Модели кластеризации

Р. Р. АЙДАГУЛОВ
Московский государственный университет им. М. В. Ломоносова
e-mail: a_rust@bk.ru

С. Т. ГЛАВАЦКИЙ
Московский государственный университет им. М. В. Ломоносова, Московский центр фундаментальной и прикладной математики
e-mail: glavatsky_st@mail.ru

А. В. МИХАЛЁВ
Московский государственный университет им. М. В. Ломоносова, Московский центр фундаментальной и прикладной математики
e-mail: aamikhalev@mail.ru

УДК 004.021

Ключевые слова: кластер, алгоритм, плотность, метод осреднения.

Аннотация

Принято считать, что термин «кластеризация» (сгусток, пучок) был предложен математиком Р. Трионом. Впоследствии возник целый ряд терминов, которые рассматриваются как синонимы термина «кластерный анализ» или «автоматическая классификация». У кластерного анализа очень широкий спектр применения, его методы используются в медицине, химии, археологии, маркетинге, геологии и других дисциплинах. Кластеризация состоит в объединении в группы схожих объектов, и эта задача является одной из фундаментальных в области анализа данных. Обычно под кластеризацией понимается разбиение заданного множества точек некоторого метрического пространства на подмножества таким образом, чтобы близкие точки попали в одну группу, а дальние — в разные. Как мы покажем ниже, это требование является довольно противоречивым. Интуитивное разбиение «на глаз» использует соображение связности получаемых групп, исходя из плотности распределения точек. В данной работе предлагается метод кластеризации, основанный на этой идее.

Abstract

R. R. Aidagulov, S. T. Glavatsky, A. V. Mikhalev, Clustering models, Fundamentalnaya i prikladnaya matematika, vol. 23 (2020), no. 2, pp. 17—36.

It is generally accepted that the term “clusterization” (bunch, bundle) was offered by the mathematician R. Trion. Subsequently, a number of terms emerged that are considered synonymous with the term “cluster analysis” or “automatic classification.” Cluster analysis has a very wide range of applications, its methods are used in medicine, chemistry, archeology, marketing, geology, and other disciplines. Clustering consists in grouping similar objects into groups, and this problem is one of the fundamental problems in the field of data analysis. Usually, clustering means the partitioning of a given set of points of a certain metric space into subsets in such a way that close points fall into one group, and distant ones fall into different groups. As will be shown below, this requirement is rather contradictory. Intuitive partitioning “by eye” uses the connectivity of the resulting groups, based on the density of distribution of points. In this paper, we offer a method of clusterization based on this idea.

Фундаментальная и прикладная математика, 2020, том 23, № 2, с. 17—36.

1. Постановка задачи

Согласно [4] кластеризация — это процесс аналитического рассмотрения заданного множества точек и дальнейшей группировки точек в кластеры согласно некоторой метрике. При этом предполагается, что точки, попадающие в один кластер, должны быть расположены недалеко друг от друга, а попадающие в разные кластеры — далеко. Подчас исследователи под кластеризацией набора точек понимают разбиение этого множества (набора) на подмножества таким образом, чтобы близкие точки попали в одну группу, а дальние — в разные. Несложно понять, что такое требование противоречиво.

Действительно, пусть самые далёкие друг от друга точки x, y (для всех z, t справедливо $\rho(x,y) \geqslant \rho(z,t)$) могут быть соединены ε-путём $x_0 = x, \ldots, x_n = y$ так, что $\rho(x_i, x_{i+1}) \leqslant \varepsilon$. Наличие такой связи назовём ε-связностью. В многомерном (неодномерном) пространстве ε-связность самых далёких друг от друга точек не даёт однозначного ответа, попадут они в один кластер или нет. Это будет зависеть от геометрического расположения всего набора точек. В случае попадания самых отдалённых точек в один кластер не будет выполнено условие близости точек из одного кластера, а в случае их попадания в разные кластеры для некоторого $1 \leqslant i \leqslant n$ близкие точки x_i и x_{i+1} ($\rho(x_i, x_{i+1}) \leqslant \varepsilon$) попадут в разные кластеры. Таким образом, приведённое выше требование является противоречивым.

Пример. В конфигурации точек, приведённой на рис. 1, человек разобьёт множество точек на два кластера, проведя разделительную границу около точки D. В то же время большинство алгоритмов кластеризации включат точки C, D, C' в один кластер, используя принцип ε-связности. Такое применение ε-связности напоминает игру «Из мухи сделать слона»:

МУХА—МУНА—МИНА—ЛИНА—ЛИНН—ЛИОН—СИОН—СЛОН

и почти неприменимо к произвольному расположению точек.

Человек интуитивно группирует точки согласно плотности их распределения. Когда астрономы наблюдают дальние галактики в телескоп, они не видят отдельные звёзды и относят их к различным галактикам согласно распределению яркости (плотности).

Лишь в одномерном случае ε-связность характеризует плотность расположения точек на заданном отрезке. Уже в двумерном пространстве (как видно из рис. 1) это не так.

Плотность распределения существенно зависит от реальной размерности совокупности точек. Реальная размерность определяется из матрицы расстояний:

$$R = (\rho_{ij}),\ \rho_{ij} = \rho(x_i, x_j),\ i,j = 1,2,\ldots,n.$$

Прежде чем определять реальную размерность совокупности n точек с заданной матрицей расстояний, сделаем несколько замечаний о метриках. Метрика на множестве X задаётся функцией

$$\rho\colon X \times X \to \mathbb{R}_+,$$

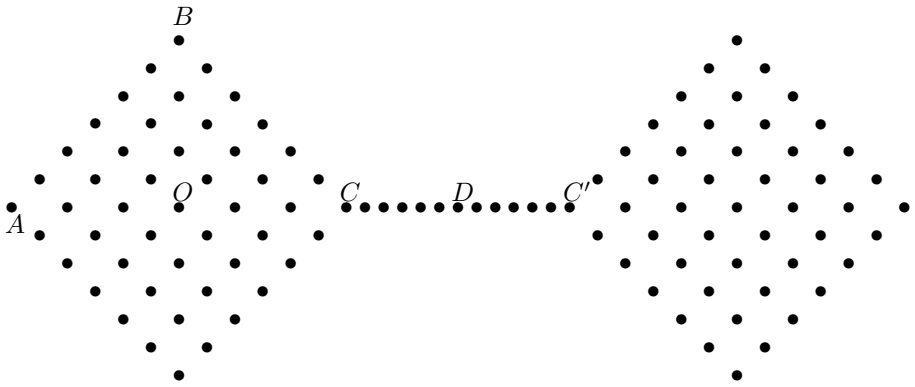

Рис. 1

удовлетворяющей следующим условиям:

1) $\rho(x,y)=0$ тогда и только тогда, когда $x=y$;
2) $\rho(x,y)=\rho(y,x)$;
3) $\rho(x,z)\leqslant\rho(x,y)+\rho(y,z)$.

