УДК 519.23

Пошаговое улучшение внутрикластерного рассеивания

Дронов С.В., Титова В.Е.
Алтайский государственный университет, г. Барнаул
dsv@math.asu.ru, vika566384@mail.ru

Аннотация

В работе детально описан алгоритм оптимизации кластерного разбиения. Критерием качества выбрано суммарное внутрикластерное рассеивание по всем вновь организуемым кластерам. Уменьшение этого рассеивание достигается направленным пошаговым перемещением отдельных объектов между кластерами. Алгоритм реализован в виде компьютерной программы. Приведены примеры его работы на реальных данных.

Ключевые слова: кластерные разбиения, сравнение разбиений, пошаговая оптимизация разбиений.

1. Оценка качества разбиения. Постановка задачи

В алгоритмах, применяемых для решения задач кластерного анализа, заданное на входе в алгоритм множество U разбивается на кластеры – такие его части, что элементы одного кластера оказываются более близкими, чем элементы различных. В силу такой постановки задачи можно считать, что построенное разбиение тем лучше, чем более близкими друг к другу являются объекты в каждом кластере, то есть чем меньше взаимные различия элементов внутри каждого из кластеров. Сформулируем точнее.

Предположим, что конечное множество объектов представлено в виде точек, координаты которых равны значениям показателей, в соответствии с которыми строится некоторое кластерное разбиение $K_1, \ldots, K_M$. В роли рассеивания j-го кластера, т.е. основного критерия качества кластера, примем величину Q_j:

$$Q_j = \sum_{a \in K_j} \rho^2 \left(\bar{u_j}, a\right),$$

где a – элемент j-го кластера, $\bar{u_j}$ – вектор средних значений показателей объектов j-го кластера, $\rho^2(\bar{u_j}, a)$– расстояние (например, евклидова метрика) между элементами $\bar{u_j}$, a_i.

Поскольку алгоритмы кластеризации имеются сегодня в большом количестве (см., например, [1]), и в результате их работы могут получаться различные наборы кластеров, то вопрос об оценивании качества конкретного кластерного весьма актуален. В [2, 3], в частности, предлагаются разные подходы к решению этого вопроса.

Основным критерием качества кластеризации в целом будем считать сумму внутрикластерных рассеиваний Q_j:

$$Q = \sum_{j=1}^{M} Q_j,$$

Чем меньше это значение, тем более плотно объекты внутри кластеров находятся по отношению к центрам этих кластеров, тем, соответственно, лучше разбиение. Разбиение, для которого Q минимально, будем называть оптимальным.

Основная задача работы – реализация и практическое тестирование алгоритма оптимизации заданного кластерного разбиения из [4] с точки зрения предложенного критерия качества.

2. Основной алгоритм улучшения внутрикластерного рассеивания

Отметим, что в случае, когда все кластеры одноэлементны, дальнейшее улучшение разбиения в указанном выше смысле невозможно, но и полезность такого разбиения нельзя назвать высокой, так как кластеризация нужна для того, чтобы обеспечить возможность работы не с множеством разнообразных, не связанных друг с другом данных, а со сгруппированными данными.

Исходя из этого, запретим изменение числа кластеров в разбиении, с которого стартует алгоритм, так как отсутствие ограничения на уменьшение числа кластеров может привести к разбиению, в котором все объекты содержатся в одном кластере. В обоих крайних случаях кластеризация, видимо, не является целесообразной.

Для описания алгоритма нам понадобятся перечисленные далее некоторые факты и обозначения из [4]. Рассмотрим кластеры K_j и K_i, где n_{K_j}, n_{K_i} – количества их элементов, центральные объекты $\bar{u_j}$, $\bar{u_i}$ и рассеивания Q_j, Q_i соответственно.

Лемма 1. *При добавлении к кластеру K_j элемента A величина рассеивания этого кластера изменится на*

$$\triangle^+(K_j) = \frac{n_{K_j}}{(n_{K_j}+1)^2}||A - \bar{u_j}\ ||^2 - \frac{Q_j}{n_{K_j}+1}. \tag{1}$$

Лемма 2. *При исключении же из кластера элемента A изменение рассеивания данного кластера произойдет на величину*

$$\triangle^-(K_j) = \frac{Q_j}{n_{K_j}-1} - \frac{n_{K_j}}{(n_{K_j}-1)^2}\|A - \bar{u_j}\|^2\,,\ \left(2 \le n_{K_j}\right). \tag{2}$$

Лемма 3. *Если один элемент из кластера K_j, содержащего не менее двух элементов, перенесется в кластер K_i, то рассеивания этих двух кластеров изменятся на величину:*

$$\triangle\left(K_j, K_i\right) = \frac{n_{K_i}}{\left(n_{K_i}+1\right)^2}\|A - \bar{u_j}\|^2 - \frac{n_{K_j}}{\left(n_{K_j}-1\right)^2}\|A - \bar{u_j}\|^2 + \frac{Q_j}{n_{K_j}-1} - \frac{Q_i}{n_{K_i}+1}. \tag{3}$$

Теперь опишем предлагаемый алгоритм по шагам, детализируя предложенную в [4] общую схему.

