}; \mathbf{B}]$, получается из основной леммы 6. Теорема полностью доказана.

Следствие. *Если* $\mathbf{C}$, $\mathbf{D}$ *лежат на отрезке* $[\mathbf{A}; \mathbf{B}]$, *то любое разбиение* $\mathbf{E} \in [\mathbf{C}; \mathbf{D}]$ *также лежит на этом отрезке.*

Доказательство. Рассмотрим четыре случая, которыми исчерпываются все возможности согласно теореме. Пусть сначала $\mathbf{AB} \subseteq \mathbf{C}, \mathbf{D} \subseteq \mathbf{A}$.

Тогда

$$\mathbf{AB} \subseteq \mathbf{CD} \subseteq \mathbf{E} \subseteq \mathbf{C} \cup \mathbf{D} \subseteq \mathbf{A} \;\Rightarrow\; \mathbf{E} \in [\mathbf{A};\, \mathbf{B}].$$

Ясно, что случай, когда оба **C, D** лежат между **AB** и **B**, полностью аналогичен.

Пусть теперь $\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{A}, \mathbf{AB} \subseteq \mathbf{D} \subseteq \mathbf{B}$. Тогда из двух правых включений следует, что $\mathbf{CD} \subseteq \mathbf{AB}$, и значит, с учетом включений левых, $\mathbf{CD} = \mathbf{AB}$. Окончательно, если $\mathbf{CD} \subseteq \mathbf{E} \subseteq \mathbf{D}$, то

$$\mathbf{AB} = \mathbf{CD} \subseteq \mathbf{E} \subseteq \mathbf{D} \subseteq \mathbf{B} \;\Rightarrow\; \mathbf{E} \in [\mathbf{A};\, \mathbf{B}].$$

Ситуация, когда **E** располагается между **CD** и **C**, и случай, при котором **C, D** в своем расположении в разных частях [**A**; **B**] меняются местами, в своем рассмотрении от уже обоснованных не отличаются. Следствие доказано.

Сформулированное следствие показывает, в чем рассматриваемый отрезок разбиений конечного множества похож на отрезок прямой в евклидовом пространстве. Но отличий у этих отрезков гораздо больше. Например, в отрезке [**A**; **B**] невозможно указать начало и конец в силу симметричности определения (3). Приведенный ниже пример показывает, что если $\mathbf{C}, \mathbf{D} \in [\mathbf{A};\, \mathbf{B}]$, то не всегда $\mathbf{C} \in [\mathbf{A};\, \mathbf{D}]$ или $\mathbf{D} \in [\mathbf{A};\, \mathbf{C}]$.

Рассмотрим $U = \{1, 2, 3, 4, 5\}$,

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

Тогда $\mathbf{AB} = \{\{1, 2\}; \{3\}; \{4\}; \{5\}\}$,

$$\mathbf{AC} = \mathbf{C},\ \mathbf{AD} = \mathbf{D};\ \mathbf{BC} = \mathbf{BD} = \mathbf{CD} = \mathbf{AB}.$$

Простые вычисления дают

$$sq_{\mathbf{A}} = 13; sq_{\mathbf{B}} = sq_{\mathbf{C}} = 9, sq_{\mathbf{D}} = 11, sq_{\mathbf{AB}} = 7,$$

что при использовании основной леммы 6 позволяет убедиться в справедливости $\mathbf{C} \notin [\mathbf{A};\, \mathbf{D}]$ и $\mathbf{D} \notin [\mathbf{A};\, \mathbf{C}]$, хотя **C, D** лежат на отрезке [**A**; **B**].

Заметим, что в этом примере $\mathbf{AB} \subseteq \mathbf{C}, \mathbf{D} \subseteq \mathbf{A}$. Тем не менее если, например,

$$\mathbf{AB} \subseteq \mathbf{C} \subseteq \mathbf{A}, \quad \mathbf{AB} \subseteq \mathbf{D} \subseteq \mathbf{B},$$

то **C** обязательно лежит на отрезке [**A**; **D**].