Здесь самым содержательным для нас является последнее условие (неравенство треугольника); оно означает выпуклость шаров. В линейном нормированном пространстве отрезок, соединяющий две точки x, y, является пересечением всех шаров, содержащих эти точки. Соответственно, на метрические пространства без свойства линейности понятие выпуклости множества обобщается, исходя из такого определения.

Понятия выпуклости и вогнутости в метрических пространствах связаны с преобразованиями Лежандра, с вариационным исчислением. В частности, значение расстояния между двумя точками удовлетворяет вариационному принципу

$$\rho(x,y)=\inf\sum_{i=0}^{i=l}\rho(x_i,x_{i+1}),\quad x_0=x,\quad x_l=y.$$

Здесь inf берётся по всем путям с l звеньями, соединяющим точки x и y, при $l=1,2,\ldots$. Это соотношение может быть использовано для определения метрики по нечётко заданным оценкам экспертов. Например, пусть на множестве картин эксперты определили оценки схожести для всех пар картин (x_i,x_j), где $i,j=1,2,\ldots,n$, ставя в соответствие этим парам числа из отрезка $[0,1]$. Оценка 1 ставится в случае полной идентичности, 0 — в случае полного несходства (бесконечной удалённости). Пусть $S(x_i,\ x_j)$ — некоторое среднее значение для данной пары (по всем экспертам) выставленных оценок. Это симметричная функция от двух переменных. Определим вначале функцию

$$s(x_i,x_j)=\log_c S(x_i,x_j),\quad 0<c<1.$$

Эта функция близка к определению расстояния, однако может не удовлетворять неравенству треугольника даже в случае выполнения для оценок схожести каждого эксперта следующего условия:

$$S_*(x_i, x_k) \geqslant S_*(x_i, x_j)S_*(x_j, x_k).$$

Среднее значение (за исключением среднего геометрического) может нарушить выполнение неравенства треугольника. Однако мы можем (как выше) определить расстояние по вариационному принципу:

$$\rho(x, y) = \inf \sum_{i=0}^{i=l} s(x_i, x_{i+1}), \quad x_0 = x, \quad x_l = y.$$

Теперь все условия метричности выполнены. Такой (вариационный) подход работает и в случае гиперболических метрик, когда неравенство в третьей аксиоме метрики меняется на противоположное.

Далее мы изложим метод кластеризации, основанный на принципах плотностной связи между совокупностями точек. Плотность расположения зависит от реальной размерности набора точек. Величина этой размерности — положительное, но не обязательно целое число (наподобие размерности Хаусдорфа), она будет определена ниже. Отличие реальной размерности от размерности вмещающего евклидова пространства хорошо видно на следующем примере n точек на кривой Веронезе.

Пусть заданы n точек на кривой в d-мерном евклидовом пространстве $x_i = (x_{i1}, \ldots, x_{id})$, $i = 1, \ldots, n$, где $x_{ij} = \phi_j(y_i)$, $\phi_j(y) \in C^\infty$. Кривая Веронезе определяется вложением $\phi_j(x) = x^j$, $x \geqslant 0$, и характеризуется тем, что $n > d$ точек на кривой не могут быть изометрично вложены в евклидово пространство размерности меньше d. Эти n точек образуют конфигурацию с реальной размерностью, близкой к 1, но в то же время не вложимы изометрично в пространство размерности меньше d.

Прежде чем перейти к изложению нашего метода, рассмотрим ряд распространённых алгоритмов кластеризации.

2. Разделение алгоритмов на типы. Особенности большой размерности

Известные алгоритмы кластеризации по своей стратегии разделяются на два основных типа: иерархические и алгоритмы включения. Они, в свою очередь, разделяются на алгоритмы, использующие евклидову и неевклидову метрику. Разделяются также случаи, когда все данные умещаются в оперативной памяти и когда данные хранятся на диске из-за их чрезмерной громоздкости.

В алгоритмах кластеризации известен эффект «проклятия размерности», заключающийся во внесении столь качественных изменений высокими размерностями пространств точек, что это приводит к нежелательным эффектам в работе алгоритмов.

Один из таких эффектов сводится к тому, что при равномерном распределении точек почти все точки будут находится примерно на одном расстоянии друг от друга. Для обоснования этого заключения можно рассмотреть равномерное распределение точек в кубе [4, с. 261], хотя это верно и для других форм объёма, например для шара.

Другой особенностью [4, с. 262] является то, что радиус-векторы положений точек будут почти перпендикулярными. На самом деле это утверждение верно только для случая, когда за точку отсчёта (начало координат) берётся центральная точка, либо точка, близкая к центру рассматриваемого объёмного тела. Для случайно же выбираемой точки это утверждение не всегда верно. Например, если точки равномерно распределены в многомерном кубе и за точку отсчёта взять одну из вершин куба, то радиус-векторы точек будут иметь малые углы между собой (в среднем шестьдесят градусов), и они не будут почти перпендикулярными. Однако этот случай является исключительным.

Имеются и более важные эффекты «проклятия размерности». Основным из них является малость данных для получения статистических выводов из их распределения. В качестве примера рассмотрим 10^{10} точек в 100-мерном пространстве. Тогда на каждую размерность приходится всего $10^{0,1} \approx 1{,}2$ точек. По этой причине нельзя прийти к статистически обоснованным выводам относительно того, является ли полученное сгущение точек реальным или же это всего лишь эффект отсутствия достаточного количества данных для получения надёжных статистических выводов. Когда количество точек, приходящихся на одну размерность, достаточно большое, мы можем каждую точку окружить ε-окрестностью, где $\varepsilon = (1/2)R_0$ и R_0 — среднее расстояние между точками совокупности. Связные компоненты можно объявить кластерами. В пространстве же высокой размерности даже если точек достаточно много (больше 2^d), всё равно такой алгоритм не будет работать. Во-первых, объём шара половинного радиуса в 2^d раз меньше объёма шара единичного радиуса. Во-вторых, для покрытия единичного шара шарами половинного радиуса требуется не 2^d шаров, а гораздо больше (примерно квадрат этого числа) ввиду того, что при покрытии возникнет многократное пересечение.

