УДК 519.237

Элементы отдаленного горизонта в семействе разбиений конечного множества

Дронов С.В.
Алтайский государственный университет, г. Барнаул
dronovsv@math.asu.ru

Аннотация

В работе предложен алгоритм, который по заданному кластерному разбиению конечного множества позволяет построить его же разбиение, наиболее сильно отличающееся от заданного в смысле специальной кластерной метрики. Реализация такого построения дает возможность оценивать степень различия двух кластерных разбиений, одно из которых рассматривается как эталонное. Предлагаемый в работе для оценки степени подобного различия коэффициент обладает рядом преимуществ по сравнению с коэффициентом кластерных различий, предложенным автором ранее. Эти преимущества позволяют, в частности, боле аргументированно принимать решения о возможности внедрения в практику новых методик классификации.

Ключевые слова: Семейство кластерных разбиений, степень различия разбиений, коэффициент кластерных различий, удаленные разбиения

1. Обоснование основной задачи

Одним из эффективных способов предварительного анализа больших объемов данных, безусловно, является разбиение их набора на группы, внутри которых они в каком-либо смысле похожи. На сегодняшний день разработано большое количество алгоритмов для совершения подобных действий, например, в рамках так называемого кластерного анализа. Набор этих методик расширяется настолько бурно, что даже перечисление наиболее популярных сегодня подобных алгоритмов может составить приличный объем текста, см., например, обзор [1].

Используя различные алгоритмы разбиения набора объектов на группы, исследователь получает каждый раз, вообще говоря, различные результаты. При этом обнаруженная близость друг к другу большинства этих результатов позволяет, видимо, утверждать, что для изучаемого набора данных существует некоторое идеальное, объективно верное разбиение, на которое все получаемые результаты в каком-то смысле должны быть похожи. Одна из практических задач, остро нуждающаяся в аппарате для сравнения разбиений, это медицинская задача дифференциальной диагностики. Допустим, у нас есть классическая методика разбиения множества пациентов по установленным диагнозам. Предлагается новая (более быстрая, простая или использующая новое оборудование) методика. Если различия в результатах работы классической и новой методик можно считать небольшими, то, видимо, новая методика может быть рекомендована к внедрению. Потребность сравнивать между собой разные разбиения влечет необходимость введения на семействе всех разбиений расстояния или метрики. Подобные метрики известны в достаточном количестве, далеко не полный список используемых сегодня подходов можно найти в [2–9]. Но знание численного значения метрики не дает строгого понимания, существенны ли различия между рассматриваемыми разбиениями. Правильней, на взгляд автора, рассматривать

не абсолютные, а относительные ее значения. Точнее, предлагается исследовать отношение значения расстояния между основным и предлагаемым разбиениями к максимально возможному такому расстоянию, и его величину принять за меру сходства. Разумеется, предварительно для данного (основного) разбиения нужно уметь находить наиболее удаленные от него. Перейдем к точным формулировкам. Рассмотрим конечное множество U, состоящее из n объектов. Его разбиением будем называть набор непустых подмножеств $A_1, \ldots, A_m$ с условиями

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

Обозначать такое разбиение будем $A = \{A_1|A_2|\ldots|A_m\}$, а множества, его составляющие, условимся называть его элементами или кластерами. Строго говоря, не любое разбиение описанного вида обязательно является кластерным, но произвольное кластерное разбиение удовлетворяет выдвинутым требованиям, поэтому никаких интуитивных препятствий для употребления подобной терминологии не возникнет. Семейство всех возможных разбиений данного множества обозначим Ξ.

Допустим, выбранное разбиение A является в каком-то смысле основным. Целью работы является изучение тех разбиений, которые в наибольшей степени отличаются от него, образуя тем самым в интуитивном смысле горизонт в семействе всех разбиений, наиболее отдаленный от этого основного разбиения. Близость или удаленность двух разбиений определим, используя кластерную метрику, предложенную в [10]. Как показано в [11], для ее вычисления удобно использовать следующую методику. Пусть имеются два разбиения $A = \{A_1|A_2|\ldots|A_m\},\ B = \{B_1|B_2|\ldots|B_k\}$, а еще одно разбиение, AB, образовано из всевозможных непустых пересечений $A_i \cap B_j$. $i = 1, \ldots, m;\ j = 1, \ldots, k$. Это разбиение называют пересечением A и B. Через $sq(C)$ обозначим сумму квадратов всех количеств объектов элементов разбиения C, например

