A Study of Similarity Coefficients

Susan Laflin.

Computer Centre,
University of Birmingham,
Birmingham, England.

Abstract

The CLUSTAN package provides a choice of 38 similarity coefficients, but the description of these varies from inadequate to non-existant. Also they are subject to severe limitations for archaeological applications since they prohibit the use of a mixture of modes of attribute and make no allowance for the problem of missing data. The paper starts with a survey of coefficients within the package, some of which may be applied to binary data and others to numeric. A separate program has been written to deal with a mixture of attributes and data where some values may be missing and is described here. The output from the program is a card file which contains control cards and a similarity matrix for input to the CLUSTAN package, making use of the option to start the CLUSTAN run from a calculated similarity matrix.

The CLUSTAN package

The University of Birmingham has access to a version of the CLUSTAN package which allows the input of numeric and binary data, the calculation of a similarity matrix according to any one of 38 similarity coefficient followed by a Cluster Analysis according to any of the 8 SAHN methods provided and a choice of representation of the results. This is a very powerful and useful facility and one might have expected the archaeology department to have made extensive use of it. In fact it has been completely neglected and on investigation the reasons for this become clear.

Firstly it is not easy to find out exactly what similarity coefficients are calculated and what implications a particular choice of coefficient has. The description in the package is extremely meagre and requires close comparison with the first textbook by Sneath and Sokal, now obsolete and out of print. Once the equation has been located, there is no indication which is preferable in any given circumstance.

Secondly, although both binary and numeric data may be read into the package, only one type of attribute may be used in the calculation of similarity, all data of the other type being masked off during the calculation. To obtain useful comparisons of archaeological data, it is often necessary to use a variety of attributes of different types.

Finally, archaeological data is often incomplete and there is no facility within the existing package for ignoring comparisons involving missing data.

For all these reasons, I decided to write a program which would make use of the option of entering the CLUSTAN package with the similarity matrix already calculated and would satisfy the specifal requirements of archaeological data.

The problem of missing data

This is particularly obvious when we are using binary data, since a value of 1 indicates presence of an attribute while a value zero indicates either absence or a missing value. To make this distinction clear, let us assume we are recording pottery mugs and one of our attributes is the presence or absence of decoration on the handle. Presence gives us no problems; if the attribute is 1 the mug has a handle and it is decorated. Absence, denoted by 0, means that the mug has an undecorated handle. Missing means that the handle was broken off in antiquity and we have no indication whether it was decorated or not, or that the mug was made without a handle and the question of decoration becomes meaningless. Both these cases are also denoted by a value of 0.

This uncertainty is reflected in the evaluation of the similarity coefficients. While mutual presence of an attribute counts towards its similarity, mutual absence is less certain. In our above case, a mug without a handle and one with an undecorated handle are certainly not similar in this respect, but two mugs with undecorated handles should be. In the standard CLUSTAN package, it is not possible to make this distinction, but in my

program I have done this. In the package, the coefficient for any binary attribute is expressed in terms of the following sums, taken over all binary attributes for each pair of objects.
a = number of cases for which object I=1 & object J=1
b = number of cases for which object I=1 & object J=0
c = number of cases for which object I=0 & object J=1
d = number of cases for which object I=0 & object J=0

Then for a number of coefficients we have two versions of each of the formulae. For example:
Dice-Sorenson.
ICOEF=6 S = 2a/(2a+b+c)
ICOEF=7 S = 2(a+d)/(2(a+d)+b+c).
Rogers-Tanimoto.
ICOEF=8 S = a/(a+2(b+c))
ICOEF=9 S = (a+d)/(a+d+2(b+c)).
Kulczynski.
ICOEF=10 S = a/(b+c)
ICOEF=11 S = (a+d)/(b+c).

Here we may assume that all the 0 values denote missing data and take "a" as the number of agreements in the similarity measure, or we may assume that all the 0 values denote absent data and take "a+d" as the number of agreements, but we cannot distinguish between absent and missing data.

