Методы и алгоритмы мультимодальной кластеризации

Дмитрий Гнатышак,
Научный руководитель: д.ф.-м.н. Кузнецов С. О.

НИУ ВШЭ

9 апреля 2015

Введение

Анализ генной экспрессии: объекты — гены, признаки — образцы, значения — генная экспрессия.

Пример

	s0	s1	s2	s3	s4	s5	s6
g0	3.6	1.0	1.0	0.0	1.0	1.0	1.0
g1	3.0	2.5	0.0	0.0	2.0	0.0	1.0
g2	0.0	5.0	0.0	0.0	5.0	0.0	5.0
g3	6.6	5.5	0.0	0.0	0.0	0.0	2.0
g4	9.0	7.5	0.0	0.0	6.0	0.0	3.0
g5	6.6	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	3.0	0.0	0.0	3.0	0.0	3.0
g7	0.0	8.0	8.0	0.0	8.0	8.0	0.0
g8	6.0	5.0	0.0	0.0	4.0	0.0	2.0
g9	0.0	4.0	4.0	0.0	4.0	4.0	4.0

Введение

Анализ генной экспрессии: объекты — гены, признаки — образцы, значения — генная экспрессия.

Пример

Классическая кластеризация

	s0	s1	s2	s3	s4	s5	s6
g0	3.6	1.0	1.0	0.0	1.0	1.0	1.0
g1	3.0	2.5	0.0	0.0	2.0	0.0	1.0
g2	0.0	5.0	0.0	0.0	5.0	0.0	5.0
g3	6.6	5.5	0.0	0.0	0.0	0.0	2.0
g4	9.0	7.5	0.0	0.0	6.0	0.0	3.0
g5	6.6	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	3.0	0.0	0.0	3.0	0.0	3.0
g7	0.0	8.0	8.0	0.0	8.0	8.0	0.0
g8	6.0	5.0	0.0	0.0	4.0	0.0	2.0
g9	0.0	4.0	4.0	0.0	4.0	4.0	4.0

Введение

Анализ генной экспрессии: объекты — гены, признаки — образцы, значения — генная экспрессия.

Пример

Классическая кластеризация

	s0	s1	s2	s3	s4	s5	s6
g0	3.6	1.0	1.0	0.0	1.0	1.0	1.0
g1	3.0	2.5	0.0	0.0	2.0	0.0	1.0
g2	0.0	5.0	0.0	0.0	5.0	0.0	5.0
g3	6.6	5.5	0.0	0.0	0.0	0.0	2.0
g4	9.0	7.5	0.0	0.0	6.0	0.0	3.0
g5	6.6	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	3.0	0.0	0.0	3.0	0.0	3.0
g7	0.0	8.0	8.0	0.0	8.0	8.0	0.0
g8	6.0	5.0	0.0	0.0	4.0	0.0	2.0
g9	0.0	4.0	4.0	0.0	4.0	4.0	4.0

Введение

Анализ генной экспрессии: объекты — гены, признаки — образцы, значения — генная экспрессия.

Пример

Бикластеризация

	s0	s1	s2	s3	s4	s5	s6
g0	3.6	1.0	1.0	0.0	1.0	1.0	1.0
g1	3.0	2.5	0.0	0.0	2.0	0.0	1.0
g2	0.0	5.0	0.0	0.0	5.0	0.0	5.0
g3	6.6	5.5	0.0	0.0	0.0	0.0	2.0
g4	9.0	7.5	0.0	0.0	6.0	0.0	3.0
g5	6.6	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	3.0	0.0	0.0	3.0	0.0	3.0
g7	0.0	8.0	8.0	0.0	8.0	8.0	0.0
g8	6.0	5.0	0.0	0.0	4.0	0.0	2.0
g9	0.0	4.0	4.0	0.0	4.0	4.0	4.0

Введение

Анализ генной экспрессии: объекты — гены, признаки — образцы, значения — генная экспрессия.

