УДК 519.23

# Оптимизация кластерных разбиений на основе анализа латентных показателей

Свеженцев М.Е.
*Алтайский государственный университет, г. Барнаул*
*byvaet54goda@mail.ru*

**Аннотация**

В работе рассматривается один из подходов к улучшению имеющегося кластерного разбиения. При этом, используя в качестве естественного критерия качества разбиения внутреннюю согласованность составляющих его кластеров, для улучшения этого качества предлагается использовать современную технику анализа латентных классов. Модифицирован и реализован в виде компьютерной программы алгоритм оптимизации латентного показателя. Рассмотрен численный пример.

*Ключевые слова:* Алгоритмы кластеризации, латентный анализ классов, оптимизация кластерных разбиений

## 1. Сравнение кластеризаций. Основная задача работы

Кластерным разбиением конечного множества $A$ принято называть его представление в виде объединения произвольного количества попарно непересекающихся непустых подмножеств, а получившиеся группы кластерами.

Существует много разных алгоритмов для разбиения множества объектов на кластеры. Видимо, наиболее часто используемыми из них алгоритм $k$-средних, иерархические древовидные алгоритмы и алгоритм FOREL, см., например, [1–3].

Разными методами можно получить разные кластерные разбиения. В связи с этим возникает необходимость сравнивать разбиения как между собой, так и по степени их отличия от некоторого эталонного, в каком-то смысле оптимального разбиения, если оно известно. В задачах классической математики степень различия объектов, как правило, устанавливают с помощью расстояния (метрики). Многочисленные способы введения метрик на семействе всех возможных кластерных разбиений можно найти в [4–8].

Поскольку объекты в рамках одного кластера предполагаются более близкими к друг другу, чем объекты разных кластеров, то для улучшения кластерного разбиения необходимо переформатировать кластеры так, чтобы внутри каждого из них степень схожести объектов повысилась. Основной задачей работы является детальная проработка и реализация одного из способов улучшения имеющегося кластерного разбиения в указанном смысле.

При этом в данной работе мы ограничимся наиболее наглядным случаем, когда каждый из объектов основного множества задан лишь двумя формирующими показателями.

## 2. Разбиение на кластеры на основе латентного показателя

В поисках оптимального разбиения множества объектов на кластеры мы пытаемся достичь максимальной схожести объектов внутри каждого кластера, основываясь на значениях показателей этих объектов. Для этого предположим, что существует некая, непосредственно не наблюдаемая, скрытая характеристика, которая влияет на конечный результат разбиения на кластеры в большей степени, чем наблюдаемые. Будем, следуя [9], называть её латентной кластерной переменной.

Если мы каким-то образом исключим вклад такой переменной в показатели объектов, то внутри каждого из кластеров оставшиеся их части должны оказаться практически независимыми, если только кластеры построены оптимально. Поэтому основной идеей для построения кластеров наилучшим образом может служить сильная корреляция показателей объектов внутри этих кластеров. Латентной кластерной переменной тогда можно будет считать ту общую часть показателей, за счет наличия которой показатели и оказываются сильно коррелированы. Таким образом, изменяя кластер так, что корреляция между признаками объектов внутри него увеличивается, мы улучшаем качество разбиения.

Имеем начальное кластерное разбиение $n$ объектов на $k$ кластеров. Пусть $R_j$ – квадрат оценки коэффициента корреляции между показателями внутри $j$-го кластера, $j = 1, \ldots, k$. Использование квадрата коэффициента позволит одновременно учесть его значимые варианты независимо от знака и облегчит поиск аналитического решения. Таким образом, мы стремимся найти разбиение множества объектов на $k$ непересекающихся кластеров, при котором критерий

$$R = \sum_{j=1}^{k} R_j$$

будет иметь максимально большое значение.

Методы оценивания латентной кластерной переменной и ее интерпретации, впервые, видимо, последовательно изложенные в [10], формируют ядро относительно нового раздела анализа данных, называемого латентным кластерным анализом или анализом латентных классов.

## 3. Алгоритм улучшения кластерного разбиения

Опишем алгоритм модифицирования оптимального разбиения и назовем его трансформирующим алгоритмом или алгоритмом K2. Идея алгоритма заимствована из [9].

Имеем некоторое кластерное разбиение из $k$ кластеров. Объявим это разбиение текущим.

**1.** Для каждого $i = 1, 2, \ldots, k$ вычисляем оценку коэффициента корреляции $\rho_i$ между показателями на наборе объектов $i$-го кластера текущего разбиения.