In my program I allow missing data to be coded as a blank column on the data card rather than a 0 and provide the choice of treating this as zero in the same way as the CLUSTAN package or of giving it the value of -1 within the program and omitting any comparisons for which either or both the objects have a missing value. This means that my data is no longer binary, since it may take 3 possible values, and the extension to multistate data is obvious. Here I refer to unordered multistate data, where it is possible to test whether values are all completely different and this difference cannot be quantified. For example, the shape of a cross-section may be circular, triangular or square. I provide a limited selection of similarity coefficients expressed in terms of the following sums.

x = number of valid agreements.
y = number of valid disagreements.
z = number of invalid comparisons (since one or both of the attributes is missing).
n = number of attributes = x+y+z

The variable L is used to indicate the choice of attribute.

L=1.	$S = 1.0 - y/n.$	Manhatten metric.
L=2.	$S = 1.0 - \sqrt{y/n.}$	Euclidean metric.
L=3.	$S = x/n.$	Simple matching coefficient.
L=4.	$S = 2x/(2x+y).$	Dice-Sorensen.
L=5.	$S = x/(x+2y).$	Rogers-Tanimoto.

Numeric Data

In my program, I accept ordered multistate data and treat it as a special case of numeric data since the values are ordered and distances between them have the same significance as for the numeric, but they are limited to certain discrete values. Thus the selection of coefficients provided for the numeric data also apply to the ordered multistate data. Once again, the variable L is used to indicate choice of attribute and in this case x_i is the value of the attribute for object i and x_j the value for object j. The summation is taken over all the attributes.

L=1.	$S = 1.0 - \Sigma\lvert x_i - x_j\rvert / x_{max}$	Manhattan metric.
L=2.	$S = 1.0 - 1/n \sqrt{\Sigma(x_i-x_j)^2/x_{max}^2}$	Euclidean metric.
L=3.	$S = 1.0 - \left[1/n\ \Sigma(x_i-x_j)/x_{max}\right]$	size difference.
L=4.	$S = 1.0 - \left[\Sigma 1/n(x_i-x_j)^2/x^2_{max} - \left(1/n\ \Sigma(x_i-x_j)/x_{max}\right)^2\right]$	shape difference.
L=5.	$S = \Sigma(x_i.x_j)/\sqrt{(\Sigma x_i^2)(\ x_j^2)}$	Cosine.

These 5 coefficients are the only ones provided at present, but others, such as the product moment correlation, may be added later.

All the coefficients, both binary and numeric, are scaled to lie in the range 0.0 to 1.0. The variable x_{max} used to scale the numeric data, is calculated from the data matrix and is simply the maximum range of values for the attribute being evaluated.

The FORTRAN Program

The program accepts as input, a data matrix containing the values of the attributes for each of the objects in the study. It assumes that for each object in turn, the data cards will contain the values of N1 numeric attributes, followed by N2 binary attributes, followed by N3 multistate attributes, followed by N4 unordered multistate attributes. At present, the total number of attributes, N1+N2+N3+N4, must not exceed 40 and the total number of objects must not exceed 99. It is easy to increase either or both of these limits, but the assumption that all the attributes of a particular type are gathered together in a single block is part of the design of the program and could not be changed without extensive rewriting and a loss of efficiency.

The first data card is a control card specifying values of the following variables; N1,N2,N3,N4,MISS,K, IPSIM

N1 to N4 are the number of each type of attribute.

MISS=1 if missing values are to be distinguished from zero ones and MISS=0 if this facility is to be suppressed.

