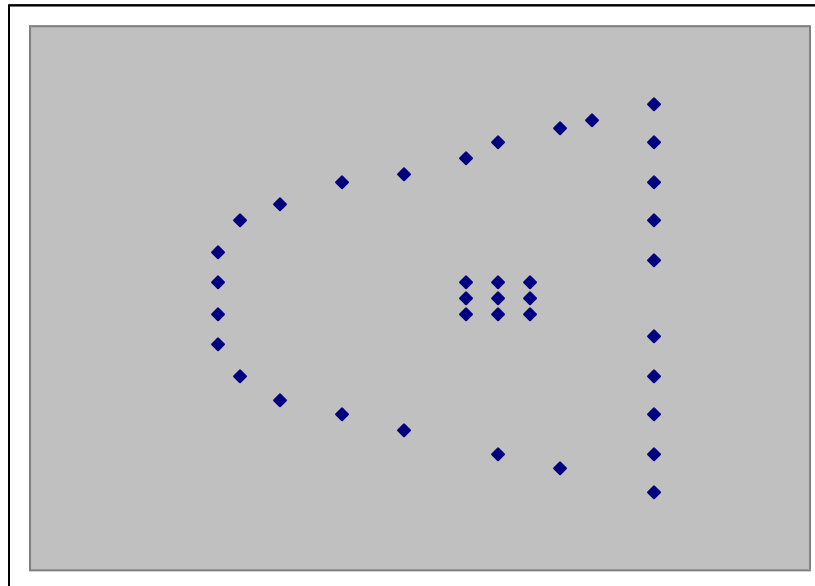


Кластеризация на основе нечетких отношений. Алгоритм Fuzzy Relation Clustering

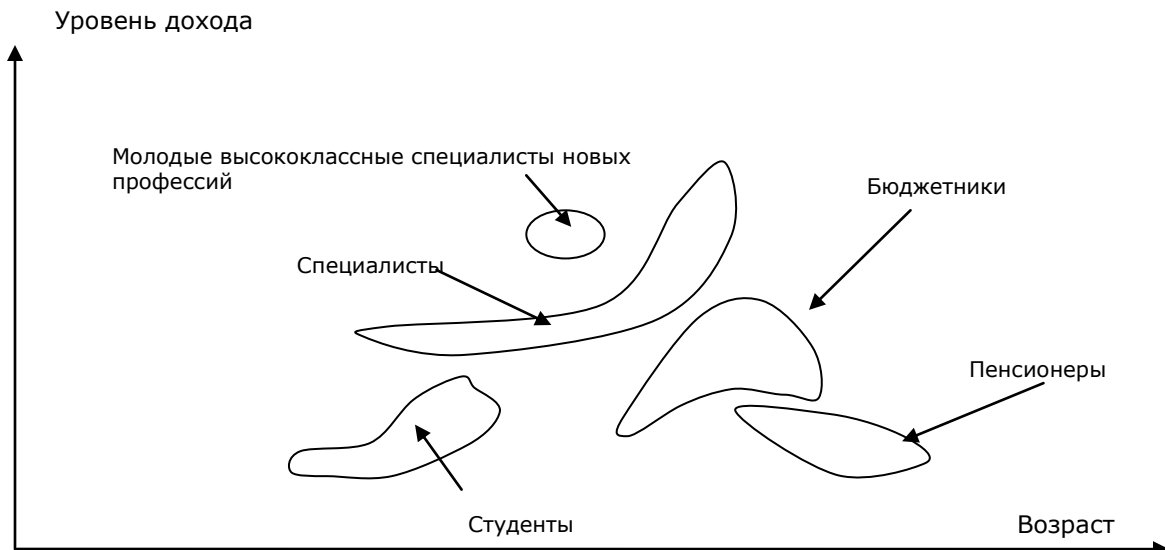
© Максим Гончаров, 2005 - **spellabs**
email: maxim.goncharov@spellabs.ru

