ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРИКЛАДНОЙ ДИСКРЕТНОЙ МАТЕМАТИКИ

УДК 519.14 + 519.23

СТРУКТУРА КОЛЛЕКТИВА БЛИЖАЙШИХ СОСЕДЕЙ В СЕМЕЙСТВЕ РАЗБИЕНИЙ КОНЕЧНОГО МНОЖЕСТВА

С. В. Дронов

Алтайский государственный университет, г. Барнаул, Россия

Рассмотрены разбиения конечного множества на дизъюнктные подмножества, ближайшие к заданному (основному) разбиению в специальной кластерной метрике. Для фиксированного основного разбиения найдены вид ближайших к нему разбиений и их количество. На основе этого предложен статистический критерий для определения значимости отличий двух разбиений. Приводится пример обработки медицинских данных с помощью этого критерия.

Ключевые слова: *разбиения конечных множеств, кластерная метрика, статистическая значимость различия разбиений.*

DOI 10.17223/20710410/47/1

A STRUCTURE OF THE NEAREST NEIGHBORS COLLECTIVE IN A FAMILY OF PARTITIONS OF A FINITE SET

S. V. Dronov

Altai State University, Barnaul, Russia

E-mail: dsv@math.asu.ru

In this paper, we study partitions of a finite set of some objects into disjoint subsets closest to a given (main) partition. The distance between two partitions is taken equal to the sum of squares of numbers of the elements of sets that make up each of the partitions minus twice the sum of squares of the values of the sets forming the intersection of the partitions. For a fixed main partition, all the closest partitions and their number are found. The closest neighbors are always obtained by picking out one of the objects into a new set or by merging two single-element sets of the main partition (Theorem 1). The nearest neighbor here is $2(m-1)$ from the main partition, where m is the number of objects of the minimum non-singleton of the main partition, if one exists. Otherwise, this distance equals 2. Theorem 2 describes a situation where the number of elements of partitions must be the same. This happens, for example, when both partitions are constructed by the method of k -means for the same k. Here, to construct the nearest neighbor, one of the objects moves between the smallest sets of the main partition. Wherein, at least one of them must contain at least two objects. The corollaries of both theorems, obtained by accurately calculating the possible number of operations of the described type, give the exact quantities of nearest neighbors of the main partition, depending on its structure. We propose an application of the

obtained results to the construction of a statistical criterion for the significance of the difference between two partitions. An example of medical data processing using this criterion is given.

Keywords: *partitions of finite sets, cluster metric, statistical significance of differences of partitions.*

1. Основная задача работы

При рассмотрении разных способов разбиения конечного множества на части возникает ряд комбинаторных проблем [1; 2, Example 11.7]. На семействе всех возможных разбиений данного множества имеется естественная структура решётки, подробно изученная в [3]. Исследование одновременно нескольких разбиений множества на непустые части может потребоваться, например, в задачах анализа данных, в частности, подобное разбиение всегда появляется в результате применения некоторого алгоритма кластерного анализа. При применении к изучаемому множеству разных кластерных алгоритмов, равно как и при попытках разделения этого множества на группы по степени близости различных наборов характеризующих его элементы признаков, мы приходим, вообще говоря, к разным разбиениям. Сравнение получившихся разбиений может привести к заключению о степени различия применявшихся алгоритмов или о силе взаимного влияния и связей двух наборов признаков.

Рассмотрим множество U, состоящее из конечного числа n объектов. Набор непустых его подмножеств $A = \{a_1, \ldots, a_k\}$ будем называть разбиением U, если эти подмножества попарно дизъюнктны, а их объединение совпадает со всем множеством U. Сделанные предположения означают, в частности, что для каждого $x \in U$ найдется единственное множество $a_{i(x)}$ из набора A, для которого справедливо $x \in a_{i(x)}$.

Заметим, что любое кластерное разбиение множества U удовлетворяет данному определению. Мы не употребляем термин «кластерное разбиение» для изучаемых далее наборов множеств только потому, что при построении этих множеств не предполагается близости элементов каждого из них в каком-либо смысле, что является обязательным для кластеров. Хотя, конечно, можно считать признаком близости двух элементов сам факт попадания их в одно и то же множество разбиения.

