1. Proximity Measure for Nominal Attributes formula and example in data mining

Dissimilarity between i and j = d(i, j) $d(i,j) = \frac{p-m}{p}$ **p= number of attributes**
m= number of matches between i and j

Colour code	Grade	Code
Blue	A	CODE A
Red	B	CODE B
Green	C	CODE B
Green	A	CODE C

Id	Test Result
1	A
2	B
3	C
4	A

In this example, p is 1. Find the dissimilarity matrix,
m= number of matches between i and j

$$d(2,1) = \frac{1-0}{1} = 1,$$

$as\ there\ is\ no\ match\ between\ object\ 2\ and\ object\ 1$

$$d(3,1) = \frac{1-0}{1} = 1,$$

$as\ there\ is\ no\ match\ between\ object\ 3\ and\ object\ 1$

$$d(4,1) = \frac{1-1}{1} = 0,$$

$as\ there\ is\ match\ between\ object\ 4\ and\ object\ 1$

	1	2	3	4
1	0			
2	d(2,1)	0		
3	d(3,1)	d(3,2)	0	
4	d(4,1)	d(4,2)	d(4,3)	0

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

Except object 4 and 1, all other objects are dissimilar to each other

Proximity Measure for Nominal Attributes formula and example in data mining

RollNo	Marks	Grade
1	90	A
2	80	B
3	82	B
4	90	A

distance(object1, Object2) = (P – M) / P

P is total number of attributes

M is total number of matches

So in our case we have four objects RollNo1, RollNo2, RollNo3, RollNo4

How to calculate Proximity Measure for Nominal Attributes?

d(RollNo1,RollNo1)	d(RollNo1,RollNo2)	d(RollNo1,RollNo3)	d(RollNo1,RollNo4)
d(RollNo2,RollNo1)	d(RollNo2,RollNo2)	d(RollNo2,RollNo3)	d(RollNo2,RollNo4)
d(RollNo3,RollNo1)	d(RollNo3,RollNo2)	d(RollNo3,RollNo3)	d(RollNo3,RollNo4)
d(RollNo4,RollNo1)	d(RollNo4,RollNo2)	d(RollNo4,RollNo4)	d(RollNo3,RollNo4)

Proximity Measure for Nominal Attributes formula and example in data mining

RollNo	Marks	Grade
1	90	A
2	80	B
3	82	B
4	90	A

distance(object1, Object2) = (P – M) / P

P is total number of attributes

M is total number of matches

So in our case we have four objects RollNo1, RollNo2, RollNo3, RollNo4

Object 4 and 1 matches with all attributes

Object 3 and 2 matches with Grade alone

How to calculate Proximity Measure for Nominal Attributes?

d(RollNo1,RollNo1)	d(RollNo1,RollNo2)	d(RollNo1,RollNo3)	d(RollNo1,RollNo4)
d(RollNo2,RollNo1)	d(RollNo2,RollNo2)	d(RollNo2,RollNo3)	d(RollNo2,RollNo4)
d(RollNo3,RollNo1)	d(RollNo3,RollNo2)	d(RollNo3,RollNo3)	d(RollNo3,RollNo4)
d(RollNo4,RollNo1)	d(RollNo4,RollNo2)	d(RollNo4,RollNo4)	d(RollNo3,RollNo4)

d(1,1) = P – M / P = 2 – 2 / 2 = 0	d(RollNo1,RollNo2)	d(RollNo1,RollNo3)	d(RollNo1,RollNo4)
(2,1) = P – M / P = (2 – 0) / 2 = 1	**(2,2) = P – M / P = (2 – 2) / 2 = 0**	d(RollNo2,RollNo3)	d(RollNo2,RollNo4)
(3,1) = P – M / P = (2 – 0) / 2 = 1	**(3,2) = P – M / P, as grade matches with Roll No 3 and 2 = (2 – 1)/ 2 = 0.5**	**(3,3) = P – M / P = (2 – 2)/ 2 = 0**	d(RollNo3,RollNo4)
(4,1) = P – M / P , as marks and grade matches with Roll no 4 and 1 = (2 – 2) / 2 = 0	**(4,2) = P – M / P = (2 – 0) / 2 = 1**	**(4,3) = P – M / P =(2 – 0)/ 2 = 1**	**(4,4) = P – M / P =(2 – 2) / 2 = 0**

