Measures of Equivalence

Lecture delivered by: Prof. Niloy Ganguly
Scribed by: Himika Biswas (05CS6027)

02-03-2006

1 Measuring Similarity of Equivalence

By structural equivalence we can mean interchangeability. In real life complete structural equivalence is not feasible. Consequently we measure the degree of similarity between two nodes. We can measure degree of structural equivalence from the adjacency matrix of the network. For example, consider the graph shown in Figure1. The corresponding adjacency matrix is shown in Table1. From this matrix we can see that rows for B & D are similar; so we can conclude that B & D are equivalent but not exactly similar as their columns differ.

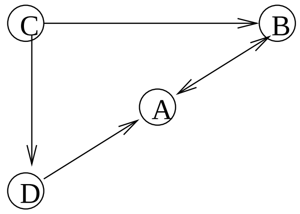

Figure 1: A graph showing relation between nodes A, B, C and D.

	A	B	C	D
A	0	1	0	0
B	1	0	0	0
C	0	1	0	1
D	1	0	0	0

Table 1: Adjacency Matrix for graph of Figure1.

We can measure similarity between two nodes simply by **Pearson's Correlation Coefficient**. Pearson's correlation coefficient between two nodes = 1 means those two nodes are exactly similar.

1.1 Pearson's Correlation Coefficient

Pearson's correlation coefficient between two nodes x and y can be measured as:

$$\gamma = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(\sum y)^2}{n})}} \tag{1}$$

If we calculate Pearson's correlation coefficient (PCC) for the graph of Figure1, the obtained result will be as shown in Table 2. If we perform hierarchical clustering based on minimum coefficient value on this graph then we shall get corresponding coefficients as shown in Tables 3 & 4. The corresponding structure is shown in Figure2.

	A	B	C	D
A	1			
B	0.5	1		
C	0.3	0.8	1	
D	0.7	0.4	0.1	1

Table 2: PCC between nodes of the graph of Figure1

	A	B	$C\|D$
A	1		
B	0.5	1	
$C\|D$	0.3	0.4	1

Table 3: PCC between nodes of the graph of Figure1 after merging C & D

	A	$B\|C\|D$
A	1	
$B\|C\|D$	0.4	1

Table 4: PCC between nodes of the graph of Figure1 after merging B, C & D.

1.2 Euclidian Distance

Another measure is Euclidian distance. It is a related but less sensitive measure of similarity. This is basically measure of dissimilarity. We can measure Euclidian distance between two nodes A & B by:

$$ED(A, B) = \sqrt{\sum_i (A_i - B_i)^2} \tag{2}$$

A	○	B	○	C	○	D
○	○	○	○	○	○	○
○	○	○	○	x	x	x
○	○	x	x	x	x	x
x	x	x	x	x	x	x

Figure 2: Hierarchical clustering performed on graph of Figure1.

Pearson's correlation coefficient may be of negative value, but Euclidian distance can not be negative. Generally

$$\frac{GreatestED}{SmallestED} < \frac{GreatestCorrelation}{SmallestCorrelation} \tag{3}$$

Moreover, Pearson's correlation coefficients are more sparsed, it takes care of marginal cases.

1.3 Percent of Exact Matches

In some cases the nodes we are measuring may be measured at the nominal level. In such data applying Euclidian distance or Pearson's correlation coefficient will be miss leading. For binary/categorical/discrete data this measure is useful. Here if we want to measure percent exact match between two nodes A an B, then we shall see how many row values of these columns are matching. For example, let column corresponding to A is 0 1 0 1 00 0 0 0 0 and column corresponding to B is 1 0 1 1 0 0 0 0 0 0. So total match is 70% out of which basically 60% no match(0s). Problem of percent match is if the data is too sparse then more absence (0) indicates percent match higher. For solving this problem we can use Jaccard Coefficients.

1.4 Jaccard Coefficients

We can measure Jaccard coefficient as:

$$JC = \frac{S}{S + P1 + P2} \tag{4}$$

Where, S: positive match
$P1$: non-null elements in A which has not matched
$P2$: non-null elements in B which has the match.
In real life $AB < -CAC < -B$.