Задача оценки степени различия разных разбиений одного и того же конечного множества имеет довольно широкий спектр приложений. При решении подобных задач можно использовать такие характеристики, как расстояние Кульбака — Лейблера (см., например, [4, гл. 14]) или взаимная информация разбиений [5, с. 104–105]. Хотя эти характеристики и не являются метриками на семействе разбиений, но с помощью специальных приёмов (симметризации и т. п.) на их основе можно построить метрики. Несколько в стороне лежат методы, основанные на так называемых редакционных расстояниях, схожих с расстоянием Левенштейна [6]. Здесь принадлежность элементов множества разным элементам разбиения кодируется с помощью набора букв, в котором элементам одного множества присваиваются одинаковые буквы, а далее рассчитывается количество замен букв, путём совершения которых набор букв одного разбиения может быть самым быстрым способом переведён в буквы другого разбиения. Алгоритмы для вычисления таких расстояний можно найти в [7]. К подобным метрикам можно также отнести расстояния Джаро и Джаро — Винклера (см. [8]). Общие подходы к определению метрик на семействе разбиений рассматриваются в [9]. Там же обсуждаются и вероятностные интерпретации различных метрик, в том числе и основной метрики настоящей работы.

Для оценки степени близости разбиений A и B одного и того же множества далее будем использовать кластерную метрику d [10]:

$$d(A,B) = \sum_{x\in U} |a_{i(x)} \Delta b_{j(x)}|. \tag{1}$$

Здесь символом Δ обозначена симметрическая разность множеств

$$a\Delta b = (a \setminus b) \cup (b \setminus a),$$

под $b_{j(x)}$ понимается то из множеств набора B, в котором оказывается x, а через $|c|$ обозначается число элементов конечного множества c.

В силу дискретного характера рассматриваемой задачи понятно, что множество всех возможных значений метрики d на семействе разбиений U конечно. В [10] замечено, что наибольшее возможное её значение равно $n(n-1)$ и достигается лишь в том случае, когда одно из разбиений каждый элемент U объявляет отдельным множеством, а второе является одноэлементным набором. В [10] для этих двух разбиений введены следующие обозначения:

$$\underline{U} = \{\{x_1\}, \ldots, \{x_n\}\}, \quad \overline{U} = \{U\}.$$

Из определения (1) ясно, что значениями d могут служить только целые неотрицательные числа. Основной целью работы является изучение возможных минимальных ненулевых значений данной метрики, которую иногда будем называть просто расстоянием, в случае, когда одно из разбиений фиксировано, а также выяснение всех возможных вариантов строения второго разбиения, которое удалено от первого на такое минимальное расстояние.

2. Несколько предварительных и технических результатов

В [10] приведена и более простая в применении формула, чем (1):

$$d(A,B) = \sum_{i,j} |a_i \cap b_j| \cdot |a_i \Delta b_j|.$$

Здесь сумма берётся по всем возможным парам a_i, b_j, которые можно составить из множеств двух изучаемых наборов.

Если имеются два разбиения $A = \{a_1, \ldots, a_k\}$, $B = \{b_1, \ldots, b_m\}$ множества U, то набор множеств

$$\{a_i \cap b_j : i = 1, \ldots, k,\ j = 1, \ldots, m\},$$

из которого исключены все пустые пересечения, также является разбиением U. Полученный таким образом набор обозначим AB и назовём пересечением разбиений A и B. Для произвольного набора конечных множеств $A = \{a_1, \ldots, a_k\}$ пусть

$$\mathrm{sq}(A) = \sum_{i=1}^{k} |a_j|^2.$$

В [11] для вычисления расстояния d получена следующая формула:

$$d(A,B) = \mathrm{sq}(A) + \mathrm{sq}(B) - 2\mathrm{sq}(AB). \tag{2}$$

Оказывается, что изучение значений, которые может принимать сумма квадратов натуральных чисел, для нашей задачи весьма важно.

Лемма 1 (о максимуме суммы квадратов). Пусть натуральные числа n, f, z_1, $\ldots$, z_f таковы, что $z_1 + \ldots + z_f = n$, $f \leqslant n$. Тогда величина

$$S(f, z_1, \ldots, z_f) = \sum_{i=1}^{f} z_i^2$$

при каждом фиксированном f достигает своего максимума тогда и только тогда, когда все z_i, кроме, возможно, одного из них, равны 1. Этот максимум равен

$$M(f) = (n - f + 1)^2 + f - 1$$

и монотонно убывает с ростом f.