$$sq(A) \ = \ \sum_{i=1}^{m} |A_i|^2.$$

Расстояние между разбиениями тогда вычислим по формуле

$$d(A, B) \ = \ sq(A) + sq(B) - 2sq(AB).$$

Теперь сформулируем главную нашу задачу строго. Задано основное разбиение A множества U. Требуется найти наибольшее возможное расстояние $d^*(A) = \max\limits_{B} d(A, B)$, а также по возможности описать структуру набора всех максимально удаленных от A разбиений B.

Такая постановка задачи связана с тем, что для различных разбиений A число $d^*(A)$ не может быть всегда одинаковым, и, например, идею характеризовать степень различия двух разбиений относительной величиной $d(A, B)$ в долях максимально возможной такой величины $n(n-1)$, как это было предложено ранее в [10], нельзя считать достаточно плодотворной, поскольку эта доля достигает значения 1 лишь для единственной пары разбиений из Ξ. Далее будем называть число $d^*(A)$ отдалением разбиения A от горизонта.

2. Структура решетки на семействе Ξ

Определим на семействе всех разбиений множества U отношение включения, на основе которого построим частичный порядок. Если даны два разбиения $A = \{A_1|A_2|\ldots|A_m\},\ B = \{B_1|B_2|\ldots|B_k\}$, то условимся называть A меньшим, чем B, и писать

$A \subset B$, если $(\forall i)(\exists j)\ A_i \subset B_j$. В [12] отмечено, что Ξ в этом частичном порядке образует решетку. В ней в качестве инфимума двух разбиений выступает их пересечение, а в качестве супремума подобным же образом определяемое объединение разбиений. В этой решетке имеется наибольшее разбиение $\overline{U}$, единственным кластером которого является само U, и наименьшее, $\underline{U}$, в котором каждый из n объектов U расположен в отдельном кластере.

Для $A \subset B$ очевидно, что $d(A,B) = sq(B) - sq(A)$, ведь $AB = A$. Из этого уже нетрудно вывести, что для произвольных разбиений A, B

$$d(A,B) = d(A,AB) + d(AB,A), \tag{1}$$

а отсюда уже немедленно следует согласованность рассматриваемой метрики со структурой решетки: кратчайший путь по ребрам решетки из A в B имеет длину, равную $d(A,B)$ и обязательно проходит через пересечение этих двух разбиений (подробности см. [12], с. 286 – 289).

Удобным будет считать решетку всех разбиений как бы изображенной вертикально, причем каждому конкретному разбиению (скажем, B) мы поставим в соответствие точку в этом изображении, лежащую на высоте $sq(B)$. Таким образом, эту характеристику можно назвать высотой разбиения B. Концы каждого из ребер будут обязательно расположены на разных высотах, т.е., двигаясь по любому ребру, мы перемещаемся либо вверх, либо вниз. Самой нижней точкой изображения будет $\underline{U}$, лежащая на высоте n, а самой высокой, расположенной на высоте n^2, окажется $\overline{U}$.

В этом смысле заявленный поиск наиболее удаленных от A разбиений равнозначен поиску самого длинного из кратчайших путей по ребрам решетки, стартующего в основном разбиении и такого, что он должен обязательно проходить через некоторое разбиение C, лежащее не выше A, и на участке от A до C идти по решетке вниз, а затем постоянно подниматься выше в конечный пункт пути B. Промежуточное разбиение C, точка смены направления движения, это пересечение AB, как немедленно следует из (1). При этом не исключается случай, когда $C = A$, и путь сразу идет вверх – этот случай мы рассмотрим далее отдельно. Путь нужного нам вида, начинающийся ребром, ведущим вниз по решетке, назовем нижним, а начинающийся ребром вверх – верхним.

Для понимания механизма работы алгоритма построения самого длинного из кратчайших путей важным для нас будет следующее элементарное утверждение.

