Pak.J.Bot., 24(1): 119-124, 1992.

ASSIGNMENT OF TAXONOMIC UNITS WITH PROBABILISTIC CHARACTERS TO GROUPS: A MONTE CARLO METHOD

S. SHAHID SHAUKAT AND S. Z. HUSAIN *

Department of Botany,
University of Karachi, Karachi-75270, Pakistan.

Abstract

A Monte Carlo algorithm is proposed to assign a new or unclassified operational taxonomic unit (O.T.U.), with variable characters, to the group it is closest to. The method involves a bootstrap resampling plan. The algorithm is limited to binary data but may easily be extended to cope with data sets involving qualitative as well as quantitative characters.

Introduction

Once a classification of the operational taxonomic units (O.T.Us.) is accomplished it is possible to assign a new O.T.U. to the group it is closest to, assuming that it has come from one of the g groups (Pankhurst, 1975; Dunn & Everitt, 1982). A variety of assignment techniques are available including diagnostic keys (Payne & Preece, 1980), Bayesian analysis (Wilcox *et al.,* 1980), discriminent functions (O'Donnell *et al.,* 1980; Ganesalingam, 1989) and I-divergence information (Orloci & Kenkel, 1985).

For most data sets each character of an O.T.U. is assigned one value; in case of binary data 0 or 1. Problem arises when an O.T.U. whose assignment is sought has variable (probabilistic) characters. The method proposed in this paper addresses the problem of assignment of such O.T.Us. to their appropriate group using a Monte Carlo simulation.

THE ALGORITHM

Suppose that a sample of O.T.U's with binary data is partitioned into g groups using Euclidean distance as resemblance function and one of the SAHN clustering strategy (Sneath & Sokal, 1973). Let $\underset{\sim}{X}$ be an O.T.U. to be assigned having elements $X_1, X_2, \ldots\ldots, X_t$ with X_i being binary (0, 1). One or more elements of $\underset{\sim}{X}$ can take up values 0 or 1 with probability P (X_i). Let $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ be the groups of k and l O.T.U's respectively (g = 2). The assignment of $\underset{\sim}{X}$ to either $\underset{\sim}{Y}$ or $\underset{\sim}{Z}$ is tested as follows:

1. Define $\underset{\sim}{X}$ with $X_i = 0$ or 1 according to whether the character i is absent or present in all the specimens or by generating the values 0 or 1 with probability P (X_i) using random numbers from a uniform distribution. If character i is present in r out of s specimens then $P(X_i)|X = 1$ is r/s and conversely $P(X_i)|X_i = 0$ is 1-r/s. Thus when RND $\leq$ r/s then X_i is set equal to 1, when RND $>$ r/s then X_i is set equal to 0, where $0 \leq RND \leq 1$.
2. Generate k random integers j such that $1 \leq j \leq k$. Each j is the label of an O.T.U. belonging to group $\underset{\sim}{Y}$.

* Department of Botany, University of Reading, Reading, U.K.

3. Calculate Euclidean distance between $\underset{\sim}{X}$ and each of the k randomly selected O.T.U. of group $\underset{\sim}{Y}$.

$$D_{YX}_{j} = [\sum_{i=1}^{t} (Y_{ij}-X_i)^2]^{1/2} \quad 1 \leq j \leq k$$

4. Calculate mean Euclidean distance between O.T.U. $\underset{\sim}{X}$ and group $\underset{\sim}{Y}$.

$$\bar{D}_{YX} = \sum_{1 \leq j \leq k}^{k} D_{YX}/k$$

5. Repeat steps 1 to 4 a large number of times (B) say 1000 and find mean and variance as;

$$\bar{\bar{D}}_{YX} = \sum_{i=1}^{B} \bar{D}_{YX}^{(i)} / B$$

$$Var(\bar{D}_{YX}) = 1/B \left[\sum_{i=1}^{B} \bar{D}_{YX}^{(i)} - \bar{\bar{D}}_{YX})^2\right]$$

6. Perform similar procedure with group $\underset{\sim}{Z}$ replacing $\underset{\sim}{Y}$ and find $\bar{\bar{D}}_{ZX}$ and Var $(\bar{D}_{ZX})$.
7. Assign $\underset{\sim}{X}$ to either $\underset{\sim}{Y}$ or $\underset{\sim}{Z}$ as follows: If $\bar{\bar{D}}_{YX} < \bar{\bar{D}}_{ZX}$ then $\underset{\sim}{X} \in \underset{\sim}{Y}$, conversely if $\bar{\bar{D}}_{ZX} < \bar{\bar{D}}_{YX}$ then $\underset{\sim}{X} \in \underset{\sim}{Z}$. Because the variances are known the significance of difference in the mean distances $\bar{\bar{D}}_{YX}$ and $\bar{\bar{D}}_{ZX}$ can be tested.