На самом деле это верно не только для покрытия шарами половинного радиуса. Известная проблема Борсука состоит в определении минимального числа подобных, чуть меньшего диаметра, множеств, покрывающих исходное. Так, для покрытия куба со стороной 1 кубами со стороной 0,99999 требуется 2^d кубов, т. е. в общем случае для покрытия меньшими подобными множествами требуется экспоненциальное значение их количества.

Краткий вывод из этого анализа: при размерности больше 10 количество точек должно исчисляться триллионами и нормальный анализ положения точек невыполним на современных компьютерах за разумное время.

На самом же деле реальная размерность данных может быть небольшая. В приведённом примере кривой Веронезе при $n > d$ реальная размерность близка к 1. Реальная размерность определима приближённо даже в случае общего метрического пространства по распределению элементов матрицы расстояний $a_{ij} = \rho(x_i, x_j)$. Фрактальная размерность определяется через количество N_ε покрытия множества шарами радиуса ε как предел отношения $\log N_\varepsilon / \log \varepsilon$ при $\varepsilon \to 0$. Здесь N_ε — минимальное количество шаров радиуса ε, необходимых для полного покрытия множества. Похожим образом далее будет предложено определение реальной размерности данных.

3. Иерархические алгоритмы

Иерархические алгоритмы первоначально рассматривают каждую точку как один малый кластер и на очередном шаге сливают два близких кластера в один. Соответственно, выполнение алгоритма определяется ответами на следующие вопросы:

1) как и чем представляются кластеры?
2) как мы выбираем два кластера для слияния?
3) когда мы заканчиваем слияние кластеров?

При наличии ответов на эти вопросы алгоритм работает следующим образом: пока не наступило время остановиться, мы выбираем для объединения два кластера и сливаем их в один.

В случае евклидова пространства кластеры представляются средними значениями координат как центров кластеров. Соответственно, правило слияния определяется через расстояния между центрами. При этом в процессе выполнения алгоритма расстояния между кластерами могут вычисляться не только как расстояния между их центрами, здесь могут применяться даже неевклидовы расстояния, а также используется возможность учёта количества точек в кластере, чтобы не объединять два «тяжёлых» кластера (при объединении которых масса включённых точек превысит некий заданный допустимый предел).

Когда в процессе выполнения алгоритма динамически формулируются кластеры и их характеристики, то результат работы алгоритма зависит от порядка рассмотрения точек. Соответственно, при иной их нумерации может сформироваться иное объединение в кластеры. Например, при одной нумерации точек множество узлов квадратной сетки на плоскости будет разбиваться на горизонтальные полосы, а при другой нумерации тех же точек — на вертикальные полосы. Имеет место также некорректность решения в том смысле, что малое шевеление точек (иногда даже одной точки) может привести к получению существенно отличающейся конфигурации кластеров.

Так, на рис. 2 при определении расстояний между кластерами X_i, X_j через расстояния между их (мигрирующими в ходе построения) центрами невозможно разбить заданное множество точек правильно на невыпуклые кластеры.

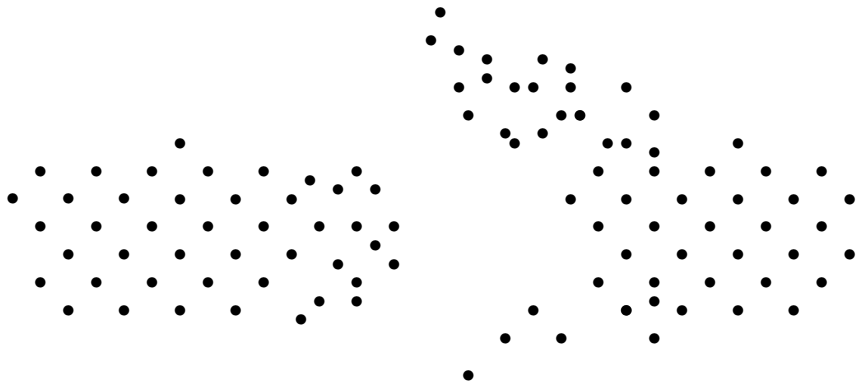

Рис. 2

Более точно, если имеется k центров и каждую точку пространства мы отнесём к ближайшему из этих k центров, то пространство разобьётся на выпуклые многогранники (некоторые из которых будут бесконечных размеров). Если центры не мигрируют в процессе выполнения алгоритма, то кластеры должны оставаться внутри этих выпуклых многогранников. Тем не менее множества точек, ограниченных этими многогранниками, не обязаны образовать выпуклые множества. Учитывая же миграцию центров, можно говорить только о тенденции, большей вероятности получения выпуклых кластеров, и о почти невыполнимом разбиении на кластеры, когда один кластер вклинивается во второй, как на рис. 2.

В алгоритме CURE сделана попытка устранить недостаток, связанный с невыпуклостью кластеров. Однако приведённое исправление через ε-связность, вообще говоря, не работает (см. рис. 1).

Часто при слиянии кластеров X_i,X_j растояние между ними определяется по формуле

$$S(X_i, X_j) = \inf(\rho(x, y) \mid x \in X_i,\ y \in X_j).$$

В таком случае метрика для кластеров динамически не меняется. Однако при слиянии кластеров, расстояние между которыми меньше ε, приводит к кластеризации компонент ε-связности. При этом возможно получение невыпуклых кластеров, но появляются недостатки, указанные на рис. 1.

4. Алгоритмы с включением

Основным (лучшим) типом алгоритма с включением считается k-значный алгоритм. Он работает в предположении евклидовости пространства и с заранее заданным количеством кластеров. Считается, что точное значение k может быть

получено повторением работы алгоритма при различных значениях k путём проб и ошибок и выбора наилучшего из значений по некоторому критерию. В нашей работе подробности работы такого критерия не рассматривается и фактически описание алгоритма рассматривается с заданным значением k.

