

Сравнение бикластеризации расстоянием  
Кульбака-Лейблера (KL) и взвешенным  
расстоянием Брэгмана

Выполнил:  
студент группы 20.Б06-мм  
Коротченок Остап Андреевич

Научный руководитель:  
Старший научный сотрудник, Ph.D.  
Мокаев Руслан Назирович

Отметка о зачете:

# Содержание

|     |  |   |
|-----|--|---|
| 1   | Введение   | 1 |
| 1.1 | Кластерный анализ . . . . .                                | 1 |
| 1.2 | Бикластеризация . . . . .                                  | 1 |
| 1.3 | Цели и задачи работы . . . . .                             | 2 |
| 2   | Расстояния и меры сходства                                 | 2 |
| 2.1 | Расстояние Брегмана . . . . .                              | 2 |
| 2.2 | Расстояние Кульбака-Лейблера (KL) . . . . .                | 3 |
| 2.3 | Взвешенное расстояние Брегмана . . . . .                   | 3 |
| 3   | Постановка задачи бикластеризации                          | 4 |
| 4   | EM-алгоритм  | 4 |
| 4.1 | Raw-реализация EM-алгоритма . . . . .                      | 4 |
| 4.2 | Soft-реализация EM-алгоритма . . . . .                     | 4 |
| 5   | Реализация   | 6 |
| 5.1 | Применение EM-алгоритма к задаче Бикластеризации . . . . . | 6 |
| 5.2 | Инициализация параметров . . . . .                         | 6 |
| 5.3 | Оценка результата алгоритма . . . . .                      | 7 |
| 5.4 | Генерация синтетических данных . . . . .                   | 7 |
| 6   | Результаты Работы  | 7 |
| 7   | Источники  | 9 |

## 1 Введение

### 1.1 Кластерный анализ

Кластерный анализ — это метод анализа данных, который позволяет выделить схожие группы (кластеры) объектов на основе их сходства или близости по определенным признакам. Он применяется во многих областях, таких как статистика, биология, социология, маркетинг, машинное обучение и другие. Он может использоваться для исследования данных, поиска закономерностей и образования групп объектов с похожими характеристиками.

### 1.2 Бикластеризация

Бикластеризация - это метод кластерного анализа, который позволяет одновременно группировать как строки, так и столбцы матрицы данных. Он используется для обнаружения подмножеств объектов и признаков, которые сильно связаны между собой, но слабо связаны с остальной частью матрицы.

Бикластеры обычно используются для анализа геномных данных, анализа текстов и других типов данных, где есть ярко выраженная структура.

### 1.3 Цели и задачи работы

Цель данной курсовой работы - исследовать и сравнить два расстояния, используемых в бикластеризации: KL-расстояние и взвешенное расстояние Брэгмана. Для этого будут решены следующие задачи:

- Изучить теоретические аспекты бикластеризации и расстояний, используемых в этом методе.
- Сравнить KL-расстояние и взвешенное расстояние Брэгмана на синтетических примерах.

## 2 Расстояния и меры сходства

Для кластерного анализа необходимо иметь меру сходства или расстояние между объектами. Эта мера позволяет определить, насколько близки или похожи два объекта или группы объектов. В бикластеризации также необходимо иметь меру сходства между признаками.

### 2.1 Расстояние Брэгмана

Дивергенция Брэгмана - это класс функций расстояния между двумя точками в выпуклом пространстве, определяемый через выпуклую функциогенератор, которая удовлетворяет некоторым свойствам.

Пусть  $\varphi$  - выпуклая дифференцируемая функция на выпуклом замкнутом множестве  $S$  в  $\mathbb{R}^n$ . Для двух точек  $x, y$  в  $S$  дивергенция Брэгмана  $D_\varphi(x, y)$  определяется следующим образом:

$$D_\varphi(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle, \quad (1)$$

где  $\nabla \varphi(y)$  обозначает градиент функции  $\varphi$  в точке  $y$ .

Некоторые свойства расстояния Брэгмана:

- Неотрицательность:  $D_\varphi(x, y) \geq 0$  для всех  $x, y \in S$ .
- Невырожденность:  $D_\varphi(x, y) = 0$  тогда и только тогда, когда  $x = y$ .