Лемма 1. *Пусть сумма набора из нескольких натуральных чисел равна заданному числу n, Sq – сумма квадратов этих чисел. Максимально возможное значение Sq равно n^2 и достигается только в том случае, когда набор состоит из единственного числа. Если же количество чисел набора фиксировано, то наибольшее значение Sq достигается лишь в случае, когда все числа, кроме одного, равны единице.*

Из этого утверждения можно сделать вывод, что сумма квадратов чисел, сумма которых задана, строго увеличивается каждый раз, когда мы добиваемся увеличения максимального из них.

3. Самый длинный путь с фиксированным пересечением

Временно допустим, что разбиение C, лежащее ниже A, в котором самый длинный путь меняет направление, далее двигаясь лишь вверх, задано. Конечной точкой искомого пути, следовательно, должно стать такое разбиение B_C, что оно лежит как можно выше в решетке, и $AB_C = C$. Опишем алгоритм построения такого разбиения, который далее будем называть борд-алгоритмом. Сначала нам понадобится

Вспомогательный алгоритм (составление списка $\Psi(C,A)$)

На входе заданы два разбиения – основное A и текущее C, $C \subset A$. Пусть за кластерами основного разбиения A закреплены произвольные метки (не обязательно числа).

Упорядочим кластеры разбиения C в порядке убывания количеств объектов в их составе и в этом порядке пронумеруем. Каждому из этих кластеров припишем метку того кластера A, частью которого он является. Кластеры C с одинаковыми метками назовем A-группой.

Образовавшийся список с приписанным каждому его элементу меткой и номером, объявим списком $\Psi(C, A)$. Конец алгоритма.

Теперь можно описать алгоритм поиска наиболее длинного пути с заранее заданной нижней точкой.

Борд-алгоритм

На входе вновь заданы два разбиения – основное A и текущее C, $C \subset A$.

- Шаг 0. Составим список $\Psi(C, A)$.
- Шаг 1. Возьмем по одному кластеру из C в каждой A-группе кластеров. При этом будем каждый раз брать тот кластер из группы, который в списке $\Psi(C, A)$ имеет наименьший номер. Это будет самый большой по количеству элементов кластер в каждой из групп. Объединив все взятые кластеры, объявим результат первым кластером разбиения B_C.
- Шаг 2. Удалим все выбранные кластеры из списка. Будем повторять шаг 1, строя последовательно второй, третий и т.д. кластер B_C. Алгоритм заканчивает работу, когда список $\Psi(C, A)$ опустеет.

Из приведенной выше основной леммы и соотношения

$$d(C, B_C) \;=\; sq(B_C) - sq(C)$$

следует, что построенное разбиение будет наиболее удаленным от C среди всех, лежащих выше него, которые в пересечении с A дают разбиение C. Для этого достаточно отметить максимальность высоты B_C среди всех таких разбиений.

Таким образом, для выявления пути, ведущего от A к наиболее удаленному от него разбиению достаточно выбрать среди всех разбиений, меньших A, оптимальную точку смены направления C (пересечение начального и конечного разбиений пути) Тогда построенное с помощью борд-алгоритма B_C будет искомой конечной точкой. Более того, любое наиболее удаленное от A разбиение может быть найдено именно таким образом. Действительно, если D не получается из AD с помощью борд-алгоритма, то B_{AD}, очевидно, окажется более удаленным от A, чем D.

Изучим возможность разделения одного из кластеров разбиения $D \subset A$ на две меньшие части. Такое деление соответствует попытке перенести точку D смены направления пути ниже по решетке. При этом ставим задачу для результирующего разбиения C добиться большего удаления B_C от A, чем от стартового разбиения было удалено B_D.

Составим список $\Psi(D, A)$. Внутри каждой из A-групп упорядочим кластеры D по убыванию количеств их элементов. Количества элементов в кластерах в i-й группе обозначим следующим образом: $w_{1,i} \geq w_{2,i} \geq \ldots$, и такую цепочку будем называть i-групповым порядком. При этом, если индекс j оказывается большим, чем имеется кластеров в группе, то условимся считать, что $w_{j,i} = 0$. Введем также обозначения

$$R_j = \sum_{i=1}^{m} w_{j,i}.$$

Эти числа представляют собой количества объектов из U в кластерах разбиения B_D, результата работы борд-алгоритма. Среди них может оказаться достаточно много нулей. В частности, когда в качестве C берется начальное разбиение A (первый шаг), то $R_1 = n,\ R_j = 0,\ j \geq 2$.