Доказательство. Пусть, скажем, $z_i > z_j$. Тогда в силу неравенства

$$(z_i + 1)^2 + (z_j - 1)^2 > z_i^2 + z_j^2$$

величина $S(f, z_1, \ldots, z_f)$ строго возрастает при перемещении единицы в сторону большего слагаемого. Таким образом, если число слагаемых менять нельзя, то максимальное значение S достигается, например, когда $z_1 = n-f+1$, $z_2 = 1, \ldots, z_f = 1$. Проверка монотонности $M(f)$ при $1 \leqslant f \leqslant n$ элементарна. ■

Условимся писать $A \subset B$ и говорить, что разбиение B содержит разбиение A, если любое из множеств, составляющих разбиение A, является подмножеством какого-то множества из B. Тогда для каждого из множеств $a \in A$ может быть выбрана часть разбиения B, являющаяся разбиением a. В частности, $AB \subset B$, $AB \subset A$. Нам понадобятся следующие два простых следствия (2):

Лемма 2. Если $A \subset B$, то $d(A, B) = \mathrm{sq}(B) - \mathrm{sq}(A)$.

Лемма 3. $d(A, B) = d(A, AB) + d(AB, B)$.

Утверждение леммы 3 можно интерпретировать как расположение пересечения двух разбиений на прямолинейном отрезке, соединяющем разбиения A и B в метрическом пространстве разбиений. Некоторое развитие такого подхода, приводящее к выводу о том, что метрика (1) согласована с частичным порядком по включению на семействе всех разбиений, реализовано в [12].

Лемма 4. Пусть разбиение B получено из разбиения $A = \{a_1, \ldots, a_k\}$ перенесением одного элемента $x \in U$ из a_t в a_s. Тогда

$$d(A, B) = 2(|a_t| + |a_s| - 1).$$

Доказательство. Применим (2):

$$d(A, B) = \mathrm{sq}(A) + \left(\mathrm{sq}(A) - |a_t|^2 - |a_s|^2 + (|a_t| - 1)^2 + (|a_s| + 1)^2\right) - \\ -2\left(\mathrm{sq}(A) - |a_t|^2 + (|a_t| - 1)^2 + 1\right).$$

После этого осталось лишь раскрыть скобки. ■

При рассмотрении таких A, B, как в лемме 4, множество a_t условимся называть донором, а a_s — реципиентом элемента x.

Рассмотрим набор неотрицательных целых чисел $g_1, \ldots, g_N$. Пусть среди них имеется ровно n ненулевых. Обозначив эти n чисел $q_1, \ldots, q_n$, положим

$$Q(g_1, \ldots, g_N) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q_i q_j, \ n \geqslant 2,$$

а при $n \leqslant 1$ будем считать, что $Q(g_1, \ldots, g_N) = 0$.

Лемма 5. Если в наборе целых неотрицательных чисел $g_1, \ldots, g_N$ имеется ровно n ненулевых, $n \geqslant 2$, то

$$1 + Q(g_1, \ldots, g_N) \geqslant \sum_{i=1}^{N} g_i. \tag{3}$$

Доказательство. Пусть список $q_1, \ldots, q_n$ содержит все ненулевые числа набора. Заметив, что $Q(g_1, \ldots, g_N) = Q(q_1, \ldots, q_n)$ и $\sum\limits_{i=1}^{N} g_i = \sum\limits_{i=1}^{n} q_i$, будем действовать индукцией по n. При $n = 2$ неравенство

$$1 + q_1 q_2 \geqslant q_1 + q_2 \tag{4}$$

очевидно, если $q_1 = 1$. Предположив, что $1 + kq_2 \geqslant k + q_2$, получаем

$$1 + (k+1)q_2 = 1 + kq_2 + q_2 \geqslant k + q_2 + 1 = (k+1) + q_2,$$

что доказывает справедливость (4). Теперь предположим, что (3) выполнено при некотором $n = k$. Тогда

$$1 + Q(q_1, \ldots, q_{k+1}) = 1 + Q(q_1, \ldots, q_k) + q_{k+1} \sum_{j=1}^{k} q_j \geqslant \sum_{j=1}^{k} q_j + q_{k+1},$$

что завершает доказательство (3). ■

Далее будем считать, что разбиение A состоит из множеств $a_1, \ldots, a_k$, а разбиение B — из множеств $b_1, \ldots, b_m$. Введём в рассмотрение $g_{i,j} = |a_i \cap b_j|$. Тогда

$$|a_i| = \sum_{j=1}^{m} g_{i,j},\ i = 1, \ldots, k;\ |b_j| = \sum_{i=1}^{k} g_{i,j},\ j = 1, \ldots, m. \tag{5}$$

Среди $g_{i,j}$ могут содержаться нули, но в каждой из сумм (5) есть по крайней мере одно ненулевое слагаемое. Обозначим количества таких слагаемых в этих суммах через $N(a_i), N(b_j)$ соответственно.