Некоторые примеры расстояния Брэгмана:

- Квадрат Евклидова расстояния:  $\varphi(x) = \|x\|^2$

$$D_\varphi(x, y) = \|x - y\|^2$$

- Расстояние Кульбака-Лейблера:  $\varphi(p) = \sum_i p_i \log p_i$

$$D_\varphi(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$$

- Расстояние Итакуры-Сайто:  $\varphi(x) = \sum_i \log x_i$

$$D_\varphi(x, y) = \sum_i \left( \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right)$$

## 2.2 Расстояние Кульбака-Лейблера (KL)

Расстояние Кульбака-Лейблера (KL) является мерой расстояния между двумя вероятностными распределениями.

Если мы имеем вероятностные распределения  $P$  и  $Q$ , то KL-расстояние между ними вычисляется по следующей формуле:

$$KL(P\|Q) = \sum_{i=1}^n P(i) \log \frac{P(i)}{Q(i)}$$

где  $n$  - количество элементов в распределении,  $P(i)$  и  $Q(i)$  - вероятности  $i$ -го элемента в распределениях  $P$  и  $Q$  соответственно.

Одним из основных свойств расстояния KL является то, что оно не является симметричным, т.е.  $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$ . Кроме того, расстояние KL не является метрикой, т.к. оно не удовлетворяет неравенству треугольника.

В контексте машинного обучения, KL-расстояние часто используется для оценки различий между распределениями вероятности, например, в задачах классификации или кластеризации. Оно также используется в задачах оптимизации, например, для построения моделей с использованием метода максимального правдоподобия.

Кроме того, KL-расстояние может быть использовано в качестве меры качества бикластеризации. Мы можем рассчитать KL-расстояние между вероятностными распределениями признаков внутри каждого бикластера и между разными бикластерами. Затем мы можем суммировать полученные значения для всех бикластеров и использовать результат как меру качества бикластеризации. Если KL-расстояние между признаками внутри бикластера мало, а между разными бикластерами велико, то это может свидетельствовать о хорошем качестве бикластеризации.

## 2.3 Взвешенное расстояние Брегмана

Взвешенное расстояние Брегмана (weighted Bregman divergence) является обобщением расстояния Брегмана для случая, когда входные данные имеют веса. В случае  $N$  точек в  $n$ -мерном пространстве, с весами  $w_1, w_2, \dots, w_N$ , и соответствующими  $n$ -мерными векторами  $x_1, x_2, \dots, x_N$ , взвешенное расстояние Брегмана между  $x_i$  и  $x_j$  задается следующим образом:

$$D_w(x_i, x_j) = \sum_{k=1}^n w_k D_{\varphi_k}(x_i^{(k)}, x_j^{(k)})$$

где  $D_{\varphi_k}(x_i^{(k)}, x_j^{(k)})$  - расстояние Брегмана между  $k$ -ой координатой  $x_i$  и  $x_j$ , заданное выпуклой и дифференцируемой функцией  $\varphi_k$ .  $x_i^{(k)}$  и  $x_j^{(k)}$  обозначают соответствующие координаты векторов  $x_i$  и  $x_j$ .

Веса  $w_k$  могут быть использованы для учета различной значимости различных координат при анализе данных.

Для задания весов в контексте задачи бикластеризации можно использовать различные подходы:

- Использовать априорную информацию о важности каждой измеряемой величины. Например, если мы знаем, что некоторые гены или образцы более важны, чем другие, мы можем задать больший вес соответствующим строкам и столбцам матрицы расстояний.

- Использовать стандартное отклонение значений каждой измеряемой величины в качестве веса. Чем больше стандартное отклонение, тем больший вес мы можем присвоить соответствующим строкам и столбцам матрицы расстояний.
- Использовать методы машинного обучения для определения оптимальных весов. Например, можно обучить модель машинного обучения для определения важности каждой из измеряемых величин.
- Использовать экспертные знания для определения весов. Например, если мы знаем, что некоторые гены играют важную роль в конкретном биологическом процессе, мы можем присвоить им больший вес, чем другим генам.