Основой идеей алгоритма является распределение точек в ближайший кластер, заданный координатами своего центра. Если представить петли (окружности) радиуса r с центрами в центрах кластеров, то точку относят к тому из кластеров, чья петля захватит её первой при одновременном увеличении значения радиуса. После захвата точки в кластер данные о кластере уточняются: вес — количество точек N, пересчёт координат центра по формуле $x_c = (Nx'_c + x_p)/(N+1)$. Здесь x_p — координаты точки p, а x_c, x'_c — координаты центра (представляемые d числами — по значению размерности пространства) после и до добавления точки p в кластер. Заметим, в частности, что центры кластеров мигрируют по ходу работы алгоритма.

Отметим сразу следующие недостатки алгоритма.

1. В типичных ситуациях (равномерное распределение точек) миграция центров зависит от порядка подачи точек, т. е. при изменении порядка нумерации точек может измениться (а при равномерном распределении это будет наверняка) разбиение на кластеры.
2. При покрытии пространства шарами одинакового радиуса большая доля (по объёму) имеет кратность покрытия больше единицы. При большой размерности пространства (десять и больше) объём кратно покрытой части (с учётом их кратности) многократно превосходит исходный покрываемый объём. Это приводит к тому, что появляется много точек, которые можно отнести одновременно к разным кластерам. Соответственно, отнесение некоторых (немалого количества) точек к тому или иному кластеру будет зависеть и от их нумерации (в каком порядке точки поступают на обработку), и от малого шевеления положения точек.

Методы выбора первоначальных k центров не устраняют указанные недостатки. Отдельно рассматривается выбор количества кластеров k с ограничением на размеры кластеров по радиусу или диаметру (радиусом кластера называется максимальное расстояние от центра до удалённой точки из этого кластера, а диаметром — максимальное расстояние среди пар точек из заданного кластера).

Однако такой подход вызывает сомнение, поскольку он заставляет объединять в кластеры разрозненные куски и разрывает на части довольно слитные протяжённые кластеры. Например, если большое количество точек расположено в одном малом сгущении, то разбросанные в пространстве первоначальные центры могут оказаться примерно на одинаковом расстоянии от него. При этом каждый центр будет захватывать преимущественно точки из нашего сгущения, разбрасывая тем самым его точки в разные кластеры.

У этого метода есть ещё один недостаток для случая невыпуклых кластеров. Два невыпуклых плотных образования точек, существенно разделённые между

собой, могут иметь один и тот же или близкие центроиды (центры этих образований). Если расстояния между кластерами определяется через расстояния между центроидами, то эти образования сольются в один.

4.1. Алгоритм Брэдли—Файяда—Рейна

Этот алгоритм считается пригодным для случая большой размерности пространства и больших (не умещающихся в оперативной памяти) объёмов данных. Алгоритм относится к рассмотренному ранее k-значному типу алгоритмов. Его особенностью является масштабирование переменных, позволяющее кластеризовать протяжённые образования точек в один кластер, когда имеется протяжённость в направлении одной (или нескольких) оси. Но при этом не допускаются повороты осей, и поэтому даже повороты на малый угол слитных образований могут разорвать кластер на несколько частей.

Как k-значный, алгоритм начинается с выбора k точек в качестве кандидатов в центры образуемых кластеров.

Группы точек разделяются на следующие классы.

1. The Discard Set. Это группы близко расположенных точек, которые определённо целиком попадают в какой-то кластер. Соответственно, их данные не сохраняются в оперативной памяти, а сразу включаются в данные какого-то кластера в виде их количества, суммы координат и суммы квадратов координат.
2. The Compressed Set. Это множества близких друг другу точек, которые сами по себе достаточно удалены от центров имеющихся кластеров, чтобы можно было на данном этапе однозначно включить их в определённый кластер. Их данные хранятся в сжатой форме, как и у кластеров, в виде
$$\sum_i x_i^a, \quad a = 0, 1, 2.$$
Так как векторы x_i являются d-мерными, то требуется хранить $2d + 1$ чисел (для сумм нульмерных степеней — одно число — их количество).
3. The Retained Set. Это отдельные точки, которые далеки от объединения их как в плотные группы, так и для отнесения их к определённому кластеру. Их данные хранятся отдельно в оперативной памяти.

Уже по этой информации можно заключить, что алгоритм не будет работать эффективно в многомерных пространствах, так как в многомерных пространствах при типичном распределении (равномерном) бóльшая часть точек будет отнесена к третьей группе.

По сохранённым $2d + 1$ числам можно определить среднеквадратичное отклонение, или вариацию:

$$\sigma_j = \sqrt{\frac{\sum\limits_i x_{i,j}^2}{N} - \left(\frac{\sum_i x_{i,j}}{N}\right)^2}, \quad N = \sum_i x_{i,j}^0.$$

Алгоритм работает по следующей схеме.

1. Точки, достаточно близкие к центрам кластеров, сразу добавляются в соответствующие кластеры, «увеличивая»(сумма первых степеней за счёт отрицательности значений может и уменьшаться при добавлении) своими данными вычисляемые суммы, и их координаты удаляются из оперативной памяти как уже использованные.
2. Точки, не являющиеся достаточно близкими к центрам кластеров, попадают либо в третий класс (отдельных точек), либо добавляются во второй класс, в какую-то группу достаточно близких точек, для которых ещё не выбран кластер. Второй тип множеств образует временные миникластеры.
3. Далее временные миникластеры либо объединяются друг с другом, либо добавляются в известные кластеры.
4. Координаты точек, определённых как входящие в кластер или миникластер, убираются из оперативной памяти, предварительно их вклад учитывается в соответствующих суммах для кластеров или миникластеров.
5. В конце работы (после прохождения по всем точкам) оставшиеся нераспределёнными точки либо добавляются в ближайший к ним кластер, либо «забраковываются» как не вошедшие никуда.

Координаты точек подаются на рассмотрение из дисковой памяти частями и обрабатываются по указанной схеме.