Введение

Основные известные алгоритмы кластеризации (например, модификации алгоритмов K-Means, Expectation Maximization, реализованные, в том числе, в Microsoft Analysis Services 2005) налагают ограничения на геометрию получаемых кластеров, в частности, требуя возможности охвата каждого кластера отдельным выпуклым множеством. Такое ограничение налагается предположениями таких алгоритмов о существовании центров кластеров (K-Means) или функции плотности вероятности для каждого кластера с соответствующим значением математического ожидания и дисперсией (Expectation Maximization). В результате, эти алгоритмы не в состоянии адекватно разбить на кластеры невыпуклые множества, тем более вложенные структуры.



Такая необходимость возникает, например, для адекватной и информативной кластеризации демографических данных, описывающих зависимость возраста от дохода:



Эту проблему решает описываемый ниже алгоритм кластеризации на основе нечетких отношений, позволяющий группировать в кластеры элементы, между которыми есть последовательность «близких» друг к другу элементов, что также соответствует интуитивному представлению о группировке.

Описание алгоритма

Описывается подход к кластеризации конечного набора элементов произвольного метрического пространства на основании разбиения множества на классы эквивалентности по нечеткому отношению.

На основании метрики определяется нечеткое отношение, обладающее свойствами четкой рефлексивности и нормальной α -симметричности. Строится транзитивное замыкание отношения, позволяющее определить для каждого значения α в диапазоне от 0 до 1 отношение эквивалентности на исходном множестве. По построению отношения два элемента входят в один класс эквивалентности тогда и только тогда, когда между ними есть последовательность попарно «близких» друг к другу элементов.

В случае если пространство кластеризуемых элементов является векторным пространством с топологией, индуцированной метрикой, то полученные при помощи описанного алгоритма кластеры могут иметь произвольную геометрическую форму, в том числе состоять из невыпуклых множеств. Это является значительным преимуществом подхода в сравнении с множеством известных алгоритмов кластеризации (например, модификациями алгоритма K-Means, Expectation Maximization и т.д.).

В числе преимуществ алгоритма можно назвать:

- Отсутствие необходимости в априорных предположениях относительно структуры данных (вид и параметры распределения вероятности по кластерам, центров плотности, числа кластеров).

- Понятная интерпретация результатов разбиения по кластерам: элементы входят в один кластер когда между ними есть последовательность близких друг к другу элементов.
- Отсутствие ограничений на геометрию кластеров.
- Время выполнения алгоритма мало зависит от числа компонент входных векторов.

Существенным недостатком алгоритма является большое время выполнения, характеризующееся порядком n^4 от числа элементов.

Математические основы

Пусть X – метрическое пространство и $d : X \rightarrow R$ определенная на нем метрика, $(X_1, \dots, X_n) \subset X$ - последовательность элементов из X .

Мы предполагаем далее, что $\forall i \in \{1, \dots, n\} \exists j \in \{1, \dots, n\} : X_i \neq X_j$ (1)

Из условия (1) следует, что $\forall i \in \{1, \dots, n\}$ справедливо:

$$\max\{d(X_i, X_k) \mid k \in \{1, \dots, n\}\} > 0 \quad (2)$$

Таким образом, для каждого индекса i мы можем определить функцию, описывающую меру сходства j -ого элемента последовательности с i -ым элементом:

Определение 1:

$$\begin{aligned} \xi_i : \{1, \dots, n\} &\rightarrow [0, 1] \\ \xi_i(j) &:= 1 - \frac{d(X_i, X_j)}{\max\{d(X_i, X_k) \mid k \in \{1, \dots, n\}\}} \end{aligned} \quad (3)$$

Для каждого индекса i определим функцию, описывающую меру сходства k -ого и l -ого элемента относительно i -ого элемента:

Определение 2:

$$\begin{aligned} \varsigma_i : \{1, \dots, n\}^2 &\rightarrow [0, 1] \\ \varsigma_i(k, l) &:= 1 - |\xi_i(X_k) - \xi_i(X_l)|, \end{aligned} \quad (4)$$