### 3 Постановка задачи бикластеризации

Пусть имеется матрица  $X$  размера  $n \times m$ , где  $n$  - число объектов,  $m$  - число признаков. Требуется разбить матрицу  $X$  на  $k$  бикластеров, то есть подматриц размера  $n_k \times m_k$ , где  $n_k \leq n$  и  $m_k \leq m$ , таких что каждый бикластер описывается своими центроидами - векторами средних значений по каждому признаку внутри бикластера. Таким образом, бикластеризация позволяет выделить группы объектов с похожими признаковыми характеристиками.

Задачу жесткой кластеризации с помощью расстояния Брэгмана можно формулировать через задачу квантования с минимизацией потери информации Брэгмана, равной ожидаемому расстоянию Брэгмана от данных до центроидов соответствующих кластеров.

## 4 ЕМ-алгоритм

ЕМ-алгоритм (Expectation-Maximization) – это итеративный алгоритм, используемый для оценки параметров статистических моделей, когда некоторые из переменных модели скрыты и не могут быть наблюдаемы напрямую.

Одним из применений ЕМ-алгоритма является решение задачи бикластеризации. В статьях в [1] и [2] было теоретически обоснованно, что подобный алгоритм будет на каждой итерации уменьшать целевую функцию, пока не окажется в локальном минимуме.

### 4.1 Raw-реализация ЕМ-алгоритма

В raw-реализации ЕМ-алгоритма на шаге Expectation каждый объект сопоставляется одному из кластеров. На шаге Maximization центроиды кластеров пересчитываются как среднее среди объектов, принадлежащих соответствующим кластерам.

Псевдокод: 1.

### 4.2 Soft-реализация ЕМ-алгоритма

В soft-реализации ЕМ-алгоритма вместо жесткого присвоения каждого элемента одному из бикластеров вычисляются вероятности принадлежности элемента к каждому бикластеру.

Псевдокод: 2.

---

**Algorithm 1 EM Raw Clustering**

---

Initialize  $\mu_h$  for  $h = 1, \dots, k$   
repeat  
    for each  $x$  do  
        Assign  $x$  to its nearest cluster  $X_h$  where  
$$h = \operatorname{argmin}_{h'} D_\varphi(x, \mu_h)$$
  
    end for  
    for each  $h$  do  
        Recompute mean  $\mu_h$  as  
$$\mu_h = \frac{\sum_{x \in X_h} x}{n_h}$$
  
    end for  
until convergence

---

---

**Algorithm 2 EM Soft Clustering**

---

Initialize  $\pi_h, \mu_h$  for  $h = 1, \dots, k$   
repeat  
    for each  $x, h$  do  
        Compute posterior probability  
$$p(x|h) = \pi_h \exp -D_\varphi(x, \mu_h)/Z(x)$$
  
    end for  
    for each  $h$  do  
        Recompute mean  $\pi_h$  and  $\mu_h$   
$$\pi_h = \frac{1}{n} \sum_x p(x|h)$$
  
$$\mu_h = \frac{\sum_x p(x|h)x}{\sum_x p(x|h)}$$
  
    end for  
until convergence

---

## 5 Реализация

К работе приложен код на языке Python с реализацией описанных выше алгоритмов для решения задачи бикластеризации. Каждый из алгоритмов можно запустить с указанием метрики - обобщенном расстоянием Кульбака-Лейблера или взвешенным расстоянием Брегмана, где для сравнения двух чисел используется квадрат евклидовой нормы, а веса считаются на основе стандартного отклонения значений каждой измеряемой величины.

Обобщенное расстояние Кульбака-Лейблера:

$$D_{KL}(x, y) = \sum_i x_i \log \frac{x_i}{y_i} - \sum_i x_i + \sum_i y_i$$

Веса рассчитываются по следующей схеме. Пусть у нас есть матрица  $X$  размера  $n \times m$ , где каждый элемент  $x_{ij}$  соответствует некоторой измеряемой величине. Мы можем вычислить стандартное отклонение для каждого столбца этой матрицы, используя следующую формулу:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \hat{x}_j)^2}{n - 1}}$$

где  $\hat{x}_j$  - среднее значение  $j$ -го столбца.

