Бикластеризация данных изображающей масс-спектрометрии

Демидов Роман

СПБАУ НОЦНТ РАН

5 июня 2014

Outline

- *Масс-спектрометрия* - распространенный метод изучения химического состава веществ.
- Технология *МАЛДИ* (матрично-активная лазерная десорбция/ионизация) - относительно новый метод ИМС (изображающей масс-спектрометрии).
- Применяется в биохимическом анализе тканей, поиске биомаркеров, фармацевтике, бактериологии.

- В каждой точке образца измеряются интенсивности наблюдения частиц с различными $\frac{m}{z}$.
- Набор значений $\frac{m}{z}$ - общий для всех точек.
- Двойственность представления данных на выходе МАЛДИ-спектрометра:
  - Как *одно многоканальное изображение*;
  - Как *набор одноканальных* $\frac{m}{z}$*-изображений*.

- Ранее решенные задачи в данной области:
  - Задача *сегментации* данных ИМС;
  - Задача *кластеризация* $\frac{m}{z}$*-изображений* по пространственной похожести.
- Задача *бикластеризации*:
  сгруппировать в несколько блоков (бикластеров) одновременно и точки, и $\frac{m}{z}$-значения так, чтобы внутри каждого блока интенсивности его $\frac{m}{z}$-значений были максимальны.
- Необходимо также учесть пространственную близость точек.

- Были разработаны 2 различных метода решения задачи.
- Реализация методов выполнена в системе MATLAB и на языке C++.
- Методы испытывались на результатах измерения MALDI-TOF спектрометром Bruker Daltonik GmbH плоского среза коронарного отдела мозга крысы.
- Данные: 3045 значений интенсивностей (в интервале масс частиц 2.5-10 кДа) для каждой из 20185 точек образца.

Outline

- Строится неориентированный взвешенный граф:
  - Вершинам графа соответствуют точки и $\frac{m}{z}$-значения.
  - Каждой вершине i сопоставляется неотрицательный вес $weight(i) \geq 0$.
  - Ребрами соединяются *точки и спектры*, а также некоторые *точки между собой*.
  - Вес e_{ij} каждого ребра между точкой и $\frac{m}{z}$-значением - измеренная интенсивность наблюдения данного $\frac{m}{z}$ в данной точке.
- Цель - найти “хорошее” разбиение вершин на 2 (пока) блока, минимизирующее некоторую меру качества.

- Мера качества разбиения вершин на 2 блока, V_1 и V_2 - **балансированный разрез**:

$$balanced\_cut(V_1, V_2) = \frac{cut(V_1, V_2)}{\sum_{v \in V_1} weight(v)} + \frac{cut(V_1, V_2)}{\sum_{v \in V_2} weight(v)}$$

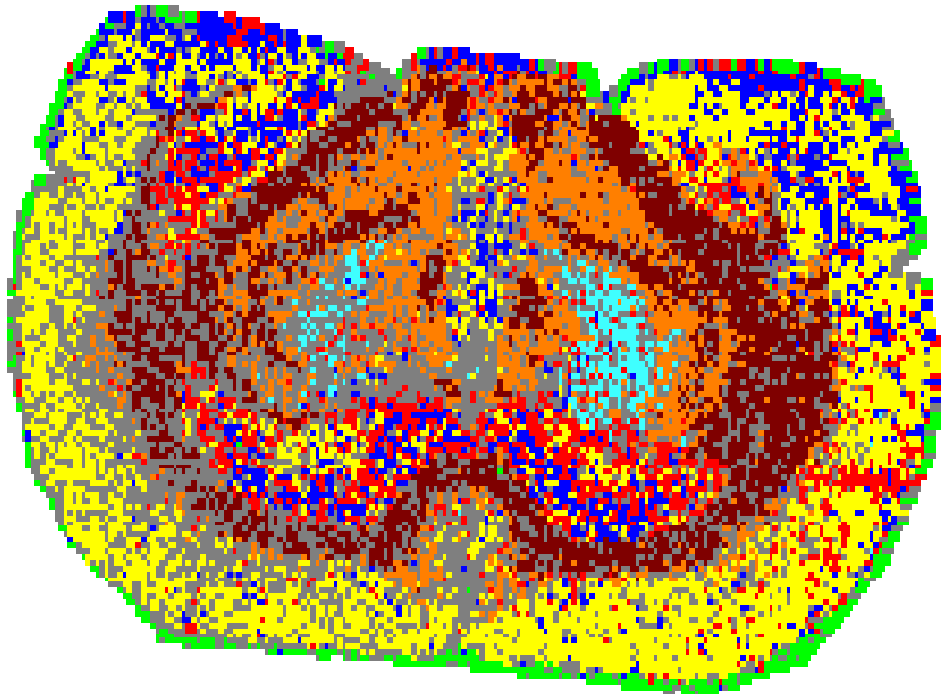
- Найти **точное** решение, минимизирующее балансированный разрез - **NP-трудная задача**.
- Ищем **приближенное** решение - оно достигается путем решения ssадачи:

$$\min_{q \neq 0} \frac{q^T L q}{q^T W q}, \; q^T W e = 0, e_i = 1 \; \forall i,$$

где L - матрица, построенная по весам ребер, W - матрица, построенная по весам вершин.