Лемма 6. Если A и B различны, то $d(A, B) \geqslant 2$.

Доказательство. Используя (2), запишем

$$d(A, B) = (\mathrm{sq}(A) - \mathrm{sq}(AB)) + (\mathrm{sq}(B) - \mathrm{sq}(AB)),$$

причём обе разности неотрицательны по лемме 2. Следовательно, хотя бы одна из них не равна нулю. Пусть это вторая разность. Тогда из

$$\mathrm{sq}(B) - \mathrm{sq}(AB) = \sum_{j=1}^{m} \left(\left(\sum_{i=1}^{k} g_{i,j} \right)^2 - \sum_{i=1}^{k} g_{i,j}^2 \right)$$

вытекает, что хотя бы одно из слагаемых в выписанной сумме положительно. Но, согласно лемме 1, это означает, что одно из $N(b_j)$ не меньше 2. Взяв в соответствующей сумме (5) два ненулевых слагаемых, скажем g_1, g_2, видим, что

$$d(A, B) \geqslant \left(\sum_{i=1}^{k} g_{i,j} \right)^2 - \sum_{i=1}^{k} g_{i,j}^2 = 2Q(g_{1,j}, \ldots, g_{1,k}) \geqslant 2g_1 g_2 \geqslant 2.$$

Лемма 6 доказана. ■

Рассмотрим произвольное разбиение A множества U. Для $x \in U$ через $\mathbf{A}(x)$ обозначим семейство всех разбиений, полученных из A перемещением x из $a_{i(x)}$ в другое множество из A или выделением x в новое одноэлементное множество.

Лемма 7. Пусть A и B — два различных разбиения U. Тогда найдётся такой $x \in U$ и такое разбиение $A' \in \mathbf{A}(x)$, что $d(A, A') \leqslant d(A, B)$.

Доказательство. Сделаем сначала два допущения:

1) среди множеств первого разбиения нашлись два, a_1 и a_2, такие, что $N(a_1), N(a_2) \geqslant 2$;
2) $\mathrm{sq}(B) - \mathrm{sq}(AB) \neq 0$.

Тогда, как показано в доказательстве леммы 6, $\mathrm{sq}(B) - \mathrm{sq}(AB) \geqslant 2$, откуда

$$d(A, B) = \mathrm{sq}(A) - \mathrm{sq}(AB) + \mathrm{sq}(B) - \mathrm{sq}(AB) \geqslant \mathrm{sq}(A) - \mathrm{sq}(AB) + 2. \tag{6}$$

Дважды применяя лемму 5, получаем

$$d(A, B) \geqslant 2Q(g_{1,1}, \ldots, g_{1,m}) + 2Q(g_{2,1}, \ldots, g_{2,m}) + 2 \geqslant 2(|a_1| + |a_2| - 1),$$

что завершает доказательство в сделанных допущениях, поскольку правая часть последнего неравенства равна расстоянию между A и разбиением, полученным из него перенесением любого из элементов a_1 в a_2 или наоборот.

Пусть нарушается первое из допущений. Если все $N(a_i) = 1$, то каждое множество из A пересекается ровно с одним множеством из B, что означает $A \subset B$. Тут каждое множество из B может быть разбито на какие-то множества из A. При этом хотя бы в одном таком разбиении должно найтись не менее двух элементов, так как иначе $A = B$. Пусть $b_1 \supset a_1 \cup a_2$. Тогда

$$d(A, B) \geqslant |b_1|^2 - |a_1|^2 - |a_2|^2 \geqslant 2|a_1| \cdot |a_2| \geqslant 2|a_1| > 2(|a_1| - 1),$$

что, согласно лемме 4, означает возможность построить A', отделяя один из элементов a_1 в новое одноэлементное множество. Если и a_1, и a_2 оказались одноэлементными, требуемый эффект достигается объединением их в одно двухэлементное множество (в этом случае $d(A, A') = 2$ и $d(A, B)$, согласно лемме 6, не меньше этого значения).

