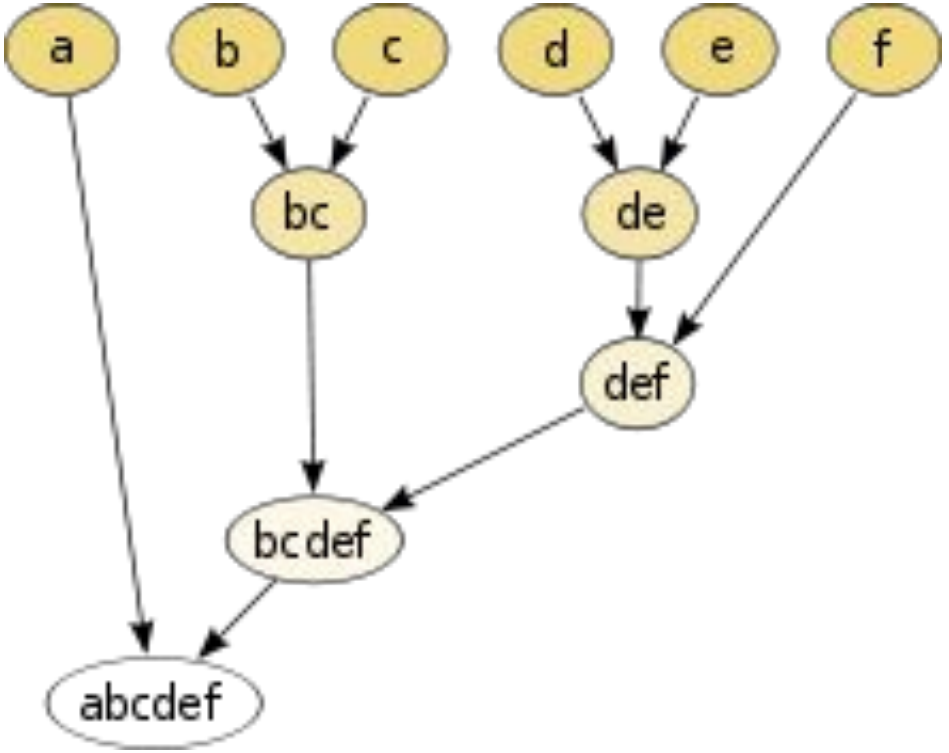
Иерархическая кластеризация

Камалов М.В.

- Иерархическая кластеризация – алгоритмы таксономии (биологическая таксономия)



- Дендограмма
- Многомерное шкалирование
- Карты Кохонена

Типы иерархической кластеризации

- Дивизимный (нисходящий)
- Алгомеративный (восходящий)

Расстояния между кластерами

- Для одноэлементных кластеров

$$R(\{x\},\{x'\}) = \rho(x,x').$$

- Универсальная формула расстояние между кластерами. Ланс и Уильямс 1967 году

$$R(U \cup V, S) = \alpha_U R(U,S) + \alpha_V R(V,S) + \beta R(U,V) + \gamma |R(U,S) - R(V,S)|,$$

Расстояние на практике

Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = -\tfrac{1}{2}.$$

Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = \tfrac{1}{2}.$$

Среднее расстояние:

$$R^{с}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0.$$

Расстояние между центрами:

$$R^{ц}(W,S) = \rho^2\left(\sum_{w\in W} \frac{w}{|W|}, \sum_{s\in S} \frac{s}{|S|}\right); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U\alpha_V,\ \gamma = 0.$$

Расстояние Уорда:

$$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}\rho^2\left(\sum_{w\in W} \frac{w}{|W|}, \sum_{s\in S} \frac{s}{|S|}\right); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

Агломеративная кластеризация Ланса-Уильямса

1: инициализировать множество кластеров C_1:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\};$
2: **для всех** $t = 2, \ldots, \ell \quad (t$ — номер итерации):
3: найти в C_{t-1} два ближайших кластера:
$(U, V) := \arg\min_{U \neq V} R(U, V);$
$R_t := R(U, V);$
4: изъять кластеры U и V, добавить слитый кластер $W = U \cup V$:
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$
5: **для всех** $S \in C_t$
6: вычислить расстояние $R(W, S)$ по формуле Ланса-Уильямса;

Свойство монотонности

$$R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell.$$

- Теорема Миллигана 1997г.

1) $\alpha_U \geqslant 0,\ \alpha_V \geqslant 0;$
2) $\alpha_U + \alpha_V + \beta \geqslant 1;$
3) $\min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$

- Из перечисленных выш $R^{\text{ц}}$ не является монотонным

Свойства растяжения и сжатия

- Растягивающие $R^{д}$ и $R^{у}$
- Сжимающие $R^{б}$
- Сохраняющие метрику пространс $R^{с}$ и $R^{ц}$
- Определяется через отношени $R_t / \rho(\mu_U, \mu_V)$
- Гибкое расстояние

$$\alpha_U = \alpha_V = (1 - \beta)/2, \quad \gamma = 0, \quad \beta < 1.$$

$\beta = -0{,}25$

Свойство редуктивности

- Ускорение алгоритма кластеризации

$$\{(U,V)\colon R(U,V) \leqslant \delta\}$$

- Определение Брюиноша 1978г.

$$\{S \mid R(U \cup V, S) < \delta,\ R(U,V) \leqslant \delta\} \subseteq \{S \mid R(S,U) < \delta \text{ или } R(S,V) < \delta\}.$$

- Теорема Диде и Моро 1984г.

$$\begin{aligned} &1)\ \alpha_U \geqslant 0,\ \alpha_V \geqslant 0;\\ &2)\ \alpha_U + \alpha_V + \min\{\beta, 0\} \geqslant 1;\\ &3)\ \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0. \end{aligned}$$

Быстрая агломеративная кластеризация на основе редуктивности

1: инициализировать множество кластеров C_1:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\}$;
2: выбрать начальное значение параметра δ;
3: $P(\delta) := \{(U, V) \mid U, V \in C_t,\ R(U, V) \leqslant \delta\}$;
4: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
5: **пока** $P(\delta) = \varnothing$
6: увеличить δ;
7: найти в $P(\delta)$ пару ближайших кластеров:
$(U, V) := \arg\min_{(U,V) \in P(\delta)} R(U, V)$;
$R_t := R(U, V)$;
8: изъять кластеры U и V, добавить слитый кластер $W = U \cup V$:
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
9: **для всех** $S \in C_t$
10: вычислить расстояние $R(W, S)$ по формуле Ланса-Уильямса;
11: **если** $R(W, S) \leqslant \delta$ **то**
12: $P(\delta) := P(\delta) \cup \{(W, S)\}$;

Определение числа кластеров

- Число кластеров $K = \ell - t + 1.$
- Ограничение $K_0 \leqslant K \leqslant K_1$
- Выбор количество t множеств

$$\ell - K_1 + 1 \leqslant t \leqslant \ell - K_0 + 1.$$

Достоинства и недостатки

- Метод ближнего соседа обладает цепочечным эффектом
- Метод дальнего соседа на раннем этапе может объединять довольно несхожие группы
- Метод расстояние между центрами масс «золотая середина»
- Метод Уорда чаще восстанавливает наилучшую кластеризацию

Источники

- http://www.ccas.ru/voron/download/Clustering.pdf
- https://yadi.sk/i/MelajPEXcG84H
- http://logic.pdmi.ras.ru/~sergey/teaching/ml/11-cluster.pdf

Спасибо за внимание!