- Доказано, что стационарные точки выражения $\frac{q^T L q}{q^T W q}$ при условиях $q^T W e = 0, Le = 0, e^T e = 1$ - собственные векторы $z_i, i \geq 2$ задачи $Lz = \lambda Wz$.
- $\lhd$ собственные векторы, соответствующие наименьшим собственным числам(кроме вектора e).
- Каждый такой собственный вектор позволяет находить разбиение вершин на 2 блока, отвечающее малому значению балансированного разреза.
- Для разбиения на N блоков - берем $r \geq \lceil log_2(N) \rceil$ собственных векторов $z_2, z_3, ..z_{r+1}$, рассматриваем их координаты для каждой вершины как точку в $\mathbb{R}^r$; алгоритмом K-средних группируем точки в N бикластеров.

- Оставляем граф двудольным(ребра только между точками и $\frac{m}{z}$-значениями) - картина бикластеризации будет искаженной.
- Причины - шум и неточности в исходных данных, погрешность приближеннного решения.



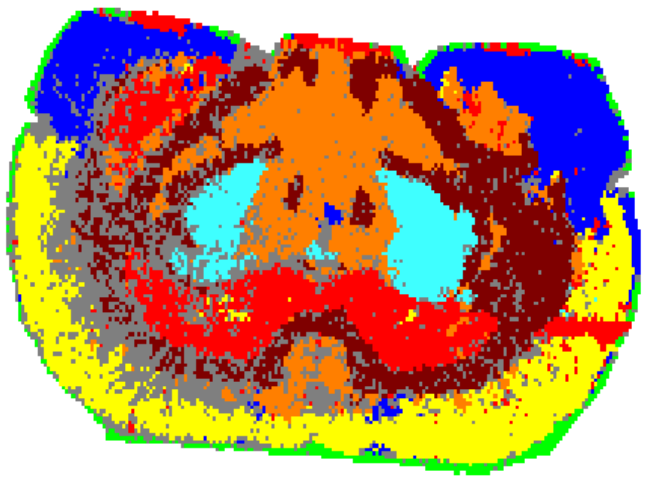
- Дополним граф G ребрами e_{ij}, связывающими пространственно близкие точки i и j графа с похожими спектрами:

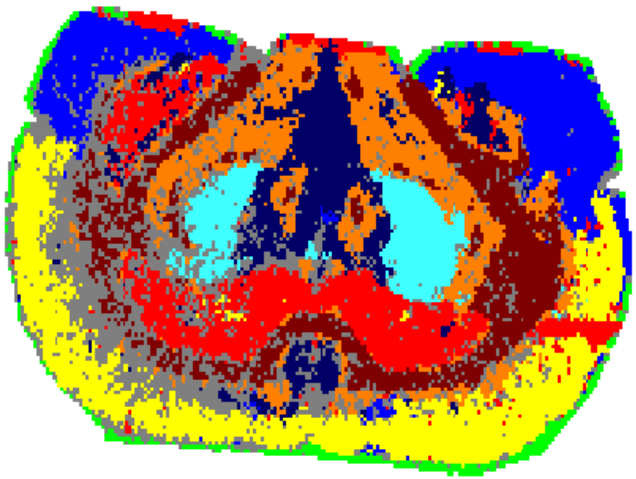
$$e_{ij} = \left. \frac{max\_weight * similarity(i,j)}{dist(i,j)^{degree}} \right|_{dist(i,j)<d}$$

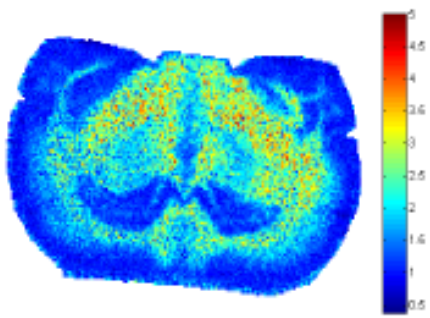
$$weight(i)_{i \in Spectors} = \sum_{j} M'_{ij}$$

$$weight(i)_{i \in Points} = \sum_{\substack{j:j \in Pixels, \\ dist(i,j) \leq d \\ similarity(i,j)<r}} M'_{ij} + \sum_{j:j \in Spectors} M'_{ij}$$

