Кластерный анализ

Многомерный анализ

Математические основы, разобранный пример, работа с инструментом

Русский

Вольфганг Бёме

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 ноября 2026 г.

Документация на этом сайте разработана при поддержке ИИ Claude (Anthropic).

Содержание

1 Введение и области применения **2**

2 Математические основы **2**
2.1 Меры расстояния . . . 2
2.2 Иерархические методы . . . 2
2.2.1 Метод Уорда (минимизация дисперсии) . . . 2
2.2.2 Полное связывание (максимальное расстояние) . . . 3
2.2.3 Среднее связывание (среднее расстояние) . . . 3
2.3 Алгоритм k-средних . . . 3
2.4 Выбор числа кластеров . . . 3
2.4.1 Метод локтя . . . 3
2.4.2 Коэффициент силуэта . . . 4

3 Разобранный пример: балансовые показатели **4**
3.1 Исходные данные . . . 4
3.2 Шаг 1: Матрица расстояний (евклидова, стандартизованная) . . . 4
3.3 Шаг 2: Иерархическая кластеризация (Уорд) . . . 5
3.4 Шаг 3: Профили кластеров ($k = 3$) . . . 5
3.5 Шаг 4: Коэффициент силуэта и выбор k . . . 6

4 Работа с интерактивным инструментом кластерного анализа **6**
4.1 Ввод данных . . . 6
4.2 Выбор параметров . . . 7
4.3 Вкладки результатов . . . 7
4.4 Типичный рабочий процесс . . . 8

5 Сводка формул **8**

6 Рекомендуемая литература **9**

Введение и области применения

Кластерный анализ — это метод обучения без учителя, который разбивает наблюдения на однородные группы (*кластеры*) по их сходству — без предварительного знания о структуре групп.

Типичные области применения:

- Сегментация рынка (типологии клиентов)
- Биологические таксономии (классификация видов)
- Кластеризация документов (интеллектуальный анализ текста)
- Обнаружение аномалий (выявление выбросов)
- Обработка изображений (квантование цвета)

Математические основы

Меры расстояния

Евклидово расстояние (стандарт):

$$d_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{1}$$

Манхэттенское расстояние (более устойчиво к выбросам):

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{2}$$

Косинусное расстояние (для пропорциональных паттернов):

$$d_C(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|} \tag{3}$$

Иерархические методы

2.2.1 Метод Уорда (минимизация дисперсии)

Метод Уорда минимизирует прирост **общей внутрикластерной дисперсии**:

$$d_W(C_i, C_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) \tag{4}$$

где $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_j$ — центроиды кластеров, n_i, n_j — их размеры.

2.2.2 Полное связывание (максимальное расстояние)

$$d_{CL}(C_i, C_j) = \max_{a \in C_i,\, b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{5}$$

2.2.3 Среднее связывание (среднее расстояние)

$$d_{AL}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{a \in C_i} \sum_{b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{6}$$

Сравнение методов:

Таблица 1: Свойства иерархических методов

Метод	Форма кластеров	Выбросы	Рекомендация
Уорд	компактные, равные	чувствителен	общий случай
Полное	округлые	умеренно	при выбросах
Среднее	компромисс	устойчиво	общий случай

Алгоритм k-средних

1. **Инициализация** (k-Means++): выбрать первый центроид случайно; последующие центроиды выбирать с вероятностью, пропорциональной $d^2(\mathbf{x}_i, \text{ближайший центроид})$
2. **Назначение**: отнести каждое наблюдение к ближайшему центроиду: $c_i = \arg\min_k d(\mathbf{x}_i, \boldsymbol{\mu}_k)$
3. **Обновление**: пересчитать центроиды: $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$
4. Повторять шаги 2–3 до сходимости

Целевая функция (WCSS — сумма квадратов внутри кластеров):

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \tag{7}$$

Выбор числа кластеров

2.4.1 Метод локтя

Строить WCSS против k; излом («локоть») на графике указывает оптимальное k.

2.4.2 Коэффициент силуэта

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad s_i \in [-1, 1] \tag{8}$$

где a_i — среднее расстояние до наблюдений в собственном кластере, b_i — минимальное среднее расстояние до ближайшего чужого кластера.

Таблица 2: Интерпретация среднего коэффициента силуэта

$\bar{s}$	Интерпретация
≥ 0.70	сильная кластерная структура
$0.50 - 0.69$	разумная структура
$0.25 - 0.49$	слабая структура
< 0.25	структуры нет

Разобранный пример: балансовые показатели

Исходные данные

Используются те же 24 синтетических предприятия U01–U24 с шестью балансовыми показателями, что и в Руководстве 1 (полная таблица данных приведена там): X_1 доля собственного капитала, X_2 ликвидность 2-й степени, X_3 покрытие внеоборотных активов II, X_4 рентабельность продаж, X_5 рентабельность активов, X_6 рост выручки. Настройки инструмента: метод Уорда, евклидово расстояние, Z-преобразование, 3 кластера (по умолчанию).