2. Proximity Measure for Binary Attributes

- A contingency table for binary data

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

- Distance measure for symmetric binary variables:

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

- Distance measure for asymmetric binary variables:

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

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

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

Proximity Measure for Binary Attributes

- A contingency table for binary data

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

- Distance measure for asymmetric binary variables:

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

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

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

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

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

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

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

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

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

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

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

$$d(jack, mary) = \text{—}$$

$$d(jack, jim) = \text{—}$$

$$d(jim, mary) = \text{—}$$

Dissimilarity measure for asymmetric binary attributes

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
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

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

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

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

2. Proximity Measure for Binary Attributes

- A contingency table for binary data

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

- Distance measure for symmetric binary variables:

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

- Distance measure for asymmetric binary variables:

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

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

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

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

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

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

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

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

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

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

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

$sim_{jaccard}$(jack, mary)=$\frac{2}{2+0+1} = \frac{2}{3} = 0.67$

$sim_{jaccard}$(jack, jim)=–

$sim_{jaccard}$(jim, mary)=–

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

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

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

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

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

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

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

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

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

$sim_{jaccard}$(jack, mary)$= \frac{2}{2+0+1} = \frac{2}{3} = 0.67$

$sim_{jaccard}$(jack, jim)$= \frac{1}{1+1+1} = \frac{1}{3} = 0.33$

$sim_{jaccard}$(jim, mary)$= \frac{1}{1+1+2} = \frac{1}{4} = 0.25$

Object *j*

Object *i*

	1	0	sum
1	q	r	$q + r$
0	s	t	$s + t$
sum	$q + s$	$r + t$	p

How similarity and distance are related here?

Similarity = 1 - distance

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
A	M	Y	N	P	N	N	N
B	M	Y	Y	N	N	N	N
C	F	Y	N	P	N	P	N

d(A,B)	
d(A,C)	
d(B,C)	

Find Jaccard Similarity

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
A	M	Y	N	P	N	N	N
B	M	Y	Y	N	N	N	N
C	F	Y	N	P	N	P	N

d(A,B)	(1+1)/(1+1+1) = 2/3 = 0.67
d(A,C)	(0+1)/(2+0+1)=0.33
d(B,C)	(1+2)/(1+1+2)=3/4=0.75

Find Jaccard Similarity

How to calculate proximity measure for symmetric binary attributes?

Object 2				
Object 1		**1 / True / Positive**	**0 / False / Negative**	**Sum**
	1 / True / Positive	**A**	**B**	**A + B**
	0 / False / Negative	**C**	**D**	**C + D**
	Sum	**A + C**	**B + D**	

Name	**Gender**	Job_Status
Akram	Male	Regular
Ali	Male	Contract

$$distance\,(object1, object2) = \frac{B + C}{A + B + C + D}$$

Contingency table for binary data: Consider 1 for positive/True and 0 for negative/False

Here we are considering Male and regular as positive and female and contract as negative.

A = Akram is positive and Ali is also positive. so A=1 because Ali and Akram both are same by gender , and it is positive.

B = Akram is positive and Ali is negative. So B=1 because Akram is regular that is positive and Ali is on contract that is negative

C = Akram is negative and Ali is 1. So C = 0 because Akram is never negative.

D = Akram is negative and Ali is also negative. So D=0 because Akram is never negative. He is always positive(Gender and regular).

How to calculate proximity measure for symmetric binary attributes?