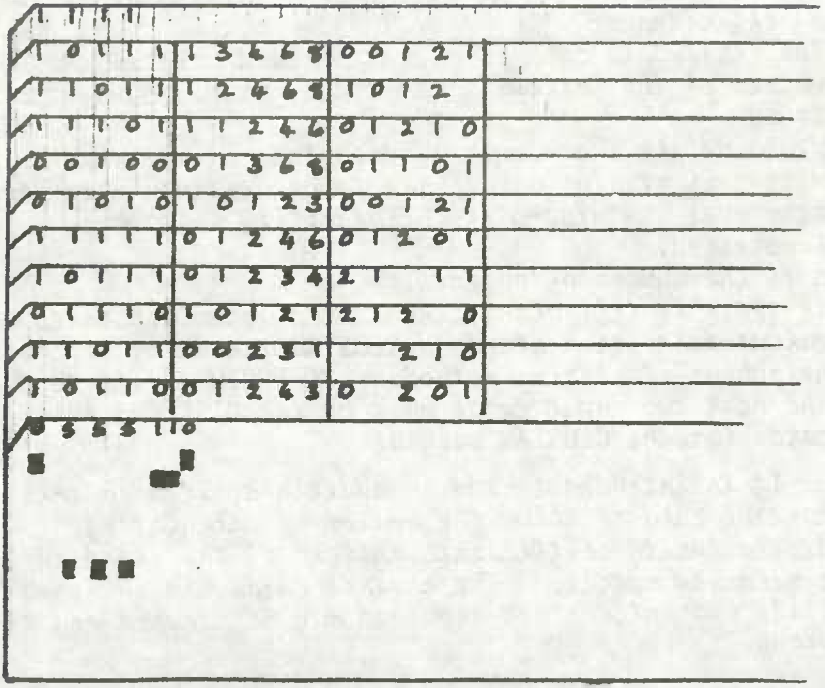
K is the number of objects.

If IPSIM is left blank, then the program generates output data for a standard CLUSTAN run using Nearest-neighbour clustering method. If IPSIM is set to 4, the next two input cards must be the DIST and ANAL data cards for the CLUSTAN package.

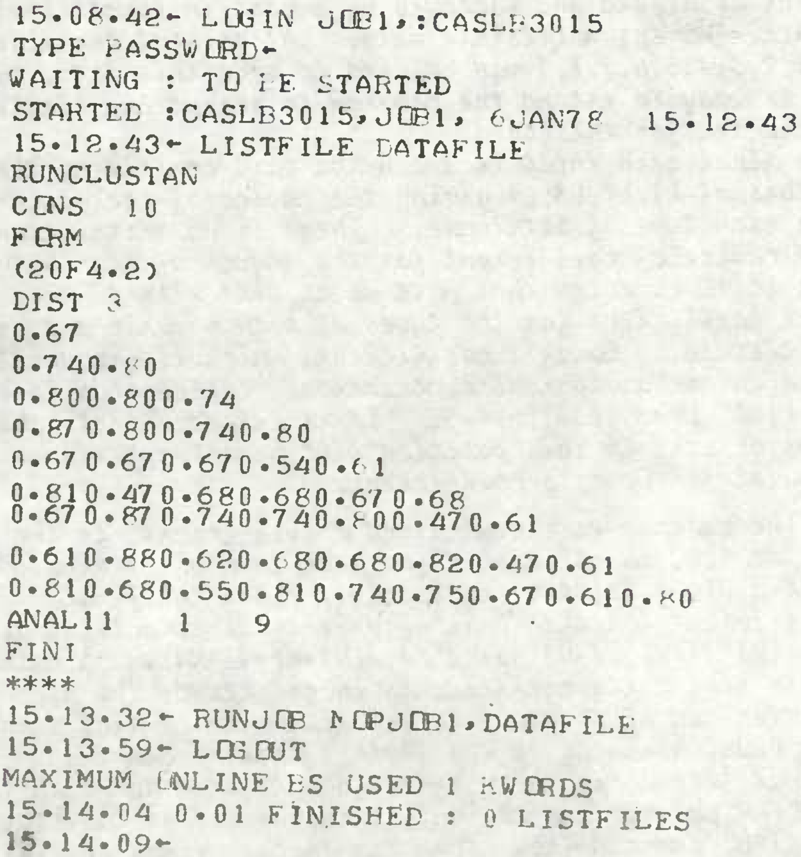
It is intended to add an additional item to this control card to allow the option of calculating dissimilarity coefficients instead of the standard similarity matrix. The control cards are in format (812) and options not required may be ignored and left blank,

Figure 1.