Цель кластерного анализа: сгруппировать предприятия в однородные типы только по показателям — *без* знания рейтинга. Лишь затем проверяется, насколько найденные кластеры совпадают с рейтингами A/B/C.

Шаг 1: Матрица расстояний (евклидова, стандартизованная)

Показатели z-стандартизуются (средние и стандартные отклонения см. в Руководстве 1); евклидова матрица расстояний $\mathbf{D}$ имеет размер 24×24. Некоторые значения:

$$d(U11, U14) = 0.726, \quad d(U01, U02) = 0.979, \quad d(U01, U12) = 3.007,$$
$$d(U01, U19) = 3.848, \quad d(U12, U19) = 5.514$$

Предприятия одного типа лежат близко друг к другу; дальше всего отстоят высокодоходное предприятие B (U12) и предприятие C в зоне риска (U19).

Шаг 2: Иерархическая кластеризация (Уорд)

Таблица 3: Первые и последние шаги слияния (Уорд)

Шаг	Кластер A	Кластер B	d_W
1	U11	U14	0.513
2	U01	U02	0.692
3	U18	U19	0.717
…	…	…	…
20	{U17, U20}	{U03, U18, U19, U21–U24}	2.468
21	{U13, U16}	{U09–U12, U14, U15}	2.523
22	{U01, U02, U04–U08}	{U09–U16}	5.972
23	{U03, U17–U24}	{U01, U02, U04–U16}	7.652

Пример, шаг 1: $d_W(U11, U14) = \sqrt{\frac{1 \cdot 1}{2}}\, d_E(U11, U14) = 0.7071 \cdot 0.726 = 0.513$.

До шага 21 расстояния слияния растут медленно (не более 2.52). Затем следует большой скачок до 5.97 и 7.65: данные **естественным образом делятся на 3 кластера**. Срез при $k = 3$ (по умолчанию) проходит между шагами 21 и 22 и выделяет именно эти три группы.

Шаг 3: Профили кластеров ($k = 3$)

Таблица 4: Средние значения кластеров (исходные данные, в %)

Кластер	X_1 Д. СК	X_2 Ликв.	X_3 Покр.	X_4 Р. пр.	X_5 Р. акт.	X_6 Рост
1 ($n = 7$)	47.01	130.29	128.71	5.33	7.59	5.30
2 ($n = 9$)	28.44	87.11	90.11	2.20	4.64	0.18
3 ($n = 8$)	34.12	99.38	96.75	8.55	11.16	11.71
Итого ($n = 24$)	35.75	103.79	103.58	5.23	7.68	5.52

Интерпретация:

- Кластер 1 (U01, U02, U04–U08) — **устойчивые предприятия с крепкой капитальной базой**: устойчивость значительно выше среднего, доходность средняя.
- Кластер 2 (U03, U17–U24) — **предприятия в зоне риска**: ниже среднего по всем шести показателям, выручка стагнирует.
- Кластер 3 (U09–U16) — **высокодоходные растущие предприятия**: наибольшая рентабельность и темпы роста при несколько пониженной устойчивости.

Таблица 5: Сравнение кластеров и рейтингов

	Рейтинг A	Рейтинг B	Рейтинг C
Кластер 1	**7**	0	0
Кластер 2	1	0	**8**
Кластер 3	0	**8**	0

23 из 24 предприятий попадают в кластер «своего» рейтинга, хотя метод рейтинга не знает. Исключение — **U03** (рейтинг A): при доле собственного капитала 32.0 %, покрытии активов 100 % и рентабельности активов всего 5.0 % оно ниже среднего почти по всем показателям. Его расстояние до центра кластера 2 (1.76 без U03) меньше, чем до центра кластера 1 (2.45). Для практики это означает: либо рейтинг A основан на информации, не отражённой в шести показателях, либо это предупреждающий сигнал о возможном понижении рейтинга.

Шаг 4: Коэффициент силуэта и выбор k

$\bar{s} = 0.441$ при $k = 3$ — по таблице 2 **слабая структура** (в инструменте оценивается как «приемлемо», пороги 0.5/0.25). Лучшие индивидуальные значения у U02 ($s_i = 0.593$) и U22 (0.588), худшие — у U07 (0.195) и U03 (0.212): это те же пограничные случаи, которые выделяются в МГК (Руководство 1) и в дискриминантном анализе (Руководство 4).

Таблица 6: Силуэты и WCSS для различных k

	$k = 2$	$k = 3$	$k = 4$
$\bar{s}$ (Уорд)	0.366	**0.441**	0.381
WCSS (k-средних)	79.45	43.79	37.31

$k = 3$ даёт наибольший коэффициент силуэта, а график локтя k-средних имеет чёткий излом при $k = 3$ (WCSS снижается с 79.45 до 43.79, затем лишь до 37.31). Метод k-средних при $k = 3$ даёт то же разбиение, что и метод Уорда. Все критерии согласованно указывают на три типа предприятий.

