DOI 10.37882/2223-2966.2023.09.07

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ АЛГОРИТМА КЛАССИФИКАЦИИ K-MEANS ДЛЯ ДАННЫХ, ПОДЧИНЕННЫХ СТЕПЕННОМУ ЗАКОНУ РАСПРЕДЕЛЕНИЯ

FEATURES OF USING THE K-MEANS CLASSIFICATION ALGORITHM FOR DATA SUBJECT TO THE POWER LAW OF DISTRIBUTION

A. Egorkin

Summary. The paper is devoted to the application of the k-means clustering method for power law distributed data. On the example of an array of data on financial transactions, clustering was carried out using the k-means method, the number of clusters was determined by optimizing the silhouette coefficient.

The article shows that when logarithms of the source data are used as input data for the k-means algorithm, the clustering quality improves, clusters become homogeneous, and the intra-class variance decreases. It is proved that in the one-dimensional case, when using logarithmic data, clustering is carried out around the geometric mean values. At the same time, the clustering results do not depend on the base of the logarithm, according to which the logarithm of the source data is performed. It was also demonstrated the need for other quality metrics, clustering, not based on the Euclidean distance or the distance of city blocks, when working with data distributed according to a power law.

Keywords: clustering, k-means algorithm, power law of distribution, silhouette coefficient.

Егоркин Антон Александрович
Аспирант, Российский Государственный Социальный Университет (РГСУ)
2-5@bk.ru

Аннотация. Работа посвящена применению метода кластеризации k-means для данных, распределенных по степенному закону. На примере массива данных по финансовым операциям была проведена кластеризация методом k-means, количество кластеров определялось путем оптимизации коэффициент силуэта.

В статье показано, что при использовании в качестве входных данных для алгоритма k-means логарифмов исходных данных, качество кластеризации улучшается, кластеры становятся однородными, внутриклассовая дисперсия снижается. Доказано, что в одномерном случае при использовании логарифмированных данных кластеризация осуществляется вокруг среднегеометрических значений. При этом результаты кластеризации не зависят от основания логарифма, по которому осуществляется логарифмирование исходных данных. Также была продемонстрирована необходимость в иных метриках качества, кластеризации, не базирующихся на евклидовом расстоянии или расстоянии городских кварталов, при работе с данными, распределенными по степенному закону.

Ключевые слова: кластеризация, алгоритм k-means, степенной закон распределения, коэффициент силуэта.

Введение

Степенной закон распределения — это математический закон, описывающий распределение случайной величины, которое удивительным образом встречается в различных областях жизни. Этот закон применяется для анализа неравномерности распределения, характеризующейся тем, что небольшое число объектов обладают очень большими значениями (так называемые «тяжелые хвосты»), в то время как большинство объектов имеют низкие значения.

Примерами степенного закона распределения являются: распределение доходов и богатства в обществе, распределение ссылок в интернете или количество друзей в социальных сетях. Так есть единицы интернет-пользователей, имеющие миллионные аудитории, и огромное количество пользователей с небольшим количеством друзей.

Указанное распределение может быть описано формулой:

$$p(y) = Cy^{-\alpha} \quad (1)$$

где C — нормирующий коэффициент, α — показатель распределения.

Нормирующий коэффициент определяется исходя из требования того, что сумма всех вероятностей должна быть равна 1:

$$1 = \int_{y_{min}}^{\infty} p(y)dy = \frac{C}{1-\alpha} y^{1-\alpha}\Big|_{y_{min}}^{\infty} \quad (2)$$

Тогда распределение существует при $\alpha>1$, а нормирующий коэффициент равен:

$$C = (\alpha - 1) y_{min}^{\alpha-1} \quad (3)$$

Показатель распределения α определяется методом максимального правдоподобия [4]:

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{y_i}{y_{min}}\right]^{-1} \quad (4)$$

Также в [3] было показано, что степенной закон распределения имеет устойчивое среднее значение при α>2 и обладает свойством масштабной инвариантности [1].

Кластеризация данных, распределенных по степенному закону, методами, базирующимися на линейных расстояниях, приводит к тому, что, как правило, основная масса переменных будет сосредоточена в первом кластере, а оставшиеся кластеры имеют единичные вхождения. Такую кластеризацию сложно использовать для дальнейших исследований. Настоящая работа посвящена тому, как на примере одномерного случая, можно улучшить кластеризацию и какая математическая логика стоит за этим.

Метод кластеризации k-means

Одним из наиболее популярных методов кластеризации является алгоритм k-means [2]. Как правило, оптимизирующей функцией в данном методе является евклидово расстояние между элементами:

$$\arg_s \min \sum_{i=1}^{k} \sum_{x \in S_i} \left(\overline{x_i} - x \right)^2 \quad (5)$$

где k — количество кластеров, S_i — состав i-го кластера, $\overline{x_i}$ — центр i-го кластера.