При определении близости точки к кластеру пользуются расстоянием Махалонобиса, задаваемым формулой

$$\sqrt{\sum_{j=1}^{d}\left(\frac{x_{i,j}-c_j}{\sigma_j}\right)^2}.$$

Эта формула для определения расстояния более приемлема в задачах кластеризации из-за нормирования переменных. Однако при наличии существенных корреляций между переменными она становится столь же малоэффективной, как и без нормирования.

Последнее обстоятельство приводит к неустойчивому росту кластеров преимущественно в одном направлении.

Некорректность в смысле зависимости от нумерации точек и существенного изменения разбиения на кластеры при малом шевелении точек является характерной особенностью для всех алгоритмов, когда в самих данных нет чёткого разделения, либо есть разделение на кластеры, но их число не совпадает с требуемым в постановке задачи. Поэтому весьма желательно определить предварительно, какую геометрическую конфигурацию имеет множество заданных точек и можно ли его разбить на несколько отдельных конфигураций.

5. Одномерный случай

Как узнать по заданной матрице расстояний $a_{ij} = \rho(x_i, x_j)$, находятся точки в одномерном пространстве или в пространстве большей размерности?

Для статистической значимости ответа будем полагать, что количество точек n — достаточно большое число. Пусть $D = \rho(x_1, \ldots, x_n)$ — диаметр множества точек. Если точки можно пронумеровать так, чтобы сумма длин

$$L = \sum_{i=1}^{n-1} \rho(x_i, x_{i+1})$$

совпала с D, то мы можем утверждать, что точки находятся на одномерной прямой. Действительно, в этом случае существует изометрия (отображение, сохраняющее взаимные расстояния) точек множества на точки прямой линии. В случае когда L ненамного (например, не более чем в 2 раза) превосходит D, также можно утверждать, что точки множества лежат на одномерной кривой. При этом сама кривая изометрически может быть вложена в евклидово пространство только очень большой размерности.

Точки могут находиться и на нескольких кривых с длинами L_i так, что сумма их длин ненамного превосходит диаметр. Одной из постановок задачи разбиения на кластеры в одномерном случае является разбиение множества точек на несколько кривых так, чтобы сумма длин L_i была минимальной.

Когда точки лежат на прямой линии, этот вопрос решается просто. Нужно проводить границу между кластерами там, где наличествуют большие расстояния между соседними точками.

Для определения того, какие расстояния между соседними точками следует считать большими, рассмотрим задачу случайного равномерного разделения отрезка длины L на n частей. Для этого с помощью генератора случайных чисел, распределённых равномерно на отрезке $[0, 1]$, выработаем $n - 1$ чисел y_i и разрежем отрезок $[0, L]$ в точках с координатами $x_i = Ly_i$. Вычислим теперь математическое ожидание длины самого большого куска l_1, следующего по длине — l_2 и т. д., k-го по длине — l_k. Пусть $l = L/n$ — средняя длина полученных кусков. Величины l_i выразим в безразмерных величинах, относя их к средней длине l.

При $n = 2$ задача решается несложно. С вероятностью $1/2$ имеем $y_1 < 1/2$, и средняя длина меньшего отрезка равна

$$\int_0^{1/2} y\, dy = \frac{1}{4}.$$

Случай $y_1 > 1/2$ рассматривается аналогично. Таким образом, длина максимального куска (при масштабе длины l) есть $l_1 = 3/2 = 1 + 1/2$, длина следующего отрезка — $1/2$. В общем случае имеет место следующее утверждение.

Теорема 1. *Математические ожидания длин отрезков при случайном равномерном делении отрезка на n частей выражаются формулами*

$$\begin{aligned}
\bar{l}_1 &= 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} = H_n,\\
\bar{l}_2 &= \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} = H_n - H_1,\\
\bar{l}_3 &= \frac{1}{3} + \ldots + \frac{1}{n} = H_n - H_2,\\
&\ldots\\
\bar{l}_k &= \frac{1}{k} + \ldots + \frac{1}{n} = H_n - H_{k-1}.
\end{aligned}$$

Доказательство. Обозначим приращения длин интервалов через

$$r_n = l_n, \quad r_{n-1} = l_{n-1} - l_n, \ldots, \quad r_2 = l_2 - l_3, \quad r_1 = l_1 - l_2.$$

Последнее приращение (l_1) можно определить из соотношения

$$r_1 = M - 2r_2 - 3r_3 - \ldots - nr_n, \quad M = \frac{L}{l} = n.$$

Математическое ожидание длины l_k выражается формулой

$$\bar{l}_k = \frac{\int\limits_0^{M_n/n}\Big(\int\limits_0^{M_{n-1}/(n-1)}\Big(\ldots\Big(\int\limits_0^{M_2/2}(r_n + r_{n-1} + \ldots + r_k)\,dr_2\Big)\ldots\Big)\,dr_{n-1}\Big)\,dr_n}{\int\limits_0^{M_n/n}\Big(\int\limits_0^{M_{n-1}/(n-1)}\Big(\ldots\Big(\int\limits_0^{M_2/2}dr_2\Big)\ldots\Big)\,dr_{n-1}\Big)dr_n}.$$

Здесь

$$M_n = M = n, \quad M_{n-1} = M_n - nr_n, \ldots, \quad M_{k-1} = M_k - kr_k, \ldots .$$

Отсюда индукцией по n получаем $\bar{l}_n = \bar{r}_n = l/n$, соответственно, $\bar{l}_{n-1} = 1/(n-1) + 1/n, \ldots$. Это доказывает формулы, указанные в теореме. □

Видно, что сумма всех длин равна $nl = L$. По этому разделению можно судить, оставлять все точки на отрезке в одном кластере или отрезок с точками нужно разделять на несколько кластеров в тех местах, где расстояния между соседними точками существенно превосходят lH_n.