Если только $N(a_1) \geqslant 2$, отделим в новое одноэлементное множество один из элементов a_1, результат приняв за A'. При этом из леммы 5 вытекает

$$d(A, B) \geqslant 2Q(g_{1,1}, \ldots, g_{1,m}) \geqslant 2(|a_1| - 1) = d(A, A').$$

Наконец, предположим, что $\mathrm{sq}(B) - \mathrm{sq}(AB) = 0$. Привлекая лемму 2, приходим к выводу, что тогда $B = AB$, т. е. $B \subset A$. На этот раз каждое из множеств A образуется объединением некоторых из $b_1, \ldots, b_m$, причём, чтобы не допустить совпадения разбиений, хотя бы одно a_i должно содержать не менее двух таких множеств. Тогда $N(a_i) \geqslant 2$, а этот случай только что был рассмотрен. ■

3. Ближайшее разбиение

Сначала займёмся поиском ближайшего к A среди тех разбиений B, для которых $A \not\subset B$. Из этого условия, в частности, следует, что $d(A, AB) \neq 0$. Поэтому для рассматриваемого случая из леммы 3 вытекает строгое неравенство $d(A, B) > d(A, AB)$. Заменим любое найденное B на AB. При этом расстояние уменьшится, следовательно, нужное B обязательно таково, что $B \subset A$. Таким образом, для его построения следует разбить какие-то из множеств, составляющих A, на дизъюнктные части. Лемма 7 показывает, что ближайшее разбиение всегда получается перемещением одного элемента. При этом, согласно лемме 4, множества, между которыми перемещается этот единственный элемент, должны содержать минимально возможные количества элементов.

Таким образом, реципиент должен быть пустым, а вот донор не может содержать менее двух элементов. Заметим, что требуемого донора не существует только в случае, когда $A = \underline{U}$, а для такого разбиения, очевидно, нет ни одного B с нужным условием. Согласно лемме 4, расстояние между найденным и исходным разбиением равно

$$d_0 = 2(n_A - 1), \tag{7}$$

где n_A — минимальное число элементов множеств A, большее 1.

Перейдём к поиску ближайшего разбиения с условием $B \supset A$. Из проведённого рассуждения понятно, что A должно получаться отделением одного элемента от некоторого множества из B, если только это возможно. Итак, если в A имеется хотя бы одно одноэлементное множество, то B получается объединением этого множества с тем из остальных множеств A, которое имело наименьшее число элементов. Если это множество также было одноэлементным, то искомое разбиение по лемме 4 удалено от исходного на расстояние 2, иначе, с учётом (7) и того, что новый донор содержит $n_A + 1$ элемент, на величину $2n_A$, что больше, чем d_0 при $n_A > 2$, поэтому не является минимальным. Тем не менее значение d_0 в (7) может быть равным 2, как и при объединении двух одноэлементных множеств, в случае, когда $n_A = 2$.

Если в A нет ни одного одноэлементного множества, то любое разбиение $B \supset A$ будет заведомо дальше от A, чем построенное «внутреннее» разбиение. Действительно, построение требуемого B должно будет сопровождаться перемещением более чем одного элемента, тогда как ранее перемещался единственный $x \in U$. Вывод следует из леммы 7.

Пусть $M(k; A)$ — количество тех множеств в разбиении A, которые состоят ровно из k элементов. Резюмируем проведённые рассуждения.

Теорема 1. Пусть $M(1; A) \geqslant 2$. Если $M(2; A) = 0$, то ближайшее к A разбиение B получается заменой двух любых одноэлементных множеств на их объединение. Если $M(2; A) \geqslant 1$, то, кроме описанного способа, можно разбить любое двухэлементное множество A на два одноэлементных. Во всех указанных вариантах $d(A, B) = 2$.

Пусть $M(1; A) \leqslant 1$, a — любое из минимальных по числу элементов неодноэлементных множеств A. Тогда ближайшее к A разбиение B получается выделением одного из элементов a в новое множество и $d(A, B) = 2(n_A - 1)$, где n_A — число элементов a. С помощью этой процедуры могут быть построены все разбиения, ближайшие к A.

Заметим, что ближайшее к A разбиение обязательно либо строго содержится в нём, либо содержит его. Это вновь подтверждает заключение [12] о том, что метрика d согласована со структурой решётки на семействе всех разбиений.

