Методы осреднения в задачах кластеризации больших данных

Р. Р. Айдагулов[1]
С. Т. Главацкий[2]
А. В. Михалев[3]

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 интеллектуального 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. В данной работе предлагается метод локального осреднения для вычисления плотности распределения данных как точек в метрическом пространстве. Выбирая далее срезы множества точек по определенному уровню плотности, мы получаем разбиение на кластеры. Предложенный метод предлагает устойчивое разбиение на кластеры и свободен от ряда недостатков, присущих известным методам кластеризации.

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

[1] *Айдагулов Рустем Римович* — старший научный сотрудник, Кафедра теоретической информатики, механико-математический факультет, Московский государственный университет имени М.В. Ломоносова, Ленинские горы 1, Москва, 119991, Россия, a_rust@bk.ru.

Aidagulov Rustem Rimovich — Senior Researcher, Department of Theoretical Informatics, Faculty of Mechanics and Mathematics, Moscow Lomonosov State University, Leninskiye Gory 1, Moscow, 119991, Russia, a_rust@bk.ru.

[2] *Главацкий Сергей Тимофеевич* — доцент, Кафедра теоретической информатики, механико-математический факультет, Московский государственный университет имени М.В. Ломоносова, Ленинские горы 1, Москва, 119991, Россия, glavatsky_st@mail.ru.

Glavatsky Sergey Timoifeevich — Associate professor, Department of Theoretical Informatics, Faculty of Mechanics and Mathematics, Moscow Lomonosov State University, Leninskiye Gory 1, Moscow, 119991, Russia, glavatsky_st@mail.ru.

[3] *Михалев Александр Васильевич* — заведующий кафедрой, Кафедра теоретической информатики, механико-математический факультет, Московский государственный университет имени М.В. Ломоносова, Ленинские горы 1, Москва, 119991, Россия, aamikhalev@mail.ru.

Mikhalev Alexandr Vasilyevich — Head of Department, Department of Theoretical Informatics, Faculty of Mechanics and Mathematics, Moscow Lomonosov State University, Leninskiye Gory 1, Moscow, 119991, Russia, aamikhalev@mail.ru.

1. 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 набора (совокупности) на подмножества таким образом, чтобы "близкие" точки попали в одну группу, а "дальние" – в разные. Несложно понять, что в буквальном понимании такое требование противоречиво.

Пример. В конфигурации точек, приведенной на рис. 1, естественно разделить множество точек на 2 кластера, проведя разделительную границу около точки D.

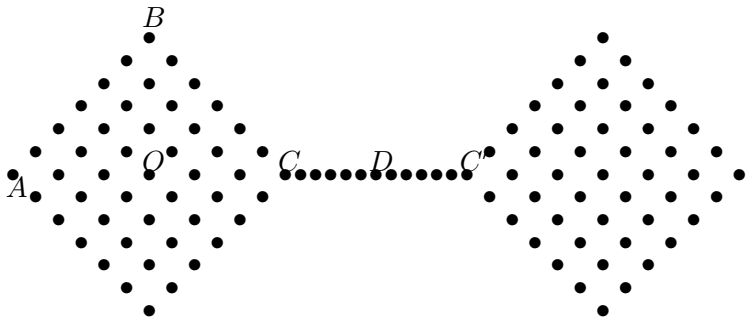

Рис.1

Аргументированной представляется группировка точек 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).

Машинное обучение, машинное распознавание образов должны, прежде всего, оперировать качественными характеристиками. Взаимные расстояния, вообще говоря, не являются таковыми. Например, расстояние между точками C и A на рис. 1 больше расстояния между точками C и C', однако более естественно поместить C и A в один кластер, а точку C' – в другой. При проверке текста на плагиат качественными характеристиками являются не только используемые наборы слов, но и смысл текста, который может передаваться употреблением совершенно иного множества терминов. В топологии качественными характеристиками являются группы гомотопий, гомологий и т.д., которые не меняются при гомотетиях. В геометрии качественными характеристиками являются такие характеристики, которые не меняются при гомотетиях, мало

изменяющих взаимные расстояния для близких точек. Такими характеристиками, в частности, являются реальная размерность распределения набора точек и распределение плотности набора точек. В качестве примера рассмотрим набор точек на кривой Веронезе:

$$x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d}) \in \mathbb{R}^d,\ x_{i,j} = \phi_j(x_{i,1}),\ \phi_j(t) = t^j, t \in [0,1].$$