Пример

Трикластеризация

time0

	s0	s1	s2	s3	s4	s5	s6
g0	3.6	1.0	1.0	0.0	1.0	1.0	1.0
g1	3.0	2.5	0.0	0.0	2.0	0.0	1.0
g2	0.0	5.0	0.0	0.0	5.0	0.0	5.0
g3	6.6	5.5	0.0	0.0	0.0	0.0	2.0
g4	9.0	7.5	0.0	0.0	6.0	0.0	3.0
g5	6.6	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	3.0	0.0	0.0	3.0	0.0	3.0
g7	0.0	8.0	8.0	0.0	8.0	8.0	0.0
g8	6.0	5.0	0.0	0.0	4.0	0.0	2.0
g9	0.0	4.0	4.0	0.0	4.0	4.0	4.0

time1

	s0	s1	s2	s3	s4	s5	s6
g0	0.0	0.5	0.5	0.0	0.5	0.5	0.5
g1	0.0	3.0	0.0	0.0	2.4	0.0	1.2
g2	0.0	2.5	0.0	0.0	2.5	0.0	2.5
g3	0.0	5.5	0.0	0.0	0.0	0.0	2.0
g4	0.0	9.0	0.0	0.0	7.2	0.0	3.6
g5	0.0	0.0	0.0	0.0	4.4	0.0	2.0
g6	0.0	1.5	0.0	0.0	1.5	0.0	1.5
g7	0.0	4.0	4.0	0.0	4.0	4.0	0.0
g8	0.0	6.0	0.0	0.0	4.8	0.0	2.4
g9	0.0	2.0	2.0	0.0	2.0	2.0	2.0

Основные случаи

- Бинарные данные. Их можно рассматривать как n-арное отношение:

$$(D_1, \ldots, D_n, \mathcal{R})$$

- Зашумлённые бинарные данные
- Действительные данные:

$$(D_1, \ldots, D_n, \mathcal{R}, V), V : R \to \mathbb{R}$$

Определение

Формальным контекстом $\mathbb{K}$ называется тройка $(G, M, \mathcal{R})$, где G — множество объектов, M — множество признаков, $\mathcal{R} \subseteq G \times M$ — бинарное отношение.

Определение

Операторы Галуа (штрих-операторы): Пусть $X \subseteq G$, $Y \subseteq M$.
$X' := \{m \in M \mid g\mathcal{R}m$ для всех $g \in X\}$ множество признаков, общих для всех объектов из X.
$Y' := \{g \in G \mid g\mathcal{R}m$ для всех $m \in Y\}$ множество объектов, обладающих всеми признаками из Y.

Определение

Пара (X, Y) называется ***формальным понятием*** контекста $\mathbb{K} = (G, M, \mathcal{R})$, если $X \subseteq G$, $Y \subseteq M$, $X' = Y$, $Y' = X$. X называется ***(формальным) объёмом***, Y — ***(формальным) содержанием***.

Определение

***n*-адическим формальным контекстом** $\mathbb{K}$ называется кортеж $(D^1, D^2, \ldots, D^n, \mathcal{R})$, где D^i — множество значений категории «признаков» i, $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$ — n-арное отношение.

Определение

***n*-адическим формальным контекстом** $\mathbb{K}$ называется кортеж $(D^1, D^2, \ldots, D^n, \mathcal{R})$, где D^i — множество значений категории «признаков» i, $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$ — n-арное отношение.

Определение

Пусть $H = (X^1, \ldots, X^n)$, $\forall i = 1, \ldots, n, X^i \subseteq D^i$.
H удовлетворяет ***условию связности*** $(\mathcal{C}_{connected})$ в $\mathcal{R} \iff$
$\forall u = (x^1, \ldots, x^n) \in X^1 \times \cdots \times X^n, u \in \mathcal{R}$.

Определение

***n*-адическим формальным контекстом** $\mathbb{K}$ называется кортеж $(D^1, D^2, \ldots, D^n, \mathcal{R})$, где D^i — множество значений категории «признаков» i, $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$ — n-арное отношение.

Определение

Пусть $H = (X^1, \ldots, X^n)$, $\forall i = 1, \ldots, n, X^i \subseteq D^i$.
H удовлетворяет ***условию связности*** $(\mathcal{C}_{connected})$ в $\mathcal{R} \iff$
$\forall u = (x^1, \ldots, x^n) \in X^1 \times \cdots \times X^n, u \in \mathcal{R}$.

Определение

H удовлетворяет ***условию замкнутости*** $(\mathcal{C}_{closed})$ в $\mathcal{R} \iff$
$\forall j = 1, \ldots, n, \forall x^j \in D^j \setminus X^j, \left(X^1, \ldots, X^j \cup \{x^j\}, \ldots, X^n\right)$ не удовлетворяет $\mathcal{C}_{connected}$.

Определение

Пусть $H = (X^1, \ldots, X^n)$, $\forall i = 1, \ldots, n, X^i \subseteq D^i$.
H удовлетворяет ***условию связности*** $(\mathcal{C}_{connected})$ в $\mathcal{R} \iff$
$\forall u = (x^1, \ldots, x^n) \in X^1 \times \cdots \times X^n, u \in \mathcal{R}$.

