

The background of the slide is a complex, abstract composition. It features a network of thin, reddish-brown lines forming a web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. On the left side, there is a vertical strip containing a grid of small, light-colored squares. The overall aesthetic is technical and data-oriented.

Proximity Measure for Symmetric vs. Asymmetric Binary Variables

Proximity Measure for Binary Attributes

- A contingency table for binary data

1: 出现, 0: 未出现

		Object j		sum
		1	0	
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
	sum	$q + s$	$r + t$	p

越大, 相似度越低

- Distance measure for symmetric binary variables:

对称

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distance measure for asymmetric binary variables:

不对称

$$d(i, j) = \frac{r + s}{q + r + s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard coefficient is the same as "coherence": (a concept discussed in Pattern Discovery)

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

Example: Dissimilarity between Asymmetric Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

Distance: $d(i, j) = \frac{r + s}{q + r + s}$

$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$

$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$

$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$

		Mary		
		1	0	Σ_{row}
Jack	1	2	0	2
	0	1	3	4
	Σ_{col}	3	3	6

		Jim		
		1	0	Σ_{row}
Jack	1	1	1	2
	0	1	3	4
	Σ_{col}	2	4	6

		Mary		
		1	0	Σ_{row}
Jim	1	1	1	2
	0	2	2	4
	Σ_{col}	3	3	6