Кроме евклидова расстояния возможно использование иных оптимизирующих метрик: квадрат евклидова расстояния, расстояние городских кварталов — модуль разницы между переменными, расстояние Чебышева и многие другие.

Как видно из формулы (5), ключевым параметром метода k-means является количество кластеров, которое задается перед началом работы алгоритма. Существуют множество способов поиска оптимального количества кластеров, но все они, как правило, базируются на эвристических подходах. Одним из таких методов является выбор количества кластеров, оптимизирующих одну из функций качества разбиения данных на кластеры. При каком количестве кластеров функция качества достигает локального оптимума, то количество кластеров и является оптимальным. В данной работе используется функция качества — силуэт [5]:

$$s_j = \frac{b_j - a_j}{max\left(b_j, a_j\right)} \quad (6)$$

где s_j — силуэт узла j, принадлежащего i-му кластеру, b_j — среднее расстояние от узла j до ближайшего соседнего кластера, a_j — среднее расстояние от узла j до всех элементов кластера, куда входит данный узел.

Силуэт всей выборки определяется как среднее значение силуэтов всех узлов, входящих в нее. Чем ближе будет показатель силуэта к 1, тем более качественной является кластеризация.

Проблема использования алгоритма k-means для данных, распределенных по степенному закону

В качестве данных для дальнейшего исследования используем информацию о финансовых транзакциях клиентов коммерческого банка. Указанные данные подчинены степенному закону распределения. Т.к. используемая информация представляет собой коммерческую тайну, то для анализа используются нормированные данные, путем деления на максимальное значение, т.е. максимальный элемент в дата-сете будет равен 1. Данные представляют собой одномерный массив, где каждый элемент — это суммарный объем ненулевых платежей одного клиента другому за определенный период времени.

Применение алгоритма k-means к описанным выше данным дает следующий результат.

С точки зрения количество кластеров, оптимальным представляется 7 штук, т.к. именно при 7 кластерах график коэффициента силуэта имеет локальным максимум.

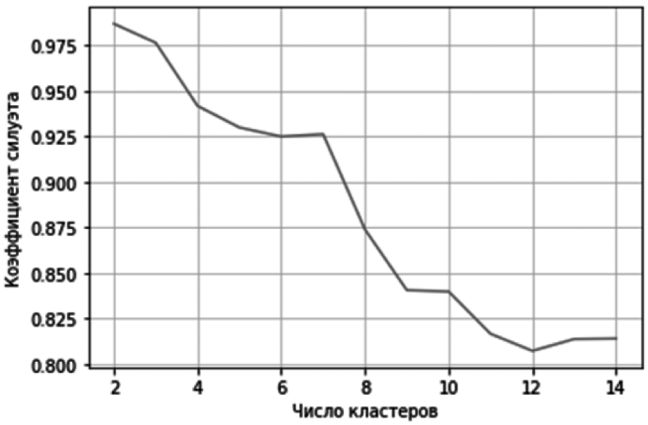

Рис. 1. Зависимость коэффициента силуэта от количества кластеров

Таблица 1.

Распределение данных по кластерам

номер кластера	1	2	3	4	5	6	7
количество элементов в кластере	7468	183	27	13	5	5	1
среднее арифметическое	0,002	0,037	0,124	0,255	0,422	0,665	1,000
среднее геометрическое	0,001	0,035	0,120	0,253	0,420	0,663	1,000
медиана	0,001	0,035	0,119	0,252	0,434	0,679	1,000
стандартное отклонение	0,003	0,015	0,029	0,031	0,036	0,058	—

Несмотря на высокие показатели качества кластеризации, коэффициент силуэта близок к 1, такую кластеризацию нельзя назвать успешной. Т.к. фактически все элементы были отнесены в первый кластер (в первый кластер попали 97 % данных выборки). При этом в последнем кластере находится всего один элемент. Такой результат работы k-means связан со структурой данных, т.к. есть очень много небольших значений и немного очень больших значений.

Как можно улучшить кластеризацию для данных, распределенных по степенному закону

Если прологарифмировать выражение (1) то получится линейная функция в логарифмических координатах:

$$\log p(y) = \log C - \alpha \log y \tag{7}$$

Далее предлагается использовать в качестве входных переменных для алгоритма k-means логарифмы исходных данных. Тогда распределение становится более равномерным и линейные расстояния, используемые в методе k-means должны давать приемлемый результат.