Определим теперь функцию, описывающую меру сходства любых двух элементов последовательности относительно всех элементов последовательности:

Определение 3:

$$\begin{aligned} \mu : \{1, \dots, n\}^2 &\rightarrow [0, 1], \\ \mu(i, j) &:= \min\{\varsigma_k(i, j) \mid k \in \{1, \dots, n\}\} \end{aligned} \quad (5)$$

$$\text{Утверждение 1: } \mu(i, i) = 1, \forall i \quad (6)$$

Доказательство: $\varsigma_k(i, i) = 1 - |\xi_k(X_i) - \xi_k(X_i)| = 1$ для всех k и i , следовательно
 $\mu(i, i) := \min\{\varsigma_k(i, i) \mid k \in \{1, \dots, n\}\} = 1$

Утверждение 2: $\mu(i, j) = \mu(j, i), \forall i, j$ (7)

Доказательство: $\varsigma_k(i, j) = 1 - |\xi_k(X_i) - \xi_k(X_j)| = 1 - |\xi_k(X_j) - \xi_k(X_i)| = \varsigma_k(j, i)$,
 следовательно $\mu(i, j) := \min\{\varsigma_k(i, j) \mid k \in \{1, \dots, n\}\} = \min\{\varsigma_k(j, i) \mid k \in \{1, \dots, n\}\} = \mu(j, i)$

Определение 4:

Для $k=1, 2, \dots, n$ определим рекурсивно функции $\mu^{(k)} : \{1, \dots, n\}^2 \rightarrow [0, 1]$

$$\begin{cases} \mu^{(1)}(i, j) := \mu(i, j) \\ \mu^{(k)}(i, j) := \max\{\min\{\mu^{(k-1)}(i, s), \mu^{(k-1)}(s, j)\} \mid s \in \{1, \dots, n\}\} \end{cases} \quad (8)$$

Утверждение 3: $\mu^{(k)}(i, i) = 1, \forall i, j, k$ (9)

Доказательство: Доказательство по индукции по k .

При $k = 1$ утверждение следует из определения $\mu^{(1)}$ и (6).

Пусть утверждение справедливо для $k-1$. Тогда из (8) следует

$$\begin{aligned} 1 &\geq \mu^{(k)}(i, i) = \max\{\min\{\mu^{(k-1)}(i, s), \mu^{(k-1)}(s, i)\} \mid s \in \{1, \dots, n\}\} \geq \\ &\geq \min\{\mu^{(k-1)}(i, i), \mu^{(k-1)}(i, i)\} = \mu^{(k-1)}(i, i) = 1 \end{aligned}$$

Т.е. утверждение доказано.

Утверждение 4: $\mu^{(k)}(i, j) = \mu^{(k)}(j, i), \forall i, j, k$ (10)

Доказательство: Доказательство по индукции по k .

При $k = 1$ утверждение следует из определения $\mu^{(1)}$ и (7).

Пусть утверждение справедливо для $k-1$. Тогда из (8) следует

$$\begin{aligned} \mu^{(k)}(i, j) &= \max\{\min\{\mu^{(k-1)}(i, s), \mu^{(k-1)}(s, j)\} \mid s \in \{1, \dots, n\}\} = \\ &= \max\{\min\{\mu^{(k-1)}(s, j), \mu^{(k-1)}(i, s)\} \mid s \in \{1, \dots, n\}\} = \\ &= \max\{\min\{\mu^{(k-1)}(j, s), \mu^{(k-1)}(s, i)\} \mid s \in \{1, \dots, n\}\} = \mu^{(k)}(j, i) \end{aligned}$$

Т.е. утверждение доказано.

