

Proximity Measure for Binary Attributes

■ A contingency table for binary data

Object
$$j$$

Object j

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

Distance measure for symmetric binary variables: マナイヤ

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

 \square Distance measure for asymmetric binary variables: $d(i,j) = \frac{r+s}{q+r+s}$

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

Jaccard coefficient (similarity measure for asymmetric

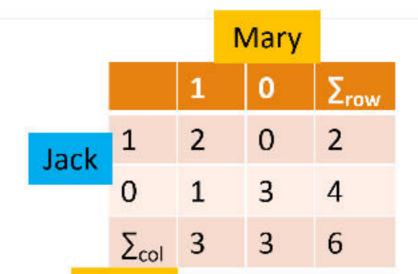
$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

□ 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



 Σ_{row}

2

4

6

Jim

1

1

- 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:	$d(i \ i)$	r+s
	a(i,j) –	q+r+s

$d(jack, mary) = \frac{0+1}{2+0}$	_ 11 2 2
$d(jack, jim) = \frac{1+1}{1+1+1}$	$\frac{1}{1} = 0.67$
$d(iim\ marv) = \frac{1+2}{1+2}$	0.75

1 + 1 + 2

				Σ_{col}	2	4
		М	ary			
		1	0	Σr	ow	
	1	1	1	2		
Jim	0	2	2	4		
	Σ_{col}	3	3	6		

Jack