**2.** Выбираем текущий кластер $(A_i)$. Во всех остальных кластерах ищем потенциальные объекты $(x, y)$ для перемещения в текущий кластер и заносим их в список. Согласно [9], это будут объекты, удовлетворяющие условию

$$\rho_i(x - \overline{X})(y - \overline{Y}) < 0.$$

**3.** Выбираем некоторый объект списка, пусть значения его показателей $x$ и $y$. Вычисляем величину изменения критерия $R$ при его перемещении в текущий кластер следующим образом

$$Q_{i,j}(x, y) = \rho_i^2 + \rho_j^2 - \rho_{new,i}^2 - \rho_{new,j}^2.$$

При этом (см. [9])

$$\rho_{new,i} = \frac{\rho_i + a_x a_y}{\sqrt{(1 + a_x^2)(1 + a_y^2)}}$$

для кластера, в который перемещаем объект, и

$$\rho_{new,j} = \frac{\rho_j - b_x b_y}{\sqrt{(1 + b_x^2)(1 + b_y^2)}}$$

для кластера, из которого объект убирается. В последних формулах

$$a_x = \frac{x - \overline{X}}{S_x\sqrt{n+1}}, \qquad a_y = \frac{y - \overline{Y}}{S_y\sqrt{n+1}},$$
$$b_x = \frac{x - \overline{X}}{S_x\sqrt{n-1}}, \qquad b_y = \frac{y - \overline{Y}}{S_y\sqrt{n-1}}.$$

**4.** Повторяем шаг 3 для всех объектов текущего кластера и всех элементов списка, запоминая соответствующие значение $Q_{i,j}(x, y)$.

**5.** Находим $Q_i = \max\limits_{j,x,y} Q_{i,j}(x, y)$. Запоминаем те $j, x, y$ для данного $i$, на которых этот максимум достигается.

**6.** Переходим к следующему кластеру, объявляя его текущим, и выполняем для него шаги 3, 4 и 5.

**7.** Среди всех $i$ выбираем то, у которой значение $Q_i$ максимально.

**8.** Если найденное максимальное значение отрицательно, то перемещать больше нечего – конец алгоритма. Текущие разбиение и есть оптимальное.

**9.** Иначе, возвращая значения $j, x, y$ для найденного на шаге 7 номера кластера $i$, производим перемещение объекта $(x, y)$ из $A_i$ в $A_j$ и объявляем полученное разбиение текущим. Возвращаемся к шагу 1.

Алгоритм K2 был реализован в виде компьютерной программы на языке Python.

## 4. Один практический пример

Продемонстрируем работу алгоритма на практических данных. воспользуемся данными исследований печени, предоставленными врачом-радиологом Санкт-Петербургского городского центра КТ/МРТ к.м.н. Жуковой О.В.

Таблица 1

Данные МРТ печени 30 пациентов

| № пациента | T2*печени | Ферритин | Степень вырожденности |
|---|---|---|---|
| 1 | 1 | 4422 | 3 |
| 2 | 2,3 | 2967,8 | 2 |
| 3 | 1,6 | 1882 | 3 |
| 4 | 4,2 | 2560 | 2 |
| 5 | 2,1 | 3924 | 2 |
| 6 | 3 | 803,4 | 2 |
| 7 | 7,2 | 847,3 | 1 |
| 8 | 3,9 | 2046 | 2 |
| 9 | 2,7 | 4483 | 2 |
| 10 | 3,3 | 1508 | 2 |
| 11 | 4,5 | 1060,27 | 2 |
| 12 | 3 | 1598 | 2 |
| 13 | 2,1 | 2083,5 | 3 |
| 14 | 2,2 | 913,9 | 2 |
| 15 | 4,9 | 1486,3 | 1 |
| 16 | 7,2 | 1154 | 1 |
| 17 | 3,8 | 1065 | 2 |
| 18 | 6,5 | 1690 | 1 |
| 19 | 7,6 | 790 | 1 |

| | | | |
|---|---|---|---|
| 20 | 1,2 | 6000 | 3 |
| 21 | 2,8 | 1640 | 2 |
| 22 | 4 | 969 | 2 |
| 23 | 3,5 | 629 | 2 |
| 24 | 1,6 | 6839 | 3 |
| 25 | 10,4 | 736 | 1 |
| 26 | 3,6 | 632 | 2 |
| 27 | 4,4 | 1600 | 2 |
| 28 | 4,7 | 1386 | 1 |
| 29 | 6,6 | 25744 | 1 |
| 30 | 4,5 | 6450 | 2 |

