DENCOMPAR

User's Manual

By J. Podani

Introduction

The WINDOWS application DENCOMPAR has been written to compare two dendrograms based on the exhaustive search procedure proposed by Podani et al. (submitted). Each of the two dendrograms is described in terms partitions (cut levels). Then, all partitions that can be derived from dendrogram 1 are compared with all possible partitions from dendrogram 2. Five different coefficients are calculated, as described below.

A partition P can be described unequivocally in terms of a symmetric incidence matrix **X** in which $x_{ij} = 1$ if objects i and j belong to the same class in P, otherwise $x_{ij} = 0$. Then, the dissimilarity between two partitions, P and Q, of the same set of objects is calculated based on the cells of a 2×2 contingency table defined as follows: a is the number of object pairs that are in the same class in both partitions, b is the number of object pairs classified together only in P, c is the number of object pairs appearing in the same class only in Q, and d is the number of object pairs that do not occur in the same class in both partitions being compared. Clearly, $a+b+c+d = \binom{m}{2}$. These values are used by a number of distance or dissimilarity coefficients suggested for general use in classification of presence-absence data, and under different names for use in comparison of partitions. A short list of these functions, generally abbreviated as D_{PQ} is given below:

$D(\mathrm{R})_{\mathrm{PQ}} = (b+c) / m$ complement of simple matching coefficient (= 1 – "Rand index", Rand 1971),

$D(\mathrm{E})_{\mathrm{PQ}} = \sqrt{b+c}$ Euclidean distance (="PAIRBONDS", Arabie and Boorman 1973),

$D(\mathrm{J})_{\mathrm{PQ}} = (b + c) / (a+b+c)$ 1 – Jaccard index (Downton and Brennan 1980),

$D(\mathrm{S})_{\mathrm{PQ}} = (b + c) / (2a+b+c)$ 1 – Sorensen index ("percent mutual matches", Arabie and Boorman 1973).

Indices R, J and S do not use their ranges of [0,1] completely, and the minimum of E is rarely zero, because there are unavoidable agreements between partitions even if they are maximally dissimilar. Thus, some form of normalization is necessary (Morey and Agresti 1984, Hubert and Arabie 1985). Milligan and Cooper (1986) found that the adjusted version of the Rand index, as proposed by Hubert and Arabie (1985) outperforms the other coefficients, and therefore suggested it for general use. Its complement is also calculated by DENCOMPAR.

Input files – a sample run

In the example, we compare two dendrograms according to the data stored in files test1.dat

```
Dendrogram 1
8
1 2 1 1 0
4 5 1 1 0
1 3 2 1 0
1 4 3 2 1
1 6 5 1 2
1 7 6 1 3
1 8 7 1 4
```

and test2.dat

```
Dendrogram 2
8
1 2 1 1 1
1 4 2 1 2
5 6 1 1 2
1 3 3 1 3
1 5 4 1 4
7 8 1 1 5
1 7 6 2 6
```

These dendrogram files follow the SYNTAX 2000 dendrogram format. The first line in each file is a title, the second line contains the number of objects (m) classified. Then follow m-1 lines, each corresponding to a given fusion step in clustering: the first two numbers are cluster identifiers, the second two are the respective cluster sizes and the last value is the dissimilarity level.

Sample run and output

After double-clicking the icon of DENCOMPAR, a dialog screen appears on the monitor. Only two filenames have to be specified by the user. Then, the program outputs the number of distinct levels in the dendrograms and after the calculations it stops.

```
  ENTER FILENAME FOR dendrogram 1
TEST1.DAT
  ENTER FILENAME FOR dendrogram 2
TEST2.DAT
 NO OF LEVELS =  5  6

  MINIMUM VALUES

  1 - Jaccard
 VALUE =    .00000 AT NO. OF CLUSTERS     3    3
  1 - Simple match = 1 - RAND
 VALUE =    .00000 AT NO. OF CLUSTERS     3    3
  Euclidean
 VALUE =    .00000 AT NO. OF CLUSTERS     3    3
  1 - Sorensen
 VALUE =    .00000 AT NO. OF CLUSTERS     3    3
  1 - ADJUSTED RAND
 VALUE =    .00000 AT NO. OF CLUSTERS     3    3

  END OF CALCULATIONS, SEE FILE MATRICES.DAT
```

In the file MATRICES.DAT we find five output matrices. Columns correspond to partitions from dendrogram 2, rows to partitions in dendrogram 1. The numbers of clusters are shown on the top row and in the first columns.

```
   0   7   5   4   3   2
  1 - jaccard
   5  .750000  .857143  .625000  .733333  .750000
   4  .900000  .727273  .454545  .333333  .375000
   3  .933333  .733333  .533333  .000000  6.250000E-02
   2  .952381  .809524  .666667  .285714  .318182
  1 - simple match = 1 - RAND
   5  .107143  .214286  .178571  .392857  .428571
   4  .321429  .285714  .178571  .178571  .214286
   3  .500000  .392857  .285714  .000000  3.571429E-02
   2  .714286  .607143  .500000  .214286  .250000
  Euclidean
   5  1.73205  2.44949  2.23607  3.31662  3.46410
   4  3.00000  2.82843  2.23607  2.23607  2.44949
   3  3.74166  3.31662  2.82843  .000000  1.00000
   2  4.47214  4.12311  3.74166  2.44949  2.64575
  1 - Sorensen
   5  .600000  .750000  .454545  .578947  .600000
   4  .818182  .571429  .294118  .200000  .230769
   3  .875000  .578947  .363636  .000000  3.225806E-02
   2  .909091  .680000  .500000  .166667  .189189
  1 - ADJUSTED RAND
   5  .636364  .875000  .555556  .747573  .777778
   4  .875000  .717949  .416667  .350000  .411765
   3  .937799  .747573  .551724  .000000  7.216495E-02
   2  .975610  .894737  .800000  .444444  .538462
```

Zero value is given first in the first row/first columns to ensure correct export to Excel table format.

References

Arabie, P. & S. A. Boorman. 1973. Multidimensional scaling of measures of distance between partitions. *J. Math. Psychol.* 10: 148-203.

Downton, M. & T. Brennan. 1980. Comparing classifications: an evaluation of several coefficients of partition agreement. *Classification Soc. Bull.* 4 (4):53-54.

Podani, J., Á. Major & A. Engloner. Multilevel Comparison of Dendrograms: A New Method with Application to Genetic Classifications. Submitted to *Statistical Applications in Genetics and Molecular Biology.*

Fowlkes, E. B. & C. L. Mallows. 1983. A method for comparing two hierarchical clusterings. *J. Amer. Stat. Assoc.* 78: 553-584.

Hubert, L. & P. Arabie. 1985. Comparing partitions. *J. Classification* 2:193-218.

Milligan, G. W. & M. C. Cooper. 1986. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar. Behav. Res.* 21: 441-458.

Morey, L. & A. Agresti. 1984. The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educ. Psychol. Measurement* 44: 33-37.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.* 66: 846-850.