Шаг 1. Для каждого из кластеров, состоящих из двух и более объектов $K_1, \ldots, K_M$, вычисляем число его элементов n_{K_j}, центральный объект $\bar{u_j}$ и рассеивание $Q_1, \ldots, Q_M$. Это можно делать только для тех кластеров, которые изменились в процессе предыдущей работы алгоритма.

Шаг 2. Для каждого объекта основного множества вычислим его отклонения от центрального объекта каждого из кластеров, а затем для кластера K_j, в котором содержался этот объект, и каждого из кластеров $K_i \neq K_j$ найдем потенциальное изменение величины $\triangle(K_j, K_i)$ по формулам (1) и (2). Суммарное рассеивание всех кластеров изменится на величину, рассчитываемую по формуле (3).

Шаг 3. Есть ли среди чисел $\triangle(K_j, K_i)$ отрицательные? Если нет, то оптимальное разбиение построено – выход из алгоритма. Если да – перейти к шагу 4.

Шаг 4. Находим такой элемент A и такой кластер K_i, что $A \in K_j$, $\triangle(K_j, K_i)$ отрицательно и наибольшее по модулю среди всех таких отрицательных значений. Переносим объект A из K_j в K_i. Возвращаемся к шагу 1.

3. Пример оптимизации алгоритмом разбиения реальных данных

Данный алгоритм был реализован в виде программы, написанной на языке Python. Рассмотрим пример. В таблице 1 приведен фрагмент данных обследования 28 пациентов Алтайского краевого диагностического центра. Данные были предоставлены врачом-пульмонологом, к.м.н. Параевой О.С. Первый столбец таблицы – номер обследованного пациента, второй – оценка времени свертывания крови в секундах, третий – оценка количества лейкоцитов (миллиардов на литр). В последнем столбце указан условный код установленного в результате обследования диагноза. Каждый из возможных 5 диагнозов принят нами за кластер объективного разбиения.

Таблица 1

Кластерное разбиение реальных данных

№	ТВ	лейкоциты	диагноз (объективное разбиение)	№	ТВ	лейкоциты	диагноз (объективное разбиение)
1	27	9,7	1	**15**	20,2	6,8	2
2	11,4	11,8	1	**16**	15	6,6	1
3	11,6	5,5	1	**17**	10,8	9,2	5
4	25,6	10,3	5	**18**	19,6	10,6	3
5	25	6,4	2	**19**	13	17,4	4
6	13,2	11	2	**20**	25,4	6	5
7	22	9	5	**21**	15,8	6,5	1
8	24,8	4,5	5	**22**	11	6,6	5
9	14	9,7	2	**23**	13,6	6,4	3
10	18	8	5	**24**	12	9,2	3
11	24	5,4	5	**25**	22,4	8,6	1
12	14	10,9	3	**26**	16	5,7	4
13	11,6	7,5	5	**27**	20	11,1	2
14	25,6	6,9	5	**28**	13	5,1	4

На рисунке 1 изображены кластеры и коэффициент рассеивания до применения алгоритма. Здесь $Q = 146.34$.

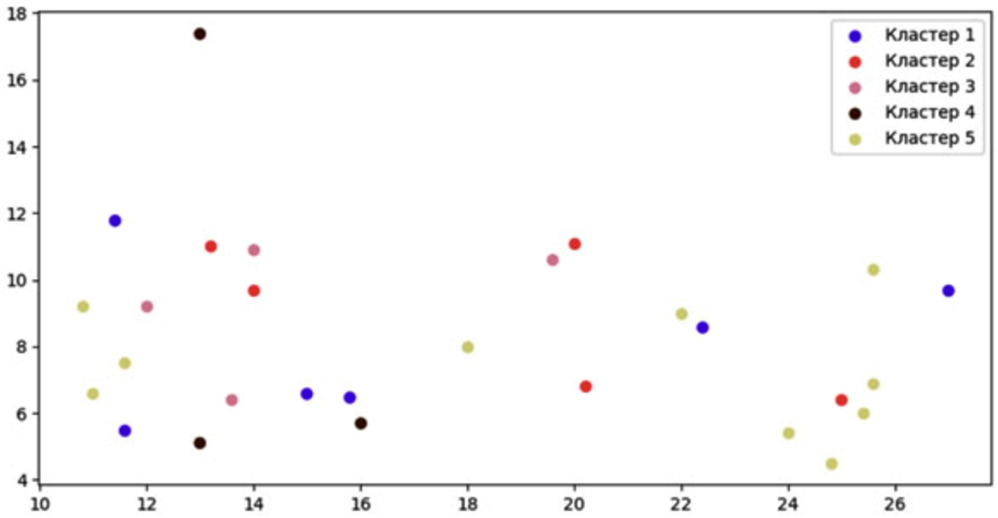

Рисунок 1. Кластеры до применения алгоритма

Можем заметить, что в данном разбиении в большинстве случаев объекты одного кластера находятся ближе к представителям других кластеров, чем к объектам из одного с ними кластера. Так что предложенное разбиение нельзя считать качественным. Кластеры после применения алгоритма оптимизации изображены на рисунке 2.