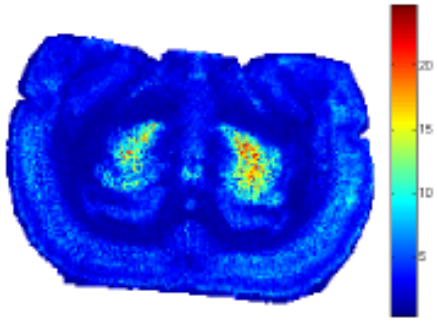
- $similarity(i, j)$ - “грубая” мера похожести спектров двух точек.
- $dist(i, j)$ - расстояние между точками.
- Результат для 8 и 9 бикластеров:



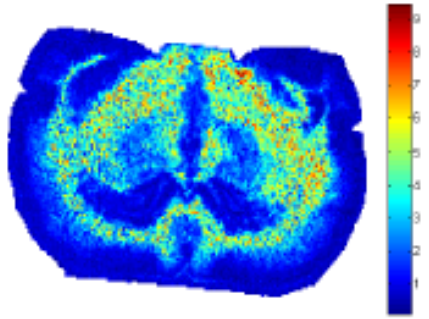




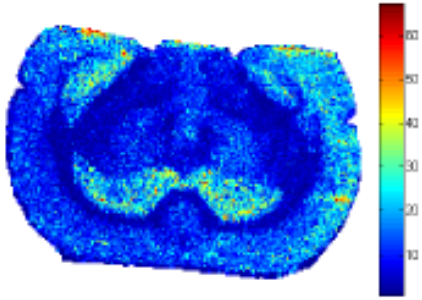
1(оранжевый)



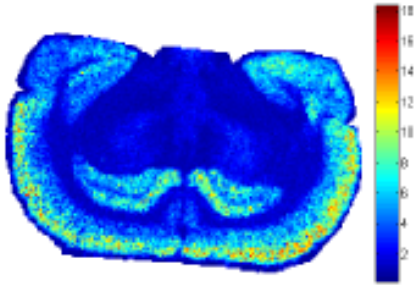
2(салатовый)



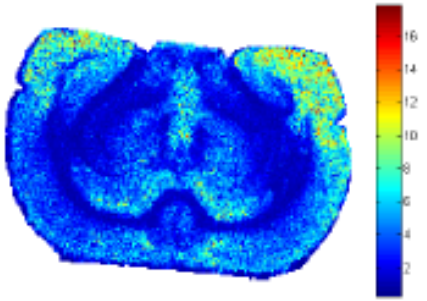
3(коричневый)



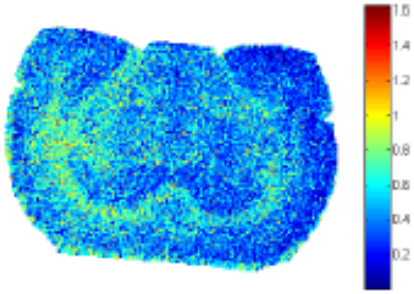
4(красный)



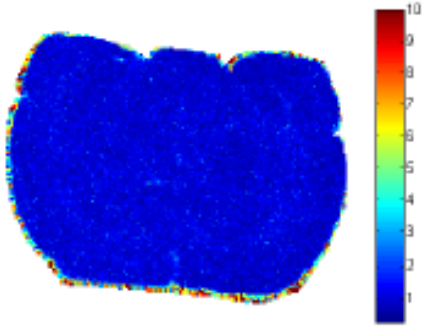
5(жёлтый)



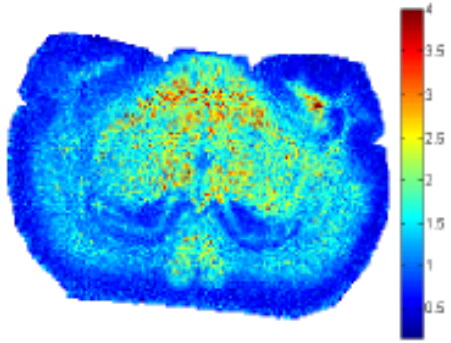
6(синий)



7(серый)



8(зелёный)



8(Темно-синий)

Outline

- Основан на графической вероятностной модели LDA(Latent Dirichlet Allocation).
- Ранее применялась для анализа текстов.
- Есть “темы”(бикластеры), их число фиксировано.
- “Документы”(точки) - неупорядоченный набор “слов”(яркие $\frac{m}{z}$ в точке).
- Документ - смесь “тем”. “Тема” - распределение на “словах”.