Set of Test Data used to
produce the datafile opposite.



```
1 0 1 1 1 | 1 3 4 6 8 | 0 0 1 2 1
1 1 0 1 1 | 1 2 4 6 8 | 1 0 1 2
1 1 1 0 1 | 1 1 2 4 6 | 0 1 2 1 0
0 0 1 0 0 | 0 1 3 6 8 | 0 1 1 0 1
0 1 0 1 0 | 1 0 1 2 3 | 0 0 1 2 1
1 1 1 1 1 | 0 1 2 4 6 | 0 1 2 0 1
1 0 1 1 1 | 0 1 2 3 4 | 2 1   1 1
0 1 1 1 0 | 1 0 1 2 1 | 2 1 2   0
1 1 0 1 1 | 0 0 2 3 1 |     2 1 0
1 0 1 1 0 | 0 1 2 4 3 | 0   2 0 1
0 5 5 5 1 1 0
```



```
15.08.42← LOGIN JOB1,:CASLB3015
TYPE PASSWORD←
WAITING : TO BE STARTED
STARTED :CASLB3015,JOB1, 6JAN78  15.12.43
15.12.43← LISTFILE DATAFILE
RUNCLUSTAN
CONS  10
FORM
(20F4.2)
DIST 3
0.67
0.740.80
0.800.800.74
0.870.800.740.80
0.670.670.670.540.61
0.810.470.680.680.670.68
0.670.870.740.740.800.470.61
0.610.880.620.680.680.820.470.61
0.810.680.550.810.740.750.670.610.80
ANAL11   1   9
FINI
****
15.13.32← RUNJOB MOPJOB1,DATAFILE
15.13.59← LOGOUT
MAXIMUM ONLINE BS USED 1 KWORDS
15.14.04 0.01 FINISHED : 0 LISTFILES
15.14.09←
```

Figure 2.

Listing of data file and
Runjob command for 1906A
installed at Birmingham

The data cards for each object in turn are then supplied, the numeric data in format (16F5.0) occupying as many cards as needed and followed by a card in format (40(1X,I1)) containing the multistate data. At present the digits 0,1,2,3,4,5,6,7,8,9 may be used to code this data, but it is easy to extend the program to accept any single character for this data.
The final card input is a control card containing the values of L1,L2,L3,L4 giving the choice of coefficient for each type of attribute. There is no restriction on the choice of coefficient for the types of attribute, but it is expected that most users will wish to have the same coefficient for the types of both numeric and ordered multistate. Gower's coefficient, which is frequently used in archaeological applications, corresponds to a choice of L1=1, L2=3, L3=1, L4=3. It is intended that a blank control card in this position will be interpreted as a request for Gower's coefficient.

The program is executed and a coefficient, in the range 0.0 to 1.0, is calculated for each type of attribute and called S1,S2,S3,S4 respectively. Then the final coefficient for this pari of objects is calculated as S = (N1*S1+N2*S2+N3*S3+N4*S4)/(N1+N2+N3+N4). A matrix of these similarity coefficients is printed on the line printer and also output to the file with control cards for the CLUSTAN run. On the 1906A system at Birmingham, the output file is in the correct format for a run with the CLUSTAN package and the only thing needed is to issue a "RUNJOB" command from a Mop terminal. Minor alterations will be needed for other installations, but as my program is written in standard FORTRAN IV for easy portability these should not be a great problem.

Examples of input and output are given in figures 1 and 2 of this paper.