Задача кластеризации

m объектов. i-й объект описывается весовым вектором $D_i=(t_{i1},t_{i2},\ldots,t_{in})$ из n значений характеристик.

Весовой вектор *центроида* Q_l составлен из средних арифметических характеристик по всем m объектам l-го кластера: $Q_l=(q_{l1},q_{l2},\ldots,q_{ln})$, где $q_{lj}=\frac{1}{m}\sum_{i=1}^{m} t_{ij}$.

Меры сходства (функции совпадения):

Коэффициент Дайса (самый распространенный) $$S\left(D_i;D_j\right)=\frac{2\left(\sum_{k=1}^{n} t_{ik}t_{jk}\right)}{\sum_{k=1}^{n} t_{ik}+\sum_{k=1}^{n} t_{jk}}$$

Коэффициент Жаккара $$S\left(D_i;D_j\right)=\frac{\sum_{k=1}^{n} t_{ik}t_{jk}}{\sum_{k=1}^{n} t_{ik}+\sum_{k=1}^{n} t_{jk}-\sum_{k=1}^{n} t_{ik}t_{jk}}$$

Коэффициент косинуса $$S\left(D_i;D_j\right)=\frac{\sum_{k=1}^{n} t_{ik}t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^2\cdot\sum_{k=1}^{n} t_{jk}^2}}$$

Коэффициент перекрытия: $$S\left(D_i;D_j\right)=\frac{\sum_{k=1}^{n} t_{ik}t_{jk}}{\min\left\{\sum_{k=1}^{n} t_{ik};\sum_{k=1}^{n} t_{jk}\right\}}$$

C.J. van Rijsbergen
«Information Retrieval»
(Ван Ризберген
«Информационный поиск»)
$\textit{сходство}=(1+\textit{несходство})^{-1}$

Кластеризация по коэффициенту покрытия (1/3)

Пусть даны m объектов $d_1, d_2, \ldots, d_m$, каждый из которых описывается вектором длины n (значения для данного объекта характеристик $t_1, t_2, \ldots, t_n$). Вход алгоритма – матрица $D = \|d_{ij}\|$ размером $m \times n$, $d_{ij} \in \{0,1\}$.

- Каждый объект описан хотя бы одной характеристикой, т.е. $\sum_{j=1}^{n} d_{ij} \geq 1 \; \forall \; i = 1, \ldots, m$
- Каждая характеристика описывает хотя бы один объект $\sum_{i=1}^{m} d_{ij} \geq 1 \; \forall \; j = 1, \ldots, n$

Построим матрицу *коэффициентов покрытия C* размера $m \times m$ (описывает степень близости объектов).

$$s_{ij} = \frac{d_{ij}}{\sum_{k=1}^{n} d_{ik}}, s'_{ij} = \frac{d_{ij}}{\sum_{k=1}^{m} d_{kj}}, \text{ для } 1 \leq i \leq m \text{ и } 1 \leq j \leq n.$$

$$C = S \times S'^{T}$$

$$c_{ij} = \sum_{k=1}^{n} s_{ik} s'^{T}_{kj} = \sum_{k=1}^{n} (\text{значимость } t_k \text{ в } d_i) \times (\text{значимость } d_i \text{ для } t_k)$$

Коэффициент покрытия c_{ij} означает
для $i \neq j$ степень покрытия объекта d_i объектом d_j, для $i = j$ – уникальность d_i

Кластеризация по коэффициенту покрытия (2/3)

Свойства матрицы *C*:

- $0 \le c_{ij} \le 1$, $c_{ii} > 0$
- $\sum_{j=1}^{m} c_{ij} = 1$ для всех i, $1 \le i \le m$
- для всех i и j, $1 \le i,j \le m$, $c_{ii} \ge c_{ij}$ (если D - двоичная)
- $c_{ij} = 0$ влечет $c_{ji} = 0$ и, аналогично, $c_{ij} > 0$ влечет $c_{ji} > 0$ (однако в общем случае $c_{ij} \ne c_{ji}$)

Диагональные элементы c_{ii} матрицы C – *коэффициенты уникальности* δ_i объекта d_i при $1 \le i \le m$.

Сумма недиагональных элементов равна $\psi_i = \sum_{j=1}^{m} c_{ij} = 1 - \delta_i$, где $i \ne j$, – *коэффициент связи* d_i с другими объектами.

Общие коэффициенты связи и уникальности для всего набора:

$$\delta = \frac{\sum_{i=1}^{m} \delta_i}{m}; \psi = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij}}{m} = 1 - \delta, i \ne j$$

Собирательная способность (cluster seed power) для d_i: $p_i = \delta_i \psi_i \sigma_i$, где σ_i - число ненулевых компонент весового вектора данного объекта.

Число кластеров

$$\eta_C = [(\text{Коэффициент уникальности для набора}) \times (\text{Число объектов})] =$$

$$= [\delta \times m] = \left[\sum_{i=1}^{m} \delta_i\right]$$

$$\left[\left| d_{ij} - \frac{\sum_{j=1}^{n} d_{ij}}{n} \cdot \theta \right|\right] > 0$$

Среднее число элементов в кластере $\delta_C = 1/\delta$

Кластеризация по коэффициенту покрытия (3/3)

Однопроходный алгоритм.

1. Определить первые η_C ядер кластеров - объекты $\mathrm{d_i}$, имеющие наибольшие значения собирательной способности $\mathrm{p_i}$.
2. Перебираем объекты, не являющиеся ядром. Объект $\mathrm{d_i}$ присоединяется к кластеру, ядро которого максимально покрывает $\mathrm{d_i}$, т.е. $c_{is_j} = \max\left\{c_{is_1}, c_{is_2}, \dots, c_{is_{\eta_C}}\right\}$. (несколько – $\max\{\mathrm{p_{s_j}}\}$)
3. Объекты, для которых $c_{ij} = 0$ при $i \neq \mathrm{j}$, не присоединены. Либо отдельный кластер, либо к кластеру максимально покрывающего объекта и повторять до стабилизации размера.

Ожидаемое число объектов в кластере i с ядром d_{s_i}

$$\eta_{iC} = \frac{p_i}{\sum_{k=1}^{\eta_C} p_k} \times m \text{ для } 1 \leq i \leq \eta_C$$

Функция связи. Покрытие нового вектора q вектором $\mathrm{d_i}$ $C(q, d_i)$ и наоборот $C(d_i, q)$:

$$C(q, d_i) = \frac{1}{\sum_{j=1}^{n} q_j} \left[\sum_{j=1}^{n} q_j d_{ik} \frac{1}{\sum_{i=1}^{m} d_{ij} + q_j}\right]$$

$$C(d_i, q) = \frac{1}{\sum_{j=1}^{n} d_{ij}} \left[\sum_{j=1}^{n} d_{ij}\, q_j \frac{1}{\sum_{i=1}^{m} d_{ij} + q_j}\right]$$

Устойчивость

Коэффициент Рэнда.

Два разбиения: P_1 и P_2, и пусть a_{ij} – количество пар объектов, которые

- в разбиении P_1 попадают в один кластер при $i=1$ и в разные кластеры при $i=0$;
- в разбиении P_2 попадают в один кластер при $j=1$ и в разные кластеры при $j=0$.

$$c(P_1, P_2) = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{00}}$$

Значение *c* лежит в между 0 и 1. Оно равно 1, если разбиения подобны, и 0 в противном случае.