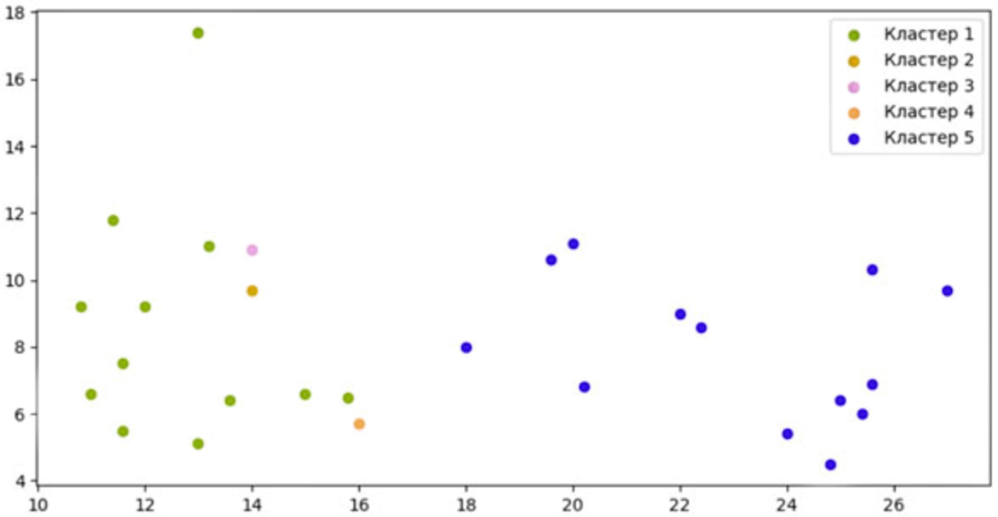

Рисунок 2. После применения алгоритма

На рисунке 2 можно увидеть тенденцию алгоритма к тому, чтобы разбить множество всех объектов на 2 кластера. Но такое изменение разбиения невозможно из-за введенного ранее нами условия на ограничение уменьшения количества кластеров. Тем не менее, коэффициент рассеивания значительно уменьшился, $Q = 25.19$.

Работу программы можно проследить по таблице 2, в которой показано, какой объект мы перемещаем (значение столбца i), из какого кластера в какой кластер мы убираем или добавляем этот объект (соответствующие значения столбцов A и B), и как при таком перемещении меняется значение коэффициента рассеивания Q (значение столбца Q).

Таблица 2

Итерации работы программы

№	A	B	i	Q	№	A	B	i	Q
1	4	1	13 17.4	146.35	**8**	5	1	11.6 7.5	60.14
2	1	5	27 9.7	122.03	**9**	2	5	20.2 6.8	48.11
3	1	5	25 6.4	109.39	**10**	2	5	20 11.1	44.10
4	1	5	22.4 8.6	97.98	**11**	3	1	13.6 6.4	35.83
5	3	5	19.6 10.6	87.60	**12**	3	1	12 9.2	32.20
6	5	1	10.8 9.2	78.50	**13**	4	1	13 5.1	28.96
7	5	1	11 6.6	69.58	**14**	2	1	13.2 11	26.38

Таким образом, коэффициент рассеивания уменьшился примерно на 83%.

4. Обсуждения и выводы

После анализа полученных в работе результатов можно увидеть, что кластеры в новом разбиении обладают меньшим суммарным внутрикластерным рассеиванием, однако новое разбиение имеет довольно необычный вид из-за наличия в нём большого числа кластеров, содержащих всего 1 элемент. Это позволяет прогнозировать тенденцию к формированию нового разбиения, имеющего только два кластера. Такой тренд, видимо, следует расценивать, как положительную тенденцию, учитывая, что в уменьшение числа кластеров в медицине в конечном итоге приводит к более простым и однотипным лечебным процедурам, а, как следствие, к экономии средств и ресурсов.

Список литературы

1. Ezugwu A.E., Ikotun A.M., Oyelade O.O. et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects // Engineering Application of Artificial Intelligence. — 2022. — Vol. 110. — 104743.

2. Журавлева В.В., Куракина А.А. Упрощенный показатель силуэта кластерной структуры // Сборник трудов Всероссийской конференции по математике с международным участием "МАК-2019"– Барнаул. — Барнаул : Изд-во АлтГУ, 2019. — С. 254–255.

3. Renedo-Mirambell M., Arratio A. Identifying bias in network clustering quality metrics // Peer. J Comput Sci. — 2023. — Vol. 9. — PMC10495975.

4. Дронов С.В. Оптимизация кластерных разбиений с привлечением техники латентного анализа классов // Известия АлтГУ. — 2023. — № 1 (129). — С. 89–94.