Подсчёт количества ближайших разбиений происходит параллельно алгоритму их формирования. Следует только учесть, что отделение любого из элементов двухэлементного множества всегда даёт один и тот же результат, тогда как для множества из большего числа элементов число разных результатов отделения равно числу элементов множества.

Следствие 1. Число $l(A)$ разбиений, ближайших к разбиению A, вычисляется следующим образом:

$$l(A) = \begin{cases} C^2_{M(1;A)} + M(2; A), & M(1; A) \geqslant 2; \\ n_A M(n_A; A), & M(1; A) \leqslant 1, M(2; A) = 0; \\ M(2; A), & M(1; A) \leqslant 1, M(2; A) \neq 0. \end{cases}$$

В некоторых исследованиях в качестве допустимых могут рассматриваться только те разбиения, которые содержат фиксированное число множеств. Так бывает, например, в случае, когда каждое из рассматриваемых разбиений исследователь получает при помощи алгоритма k-средних, который часто применяется в приложениях и постоянно совершенствуется [13]. При таком предположении полученный результат нуждается в пересмотре — каждое из найденных ближайших разбиений имеет иное количество составляющих его множеств, чем исходное. Но леммы 4 и 7 позволяют произвести требуемый пересмотр довольно легко: для построения ближайшего к данному разбиению необходимо переместить один элемент между двумя множествами исходного разбиения с минимальными количествами элементов. Запрещённым оказывается лишь случай, когда реципиент оказывается пустым. Сформулируем результат.

Упорядочим множества, составляющие разбиение A, по возрастанию числа их элементов. Каждому множеству присвоим ранг, считая, что множества, имеющие одинаковое число элементов, получают одинаковые ранги. Через k_j будем обозначать количество элементов в множестве с рангом j. С учётом введённых ранее обозначений количество множеств разбиения A, имеющих ранг j, равно $M(k_j; A)$.

Теорема 2. Пусть $A \neq \underline{U}, \overline{U}$. Если $k_1 = 1$ или $M(k_1; A) = 1$, то произвольное B, ближайшее к A среди разбиений, состоящее из такого же числа множеств, образуется перенесением одного элемента из множества ранга 2 в произвольное множество ранга 1. При этом

$$d(A, B) = 2(k_2 + k_1 - 1). \tag{8}$$

Иначе ближайшее разбиение требуемого типа образуется перемещением одного элемента между произвольными двумя множествами ранга 1, причём

$$d(A, B) = 2(2k_1 - 1). \tag{9}$$

Ясно, что для разбиений $\underline{U}, \overline{U}$ задача не имеет решений. Поскольку в (8) $k_2 \geqslant 2$, а расстояние (9) возникает лишь при $k_1 \geqslant 2$, то из теоремы следует, что $d(A, B) \geqslant 4$. Это совпадает с результатом, полученным ранее в [10].

Следствие 2. Пусть $A \neq \underline{U}, \overline{U}$. Число разбиений $l_k(A)$, удаленных от заданного разбиения A на минимальное расстояние и имеющее то же число множеств k, может быть вычислено следующим образом:

$$l_k(A) = \begin{cases} k_2 M(k_2; A) M(k_1; A), & k_1 = 1 \text{ или } M(k_1; A) = 1; \\ k_1^2 C^2_{M(k_1;A)}, & k_1 \neq 1,\ M(k_1; A) \geqslant 2. \end{cases}$$

4. Обсуждение. Пример применения

Способ получения ближайшего к некоторому фиксированному разбиению, а также полученные в п. 3 количества ближайших разбиений можно рассматривать как первый шаг к построению статистического критерия для определения значимости отличий друг от друга разных разбиений. Такой критерий можно использовать для решения многих практических задач.

Предположим, задано некоторое разбиение A, которое назовём основным. Например, оно было получено методами, которым мы доверяем, или предложено квалифицированными экспертами. Пусть в результате применения новых методов к тем же данным получено другое разбиение B. Если различия разбиений A и B оказываются статистически незначимыми, то это может являться основанием для внедрения новых методов в исследовательскую практику.

Можно предложить применение подобного рассуждения и для нового решения задач сокращения размерности в задачах кластерного анализа. Если исключение одного или нескольких формирующих показателей не приводит к существенному изменению итогового разбиения, то эти показатели можно исключить без существенных потерь информации.

