УДК 004.67

Айдагулов Р.Р., Главацкий С.Т.

Московский государственный университет имени М.В. Ломоносова, г. Москва, Россия

АЛГОРИТМ ОСРЕДНЕНИЯ В КЛАСТЕРИЗАЦИИ ДАННЫХ*

Аннотация

Кластеризация данных состоит в объединении в группы схожих элементов, и эта задача является одной из фундаментальных в области анализа данных. Обычно под кластеризацией понимается разбиение заданного множества точек на подмножества так, чтобы близкие точки попали в одну группу, а дальние — в разные. Это требование является достаточно противоречивым. Интуитивное разбиение «на глаз» использует соображение связности получаемых групп, исходя из плотности распределения точек. Известный алгоритм DBSCAN использует два параметра — радиус окрестности ϵ и количество соседей m. Принципы выбора значений этих параметров остаются загадочными, а без строгого их определения сложно понять границы применимости алгоритма. В данной статье предложен новый метод кластеризации больших данных, также основанный на идее рассмотрения плотности распределения точек заданного множества в многомерном пространстве. Поскольку плотность зависит от реальной размерности набора точек, то предложен способ определения этой размерности (наподобие размерности Хаусдорфа). Предложена математически обоснованная оценка величины радиуса осреднения (аналог величины ϵ в DBSCAN). На ряде примеров показана высокая эффективность созданного алгоритма.

Ключевые слова

Большие данные, кластеризация, плотность распределения, размерность множества, метод осреднения.

Aidagulov R.R., Glavatsky S.T.

Lomonosov Moscow State University, Moscow, Russia

ALGORITHM OF AVERAGING IN DATA CLUSTERING

Abstract

Data clustering consists in grouping together similar elements, and this task is one of the fundamental in the field of data analysis. Usually, clustering is considered as the partition of a given set of points into subsets so that close points fall into one group, and distant ones into different groups. This requirement is quite contradictory. Intuitive partitioning "by eye" uses the connectivity of the resulting groups, based on the density of distribution of points. The known DBSCAN algorithm uses two parameters – the radius ε of the neighborhood and the number m of neighbors. The principles of choosing the values of these parameters remain mysterious, and without strict definition it is difficult to understand the applicability frames of the algorithm. In this paper, we propose a new method for clustering big data, also based on the idea of considering the density of distribution of points in a multidimensional space. Since the density depends on the real dimension of the set of points, a method is proposed for determining this dimension (similar to the Hausdorff dimension). A mathematically well-founded estimate of the averaging radius is proposed (an analogue of ε in DBSCAN). A number of examples show the high efficiency of the created algorithm.

Keywords

Big data, clusterization, the density of distribution, the dimension of the set, the method of averaging.

* **Труды II Международной научной конференции «Конвергентные когнитивно-информационные технологии» (Convergent'2017), Москва, 24-26 ноября, 2017**

Proceedings of the II International scientific conference "Convergent cognitive information technologies" (Convergent'2017), Moscow, Russia, November 24-26, 2017

Введение

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 достаточно противоречивым. Интуитивное разбиение «на глаз» использует соображение связности получаемых групп, исходя из плотности распределения точек. В данной работе предлагается алгоритм, основанный на этой идее.

Постановка задачи

Согласно [1], 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 удаленные друг от друга точки x, y $(\forall z,t\ \rho(x,y) \geq \rho(z,t))$ могут быть соединены ε-путем $x_0 = x, \ldots, x_n = y$ так, что $\rho(x_i, x_{i+1}) \leq \varepsilon$. Наличие такой связи назовем ε-связностью. В многомерном (не одномерном) пространстве ε-связность самых удаленных точек не дает однозначного ответа, попадут они в один кластер или нет. Это будет зависеть от геометрического расположения всего набора точек. В случае попадания самых дальних точек в один кластер не будет выполнено условие близости точек из одного кластера, а в случае их попадания в разные кластеры для некоторого $1 \leq i \leq n$ близкие точки $\rho(x_i, x_{i+1}) \leq \varepsilon$ попадут в разные кластеры. Таким образом, приведенное выше требование является противоречивым.

Пример. В конфигурации точек, приведенной на Рисунке 1, человек разобьет множество точек на 2 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 звезды, 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. 2) это не так.