Рассмотрим W, наибольший по количеству объектов кластер некоторой (ниже мы назовем ее текущей) группы списка $\Psi(D, A)$. Пусть он состоит из w объектов и для построения C разбивается на две части из s ($s \leq w/2$) и $w - s$ объектов соответственно. Можно без ограничения общности считать, что такой кластер всегда есть, поскольку разделение кластера на несколько частей можно осуществлять последовательно, а результаты будут лежать вдоль цепочки ребер строящегося пути. Далее второй индекс в обозначениях (номер группы, внутри которой мы работаем) временно опустим.

Два образовавшихся новых кластера локализуем в групповом порядке. Предположим, что больший по числу объектов кластер (из $w - s$ объектов) встанет в новом порядке на место, ранее занятое w_t, а меньший – на место w_z. При этом, разумеется, $z > t$. На самом деле, для алгоритма значения z, t не важны, важны лишь численные значения w_t и w_z, которые будут в итоге замещены в начальном A-порядке. Аккуратно отслеживая изменения всех входящих в (1) величин при замене B_D на B_C и D на C, приходим к выводу, что

$$d = d(A, B_D) = d(A, B_C) + Q - 2f(s), \tag{2}$$

где Q напрямую не зависит от s, а

$$f(s) = s^2 + s(R_t - w_t - R_z + w_z - w). \tag{3}$$

Отсюда видно, что при выполнении условия

$$R_t - w_t - R_z + w_z - w \geq 0,$$

или, иначе,

$$R_t - w_t \geq R_z - w_z + w, \tag{4}$$

увеличение s приводит к уменьшению расстояния от A до нового кандидата на самое удаленное разбиение, а значит, следует принять s равным 0. Это приводит к отказу от изменения самого большого из кластеров в рассматриваемой группе списка $\Psi(D, A)$. Отметим далее, что, в силу монотонности d по длине разбиваемого кластера w, отсюда ясно, что и остальные кластеры в этой группе также менять не следует.

В частности, если (4) окажется выполненным уже на старте основного алгоритма, когда $D = A, t = 1, z = 2$, а следовательно, (4) превращается в $w \leq n/2$, то искомым оказывается верхний маршрут, а наиболее удаленным от A разбиением – $\overline{U}$. Действительно, поскольку на этом шаге каждая A-группа состоит из одного кластера, а первый из них, который мы пытались разбить на части, был самым большим, то попытки разбить другие кластеры не могут привести к более удовлетворительным результатам.

Если же (4) нарушается, то наибольшее значение d будет достигнуто при

$$s_* = \left[\frac{w + R_z - w_z - R_t + w_t}{2}\right]. \tag{5}$$

Здесь [.] – целая часть числа. В этой точке достигается минимум $f(s)$ на целых аргументах. Для того, чтобы в (2) максимум достигался именно в точке (5) требуется, чтобы вид функции $f(s)$ в этой точке определялся формулой (3), т.е. необходимо выполнение условий

$$w_{z-1} \geq s \geq w_z, \qquad w_{t-1} \geq w - s \geq w_t,$$

иначе говоря,

$$\max\{w_{z+1}, w - w_t\} \le s_* \le \min\{w_z, w - w_{t+1}\}. \tag{6}$$

При нарушении (6) изменения в кластерах на этом этапе также не производятся, несмотря на несоблюдение (4).

4. Основной алгоритм

Этот алгоритм путем направленного перебора позволит нам найти разбиение, наиболее удаленное от основного разбиения A, и, таким образом, вычислить $d^*(A)$, отдаление A от горизонта.

На старте алгоритма в качестве текущего разбиения C выберем само A.

- Шаг 1. Построим список $\Psi(C, A)$. Выберем в нем ту группу, в которой фигурирует наибольший по числу объектов кластер (это первый кластер в выбранной группе, поэтому достаточно сравнить первые кластеры в каждой из групп). Выбранную группу объявим текущей. Если кластеров одинакового размера несколько, и они расположены в разных группах, возьмем любую из групп, а в ней – произвольный из одинаковых по размеру больших кластеров. Число объектов в выбранном большом кластере примем за w.