Допустим, что для основного разбиения найдены все значения, принимаемые $d(A, C)$ для всех возможных разбиений C, а также повторности каждого из этих значений. Это означает, что каждое возможное значение расстояния d встречается известное число n_d раз. Известно [14], что число разбиений множества из n элементов на непустые подмножества задаётся числом Белла, которое равно сумме чисел Стирлинга второго рода:

$$B_n = \sum_{m=1}^{n} S(n, m).$$

Тогда в предположении, что разбиение B могло оказаться произвольным, вероятность получить более удалённое от A разбиение равна

$$Q(B) = \frac{1}{B_n} \sum_{d>d(A,B)} n_d, \tag{10}$$

и именно это число следует рассматривать как меру значимости различия рассматриваемых двух разбиений (или вероятность того, что они близки).

Если число множеств в разбиении B может быть только тем же, что и в разбиении A, то следует составить таблицу всех возможных расстояний и их повторностей только для таких разбиений. При этом полное число допустимых разбиений равно $S(n, k)$, если основное разбиение состоит из k множеств.

Для примера рассмотрим множество из пяти пациентов с достоверно установленными тремя диагнозами (тромбоз глубоких вен, тромбоэмболия легочной артерии и их сочетание). Второй строкой в табл. 1 задано основное разбиение A множества пациентов.

Таблица 1

Диагнозы и генотип

Пациенты	A	B	C	D	E
Диагноз	1	2	1	3	2
Генотип	1	3	2	2	3

Изучим разбиение множества этих же пациентов, задаваемое их генотипом по гену F5 (фактор Лейден, свертываемость крови); обозначим: 1 — отсутствие патологического гена в обеих аллелях (нормозигота), 2 — наличие патологии в одной аллели (гетерозигота), 3 — патологический ген в обеих аллелях (гомозигота). Данные приведены в третьей строке табл. 1. Расстояние d между двумя этими разбиениями равно 4. Все возможные расстояния от основного разбиения и их повторности приведены в табл. 2.

Таблица 2

Расстояния и повторности

d	0	2	4	6	8	10	12	16	Всего
n_d	1	2	7	14	15	4	8	1	52

Отсюда формула (10) приводит к $Q(B) = 42/52 \approx 0{,}81$, что означает довольно высокую вероятность схожести двух разбиений. Это даёт повод для заключения о существенной сцепленности генотипа и диагноза. Если вместо изучения трёх возможных генотипов второе разбиение свести к констатации наличия или отсутствия патологии в генотипе, тем самым объединив множества, элементы которых соответствуют 2 и 3 в третьей строке табл. 1, то получается $d = 8$, откуда $Q(B) = 13/52 = 0{,}25$, что приводит к выводу о гораздо меньшей надёжности такой «более грубой» формы представления данных.

Конечно, уверенный вывод обычно делают, если соответствующая вероятность не менее 0,95 или не более 0,05, но в примере для наглядности взята выборка слишком малого объёма, по которой надёжные выводы сделать заведомо невозможно.

Заключение

В работе изучена структура метрического пространства на семействе всех разбиений конечного множества. Для каждого из возможных разбиений полностью описано строение разбиений, удалённых от фиксированного разбиения на минимальное расстояние в специальной кластерной метрике. Рассчитано количество таких ближайших соседей для произвольного конкретного разбиения. В качестве приложения результатов предложен новый статистический критерий для установления значимости различий двух разбиений одного и того же множества.

Автор выражает благодарность рецензенту за ценные замечания и указание ряда альтернативных способов оценки различия двух разбиений, ранее автору неизвестных.