Object 2				
Object 1		1 / True / Positive	0 / False / Negative	Sum
	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

Name	**Gender**	Job_Status
Akram	Male	Regular
Ali	Male	Contract

$$distance\,(object1, object2) = \frac{B + C}{A + B + C + D} \qquad distance\,(Akram, Ali) = \frac{1 + 0}{1 + 1 + 0 + 0} = \frac{1}{2} = 0.5$$

Contingency table for binary data: Consider 1 for positive/True and 0 for negative/False

Here we are considering Male and regular as positive and female and contract as negative.

A = Akram is positive and Ali is also positive. so A=1 because Ali and Akram both are same by gender , and it is positive.

B = Akram is positive and Ali is negative. So B=1 because Akram is regular that is positive and Ali is on contract that is negative

C = Akram is negative and Ali is 1. So C = 0 because Akram is never negative.

D = Akram is negative and Ali is also negative. So D=0 because Akram is never negative. He is always positive(Gender and regular).

Distance measure for asymmetric binary attributes in data mining

How to calculate proximity measure for asymmetric binary attributes?
IIn this example, consider 1 for positive/True and 0 for negative/False.
Table 1. Contingency Table

		Object 2		
		1 / True / Positive	0 / False / Negative	Sum
Object 1	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

In table 1 we can consider the following facts.
A represents that object 1 is True and object 2 is also True.
B represents that object 1 is True and object 2 is also False.
C represents that object 1 is False and object 2 is also True.
D represents that object 1 is False and object 2 is also False.

In table 2, Asad, Bilal and Tahir are objects. Negative values represents False and Positive represents Negative.

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Asad	Negative	Yes	Negative	Positive	Negative	Negative
Bilal	Negative	Yes	Negative	Positive	Positive	Negative
Tahir	Positive	Yes	Negative	Negative	Negative	Negative

Distance measure for asymmetric binary attributes in data mining

$$distance\,(object1, object2) = \frac{B + C}{A + B + C}$$

In this example, consider 1 for positive/True and 0 for negative/False.

Table 1. Contingency Table

		Object 2		
		1 / True / Positive	0 / False / Negative	Sum
Object 1	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

$$distance\,(Asad, Tahir) = \frac{1 + 1}{1 + 1 + 1} = \frac{2}{3} = 0.67$$

$$distance\,(Asad, Bilal) = \frac{0 + 1}{2 + 0 + 1} = \frac{1}{3} = 0.33$$

$$distance\,(Tahir, Bilal) = \frac{1 + 2}{1 + 1 + 2} = \frac{3}{4} = 0.75$$

In table 2, Asad, Bilal and Tahir are objects. Negative values represents False and Positive represents Negative.

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Asad	Negative	Yes	Negative	Positive	Negative	Negative
Bilal	Negative	Yes	Negative	Positive	Positive	Negative
Tahir	Positive	Yes	Negative	Negative	Negative	Negative

In the results, we can see the following facts;

The distance between object 1 and 2 is 0.67.

Asad is object 1 and Tahir is in object 2 and the distance between both is 0.67.

Less distance is between Asad and Bilal.

It means that Asad and Bilal are more similar to each other as compared to other objects.

Jaccard coefficient similarity measure

How to calculate proximity measure for asymmetric binary attributes?
IIn this example, consider 1 for positive/True and 0 for negative/False.
Table 1. Contingency Table

		Object 2		
		1 / True / Positive	0 / False / Negative	Sum
Object 1	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

In table 1 we can consider the following facts.
A represents that object 1 is True and object 2 is also True.
B represents that object 1 is True and object 2 is also False.
C represents that object 1 is False and object 2 is also True.
D represents that object 1 is False and object 2 is also False.

$$Sim_{jaccard} = \frac{A}{A + B + C}$$

In table 2, Asad, Bilal and Tahir are objects. Negative values represents False and Positive represents Negative.

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Asad	Negative	Yes	Negative	Positive	Negative	Negative
Bilal	Negative	Yes	Negative	Positive	Positive	Negative
Tahir	Positive	Yes	Negative	Negative	Negative	Negative

Jaccard coefficient similarity measure

In this example, consider 1 for positive/True and 0 for negative/False.
Table 1. Contingency Table

		Object 2		
Object 1		1 / True / Positive	0 / False / Negative	Sum
	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

In table 2, Asad, Bilal and Tahir are objects. Negative values represents False and Positive represents Negative.