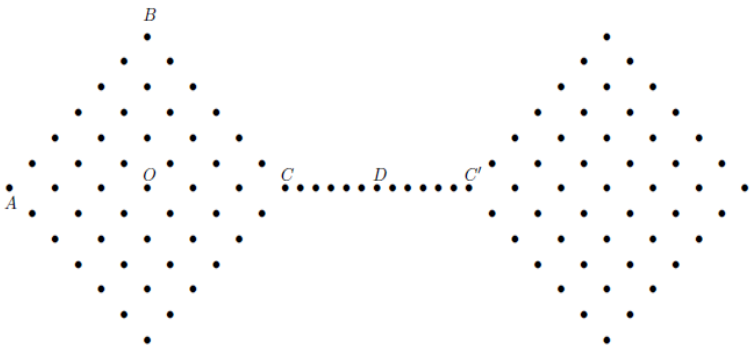

Рис. 1

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

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

В метрическом пространстве неравенство треугольника в свойствах метрики означает выпуклость

шаров. В линейном нормированном пространстве отрезок, соединяющий две точки *x, y*, является пересечением всех шаров, содержащих эти точки. Соответственно, понятие выпуклости множества обобщается на метрические пространства без свойства линейности, исходя из такого определения.

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

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

Здесь *inf* берется по всем путям с *l* звеньями, соединяющим точки *x* и *y*, при *l* = 1,2,...

Это соотношение может быть использовано для определения метрики по нечетко 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,...,n, сопоставляя этим парам числа из отрезка [0, 1]. Оценка 1 ставится в случае полной идентичности и 0 — в случае полного несходства (бесконечной удаленности). Пусть $S(x_i, x_j)$ — некоторое среднее значение (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. Среднее значение (за исключением среднего геометрического) может нарушить выполнение этого неравенства. Однако мы можем (как выше) определить расстояние по вариационному принципу:

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

Теперь все условия метричности выполнены.

Далее мы построим алгоритм кластеризации 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 $x_i = (x_{i1},...,x_{id}), \quad i = 1,...,n,$ где $x_{ij} = \phi_j(y_i), \quad \phi_j(y_i) \in C^\infty$. Кривая Веронезе определяется вложением $\phi_j(x) = x^j, \, x \geq 0,$ 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*.

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 $D = \rho(x_1,...,x_n)$ — диаметр множества точек. Если точки можно пронумеровать так, чтобы сумма длин $L = \sum_{i=1}^{n-l} \rho(x_i, x_{i+1})$ 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*, то 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 0,5 имеем $y_1 < 0{,}5$, и средняя длина меньшего

отрезка $\int_0^{0,5} ydy = 0{,}25$. Случай $y_1 > 0.5$ рассматривается аналогично. Таким образом, длина максимального куска (при масштабе длины *l*) есть $l_1 = \frac{3}{2} = 1 + \frac{1}{2}$, длина следующего отрезка — $\frac{1}{2}$. В общем случае имеет место следующее утверждение.

Теорема. Математические ожидания длин отрезков при случайном равномерном делении отрезка [0, *L*] на *n* частей выражаются формулами:

$$\bar{l}_1 = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} = H_n,$$

$$\bar{l}_2 = \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} = H_n - H_1,$$

$$\bar{l}_3 = \frac{1}{3} + \ldots + \frac{1}{n} = H_n - H_2,$$

$$\ldots$$

$$\bar{l}_k = \frac{1}{k} + \ldots + \frac{1}{n} = H_n - H_{k-1}.$$

Доказательство. Обозначим приращения длин интервалов через:

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

Последнее (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:

$$\bar{l}_k = \frac{\int_0^{M_n/n}(\int_0^{M_{n-1}/(n-1)}(\cdots\int_0^{M_2/2}(r_n + r_{n-1} + \ldots + r_k)dr_2\ldots)dr_{n-1})dr_n}{\int_0^{M_n/n}(\int_0^{M_{n-1}/(n-1)}(\cdots\int_0^{M_2/2}dr_2\ldots)dr_{n-1})dr_n}.$$

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

Отсюда, индукцией по *n* получаем $\bar{l}_n = \bar{r}_n = \frac{l}{n}$, соответственно $\bar{l}_{n-1} = \frac{1}{n-1} + \frac{1}{n}$,... Это доказывает формулы, указанные в теореме. Видно, что сумма всех длин равна *nl = L.* По этому разделению можно судить, оставлять ли 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.