4. Обсуждение и выводы. Теорема, доказанная в работе, показывает, что отрезок в семействе кластерных разбиений [**A**; **B**] как бы состоит из двух частей — [**A**; **AB**] и [**AB**; **B**]. При этом все линейно упорядоченные цепочки разбиений, начала и концы которых расположены в одной из этих частей, целиком содержатся как в этой части, так и в отрезке [**A**; **B**] — это вытекает из леммы 7 и следствия теоремы. Множество элементов частично упорядоченного множества, попарно сравнимых между собой, в котором есть наибольший и наименьший элементы, иногда также называют его отрезком (см., например, [9]). Поэтому можно сказать, что отрезки в смысле частичного порядка всегда лежат в рассмотренных нами отрезках целиком. Однако на самом деле каждая из двух частей отрезка (3) имеет более сложную, «расслоенную» структуру — не любые два ее элемента сравнимы между собой. Это было продемонстрировано примером, завершившим предыдущий раздел.

Таким образом, отрезок в пространстве кластерных разбиений совсем не похож на обычные отрезки, например, в векторных пространствах — отдельные его подотрезки внутри себя содержат не сравнимые между собой элементы. Именно поэтому определение прямой в общих метрических пространствах через понятие отрезка (точка лежит на прямой, проходящей через две другие точки, если какой-либо из возможных отрезков с концами в двух из этих точек содержит третью точку) невозможно. Упоминание об этом содержалось еще в [10].

Содержательное определение понятия прямой в изучаемом метрическом пространстве будет трудно дать также в силу конечности семейства кластерных разбиений. Так, в [11] указано на важность понятия выпуклости метрического пространства. В нашей терминологии выпуклость означает наличие внутри любого отрезка [**A**; **B**] разбиения **C**, не совпадающего ни с одним из его концов. Конечно же, **AB** всегда там лежит, но если, например, разбиения $\mathbf{A} \subseteq \mathbf{B}$ оказываются соседними в частичном порядке, то требуемого разбиения **C**, очевидно, не существует. Приведем соответствующий пример. Пусть $U = \{1, 2, 3, 4\}$,

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

Ясно, что $\mathbf{A} \subset \mathbf{B}$, поэтому, согласно теореме, любое **C** из отрезка [**A**; **B**] должно удовлетворять условию $\mathbf{A} \subseteq \mathbf{C} \subseteq \mathbf{B}$. То, что $\mathbf{C} \supseteq \mathbf{A}$, означает, что объекты 1 и 2 в **C** должны попасть в один кластер. Если 3 и 4 попадут при этом в один кластер, то $\mathbf{C} = \mathbf{B}$, иначе $\mathbf{C} = \mathbf{A}$, и строго внутри отрезка разбиений нет.

Результаты работы могут быть применены, например, для нахождения точных вероятностных распределений расстояний между двумя кластерными разбиениями некоторого множества, если одно из них фиксировано, а другое может считаться случайным. Потребность в знании таких распределений возникает в самых разных областях, например, в доказательной медицине [12] или обработке социологических данных [13].

Библиографический список

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. — М., 1989.

2. Mills P. Efficient statistical classification of satellite measurements. // International Journal of Remote Sensing. — 2011. — № 32 (21). DOI: 10.1080/01431161.2010.507795

3. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». — Новосибирск, 2008.

4. Dronov S.V., Dementjeva E.A. A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. — 2012. — Vol. 7, № 1. DOI: 10.3233/MAS-2011-02-01.

5. Биргхоф Г. Теория решеток. — М. ; 1984.

6. Gratzer G. Lattice Theory: Foundations. — N.Y.: 2011.

7. Khamsi M.A. An Introduction to Metric Spaces and Fixed Point Theory. — San Francisco, CA: 2001.

8. Бураго Д.Ю., Бураго Ю.Д., Иванов С.В. Курс метрической геометрии. — М.; Ижевск, 2004.

9. Гуров С.И. Булевы алгебры, упорядоченные множества, решетки: определения, свойства, примеры. — М., 2013.

10. Дьёдонне Ж. Линейная алгебра и элементарная геометрия. — М., 1972.

11. Kaplansky I. Set Theory and Metric Spaces. — Washington, DC: 2001.

12. Sackett D.L., Rosenberg W.M., Gray J.A., Haynes R.B., Richardson W.S. Evidence Based Medicine: What It Is and What It Isn't // BMJ — 1996. – № 312 (7023). DOI:10.1136/bmj.312.7023.71

13. Bryukhanova E.A., Dronov S.V., Chekryzhova O.I. Spatial Approach to the Analysis of the Employment Data in Siberia Based on the 1897 Census (the Experience of the Multivariate Statistical Analysis of the Districts Data) // Journal of Siberian Federal University. Humanities & Social Sciences. — 2016. — № 7. DOI: 10.17516/1997-1370-2016-9-7-1651-1660.