- LDA описывает вероятностный процесс порождения документов независимо слово за словом.
- Совместная распределение случайных переменных, отвечающих параметрам модели:

$$Pr(\{z_{nd}\},\{w_{nd}\},\{\theta_d\},\{\beta_k\}|\alpha,\eta) =$$

$$= \prod_{k=1}^{K} Pr(\beta_k|\nu) \prod_{d=1}^{D} Pr(\theta_d|\alpha) \prod_{n=1}^{N} Pr(z_{nd}|\theta_d) Pr(w_{nd}|\beta, z_{nd})$$

- $\{\theta_d\}$ - распределения на темах для каждого документа,
- $\{z_{nd}\}$ - номер темы для слова, из которой оно порождается,
- $\{w_{nd}\}$ - слова,
- β - распределения на словах для каждой темы.

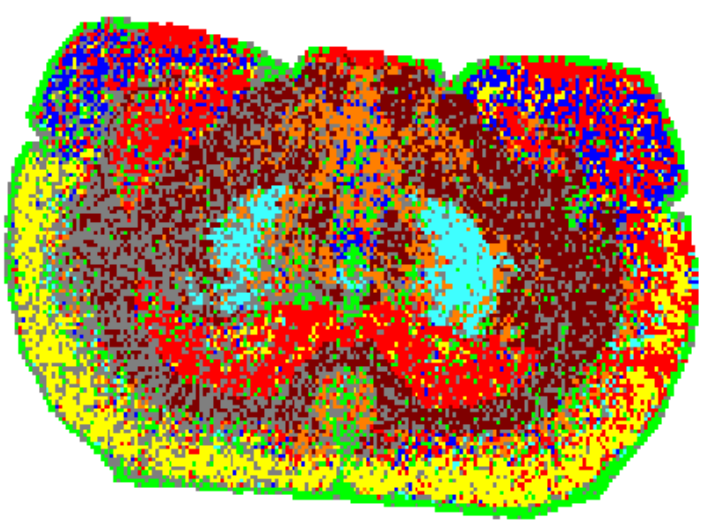
- LDA задает совместное распределение на наблюдаемых $\{w_{nd}\}$ и скрытых $\{\theta_d\}, \{\beta_k\}, \{z_{nd}\}$.
- Задача вывода - **найти максимальное апостериорное распределение скрытых переменных при условии наблюдаемых**:

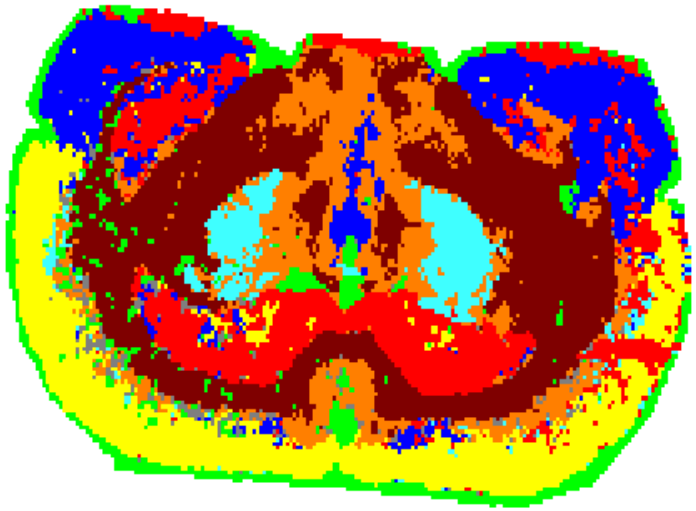
$$\beta^*, \theta^*, z^* = \underset{\beta,\theta,z}{arg\ max}\ Pr(\beta, \theta, z | w.\alpha, \eta)$$

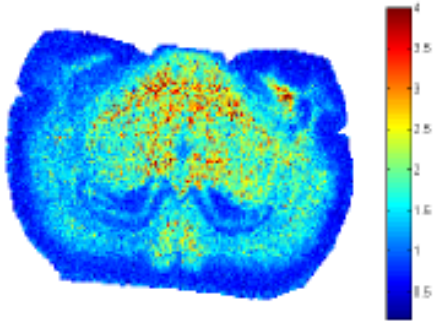
- Задача вывода - обратная по отношению к порождению документов.
- Семплирование по Гиббсу - приближенный алгоритм вывода.