В рассматриваемом частном одномерном случае справедливо следующая интерпретация алгоритма k-means: здесь эвклидово расстояние является модулем разности координат и выражение (5) принимает следующий вид:

$$\min \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left| \overline{x_i} - x_j \right| \tag{8}$$

Где k — количество кластеров, n_i — количество элементов в кластере i.

Если в качестве $x_j = \log y_j$, тогда:

$$\overline{x_i} = \frac{\sum_{j=1}^{n_i} x_j}{n_i} = \frac{\sum_{j=1}^{n_i} \log y_j}{n_i} = \\ = \frac{\log \prod_{j=1}^{n_i} y_j}{n_i} = \log \prod_{j=1}^{n_i} y_j^{1/n_i} \tag{9}$$

Выражение (8) можно записать следующим образом:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left| \log \overline{y_i} - \log y_j \right| = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left| \log \frac{\overline{y_i}}{y_j} \right| \to 0 \tag{10}$$

Потенцировав выражение (10) получаем:

$$\prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{\overline{y_i}}{y_j}^{sgn\left(\overline{y_i} - y_j\right)} \to 1 \tag{11}$$

Где $\overline{y_i} = \sqrt[n_i]{\prod_{j=1}^{n_i} y_j}$ — среднегеометрическое значение точек, попавших в кластер i,

$$sgn(x) = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases} \text{ — функция знака.}$$

Таким образом, если в алгоритм k-means в качестве исходных данных передать логарифмированный вектор данных, то результатом действия алгоритма будет выбор кластеров таким образом, чтобы его элементы объединялись вокруг своих среднегеометрических значений.

Метод k-means для логарифмированных данных и сравнение результатов

Прологарифмировав исходные данные, применим к полученным переменным алгоритм k-means.

Здесь, также как и в случае с нелогарифмированными данными, с точки зрения коэффициента силуэта оптимальным представляется 7 кластеров:

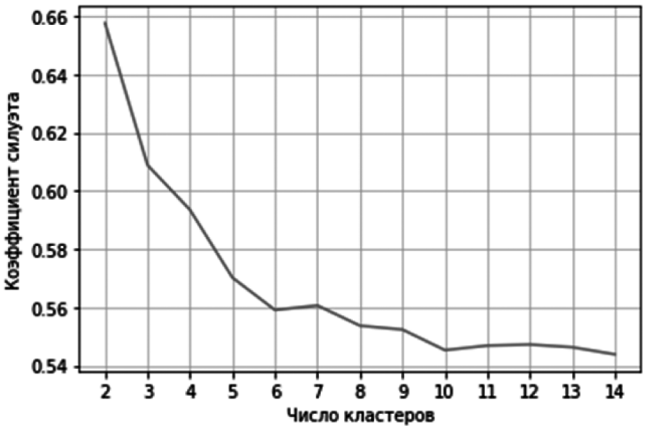

Рис. 2. Коэффициента силуэта для логарифмированных данных

Таблица 2.

Распределение логарифмированных данных по кластерам[1]

номер кластера	1	2	3	4	5	6	7
количество элементов в кластере	2637	1955	1384	869	538	251	68
среднее арифметическое	0,0004	0,0007	0,0014	0,0030	0,008	0,028	0,210
среднее геометрическое	0,0004	0,0007	0,0014	0,0029	0,007	0,026	0,155
медиана	0,0004	0,0007	0,0014	0,0028	0,007	0,024	0,131
стандартное отклонение	0,000	0,000	0,000	0,001	0,002	0,012	0,192

[1] Для сопоставимости результатов все статистические показатели кластеров рассчитывались после приведения логарифмов в исходные данные. Т.е. $y_i = a^{x_j}$, a — основание логарифма

Полученное в последнем случае разбиение имеет более равномерную структуру. Как отмечалось выше, в анализируемом случае данные группируются вокруг среднегеометрических значений. Из таблицы 2 видно, что медиана в каждом из кластеров ближе к среднегеометрическому значению, чем к среднеарифметическому. Также среднее геометрическое лучше описывает популяцию в кластере и в случае кластеризации нелогарифмированных данных (Табл. 1).

В части плотности кластеров, можно утверждать, что использование логарифмированных данных приводим к более плотным кластерам. Это видно из распределения стандартного отклонения в кластерах (Рис. 3). Седьмой кластер здесь не показан, т.к. при кластеризации исходных данных в него попадал только один элемент.