ЛИТЕРАТУРА

1. *Brualdi R. A.* Introductory Combinatorics. 5th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2017. 624 p.
2. *Bender E. A. and Williamson S. G.* Foundations of Combinatorics with Applications. Mineola, NY: Dover Publ., 2006. 480 p. `www.math.ucsd.edu/~ebender/CombText/ch-11.pdf`
3. *Birkhoff G.* Lattice Theory. 3rd ed. Providence, Rhode Island: AMS, 1991. 420 p.
4. *Press W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P.* Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge University Press, 2007. 1235 p.
5. *Яглом А. М., Яглом И. М.* Вероятность и информация. 3-е изд. М.: Наука, 1973. 513 с.
6. *Левенштейн В. И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // ДАН СССР. 1965. Т. 163. Вып. 4. С. 845–848.
7. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология. СПб.: Невский Диалект БВХ-Петербург, 2003. 654 с.
8. *Cohen W. W., Rawikumar P., and Fienberg S. E.* A comparison of string distance metrics for name-matching tasks // Proc. IIWEB'03, Acapulco, Mexico: AAAI Press, 2003. P. 73–78.
9. *Каграманян А. Г., Машталир В. П., Скляр Е. В., Шляхов В. В.* Метрические свойства разбиений множеств произвольной природы // Докл. НАН Украины. 2007. Т. 6. С. 35–39.
10. *Дронов С. В.* Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. 2011. Т. 69. № 1/2. С. 32–35.
11. *Dronov S. V. and Evdokimov E. A.* Post-hoc cluster analysis of connection between forming characteristics // Model Assisted Statistics Appl. 2018. V. 13. No. 2. P. 183–192.
12. *Дронов С. В.* Кратчайшие маршруты семейства кластерных разбиений // Труды семинара по геометрии и математическому моделированию. 2017. № 3. С. 4–12.
13. *Gribel D. and Vidal T.* HG-means: A scalable hybrid metaheuristic for minimum sum-of-squares clustering // Pattern Recognition. 2019. V. 88. No. 1. P. 569–583.

14. *Riordan J.* Introduction to Combinatorial Analysis. Mineola, NY: Dover Publ., 2006. 256 p.

REFERENCES

1. *Brualdi R. A.* Introductory Combinatorics. 5th ed. Upper Saddle River, NJ, Pearson Prentice Hall, 2017. 624 p.
2. Foundations of Combinatorics with Applications. Mineola, NY, Dover Publ., 2006. 480 p. www.math.ucsd.edu/~ebender/CombText/ch-11.pdf
3. *Birkhoff G.* Lattice Theory. 3rd ed. Providence, Rhode Island, AMS, 1991. 420 p.
4. *Press W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P.* Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge University Press, 2007. 1235 p.
5. *Yaglom A. M. and Yaglom I. M.* Veroyatnost i Informatsiya [Probability and Information], 3rd ed. Moscow, Nauka Publ., 1973. 513 p. (in Russian)
6. *Levenshteyn V. I.* Dvoichnyye kody s ispravleniyem vypadeniy, vstavok i zameshcheniy simvolov [Binary codes for correcting dropouts, inserts, and symbol substitutions]. Reports of the USSR Academy of Sciences, 1965, vol. 163, no. 4, pp. 845–848. (in Russian)
7. *Gasfild D.* Stroki, Derevia i Posledovatelnosti v Algoritmakh. Informatika i Vychislitelnaya Biologiya [Lines, Trees, and Sequences in Algorithms. Computer Science and Computational Biology]. St. Petersburg, Nevskiy Dialekt BVKh-Peterburg, 2003. 654 p. (in Russian)
8. *Cohen W. W., Rawikumar P., and Fienberg S. E.* A comparison of string distance metrics for name-matching tasks. Proc. IIWEB'03, Acapulco, Mexico, AAAI Press, 2003, pp. 73–78.
9. *Kagramanyan A. G., Mashtalir V. P., Sklyar E. V., and Shlyakhov V. V.* Metricheskiye svoystva razbiyeniy mnozhestv proizvolnoy prirody [Metric properties of partitions of sets of arbitrary nature]. Reports of the Academy of Sciences of Ukraine, 2007, vol. 6, pp. 35–39. (in Russian)
10. *Dronov S. V.* Odna klasternaya metrika i ustoychivost klasternykh algoritmov [One cluster metric and the stability of cluster algorythms]. Izvestiya AltGU, 2011, vol. 69, no. 1/2, pp. 32–35. (in Russian)
11. *Dronov S. V. and Evdokimov E. A.* Post-hoc cluster analysis of connection between forming characteristics. Model Assisted Statistics Appl., 2018, vol. 13, no. 2, pp. 183–192.
12. *Dronov S. V.* Kratchayshie marshruty semeystva klasternykh razbieniy [The shortest routes in the family of the cluster partitions]. Workshop on Geometry and Mathematical Modeling, 2017, no. 3, pp. 4–12. (in Russian)
13. *Gribel D. and Vidal T.* HG-means: A scalable hybrid metaheuristic for minimum sum-of-squares clustering. Pattern Recognition, 2019, vol. 88, no. 1, pp. 569 – 583.
14. *Riordan J.* Introduction to Combinatorial Analysis. Mineola, NY, Dover Publ., 2006. 256 p.