Работа с интерактивным инструментом кластерного анализа

Ввод данных

Помимо стандартных вкладок (Вручную, CSV, Примеры), инструмент кластерного анализа имеет **столбец меток** («Метка») слева в таблице ручного ввода. Здесь можно вводить содержательные названия наблюдений, которые отображаются в дендрограмме и диаграмме рассеяния.

По умолчанию загружены балансовые показатели (24 предприятия U01–U24, 6 пер.). Дополнительные примеры: Iris (15 набл., 4 пер.), профиль городов (12 городов, 4 пер.) и сегментация клиентов (20 клиентов, 4 пер.). Для каждого примера над таблицей ввода появляется пояснительный блок (данные, цель анализа,

ожидаемый результат); он исчезает при нажатии «Очистить» или загрузке CSV-файла.

Выбор параметров

Метод
Иерархический: Уорд (стандарт), полное связывание, среднее связывание. Неиерархический: k-средних. При выборе k-средних вкладка дендрограммы скрывается.

Мера расстояния
Евклидово (стандарт), манхэттенское, косинусное сходство. Метод Уорда использует выбранное расстояние между центрами кластеров; теоретически он обоснован только для евклидова расстояния. Метод k-средних всегда работает с евклидовым расстоянием (выбор блокируется). Коэффициент силуэта использует выбранную меру расстояния.

Стандартизация
Z-преобразование настоятельно рекомендуется при разных единицах измерения.

Число кластеров (иерархические)
Определяет уровень среза дендрограммы. Золотая пунктирная линия проходит между двумя уровнями слияния, на которых срезается дерево; ветви ниже среза зелёные, выше — красные.

k и kmax (k-средних)
k: желаемое число кластеров. $k_{\max}$: верхняя граница для графика локтя и анализа силуэта (до 12).

Вкладки результатов

Обзор
Показатели (n, p, кластеры, силуэт) + полная таблица назначений.

Дендрограмма
Дендрограмма с линией среза; наблюдения расположены в порядке дерева и окрашены по назначенному кластеру. Доступна только для иерархических методов.

Графики
k-средних: график локтя (WCSS) и столбцы силуэта по k. Оба метода: диаграмма рассеяния первых двух переменных (стандартизованных), наблюдения окрашены по кластеру.

Профиль кластеров
Таблица средних значений по кластерам (исходные данные) + радарная диаграмма со стандартизованными профилями кластеров.

Интерпретация
Автоматическая оценка: коэффициент силуэта, подсказки по дендрограмме, интерпретация локтя k-средних.

Экспорт

CSV: назначение в кластеры, профиль кластеров. PNG: диаграмма рассеяния, радарная диаграмма, дендрограмма.

Типичный рабочий процесс

1. Загрузить данные, проверить стандартизацию
2. Выбрать метод Уорда и евклидово расстояние
3. Установить число кластеров 3 (или скорректировать после осмотра дендрограммы)
4. Нажать **«Вычислить кластерный анализ»**
5. Вкладка «Дендрограмма»: большие расстояния слияния — признак естественных границ
6. Проверить значение силуэта ($\bar{s} \geq 0.5$ = хорошо)
7. При необходимости подтвердить k-средних с помощью графика локтя
8. Вкладка «Профиль кластеров»: содержательная интерпретация групп

Важные замечания:

- Требуется минимум $n = 4$ полных наблюдения.
- k-средних: 10 повторных запусков (k-Means++) предотвращают локальные минимумы.
- Метод Уорда предполагает сферические кластеры примерно одинакового размера.
- При $p > 2$ диаграмма рассеяния отображает только первые две переменные.
- Дендрограмма: скачать как PNG для публикаций.

Сводка формул

Таблица 7: Формулы кластерного анализа

Обозначение	Формула
Евклидово расстояние	$d_E = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$
Манхэттенское расстояние	$d_M = \sum_k \lvert x_{ik} - x_{jk} \rvert$
Расстояние слияния Уорда	$d_W = \sqrt{\frac{n_i n_j}{n_i + n_j}}\, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$
Косинусное расстояние	$d_C = 1 - \mathbf{x}_i^\top \mathbf{x}_j / (\lVert\mathbf{x}_i\rVert \, \lVert\mathbf{x}_j\rVert)$
WCSS (k-средних)	$\sum_k \sum_{i \in C_k} \lVert\mathbf{x}_i - \boldsymbol{\mu}_k\rVert^2$
Коэфф. силуэта	$s_i = (b_i - a_i) / \max(a_i, b_i)$

Рекомендуемая литература

- Kaufman, L. & Rousseeuw, P. J. (2005). *Finding Groups in Data*. Wiley.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.
- Arthur, D. & Vassilvitskii, S. (2007). k-Means++: The Advantages of Careful Seeding. *SODA 2007*, 1027–1035.