Любые $n > d$ точек этой кривой не могут быть вложены в евклидово пространство размерности меньше d с сохранением взаимных расстояний. Пусть набор точек упорядочен: $x_{i,1} < x_{j,1} \Leftrightarrow i < j$. Существует гомотетия, отображающая точки x_i на прямую: $x_{1,1} + \sum_{j=1}^{i-1} \rho(x_j, x_{j+1})$. При этом сохраняются только расстояния между соседними точками. Однако если две точки были близки, то относительное изменение расстояний

$$\left| \frac{r(x_i, x_j)}{\rho(x_i, x_j} - 1 \right| < \varepsilon,\ \rho(x_i, x_j) < \delta, \tag{1}$$

мало. Поэтому распределение n точек на кривой Веронезе должно считаться не d-мерным распределением, а распределением, близким к одномерному.

Геометрические свойства, почти не меняющиеся при гомотетиях типа (1), являются (должны считаться) качественными. Основным требованием при разбиении совокупности точек на кластеры должно быть свойство качественности, т.е., два распределения точек, соединямые гомотетиями типа (1), должны разбиваться на кластеры почти одинаково.

Далее мы построим алгоритм кластеризации, базируясь на принципах плотностной связи между различными совокупностями точек. Отметим, что плотность в метрическом пространстве является основной качественной характеристикой.

2. 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. [3, 4]).

Введенное А.Н. Колмогоровым осреднение относится к глобальному типу, типа нахождения центра масс, определяющему среднее значение для совокупности точек. Однако глобальное осреднение полностью стирает информацию о локальном характере распределения точек. Поэтому мы далее будем использовать локальное осреднение, обладающее свойством качественности.

Плотность распределения в точке x определяется локальным осреднением следующим образом. Пусть точка x_i включена в шар единичного объема с центром в точке x с вероятностью $P(x, x_i)$. Тогда среднее количество (математическое ожидание) точек, входящих в этот объем, 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 объема с центром в точке x). Для определения областей сгущения плотности требуется именно такое локальное осреднение. При этом, если точки x_i, x_j близки, т.е., $\rho(x_i, x_j) < \delta$, то их вклад в осредненные величины

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

почти одинаков: $\left|\frac{P(x,x_i)-P(x,x_j)}{P(x,x_i)}\right| < \varepsilon$.

Часто для определения локальных значений параметров используют осреднение $P(x, x_i) = P(x, x_j)$ (см. [5]), когда точки из радиуса осреднения берутся с одинаковым весом. Такое осреднение приводит к различным затруднениям. Малое шевеление точек около границы осреднения может заметно влиять на результат работы известных алгоритмов, базирующихся на плотности распределения точек (например, DBSCAN). В итоге такие алгоритмы могут как ложно разъединять кластер, так и ложно сливать в один несколько различных кластеров.

Мы используем осреднение Гауссового типа

$$P(x, x_i) = exp(-\pi(\frac{\rho(x, x_i)}{R})^2),$$

учитывая, что оно инвариантно относительно поворотов (отображений, не меняющих взаимные расстояния) и суммарная вероятность вхождения точки x_i в пространство равна 1:

$$\int exp(-\pi((x - x_i).(x - x_i))dV = 1,$$

и не зависит от размерности пространства.

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 $(\ln N)^d$ (здесь d – реальная размерность, определяемая ниже), а, с другой стороны, – намного меньше, чем N^ε. Это свойство выполняется для часто встречающейся функции:

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

Метод осреднения заключается в осреднении множества точек с функцией плотности распределения

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

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

3. Размерность пространства, осреднение и плотность

Размерность пространства является локальной топологической характеристикой. Для наших целей удобнее использовать размерность, аналогичную размерности по Хаусдорфу, которая определяется как степень d роста величины $O(\varepsilon^{-d})$ минимального количества шаров радиуса ε, необходимого для покрытия нашего множества точек, 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 наклона 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:

$$r_1 \le r_2 \le ... \le r_{n(n-1)/2}. \tag{2}$$

На расстоянии не более r_i от некоторой точки в среднем находится $2i/n$ точек, без учета самой точки. Пусть $y_i = \log(2i/n),\ \ x_i = \log(r_i)$. Находим наилучшее приближение (аппроксимацию) $y_i = dx_i + c$ или $x_i = \frac{y_i - c}{d}$. На значение d не влияет ни основание log, ни постоянный коэффициент $\log(2/n)$ (уходит в определение величины c). Для уменьшения влияния больших расстояний (возможных между разными малыми кластерами) в вычислении корреляции оставим только определенную часть m значений $(1 \le m \le n)$ – только тех, где $r_i < s(R) \sim R, (s = 2R),\ \ 1 \le i \le m$. Вычисляя наилучшую линейную аппроксимацию, получим реальную размерность d и коэффициент пропорциональности $exp(c)$. Размерность пространства d выражается формулой:

$$d = \frac{\sum_i (\log i)^2 - \frac{1}{m}(\sum_i \log i)^2}{\sum_i \log(r_i) \log i - \frac{1}{m}(\sum_i \log i)(\sum_i \log(r_i))}. \tag{3}$$

Здесь m – длина суммирования (в сумме участвуют только первые m членов из $\{r_i\}$). Радиус суммирования s должен быть больше радиуса осреднения R, но не больше величины, при которой весом $f(\frac{s}{R}) = exp(-4\pi) << 1, (s = 2R)$ можно пренебречь. Ограничиваясь рассмотрением только точек, отстоящих на расстоянии менее s, где среднее коли-

чество точек $n = \frac{2m}{N} \sim exp(\sqrt{\ln N})$ не мало (больше $(\ln N)^d$), мы добиваемся уменьшения количества операций, практически не теряя статистической значимости, в вычислениях локальной плотности

$$\rho(x) = \sum_{i, r(x,x_i)<s} P(x, \frac{r(x, x_i)}{R}). \tag{4}$$

При этом локальные размерности (3) и локальные радиусы осреднения, соответствующие количеству $\frac{n}{2^d}$ точек в шаре, могут различаться для разных кластеров. Выбирая вначале малое значение s и увеличивая это значение примерно в 2 раза в следующей итерации до тех пор, пока не будет выполнено условие оценки среднего количества точек $n \sim exp(\sqrt{\ln N})$, получаем, что количество операций как в упорядочении (2), так и при вычислении плотности можно уменьшить до $O(N^{1+\varepsilon})$.

Далее нам нужно только осуществить сравнение плотности в разных точках как отношение вычисленных сумм.

Список литературы

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

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

[3] Бахвалов Н.С., Панасенко Г.П. Осреднение процессов в периодических средах. Математические задачи механики композиционных материалов. М.: Наука. Главная редакция физико-математической литературы, 1984.

[4] Маслов В.П. О способе осреднения для большого числа кластеров. Фазовые переходы. // Теоретическая и математическая физика. – 2000, т. 125, № 2, с. 297–314.

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

Averaging Methods in Big Data Clustering Problems

Aidagulov R.R., Glavatsky S.T., Mikhalev A.V.

Cluster analysis has a very wide range of applications; its methods are used in medicine, chemistry, archeology, marketing, geology and other disciplines. Clustering consists of grouping similar objects together, and this task is one of the fundamental tasks in the field of data mining. Usually, clustering is understood as a partition 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 points fall into different ones. In this paper, we offer a local averaging method for calculating the distribution density of data as points in a metric space. Choosing further sections of the set of points at a certain level of density, we get a partition into clusters. The proposed method offers a stable partitioning

into clusters and is free from a number of disadvantages inherent in known clustering methods.

Keywords: cluster, algorithm, density, averaging method.

References

[1] Leskovec J., Rajaraman A., Ullman J.D., *Mining of massive datasets.*, Cambridge Univ. Press, Cambridge, 2018, 534 pp.

[2] Kolmogorov A.N., *Izbrannye trudy. Matematika i mekhanika.*, Nauka, Moscow, 1985 (In Russian)

[3] Bahvalov N.S., Panasenko G.P., *Osrednenie processov v periodicheskih sredah. Matematicheskie zadachi mekhaniki kompozicionnyh materialov*, Nauka. Glavnaya redakciya fiziko-matematicheskoj literatury, Moscow, 1984 (In Russian), 352 pp.

[4] Maslov V.P., "O sposobe osredneniya dlya bol'shogo chisla klasterov. Fazovye perekhody", *Teoreticheskaya i matematicheskaya fizika*, **125**:2 (2000), 297–314 (In Russian)

[5] Nigmatulin R.I., *Osnovy mekhaniki geterogennyh sred*, Nauka, Moscow, 1978 (In Russian), 336 pp.