Определение

H удовлетворяет ***условию замкнутости*** $(\mathcal{C}_{closed})$ в $\mathcal{R} \iff$
$\forall j = 1, \ldots, n, \forall x^j \in D^j \setminus X^j, \left(X^1, \ldots, X^j \cup \{x^j\}, \ldots, X^n\right)$ не удовлетворяет $\mathcal{C}_{connected}$.

Определение

H называется n-***адическим формальным понятием*** ($\equiv$ ***замкнутым*** n-***множеством***) $\iff$ H одновременно удовлетворяет $\mathcal{C}_{connected}$ и $\mathcal{C}_{closed}$.

n-адические формальные понятия - абсолютно плотные n-кластеры в n-адическом случае.

Анализ формальных понятий

$G \backslash M$	A	B	C	A	B	C	A	B	C
1	x	x	x	x	x	x	x	x	
2	x	x		x			x	x	
3		x				x	x		x
4			x	x		x	x	x	x
B	α			β			γ		

Анализ формальных понятий

$G \backslash M$	A	B	C	A	B	C	A	B	C
1	x	x	x	x	x	x	x	x	
2	x	x		x			x	x	
3		x				x	x		x
4			x	x		x	x	x	x
B	α			β			γ		

Пример

$(\{1,2\},\{A,B\},\{\alpha,\gamma\})$ — триадическое формальное понятие (удовлетворяет и $\mathcal{C}_{connected}$, и $\mathcal{C}_{closed}$).

Анализ формальных понятий

$G \backslash M$	A	B	C	A	B	C	A	B	C
1	x	x	x	x	x	x	x	x	
2	x	x		x			x	x	
3		x				x	x		x
4			x	x		x	x	x	x
B	α			β			γ		

Пример

$(\{1,2\},\{A,B\},\{\alpha,\beta,\gamma\})$ — ***не*** триадическое формальное понятие: нарушается $\mathcal{C}_{connected}$.

Анализ формальных понятий

$G \backslash M$	A	B	C	A	B	C	A	B	C
1	x	x	x	x	x	x	x	x	
2	x	x		x			x	x	
3		x				x	x		x
4			x	x		x	x	x	x
B	α			β			γ		

Пример

$(\{1,2\},\{A,B\},\{\alpha\})$ — ***не*** триадическое формальное понятие: нарушается $\mathcal{C}_{closed}$.

Пусть $\mathbb{K} = (D_1, \ldots, D_p, I)$ — p-мерный контекст. D_1, $\ldots$, D_p и I, а также их мощности пользователю не известны.

На каждой итерации пользователь получает $J \subseteq I$.

Требуется хранить в памяти множества результатов применения штрих-операторов:

- $PrimesX1$ — словарь с элементами вида $((d_2, \ldots, d_p), \{d_1 \in D_1\})$, $d_2 \in D_2$, $\ldots$, $d_p \in D_p$;
- $\ldots$
- $PrimesXp$ — словарь с элементами вида $((d_1, \ldots, d_{p-1}), \{ad_p \in D_p\})$, $d_1 \in D_1$, $\ldots$, $d_{p-1} \in D_{p-1}$.

На каждой итерации для каждого кортежа:

1. Добавить во множества результатов применения штрих-операторов данную тройку:
   - в $PrimesX1[(d_2, \ldots, d_p)]$ добавить d_1
   - ...
   - в $PrimesXp[(d_1, \ldots, d_{p-1})]$ добавить d_p
2. Добавить p-кластер на основе порождающего кортежа $(d_1, \ldots, d_p)$ во множество всех p-кластеров. При этом, вместо добавления «настоящих» результатов применения штрих-операторов, необходимо добавить ссылки на соответствующие элементы соответствующих множеств.

Пост-обработка:

Вход: $\mathcal{C} = \{C = (*X_1, \ldots, *X_p)\}$ — полное текущее множество p-кластеров;
Выход: $\overline{\mathcal{C}} = \{C = (*X_1, \ldots, *X_p)\}$ — обработанное хеш-множество p-кластеров;

1: **для всех** $C \in \mathcal{C}$
2: Вычислить $hash(C)$
3: **если** $hash(C) \notin \overline{\mathcal{C}}$ **то**
4: $\overline{\mathcal{C}} := \overline{\mathcal{C}} \cup C$

\+ линейное время работы алгоритма и базовой пост-обработки

\- приблизительные результаты, средняя плотность

Алгоритм Data-Peeler

Стратегия перебора

Алгоритм Data-Peeler получает на вход n-адический контекст и множество почленно (анти)монотонных ограничений и выдаёт множество всех n-адических формальных понятий, удовлетворяющих данным ограничениям.