За исследуемые показатели выберем T2* печени (время релаксации протонов на МРТ) и ферритин (лабораторный показатель насыщения крови железом). В качестве кластеров возьмем наборы пациентов с одинаковой степенью выраженности симптомов (последний столбец таблицы). Последующие шаги работы алгоритма отображены на рисунках 1–4:

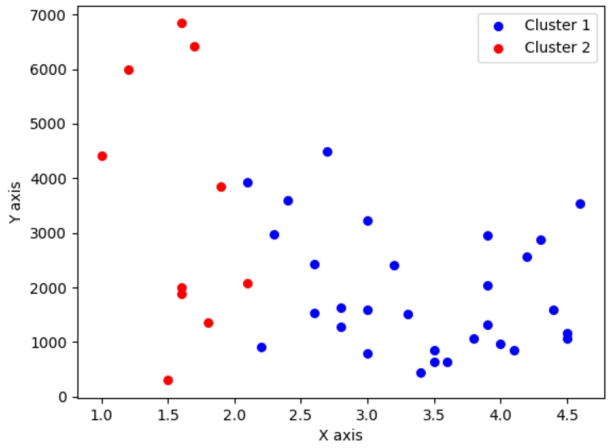


Рисунок 1. Начальное разбиение медицинских данных ($R = 0.153$)

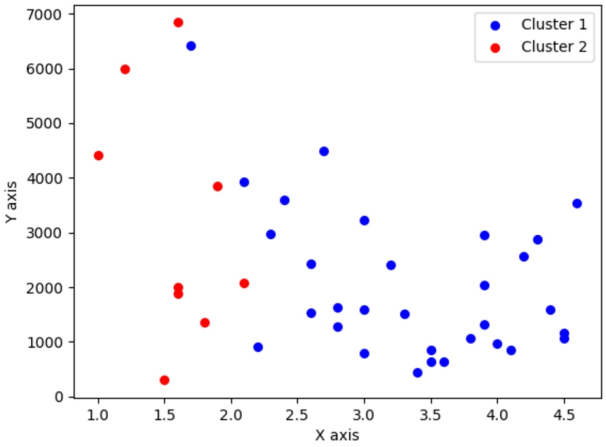


Рисунок 2. Разбиение после первого перемещения объекта ($R = 0.323$)

На рисунке 4 изображено итоговое (далее неулучшаемое) разбиение данных на два кластера.

Рассчитаем численную оценку качества стартового и конечного разбиений, основанную на средних внутрикластерных и межкластерных расстояниях. Для этого определим среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j}[A_i = A_j]d(A_i, A_j)}{\sum_{i<j}[A_i = A_j]},$$

где

$$d(A_i, A_j) = \frac{1}{C_n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho(a_i, a_j).$$

Среднее межкластерное расстояние определим по формуле

$$F_1 = \frac{\sum_{i<j}[A_i \neq A_j]q(A_i, A_j)}{\sum_{i<j}[A_i \neq A_j]},$$

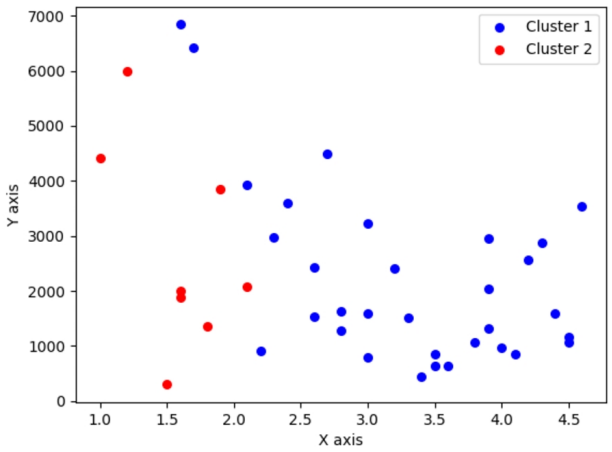


Рисунок 3. Разбиение после второго перемещения объекта ($R = 0.522$)