Утверждение 5: $\mu^{(k)}(i, j) \geq \mu^{(k-1)}(i, j), \forall k > 2$ (11)

Доказательство:

$$\begin{aligned} \mu^{(k)}(i, j) &= \max_{s=i} \{\min\{\mu^{(k-1)}(i, s), \mu^{(k-1)}(s, j)\} \mid s \in \{1, \dots, n\}\} \geq \\ &\geq \min \left\{ \underbrace{\mu^{(k-1)}(i, i)}_{=1}, \mu^{(k-1)}(i, j) \right\} \geq \min \left\{ 1, \underbrace{\mu^{(k-1)}(i, j)}_{\leq 1} \right\} = \mu^{(k-1)}(i, j) \end{aligned}$$

Т.е. утверждение доказано.

Утверждение 6: Для $\alpha \in [0,1]$ и $i, j_1, \dots, j_k, j \in \{1, \dots, n\}$

$$\text{справедливо: } \mu(i, j_1), \dots, \mu(j_k, j) \geq \alpha \Rightarrow \mu^{(k)}(i, j) \geq \alpha \quad (12)$$

Доказательство: Доказательство по индукции по k . При $k = 1$ нечего доказывать. Пусть утверждение справедливо для $k-1$. Тогда из предположения индукции следует, что $\mu^{(k-1)}(i, j_k) \geq \alpha$.

$$\text{Из (11) следует, что } \mu^{(k-1)}(j_k, j) \geq \mu^{(1)}(j_k, j) = \mu(j_k, j) \geq \alpha \quad (14)$$

Тогда из определения $\mu^{(k)}(i, j)$, (13) и (14) следует

$$\begin{aligned} \mu^{(k)}(i, j) &= \max \left\{ \min \left\{ \mu^{(k-1)}(i, s), \mu^{(k-1)}(s, j) \right\} \mid s \in \{1, \dots, n\} \right\} \geq \\ &\geq \min \left\{ \underbrace{\mu^{(k-1)}(i, j_k)}_{\geq \alpha}, \underbrace{\mu^{(k-1)}(j_k, j)}_{\geq \alpha} \right\} \geq \alpha \end{aligned}$$

Т.е. утверждение доказано.

Утверждение 7: Пусть $\alpha \in [0,1]$, $i, j, k \in \{1, \dots, n\}$ и $\mu^{(k)}(i, j) \geq \alpha$. Тогда существует такое число m и индексы $j_1, \dots, j_m \in \{1, \dots, n\}$, что $\mu(i, j_1), \dots, \mu(j_m, j) \geq \alpha$ (15)

Доказательство: Доказательство по индукции по k . При $k = 1$ утверждение справедливо для индекса j . Пусть утверждение справедливо для $k-1$. Из определения $\mu^{(k)}(i, j)$ следует существование индекса p такого, что

$$\begin{aligned} \mu^{(k)}(i, j) &= \max \left\{ \min \left\{ \mu^{(k-1)}(i, s), \mu^{(k-1)}(s, j) \right\} \mid s \in \{1, \dots, n\} \right\} = \\ &= \min \left\{ \mu^{(k-1)}(i, p), \mu^{(k-1)}(p, j) \right\} \end{aligned} \quad (16)$$

Из (16) и условия утверждения следует, что

$$\alpha \leq \mu^{(k)}(i, j) = \min \left\{ \mu^{(k-1)}(i, p), \mu^{(k-1)}(p, j) \right\} \Rightarrow \mu^{(k-1)}(i, p) \geq \alpha, \mu^{(k-1)}(p, j) \geq \alpha \quad (17)$$

Из условия $\mu^{(k-1)}(i, p) \geq \alpha$ и из условия индукции следует существование индексов $s_1, \dots, s_{m_1}$ таких, что $\mu(i, s_1), \dots, \mu(s_{m_1}, p) \geq \alpha$.

Аналогично, из условия $\mu^{(k-1)}(p, j) \geq \alpha$ и из условия индукции следует существование индексов $p_1, \dots, p_{m_2}$ таких, что $\mu(p, p_1), \dots, \mu(p_{m_2}, j) \geq \alpha$.