6. Размерность пространства

Размерность пространства является локальной топологической характеристикой. Для топологических пространств размерность пространства определил академик П. С. Александров через пересечения открытого покрытия. Для наших целей удобнее использовать размерность по Хаусдорфу, определяемую как степень d роста величины $O(\varepsilon^{-d})$ минимального количества шаров радиуса ε,

необходимых для покрытия нашего множества, при $\varepsilon \to 0$. Так как у нас конечное число точек n, то при любом ε необходимое количество шаров не превосходит n. Тем не менее нужную размерность можно определить через тангенс угла в линейной аппроксимации логарифма от количества точек в зависимости от (при увеличении) логарифма радиуса. Для этого упорядочим $n(n-1)/2$ ненулевых расстояний между n точками по возрастанию:

$$r_1 \leqslant r_2 \leqslant \ldots \leqslant r_{n(n-1)/2}.$$

На расстоянии не более r_i от некоторой точки в среднем находится $2i/n$ точек без учёта самой точки. Пусть $y_i = \log(2i/n)$, $x_i = \log(r_i)$. Находим наилучшее приближение (аппроксимацию) $y_i = dx_i + c$ или $x_i = (y_i - c)/d$. На значение d не влияет ни основание логарифма, ни постоянный коэффициент $\log(2/n)$ (уходит в определение величины c). Для уменьшения влияния граничных элементов в вычислении корреляции оставим только определённую часть $1 \leqslant m \leqslant n$ значений — только тех, где $r_i < r_{\max}/2$, $1 \leqslant i \leqslant m$. Вычисляя наилучшую линейную аппроксимацию, получаем реальную размерность d и коэффициент пропорциональности $\exp(c)$. Размерность пространства d выражается формулой

$$d = \frac{\sum\limits_i (\log i)^2 - \frac{1}{m}\Big(\sum\limits_i \log i\Big)^2}{\sum\limits_i \log(r_i)\log i - \frac{1}{m}\sum\limits_i \log(i)\log(r_i)}.$$

Здесь m — длина суммирования (в сумме участвуют только первые m членов из $\{r_i\}$).

6.1. Локальные свойства распределения точек

Размерность метрического подпространства дискретно распределённых точек, вообще говоря, является локальным свойством, как и плотность. Соответственно, она должна определяться через локальное распределение.

Будем предполагать, что пространство локально ведёт себя как линейное (вложено в линейное пространство), т. е. точки задаются своими координатами $x_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$, $x_{i_j} \in \mathbb{R}$, и их можно складывать и вычитать как векторы.

В линейном пространстве понятие объёма определяется независимо от наличия метрики. Объём представляет собой сумму объёмов элементарных параллелепипедов, и это понятие переносится даже на нелинейные пространства путём интегрирования бесконечно малых параллелепипедов, образованных векторами из касательного пространства. Для определения объёма не требуется задания метрики.

Пусть V — n-мерное линейное пространство. Объём представляет собой отображение $V^n \to \mathbb{R}$, отличное от тождественно нулевого, из тензорной степени V^n в поле действительных чисел. При этом каждому набору из n векторов пространства V ставится в соответствие число $V(v_1, v_2, \ldots, v_n)$, такое что выполнены следующие условия:

1) $V(v_1, \ldots, av_i, \ldots, v_n) = aV(v_1, v_2, \ldots, v_n)$ (умножение — увеличение или уменьшение одного вектора в a раз — приводит к увеличению или уменьшению объёма в a раз; это относится и к отрицательным значениям числа a, когда объём измеряется с учётом ориентации, характеризуемой её знаком);
2) $V(v_1, \ldots, v_i, \ldots, v_i + v_j, \ldots, v_n) = V(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_n)$ (сложение одного вектора из набора с другим не меняет значение объёма).

Из этих свойств следует, что объём является кососимметричной билинейной формой и вычисляется однозначно, если задано его значение на каком-нибудь базисе. При этом равенство нулю объёма не зависит от выбора начального базиса для масштабирования. Отношение объёмов двух наборов векторов также не зависит от выбора базиса и является инвариантом относительно произвольного выбора функции объёма и вычисляется как отношение определителей матриц, составленных из значений координат (в произвольном базисе).

Это позволяет определить реальную локальную размерность пространства данных. Посчитаем миноры порядка k, составленные от значений k переменных для k близких точек. Здесь мы не увеличиваем размерность миноров добавлением единичных строк/столбцов, как это делается в определителе Кэли—Менгера. Соответственно, значения координат берём относительно выбранного центра шара. Пусть M_k — максимальное (по абсолютной величине) значение минора k-го порядка. При этом M_k есть k-мерный объём параллелепипеда с вершинами в центре и на k точках из шара.

В каждой точке можно выбрать шар с центром в этой точке, содержащий $n \ll N$ точек. Определим минорную плотность k-го порядка в этой точке как максимальное по абсолютной величине значение среди всех миноров k-го порядка. Если

$$\frac{|M_{d+1}|}{|M_d|^{1+1/d}} \ll 1,$$

то локально наша фигура имеет размерность d. Здесь следует обратить внимание на возможное резкое падение плотности. В минорах порядка $d+1$ максимальный минор порядка d, вообще говоря, не совпадает с максимальным минором из d векторов, тем не менее при выполнении указанного свойства полученные наборы из d векторов образует «почти» одно и то же пространство, которое можно назвать касательным пространством в данной окрестности. Это пространство «почти» не зависит от избранной вначале метрики для выбора точек из небольшого шара, т. е. метрика здесь используется только в качественном отношении, определяя близость, как в топологии. Размерность пространства всегда получается не менее 1.

Из-за значительного размера окрестности для реальных данных и искривления формы величина $\sqrt[k]{M_k}$, скорее всего, будет убывать постепенно в зависимости от k. Тогда в качестве локальной размерности можно взять такое значение d, что

$$\frac{M_{d+1}}{|M_1|^{d+1}} < \varepsilon.$$

При вычислении миноров второго и третьего порядков в отношении добавляется нормализация через степень, учитывающая разные размерности объёмов (не делим кубические сантиметры на квадратные сантиметры). Справа у критерия малости должна стоять константа, зависящая от количества точек и, возможно, в некоторой слабой форме от радиуса и размерности. Однако такое уточнение лучше делать для глобального значения, получаемого интегрированием. Так как мы пока не имеем хорошей метрики и в нашей будущей метрике точки должны распределяться равномерно, мы можем суммировать по нашим точкам с вычисленными локальными значениями минорной плотности. Когда сумма таких отношений будет мала, это будет означать, что мы нашли глобальную размерность пространства данных.