В таком случае веса можно задать по формуле:

$$w_j = \frac{1}{\sigma_j}$$

Чем больше стандартное отклонение  $\sigma_j$ , тем меньше будет вес  $w_j$  соответствующего столбца. Таким образом, если значения в столбце имеют большое разнообразие, то этот столбец будет иметь меньший вес в расчете взвешенного расстояния Брегмана. Если значения в столбце более однородны, то столбец будет иметь больший вес.

### 5.1 Применение ЕМ-алгоритма к задаче Бикластеризации

В приложенном к работе коду ЕМ-алгоритм последовательно применяется сначала для кластеризации строк дата-матрицы, а затем по столбцам кластеризуются центроиды, полученные на предыдущем этапе. Таким образом, если дата-матрица содержит  $n$  строк,  $m$  столбцов,  $h$  кластеров по строкам и  $k$  кластеров по столбцам, то асимптотическая сложность реализованного алгоритма будет  $O(nmk + kmh)$ .

### 5.2 Инициализация параметров

Для raw и soft версий алгоритма центроиды инициализировались по-разному. В случае raw алгоритма сначала каждой строке(столбцу) сопоставлялся случайный кластер, затем вычислялись центроиды каждого кластера как среднее значение среди попавших в кластер строк(столбцов). В случае soft алгоритма центроиды вычислялись как случайная строка(столбец) дата-матрицы. В рамках данной работы не будет формально обоснован такой выбор задания начальных значений алгоритма, но на синтетических данных это приводило к лучшим результатам.

### 5.3 Оценка результата алгоритма

Для оценки результата работы алгоритма использовалась `consensus_score` из библиотеки `sklearn` - это функция, которая вычисляет показатель согласованности (`consensus score`) между несколькими различными бикластеризациями (исходной и полученной в результате работы алгоритма) для одних и тех же данных.

Функция вычисляет все возможные попарные комбинации бикластеризаций из списка, сравнивает бикластеры в каждой паре с помощью указанной метрики сходства и вычисляет показатель согласованности как среднее значение метрики по всем попарным сравнениям.

Значение показателя согласованности находится в диапазоне от 0 до 1, где 0 означает полное отсутствие согласованности между бикластеризациями, а 1 - полную согласованность.

### 5.4 Генерация синтетических данных

На основе `sklearn.datasets.make_checkerboard` была реализована функция для создания дата-матриц заданного размера и количества кластеров. Отличие заключается в том, что средние значение для каждой подматрицы создавалось с заданным распределением (Пуассоновым, Гауссовым, Мультиномиальным и Равномерным). Это делалось из предположения, что алгоритмы будут иметь разную эффективность в зависимости от выбранной метрики и распределения. Также матрицы не содержали неположительных значений, чтобы избежать деления на 0 или отрицательных значений в аргументе логарифма.

## 6 Результаты Работы

Для иллюстрации работы алгоритмов приведены примеры бикластеризации матриц  $300 \times 300$  с 5-ю кластерами по строкам и столбцам. Алгоритмы запускались по 20 раз для каждой матрицы, затем брался лучший результат.