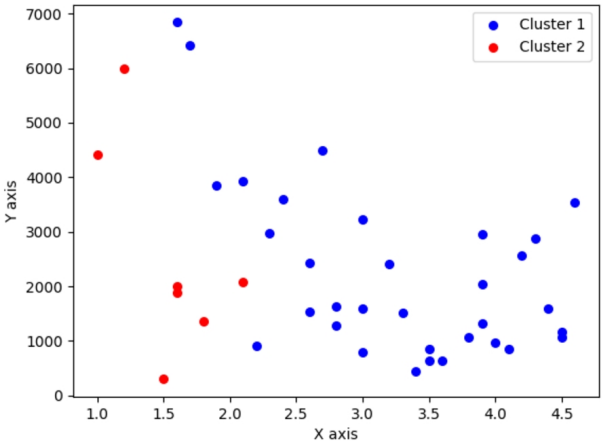


Рисунок 4. Разбиение после третьего перемещения объекта ($R = 0.708$)

где

$$q(A_i, A_j) = \frac{1}{n \cdot m} \sum_{i}^{n} \sum_{j}^{m} \rho(a_i, a_j),$$

$\rho(a_i, a_j)$ есть среднее расстояние между двумя объектами $a_i$ и $a_j$, а $[\cdot]$ – индикаторная функция в нотации Айверсона:

$$[P] = \begin{cases} 1, & \text{если } P \text{ истина;} \\ 0, & \text{если } P \text{ ложь.} \end{cases}$$

Обычно хотят добиться такого разбиения, при котором величина $F_0/F_1$ будет минимальной. Кластеризация объявляется тем лучшей, чем это отношение меньше по величине.

При стартовом разбиении величина среднего внутрикластерного расстояния ($F_0$) была равна 1403.553, а среднего межкластерного расстояния ($F_1$) 2352.518. Получили итоговую оценку начального разбиения $F_0/F_1 = 0.5966$. При выполнении алгоритма было перемещено 3 объекта из второго кластера в первый. Величина среднего внутрикластерного расстояния ($F_0$) равна 1724.196, а среднего межкластерного расстояния ($F_1$) 3081.092. Итоговая оценка качества разбиения равна $F_0/F_1 = 0.5596$.

## 5. Обсуждение

Можно ли уверенно утверждать, что новое разбиение, построенное преобразованием исходного, точно лучше? Ведь среднее внутрикластерное расстояние, на уменьшение которого в основном был направлен алгоритм, увеличилось. Но несмотря на то, что величина $F_0$ возросла, величина $F_1$ также возросла значительным образом, что и дало нам итоговое улучшение оценки качества разбиения конечного множества. Это показывает, что на самом деле механизм улучшения разбиения, рассматриваемый в работе, позволяет учесть совместное изменение этих величин, что является общепринятой практикой при оценке качества разбиения.

Ближайшая цель продолжения начатых в настоящей работе исследований состоит в том, чтобы перенести изложенные идеи на задачу, в которой участвуют более двух показателей. При этом необходимо заменить парный коэффициент корреляции на меру связи между большим числом показателей. Например, вместо максимизации суммы квадратов

коэффициентов корреляции можно попытаться использовать меру изменчивости латентной кластерной переменной внутри каждого из кластеров. Но тогда придется создать способ ее явной оценки, что, разумеется, не так просто сделать.

## Список литературы

1. Lorbeer B., Kosareva A., Deva B. et al. Variations on the Clustering Algorithms // Big Data Research. — 2018. — Vol. 11. — P. 44–53.

2. Everitt B., Landau S., Leese M., Stahl D. Cluster analysis. — Chichester, West Sussex, U.K : Wiley, 2012. — 330 p.

3. Дронов С.В. Анализ многомерных статистических данных: монография. — М. : Инфра-Инженерия, 2025. — 308 с.

4. Каграманян А.Г., Машталир В.П., Скляр Е.В., Шляхов В.В. Метрические свойства разбиений множеств произвольной природы // Доклады Национальной академии наук Украины. — 2007. — № 6. — С. 35–39.

5. Kullback S., Leibler R.A. On information and sufficiency // Annals of Mathematical Statistics. — 1951. — Vol. 22(1). — P. 79–86.

6. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. — 1965. — Т. 163, № 4. — С. 845–848.

7. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. Серия: Информатика и вычислительная биология. — СПб. : Невский Диалект БВХ-Петербург, 2003. — 654 с.

8. Cohen W.W. A comparison of string distance metrics for name-matching tasks // KDD Workshop on Data Cleaning and Object Consolidation. — 2003. — Vol. 3. — P. 73–78.

9. Дронов С.В., Шеларь А.Ю. Новый алгоритм выявления и квантификации латентных классов // Известия АлтГУ. — 2020. — Т. 4, № 11. — С. 81–85.

10. Rindskopf D. Latent Class Analysis // The SAGE Handbook of Quantitative Methods in Psychology. — N.Y. : Sage, 2009. — P. 199–216.