Идея: последовательно разбивать пространство решений таким образом, чтобы:

1. каждую его часть можно было независимо исследовать;
2. объединение n-адических формальных понятий, извлечённых из каждой части давало всё множество n-адических формальных понятий.

Строится бинарное дерево:

- Каждая вершина N — пара (U, V);
- N *представляет* множество всех n-адических формальных понятий $(X^1, \ldots, X^n)$ таких, что $\forall i = 1, \ldots, n, U^i \subseteq X^i \subseteq U^i \cup V^i$;
- Корневая вершина $((\emptyset, \ldots, \emptyset), (D^1, \ldots, D^n))$ представляет множество всех формальных понятий, вершины вида $((U^1, \ldots, U^n), (\emptyset, \ldots, \emptyset))$ представляют формальные понятия $(U^1, \ldots, U^n)$.

Алгоритм Data-Peeler

Стратегия перебора

Строится бинарное дерево:

- Каждая вершина N — пара (U, V);
- N *представляет* множество всех n-адических формальных понятий $(X^1, \ldots, X^n)$ таких, что $\forall i = 1, \ldots, n, U^i \subseteq X^i \subseteq U^i \cup V^i$;
- Корневая вершина $((\emptyset, \ldots, \emptyset), (D^1, \ldots, D^n))$ представляет множество всех формальных понятий, вершины вида $((U^1, \ldots, U^n), (\emptyset, \ldots, \emptyset))$ представляют формальные понятия $(U^1, \ldots, U^n)$.

Пример

Вершина $N = ((\{1,2\}, \{A,B\}, \{\alpha,\beta\}), (\{3\}, \{C\}, \emptyset))$ представляет все формальные понятия из множества
$\{ (\{1,2\}, \{A,B\}, \{\alpha,\beta\}), (\{1,2,3\}, \{A,B\}, \{\alpha,\beta\}), (\{1,2\}, \{A,B,C\}, \{\alpha,\beta\}), (\{1,2,3\}, \{A,B,C\}, \{\alpha,\beta\}) \}$

Алгоритм Data-Peeler

Стратегия перебора

Процедура построения потомков вершины $N = (U, V)$:

1. Выбрать элемент p из V
2. Построить двух потомков:
   - $N_L = (U_L, V_L) = (U \cup \{p\}, V \setminus \{p\})$
   - $N_R = (U_R, V_R) = (U, V \setminus \{p\})$
3. Если для U_L или U_R не выполняется $\mathcal{C}_{connected}$, остановить вычисления на ветви

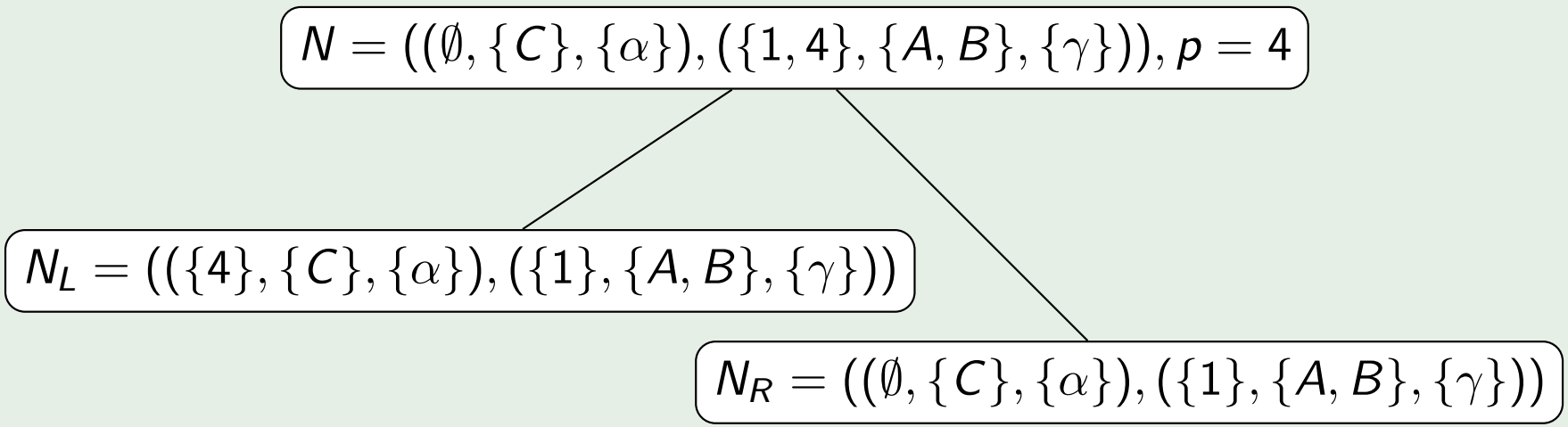
Алгоритм Data-Peeler

Стратегия перебора

Процедура построения потомков вершины $N = (U, V)$:

1. Выбрать элемент p из V
2. Построить двух потомков:
   - $N_L = (U_L, V_L) = (U \cup \{p\}, V \setminus \{p\})$
   - $N_R = (U_R, V_R) = (U, V \setminus \{p\})$
3. Если для U_L или U_R не выполняется $\mathcal{C}_{connected}$, остановить вычисления на ветви

Пример



Алгоритм Data-Peeler

Data-Peeler перебирает все формальные понятия, удовлетворяющие заданным условиям.

\+ Точная полная выдача, эффективное вычисление

\- большая временная сложность в худшем случае

Модификация для работы с зашумленным контекстом и пропущенными значениями: Multidupehack

TriBox

Метод TriBox использует оптимизационный подход для поиска трикластеров (может быть расширен на n-адический случай).
Триадический контекст $\mathbb{K} = (G, M, B, I)$ представлен трехмерным тензором:

$$r_{ijk} = \begin{cases} 1, & \text{if } (g_i, m_j, b_k) \in I; \\ 0, & \text{if } (g_i, m_j, b_k) \notin I. \end{cases}$$

Множество трикластеров $\mathcal{T} = \{T = (X, Y, Z)\}$ формирует следующую модель данных:

$$r_{ijk} = \max_{t=1,\ldots,|\mathcal{T}|} \lambda_t g_{it} m_{jt} b_{kt} + \lambda_0 + \epsilon_{ijk}$$

где:

1. λ_t — параметр (некоторая мера для трикластера $\mathcal{T}_t$)
2. g_{it} (а также m_{jt}, b_{kt}) равняется 1, если g_i (m_j, b_k) принадлежит X (Y, Z), и 0 иначе
3. λ_0 — константа
4. ϵ_{ijk} — остаток

TriBox

В результате оптимизации с помощью метода наименьших квадратов модели с одним трикластером получаем критерий для максимизации:

$$f(T) = \rho(T)^2 |X||Y||Z|$$

Схема алгоритма: начиная с каждой тройки отношения I, достраиваем её до большего трикластера, пока не находим локальным максимум.
Для этого используем следующую функцию $D(e)$: допустим мы определяем приращение значения критерия от добавления или удаления объекта g^*.
Тогда $D(g^*)$ будет равна:

$$D(g^*) = \frac{[r^2(i^*, Y, Z) + 2z_{i^*} r(X, Y, Z) r(i^*, Y, Z) - z_{i^*} r^2(X, Y, Z)/|X|]}{((|X| + z_{i^*})|Y||Z|)} \quad (1)$$

1. $z_{i^*} = 1$, если g^* добавляется к X и $z_{g^*} = -1$ иначе
2. $r(X, Y, Z)$ — число троек отношения I в трикластере

+ Плотные, хорошо интерпретируемые трикластеры, возможно использовать на действительных числах
- Крайне высокая временная сложность

Критерии качества n-кластеров

Пусть дан n-кластер $N = (X_1, \ldots, X_n)$ и контекст $\mathbb{K} = (D_1, \ldots, D_n, R)$, $X_1 \in D_1, \ldots, X_n \in D_n$.

- Минимальная поддержка n-кластера: $X_i \geq si_{min}$
- Минимальная плотность n-кластера: $\rho(T) = \frac{|X_1 \times \cdots \times X_n \cap R|}{|X_1| \ldots |X_n|} \geq \rho_{min}$
- $\rho(T)^2 |X_1| \ldots |X_n|$
- Покрытие множеством n-кластеров кортежей контекста
- Разнообразие (относительное число непересечений n-кластеров)
- Экспертная оценка
- Средняя плотность множества n-кластеров
- Число n-кластеров в выдаче
- ...

Текущие исследования

- Построение более эффективных версий существующих методов (в идеале - построение линейных версий)
- Исследование связи бикластеризации и мультимодальной кластеризации в случаях большей размерности
- Применимость методов в рамках моделей параллельных вычислений
- Разработка не зависящих от предметных областей методов на основе существующих
- Сравнение, а также выявление преимуществ и недостатков различных методов в зависимости от условий и данных

Спасибо за внимание!

Вопросы?