Объединяя эти выводы, получаем, что утверждение доказано для

$$m := m_1 + m_2 \text{ и } j_1, \dots, j_m, \text{ где } j_u := \begin{cases} s_u, & u \leq m_1 \\ p_{u-m_1}, & m_1 < u \leq m_2 \end{cases}$$

Утверждение 8: Пусть $\alpha \in [0,1]$, $i, j, m \in \{1, \dots, n\}$ и $\mu^{(n)}(i, j) \geq \alpha, \mu^{(n)}(j, m) \geq \alpha$.

$$\text{Тогда } \mu^{(n)}(i, m) \geq \alpha \quad (18)$$

Доказательство: Из условий $\mu^{(n)}(i, j) \geq \alpha$, $\mu^{(n)}(j, m) \geq \alpha$ и (15) следует существование индексов $j_1, \dots, j_p \in \{1, \dots, n\}$ и $j_{p+1}, \dots, j_{p+q} \in \{1, \dots, n\}$ таких, что $\mu(i, j_1), \dots, \mu(j_p, j) \geq \alpha$ и $\mu(j, j_{p+1}), \dots, \mu(j_{p+q}, m) \geq \alpha$ (19)

Так как $\{j_1, \dots, j_p, j, j_{p+1}, \dots, j_{p+q}\} \subset \{1, \dots, n\}$, то из этого множества индексов можно выбрать множество, состоящее из не более чем n индексов (путем удаления повторяющихся элементов) $\{u_1, \dots, u_r\} = \{j_1, \dots, j_p, j, j_{p+1}, \dots, j_{p+q}\}$, $r \leq n$ для которых также справедливо $\mu(i, u_1), \dots, \mu(u_r, m) \geq \alpha$ (20)

Из (12) следует, что $\mu^{(r)}(i, m) \geq \alpha$. Так как $r \leq n$, то из (11) следует $\mu^{(n)}(i, m) \geq \mu^{(r)}(i, m) \geq \alpha$, что и требовалось доказать.

Определение 5: Для $\alpha \in [0, 1]$ определим на множестве $\{X_1, \dots, X_n\}$ бинарное отношение $R_\alpha \subset \{X_1, \dots, X_n\}^2$ следующим образом:

$$(X_i, X_j) \in R_\alpha \Leftrightarrow \mu^{(n)}(i, j) \geq \alpha$$

Утверждение 9: Отношение R_α является отношением эквивалентности

Доказательство:

1. Рефлексивность. Из (9) следует, что $\mu^{(n)}(i, i) = 1 \geq \alpha \Rightarrow (X_i, X_i) \in R_\alpha$
2. Симметричность. Пусть $(X_i, X_j) \in R_\alpha \Rightarrow \mu^{(n)}(i, j) \geq \alpha$. В силу (10) $\mu^{(n)}(j, i) = \mu^{(n)}(i, j) \geq \alpha \Rightarrow (X_j, X_i) \in R_\alpha$
3. Транзитивность. Пусть $(X_i, X_j) \in R_\alpha, (X_j, X_m) \in R_\alpha \Rightarrow \mu^{(n)}(i, j), \mu^{(n)}(j, m) \geq \alpha$. Из (18) следует, что $\mu^{(n)}(i, m) \geq \alpha \Rightarrow (X_i, X_m) \in R_\alpha$

Результат:

Таким образом, отношение эквивалентности R_α разбивает множество $\{X_1, \dots, X_n\}$ на непересекающиеся классы эквивалентности. Два элемента X_i, X_j входят в один класс эквивалентности тогда и только тогда, когда значение функции $\mu^{(n)}$ от этих элементов велико, что на основании (12), (15) эквивалентно существованию последовательности пар элементов $(X_i, X_{j_1}), (X_{j_1}, X_{j_2}), \dots, (X_{j_r}, X_j)$, на которых значение функции μ велико. По определению μ это означает близость элементов каждой пары друг другу. Т.е. два элемента входят в один класс эквивалентности тогда и только тогда, когда между ними есть последовательность попарно близких друг к другу элементов.