- Усредняем спектры каждой точки по соседям - вносим корреляцию между близкими точками.
- Применяем алгоритм вывода LDA на документах-точках, словах-ярких $\frac{m}{z}$. Темы – бикластеры. Получаем оптимальные распределения $\{\theta_d^*\}$, $\{\beta_k^*\}$
- По $\{\beta_k^*\}$ строим распределения $\{\gamma_n^*\}$ слов по бикластерам из теоремы Байеса.
- Точкам и $\frac{m}{z}$ присваиваем номер бикластера, имеющий наибольшую вероятность в $\{\theta_d^*\}$ и $\{\gamma_n^*\}$ соответственно.

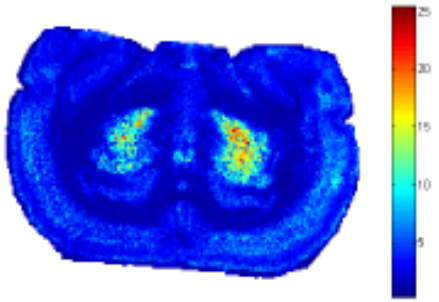
- Пространственная часть бикластеризации с применением LDA. Слева - без усреднения спектров по соседям, справа - с усреднением:



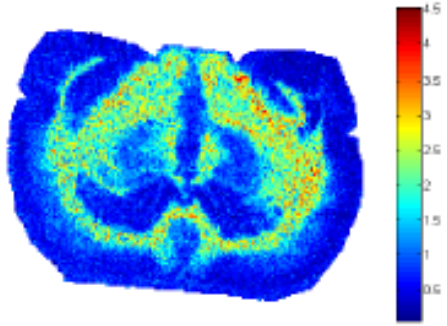




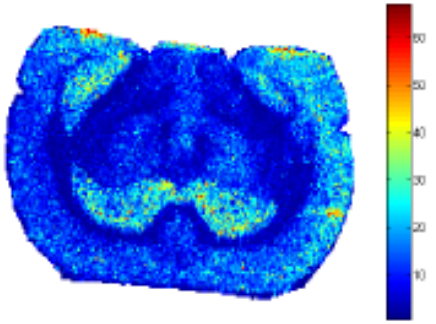
1(оранжевый)



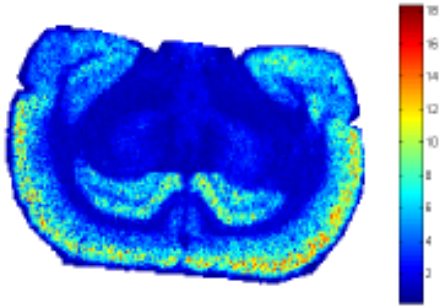
2(салатовый)



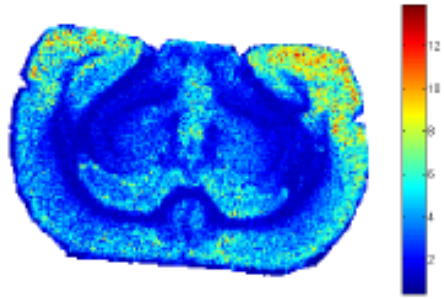
3(коричневый)



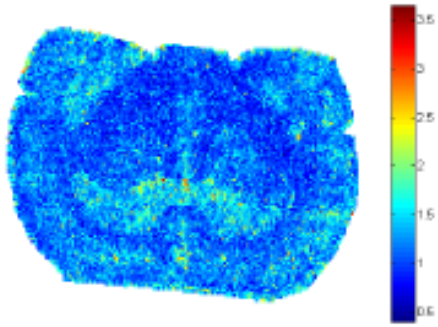
4(красный)



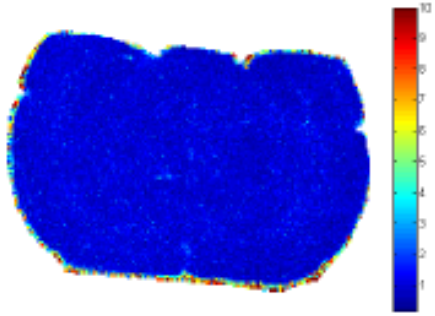
5(жёлтый)



6(синий)



7(серый)



8(зелёный)

Outline

- Предложено 2 разноплановых метода решения задачи бикластеризации данных ИМС.
- Задача бикластеризации данных с учетом пространственной близости точек до того не ставилась, существующие методы в данной ситуации неприменимы.
- Предложенные методы дают схожие результаты, каждый имеет свои плюсы и минусы.

Спасибо за внимание!