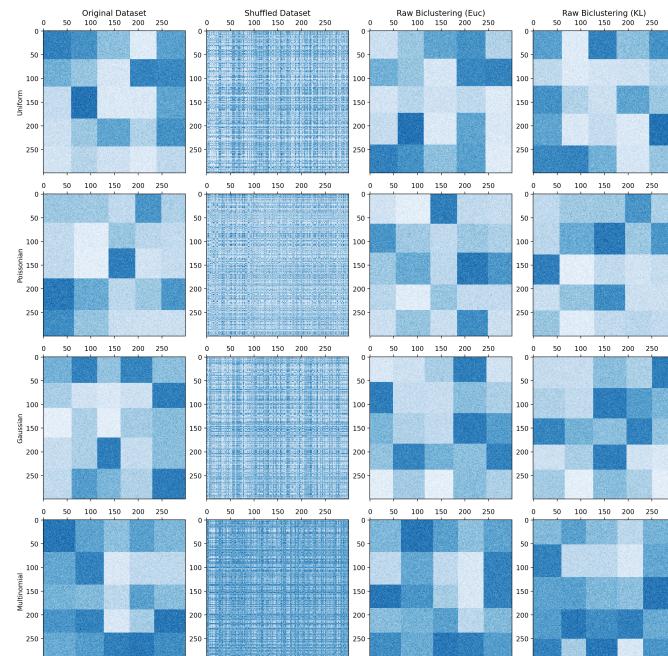


Рис. 1: Пример работы raw алгоритма

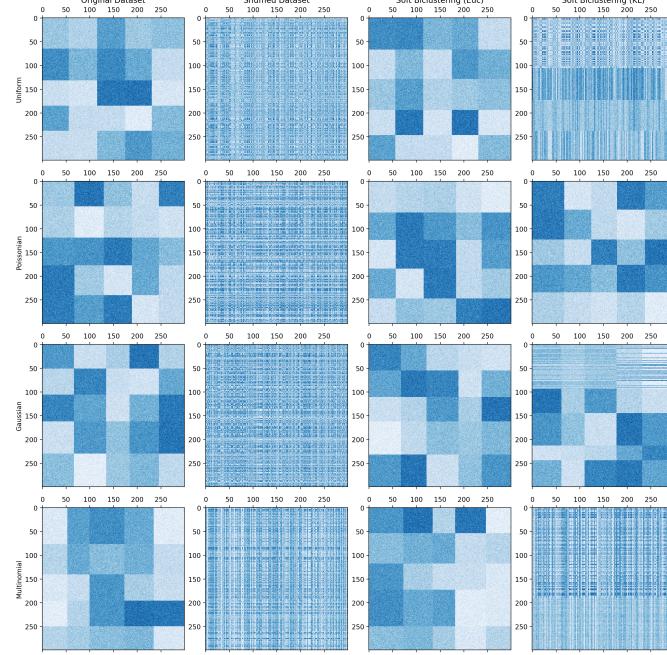


Рис. 2: Пример работы soft алгоритма

Далее представлены результаты последовательного запуска алгоритма на 50-и дата-матрицах каждого из типов с размерностью  $50 \times 50$ , 5-ю кластерами по строкам и столбцам. Для каждой матрицы данных алгоритмы запускались по 5 раз и из результатов выбирался лучший. В ячейках представлены средние значения consensus\_score за 50 экспериментов.

Таблица 1: Сравнение результатов

|            | Poissonian | Gaussian | Multinomial | Uniform |
|------------|------------|----------|-------------|---------|
| Raw (Euc)  | 0.788      | 0.971    | 0.972       | 0.948   |
| Raw (KL)   | 0.786      | 0.954    | 0.942       | 0.949   |
| Soft (Euc) | 0.769      | 0.814    | 0.822       | 0.820   |
| Soft (KL)  | 0.615      | 0.834    | 0.875       | 0.858   |

Как видно из таблицы, выбор KL-расстояния или взвешенного расстояния Брегмана не приводит заметным различиям в эффективности работы raw-версии алгоритма. В случае soft-реализации можно заметить, что KL-расстояние наиболее эффективно в том случае, когда средние значения в дата-матрице имеют мультиномиальное или равномерное распределение, в то время как взвешенное расстояние Брегмана приблизительно одинаково эффективно для вышеупомянутых и нормального распределения. Как было показано в статьях [1] и [2], кластеризация с использованием дивергенции Брегмана, соответствующей используемому распределению в генеративной модели, должно давать лучшие результаты. KL-расстоянию соответствует Мультиномиальное распределение, квадрату Евклидова расстояния соответствует Гауссово распределение. Неполное соответствие упомянутым результатам может быть связано с неправильно подобранными гиперпараметрами при генерации дата-матрицы и добавления шума.

## 7 Источники

### Список литературы

- [1] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.
- [2] Inderjit S Dhillon, Siva Mallela, and Rahul Kumar. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, 2007.