Максимальные миноры порядка d определяют локально карту из тех переменных, которые вошли в этот минор. Так как на пересечениях карт миноры порядка $d+1$ малы, то переменные из одной карты имеют хорошую функциональную зависимость от переменных другой карты и обратно. Взяв все участвующие переменные, мы получим заодно локальное погружение без пересечений, т. е. нормальное вложение в евклидово пространство. Можно определить финслерову метрику, убирая зависимости между оставшимися переменными. Это сложная процедура, и для простоты мы ограничимся только локальными квадратичными метриками, сделав наше многообразие римановым. Для этого вычислим локальные корреляции между оставшимися переменными и определим локальную метрику g_{ij} как матрицу, обратную к матрице локальных корреляций (с единицами на главной диагонали). Далее вышеописанным образом (с помощью усреднения с гауссовыми весами) распространим метрику на всю фигуру из точек. При склеивании карт останутся пустоты, связные компоненты образуют кластеры, а точки будут равномерно (с некоторой точностью пропорционально объёму) распределены в полученной фигуре.

Локализация, описанная выше, позволяет существенно сократить количество операций в алгоритме, когда количество точек N велико. Тогда требуется анализировать не все $N(N-1)/2$ взаимных расстояний, а только $sn(n-1)/2$, где s — количество карт, а n — количество точек в одной карте.

В этом заключаются основные идеи построения статистической геометрии, позволяющей не только разбить множество точек на кластеры, но и определить геометрическую форму множества дискретных данных.

7. Метод осреднения

Метод осреднения используется в решении задач широкого круга областей естествознания, связанных с изучением свойств неоднородных сред. Понятие нелинейного осреднения (называемого сейчас «осреднением по Колмогорову») было введено А. Н. Колмогоровым в [2]. Далее этот метод был развит в трудах его последователей в широком спектре приложений в механике, квантовой механике, экономике и др. (см. [1,5,6]).

Однако указанное осреднение относится к глобальному типу, определяющему среднее значение для набора точек типа центра масс. Здесь если точки расположены в аффинном пространстве, то центр набора можно найти линейным осреднением. Если же точки расположены на некоторой карте многообразия M, то центр набора можно получить отображением $\phi\colon M \to \mathbb{R}^n$ точек на карту, далее найти центр x_c набора отображений точек на карте и в итоге получить центр набора точек в многообразии как прообраз $\phi^{-1}(x_c)$. Но для задачи вычисления плотности расположения точек в произвольной области такое глобальное осреднение не подходит: плотность по определению получается не делением вычисляемой величины на количество точек (как в случае глобального среднего), а наоборот, отношением количества точек к содержащему их объёму.

Плотность в точке x определяется локальным осреднением следующим образом. Пусть точка x_i включена в шар единичного объёма с центром в точке x с вероятностью $P(x, x_i)$. Тогда среднее количество (математическое ожидание) точек, входящих в этот объём, равно

$$\sum_i P(x, x_i).$$

Если

$$\int_y P(x, y)dy = 1,$$

то полученное среднее и есть плотность (среднее количество точек в шаре единичного объёма с центром в точке x). Для определения областей сгущения плотности требуется именно такое локальное осреднение.

Глобальное осреднение может иметь веса для точек. Однако эти веса не зависят от положения. Локальное осреднение всегда имеет характер осреднения относительно некоторого заданного положения x и имеет веса точек, зависящие от положения относительно этой точки (от расстояния до этой точки). Характерное отличие между локальным и глобальным осреднением проявляется при вычислении средней скорости в физике. Глобальная средняя скорость определяется по формуле

$$v = \frac{\sum\limits_i m_i v_i(x_i)}{\sum\limits_i m_i}.$$

Здесь осредняемая функция скорости входит с весами m_i — массами точек. При локальном осреднении получаем иную среднюю скорость в точке с положением x:

$$v = \frac{\sum\limits_i m_i v_i(x_i) P(x, x_i)}{\sum\limits_i m_i P(x, x_i)}.$$

В физике законы не должны зависеть от системы координат. Это создаёт определённые ограничения относительно выбора осреднения. Осреднение, ин-

вариантное относительно сдвигов (выбора начала координатной системы в аффинном пространстве), называется осреднением по Фридрихсу. В этом случае веса должны зависеть только от положения $P(x, x_i) = K(x - x_i)$. В группе Галилея имеются и повороты системы координат. Осреднение, инвариантное относительно сдвигов и поворотов, называется осреднением по Гауссу. В этом случае ядро осреднения зависит только от расстояния $P(x, x_i) = K(|x - x_i|)$. В группе Галилея имеется также переход к другой инерциальной системе координат. Чтобы закон сохранения импульсов был верен и для осреднённых величин, требуется определить средние скорости как среднемассовые — с весами m_i. В физике Ньютона скорости определяются как точки аффинного пространства, выбор инерциальной системы соответствует выбору начала координат в аффинном пространстве. При этом закон сохранения импульсов будет выполняться именно при линейном осреднении с весами. В теории относительности скорости не являются элементами аффинного пространства. Они скорее являются элементами группы Ли, алгеброй Ли которой является алгебра ускорений с весами-силами. Поэтому для инвариантности законов относительно перехода к другой инерциальной системе мы должны отобразить пространство скоростей в аффинное пространство $\ln\colon V \to A$ (в алгебру Ли), найти там среднее a_c и получить среднюю скорость обратным отображением $v_c = \exp(a_c)$. В двумерном пространстве-времени отображение $\ln$ соответствует arcth и среднее находится по формуле

$$v_c = \operatorname{th}\left(\frac{\sum\limits_i m_i \operatorname{arcth}(v_i)}{\sum\limits_i m_i}\right).$$

В многомерном пространстве группа Лоренца некоммутативна, и поэтому в геометрии Минковского нельзя корректно определить среднюю скорость. Она корректно определяется только в пространстве Бервальда—Моора.

Для простоты изложения мы ограничимся рассмотрением случая, когда данные (точки) расположены в евклидовом пространстве. Зададим оператор осреднения на функциях распределения $a(x)$, $x \in \mathbb{R}^n$ по формуле

$$\bar{a}(x) = \int P(x - x')a(x')\,dx'.$$

Взяв в качестве $P(x)$ бесконечно дифференцируемую функцию, получим, что функция распределения после осреднения станет бесконечно дифференцируемой. Взяв неотрицательную функцию $P(x)$, интеграл от которой равен 1, получим, что оператор осреднения как оператор из L_1 в L_1 имеет норму 1 и неотрицательное распределение переводит в неотрицательное. Оператор осреднения перестановочен с трансляциями (сдвигами аргумента). Если функция $P(x)$ сферически симметрична, то оператор осреднения коммутирует с вращениями распределений. В задачах механики вращения включаются в группу симмет-

рий, для сохранения симметрии для осреднённых уравнений мы будем также использовать сферически симметричные ядра $P(x)$.