- Шаг 2. Количество кластеров текущей группы обозначим q. Будем перебирать все пары индексов $(z, t), 1 \le t < z \le q + 1$ и для каждой из них проверять (4). Если для очередной пары условие выполнено – переходим к проверке следующей пары. Если нарушается – к шагу 3. После окончания проверки всех пар переходим к шагу 4.

- Шаг 3. Для найденной пары индексов z, t вычисляем s_* согласно (5) и проверяем условие (6). Если оно нарушено, продолжаем прерванный перебор пар, возвращаясь к шагу 2. Иначе разбиваем большой кластер на произвольные кластеры из $w - s_*$ и s_* объектов. Принимаем полученное разбиение за C, составляем список $\Psi(C, A)$. Переходим к шагу 4.

- Шаг 4. Есть ли в разбиении C еще кластеры, состоящие из w объектов? Если да, объявляем группу, содержащую такой кластер, текущей (совпадение новой текущей группы с предыдущей не исключается) и переходим к шагу 2. Иначе к шагу 5.

- Шаг 5. Менялось ли текущее разбиение C? Если да, к шагу 1, иначе алгоритм заканчивает работу с результатами B_C; $d^*(A) = d(A, B_C)$.

5. Обсуждение и выводы

Способ определения $d^*(A)$ для заданного основного разбиения A, который был описан, позволяет оценить степень отличия двух различных пар разбиений друг от друга в одном и том же масштабе единиц – некоторой величиной от 0 до 1. Это иногда может оказаться важным, поскольку возникающие в конкретном исследовании основные разбиения могут по-разному быть расположены в решетке на семействе Ξ. Тогда их отдаления от горизонта могут значительно отличаться, а следовательно, одинаковые численно значения $d(A, B)$ могут характеризовать отличия B от основного разбиения в разной степени.

Если же для каждого из основных разбиений A_1, A_2 вычислить их отдаления от горизонта $d^*(A_i)$, $i = 1, 2$, то вывод о том, что некоторое разбиение B менее отличается от A_1, чем от A_2, сделанный на основании справедливости неравенства

$$\frac{d(A_1, B)}{d^*(A_1)} < \frac{d(A_2, B)}{d^*(A_2)}$$

очевидно, является более корректным, чем если бы он был основан на простом сравнении расстояний.

К сожалению, задача полного описания класса всех разбиений, максимально удаленных от основного, или даже задача подсчета количества всех таких разбиений оказалась слишком сложной, хотя и понятно, что каждое такое разбиение может быть получено с помощью приведенного в работе алгоритма.

Список литературы

1. Yazhou Ren, Jingyu Pu, Zhimeng Yang et al. Deep Clustering: A Comprehensive Survey // arXiv:2210.04142v1 [cs.LG].

2. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura // Bulletin de la Société vaudoise des sciences naturelles. — 1901. — no. 37. — P. 547–579.

3. Kullback S., Leibler R.A. On information and sufficiency // Annals of Mathematical Statistics. — 1951. — Vol. 22(1). — P. 79–86.

4. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. — 1965. — Т. 163, № 4. — С. 845–848.

5. Fowlkes E.B., Mallows C.L. A Method for Comparing Two Hierarchical Clusterings // Journal of the American Statistical Association. — 1983. — no. 78(383). — P. 553–569.

6. Hubert L.J., Arabie P. Comparing partitions // Journal of Classification. — 1985. — no. 2(1). — P. 193–218.

7. Rousseeuw Peter J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis // Computational and Applied Mathematics. — 1987. — no. 20. — P. 53–65.

8. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. Серия: Информатика и вычислительная биология. — СПб. : Невский Диалект БВХ-Петербург, 2003. — 654 с.

9. Cohen W.W. A comparison of string distance metrics for name-matching tasks // KDD Workshop on Data Cleaning and Object Consolidation. — 2003. — Vol. 3. — P. 73–78.

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

11. Dronov S.V., Evdokimov E.A. Post-hoc cluster analysis of connection between forming characteristics // Model Assisted Statistics and Applications. — 2018. — Vol. 13, no. 2. — P. 183–192.

12. Дронов С.В. Анализ многомерных статистических данных: монография. — М.; Вологда : Инфра-Инженерия, 2025. — 308 с.