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, необходимого 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: $r_1 \le r_2 \le \ldots \le r_{n(n-1)/2}$. На расстоянии не более r_i от некоторой точки в среднем находится *2i/n* точек без учета самой точки. Пусть $y_i = \log(2i/n), \quad x_i = \log(r_i)$. Находим наилучшее приближение (аппроксимацию) $y_i = dx_i + c$ или $x_i = \frac{y_i - c}{d}$. 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 определенную часть $1 \le m \le n$ значений — только тех, где $r_i < r_{\max}/2, \quad 1 \le i \le m$. Вычисляя наилучшую линейную аппроксимацию, получим реальную размерность *d* и коэффициент пропорциональности *exp(c)*. Размерность пространства *d* выражается формулой:

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

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

Размерность множества точек, подобно размерности Хаусдорфа, можно определять различными способами. Они все задают примерно одинаковую функцию плотности распределения.

Метод осреднения

Здесь, для простоты, мы ограничимся рассмотрением данных (точек) из Евклидова пространства. Зададим оператор осреднения на функциях распределения $a(x), x \in R^n$ 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 симметрий, для сохранения симметрии для осредненных уравнений мы будем также использовать сферически симметричные ядра $P(x)$.

Пусть $P(x)$ — ядро осреднения (неотрицательно и интеграл от него равен 1), тогда $\frac{1}{a^d}P(\frac{x}{a})$ также является ядром осреднения. Такое ядро назовем подобным. Было бы желательно, чтобы осреднение осредненной 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 $\frac{1}{R^d}\exp(-\pi\frac{x^2}{R^2})$ 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 осредненная 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 осредненного 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 в d-мерном пространстве из N точек 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 [2]: $n = L(N) = \exp(\sqrt{\ln N \ln\ln N})$.

Метод осреднения заключается в осреднении плотности точек $\sum_i \delta(x - x_i)$. Выбирая далее срезы множества точек по определенному уровню плотности, мы получим разбиение на кластеры. Этот метод свободен от таких отмеченных выше недостатков, как зависимость от нумерации точек, и как существенное изменение разбиения на кластеры при малом изменении позиции даже одной точки.

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, у которых в ε-окрестности имеется m точек. При этом результат кластеризации сильно зависит от параметров (m, ε). Рекомендуемые значения этих параметров не учитывают ни размерность множества, ни 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 $P(x) = \begin{cases} 1, & |x| < \varepsilon, \\ 0, & |x| \geq \varepsilon \end{cases}$.

Тогда одинаковый вес при вычислении плотности точек вплоть до удаления на расстояние $R = \varepsilon$ приводит к ложному объединению всех точек в один кластер в ситуации на Рис. 2.

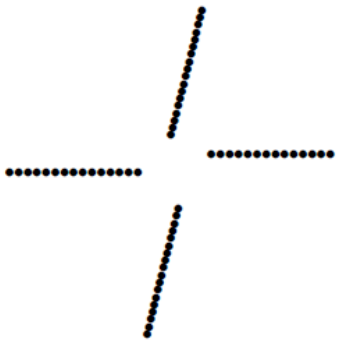

Рис. 2

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

Литература

1. Лесковец Ю., Раджараман А., Ульман Дж. Анализ больших данных. Москва, ДМК, 2016.
2. Крендалл Р., Померанс К. Простые числа. Криптографические и вычислительные аспекты, УРСС, 2011.

References

1. Leskovec J., Rajaraman A., Ullman J.D. Analiz bol'shikh dannyh. Moskva, DMK, 2016.
2. Crandall R., Pomerance C. Prostye chisla. Cryptographicheskie i vychislitel'nye aspekty. URSS, 2011.

Об авторах:

Айдагулов Рустем Римович, кандидат физико-математических наук, старший научный сотрудник кафедры теоретической информатики механико-математического факультета, Московский государственный университет имени М.В. Ломоносова, a_rust@bk.ru

Главацкий Сергей Тимофеевич, кандидат физико-математических наук, доцент, доцент кафедры теоретической информатики механико-математического факультета, Московский государственный университет имени М.В. Ломоносова, glavatsky_st@mail.ru

Note on the authors:

Aidagulov Rustem R., Candidate of Physics-Mathematical Science, Senior Researcher of Department of Theoretical Informatics of Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, a_rust@bk.ru

Glavatsky Sergei T., Candidate of Physics-Mathematical Science, Associate Professor of Department of Theoretical Informatics of Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, glavatsky_st@mail.ru