Пусть $P(x)$ — ядро осреднения (оно неотрицательно, и интеграл от него равен 1), тогда

$$\frac{1}{a^d} P\left(\frac{x}{a}\right)$$

также является ядром осреднения. Такое ядро назовём подобным. Было бы желательно, чтобы осреднение осреднённой функции ничего не меняло. Однако такое невозможно, кроме случая тождественного оператора, соответствующего ядру $P(x) = \delta(x)$. Достижимым является такое свойство, когда двойное осреднение эквивалентно одному осреднению с подобным ядром. Такое ядро единственное с точностью до подобия. Это осреднение Гаусса с ядром $P(x) = \exp(-\pi x^2)$. Для этого ядра радиус осреднения положим равным 1. Подобные осреднения

$$\frac{1}{R^d} \exp\left(-\pi \frac{x^2}{R^2}\right)$$

имеют радиус осреднения R, случаю $R = 0$ (точнее, пределу при $R \to 0$) соответствует тривиальное осреднение с ядром $P(x) = \delta(x)$. Чем меньше R, тем более осциллирующей получается осреднённая функция плотности, бывшей до осреднения функцией

$$\sum_i \delta(x - x_i).$$

Заметим, что если сделать осреднение Гаусса с радиусом осреднения R_1, а потом осреднение полученного осреднённого распределения с радиусом осреднения R_2, то получим в точности результат однократного осреднения с радиусом осреднения $\sqrt{R_1^2 + R_2^2}$.

Радиус осреднения для N точек в d-мерном пространстве следует выбирать так, чтобы, с одной стороны, среднее количество точек n в одном шаре такого радиуса было намного больше, чем $(\ln N)^d$, а с другой стороны — намного меньше, чем N^ε. Это свойство выполняется для часто встречающейся функции [3]

$$n = L(N) = \exp\bigl(\sqrt{\ln N \ln\ln N}\,\bigr).$$

Метод осреднения заключается в осреднении множества точек с функцией плотности распределения

$$\sum_i \delta(x - x_i).$$

Выбирая далее срезы множества точек по определённому уровню плотности, мы получаем разбиение на кластеры. Этот метод свободен от таких отмеченных выше недостатков, как зависимость от нумерации точек и существенное изменение разбиения на кластеры при малом изменении позиции даже одной точки.

8. Алгоритм Dbscan и сравнение его с методом осреднения

Хорошо известен алгоритм Dbscan, также позиционируемый на плотности распределения точек. В этом алгоритме достаточными для создания групп связности по плотности выделяются точки, в ε-окрестности которых имеется m точек. При этом результат кластеризации сильно зависит от параметров (m, ε). Рекомендуемые значения этих параметров не учитывают размерность множества и не соответствуют росту количества точек N. В этом алгоритме ε играет роль радиуса R (в нашем представлении), параметр m определяет пороговую плотность с точностью до постоянного множителя, зависящего от количества точек N и размерности d. Вычисление плотности соответствует методу осреднения с ядром

$$P(x) = \begin{cases} 1, & |x| < \varepsilon, \\ 0, & |x| \geqslant \varepsilon. \end{cases}$$

В задачах механики [7] такое осреднение с разрывным ядром при выводе осреднённых уравнений приводит к дополнительному предположению о совпадении среднего значения по сфере со средним по шару. Тогда одинаковый вес при вычислении плотности точек вплоть до удаления на расстояние $R = \varepsilon$ приводит к ложному объединению в один кластер в ситуации на рис. 3.

В случае гауссовского ядра веса точек, находящихся на расстоянии $R/2$, будут меньше $0,46$ от максимального веса, а у точек на расстоянии R — уже только $0{,}043$, и они вносят совсем малый вклад. А вот алгоритм Dbscan может

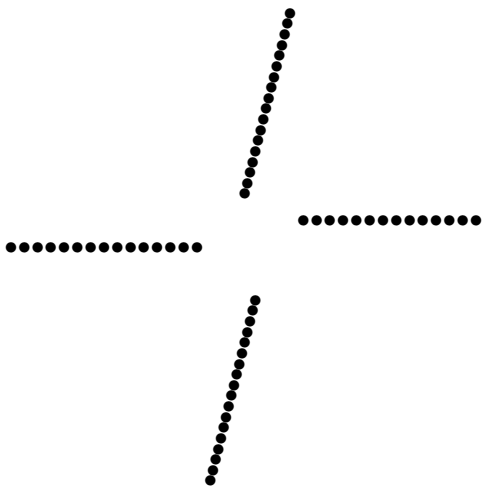

Рис. 3

не разделять множество точек на два кластера даже в ситуации на рис. 1, если длина CD будет порядка $R/2$. В нашем же методе плотность в точке D падает более существенно и происходит разбиение на два кластера (на четыре на рис. 3).

Работа второго и третьего авторов финансово поддержана грантом «Структурная теория и комбинаторно-логические методы в теории алгебраических систем» Московского центра фундаментальной и прикладной математики.

Литература

[1] Бахвалов Н. С., Панасенко Г. П. Осреднение процессов в периодических средах. Математические задачи механики композиционных материалов. — М.: Наука, 1984.

[2] Колмогоров А. Н. Избранные труды. Математика и механика. — М.: Наука, 1985. — С. 136—137.

[3] Крендалл Р., Померанс К. Простые числа. Криптографические и вычислительные аспекты. — М.: УРСС, 2011.

[4] Лесковец Ю., Раджараман А., Ульман Дж. Анализ больших наборов данных. — М.: ДМК, 2016.

[5] Маслов В. П. О способе осреднения для большого числа кластеров. Фазовые переходы // Теор. и матем. физ. — 2000. — Т. 125, № 2. — С. 297—314.

[6] Маслов В. П. Аксиомы нелинейного осреднения в финансовой математике и динамика курса акций // Теор. вероятн. и её примен. — 2003. — Т. 48, № 4. — С. 800—810.

[7] Нигматулин Р. И. Основы механики гетерогенных сред. — М.: Наука, 1978.