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Asad	Negative	Yes	Negative	Positive	Negative	Negative
Bilal	Negative	Yes	Negative	Positive	Positive	Negative
Tahir	Positive	Yes	Negative	Negative	Negative	Negative

Use manhattan distance formula

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3|$$

3. Proximity (Dissimilarity) measure for ordinal attributes

Consider the data matrix given below

Step -1 Count the states **m_f = 3 (fair, Good Excellent)**
Step-2 Replace each ordinal data of test2 by rank **Fair-1, Good-2, Excellent-3**
Step-3 Normalize the ranking using the below formula,

$$Z_{if} = \frac{R_{if} - 1}{m_f - 1}$$

Given a data matrix shown below,

Object	Test2 (Ordinal)
1	Excellent
2	Fair
3	Good
4	Excellent

Object	Test2	Rank
1	Excellent	3
2	Fair	1
3	Good	2
4	Excellent	3

Object	Z_{if}
1	1
2	0
3	0.5
4	1

$$Fair\ (1) = \frac{1-1}{3-1} = 0$$

$$\text{Good}\ (2) = \frac{2-1}{3-1} = 0.5$$

$$\text{Excellent}(3) = \frac{3-1}{3-1} = 1$$

$$d(2,1) = |0 - 1| = 1.0$$

$$d(3,1) = |0.5 - 1| = 0.5$$

$$d(4,1) = |1 - 1| = 0.0$$

	1	2	3	4
1	0			
2	1.0	0		
3	0.5	0.5	0	
4	0.0	1.0	0.5	0

$$d(x,y) = |x - y|$$

Use manhattan distance formula

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3|$$

3. Proximity (Dissimilarity) measure for ordinal attributes

Consider the data matrix given below

Step -1 Count the states **m_f = 3 (fair, Good Excellent)**
Step-2 Replace each ordinal data of test2 by rank **Fair-1, Good-2, Excellent-3**
Step-3 Normalize the ranking using the below formula,

$$Z_{if} = \frac{R_{if} - 1}{m_f - 1}$$

Given a data matrix shown below,

Object	Test2 (Ordinal)
1	Excellent
2	Fair
3	Good
4	Excellent

Object	Test2	Rank
1	Excellent	3
2	Fair	1
3	Good	2
4	Excellent	3

Object	Z_{if}
1	1
2	0
3	0.5
4	1

$$Fair\ (1) = \frac{1-1}{3-1} = 0$$

$$\text{Good}\ (2) = \frac{2-1}{3-1} = 0.5$$

$$\text{Excellent}(3) = \frac{3-1}{3-1} = 1$$

$$d(2,1) = |0 - 1| = 1.0$$

$$d(3,1) = |0.5 - 1| = 0.5$$

$$d(4,1) = |1 - 1| = 0.0$$

	1	2	3	4
1	0			
2	1.0	0		
3	0.5	0.5	0	
4	0.0	1	0.5	0

Use manhattan distance formula

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3|$$

Dissimilarity measure for ordinal attributes

Consider the data matrix given below

Step -1 Count the states **m_f = 4 (Bad, Fair, Good Excellent)**
Step-2 Replace each ordinal data of test2 by rank **Bad – 1**, **Fair-2, Good-3, Excellent-4.**
Step-3 Normalize the ranking using the below formula,

$$Z_{if} = \frac{R_{if} - 1}{m_f - 1}$$

Given a data matrix shown below,

Object	Test3 (Ordinal)
1	Excellent
2	Fair
3	Good
4	Excellent
5	Bad
6	Good

Object	Test2	Rank
1	Excellent	
2	Fair	
3	Good	
4	Excellent	

Object	Z_{if}
1	
2	
3	
4	
5	
6	