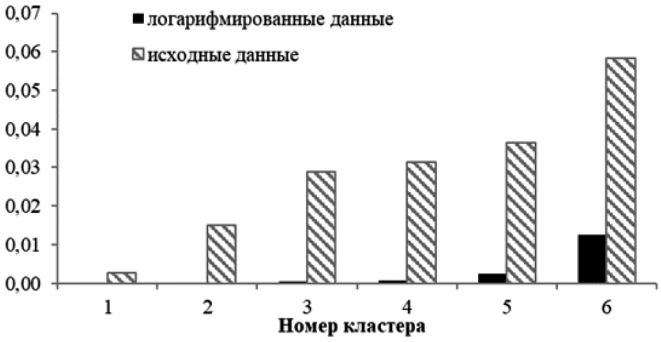

Рис. 3. Распределение дисперсии по кластерам

Из формулы (11) следует, что целевая функция в рассматриваемом случае не зависит от основания логарифма, что также было подтверждено расчетами. Результаты не изменялись при использовании логарифмов различных оснований.

Отдельно необходимо отметить, что для такой кластеризации метрика качества силуэт, базирующаяся на линейном или евклидовом расстоянии, оказывается хуже, чем при кластеризации нелогарифмированных данных. Однако, это вопрос к самим метрикам качества, которые в данном случае, по мнению автора, должны базироваться на формуле (11).

Наглядно сравнение результатов кластеризации можно представить на диаграммах в логарифмических координатах, где каждый цвет соответствует своему кластеру.

Выводы

В случае применения метода k-means для данных, подчиненных степенному закону распределения необходимо предварительно нормировать их путем логарифмирования. Тогда использование линейных метрик алгоритма k-means будет приводить к тому, что элементы будут стремиться создать кластеры вокруг своих среднегеометрических значений. Как было показано в статье, среднегеометрическое значение лучшее описывает популяцию в кластере, чем среднеарифметическое.

При этом логарифмировать исходные данные можно по любому основанию, как показано в формуле (11), итоговая целевая функция не зависит от основания логарифма.

Также необходимы иные метрики качества кластеризации. Метрики, основанные на евклидовом расстоянии или расстоянии городских кварталов, не применимы для такого распределения. При этом плотность самих кластеров в случае логарифмирования исходных данных оказывается выше. Таким образом, метрики качества кластеризации дынных, распределенных по степенному закону, должны базироваться на формуле (11).

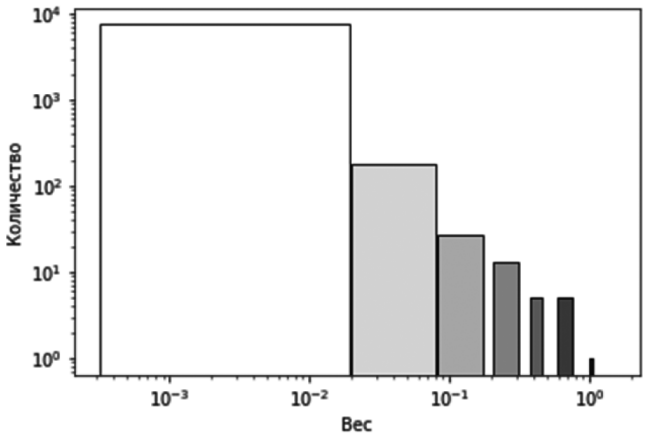

Рис. 4. Распределение исходных данных

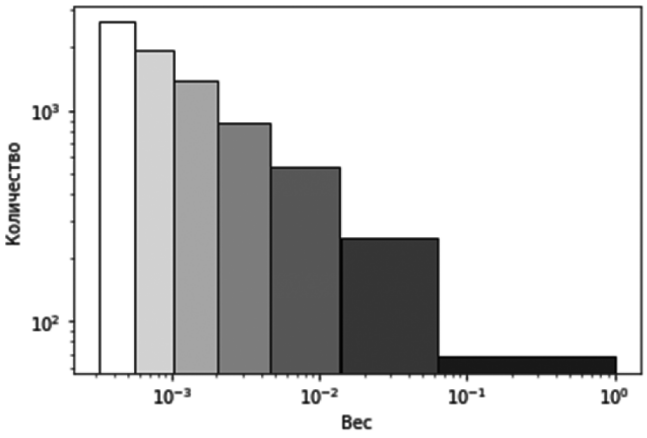

Рис. 5. Распределение логарифмированных данных

ЛИТЕРАТУРА

1. Barabasi A.L., Bonabeau E. Scale Free Networks // Scientific American, 2003, p. 50–59.
2. David Arthur and Sergei Vassilvitskii. How slow is the k-means method? In SCG '06: Proceedings of the twenty-second annual symposium on computational geometry. ACM Press, 2006.
3. Newman M. Power laws, Pareto distributions and Zipf's law// Statistical Mechanics, 2006.
4. Newman M., Aaron C., Cosma R. S. Power-law distributions in empirical data // Physics Today, 2009.
5. Rousseeuw P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis // Journal of Computational and Applied Mathematics, 1987.

Журнал «Современная наука: актуальные проблемы теории и практики»