Program ASSIGNMENT1 performs the computations in accordance with the algorithm described above.

AN APPLICATION

Based on the presence and absence of 18 flavonoids, 17 O.T.Us, belonging to genus *Thymus* (Table 1) were classified using sum of squares clustering (Orloci & Kenkel, 1985). Three main groups were obtained. One of the O.T.U. *T. longiflorus* which fell in group 2 had indicated the presence of one flavonoid (character 11) in 60% of the specimens tested (Table 1). With another exception of one flavonoid (character 14) in *T. cephalotus* (from Portugal), the rest of the flavonoids were either absent or present in all the specimens examined for a given O.T.U. Since groups 1 and 2 were closely related in the classification hierarchy the assignment of *T. longiflorus* was tested to these groups using the proposed algorithm. The grand mean distance between *T. longiflorus* (x) and group 1 was found to be $\bar{\bar{D}}_{1x}$ = 1.3769 with a variance Var $(\bar{D}_{1x})$ = 0.1182 and with group 2 $\bar{\bar{D}}_{2x}$ = 1.4479 with a variance Var $(\bar{D}_{2x})$ = 0.1651. This suggested a higher likelihood of membership of *T. longiflorus* to group 1 rather than group 2. This is contrary to the result of cluster analysis for which the character 11 was scored as 1. Another reason for the affinity of *T. longi florus* to

Table 1. Flavonoids found in the leaves of various taxa, O.T.U's (genus *Thymus*). (Data from Hussain *et al.*, manuscript).

	Flavonoids																		
Taxa (O.T.U's)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Group 1																			
T. cephalotus (Portugal)	1	1	0	0	1	1	1	0	1	1	0	1	0	1*	0	0	0	1	
T. dolopicus	1	0	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	1	
T. parnassicus	1	0	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	1	
T. cephalotus (Spain)	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	0	0	1	
T. longiflorus	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	0	0	1	
var. ciliatus																			
T. vulgaris (Spain)	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	0	1	1	
T. villosus	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0	0	0	1	
T. villosus	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0	0	0	1	
var. lusitanicus																			
Group 2																			
T. leucotrichus	1	0	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	1	
(E. Macedonia)																			
T. mastegophorus	1	1	1	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	
T. zygis	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	1	
T. membranaceus	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1	
O.T.U. to be assigned																			
T. longiflorus	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1	

* Present in 25% specimens, present in 60% specimens. 1 = Luteolin-7- glucoside; 2 = Luteolin -7-glucuronide; 3 = Luteolin-7, 4-diglucoside (a); 4 = Luteolin-7, 4-diglucoside (b); 5 = Luteolin -5-glucoside; 6 = Apigenin-7- glucoside; 7 = Apigenin-7-glucuronide; 8 = Vicenin-2; 9 = 60H-luteolin -7-glucoside; 10 = Scutellarein 6, 4-dimethyl ether; 11-18 = unidentified.

group 1 is that this group was more compact than group 2. The result of the test, nonetheless clearly demonstrates the necessity for testing the assignment of O.T.Us with probabilistic characters to their probable parent groups.

Discussion

The basic advantage of a Monte Carlo test is that the investigator is free to use a variety of informative statistics of his own choice rather than be dictated by distributional assumptions. The Monte Carlo assignment technique advocated in this paper involves resampling from the group with which membership of the unknown individual (O.T.U.) is sought. This process is essentially a bootstrap method (Efron, 1982). The algorithm is such that the members (O.T.Us) within the group are not reused equal number of times in conformity with the original bootstrap procedure. Alternatively, repeated sampling can be achieved by jack knife, half-sampling, sub-sampling, generalized jackknife and balanced bootstrap resampling plans (Efron, 1981; Gleason, 1988). The latter procedure uses sample observations exactly equally often (Davidson *et al.*, 1986) and appears attractive since such balancing can yield appreciable gains in terms of bias and variance reduction. A number of

efficient algorithms of balanced bootstrap sampling have been developed (Davidson *et al.*, 1986; Gleason, 1988). When balanced bootstrap resampling was employed in the simu lation $\bar{D}_{1x}$ was found to be 1.3533 with a variance Var $(\bar{D}_{1x}) = 0.1155$ and $\bar{D}_{2x}$ was found 1.4655 with Var $(\bar{D}_{2x}) = 0.1601$. The program ASSIGNMENT2 accomplishes the procedure. This program is available from the senior author on request.

The application of the assignment algorithm presented here is limited to binary data. However, the procedure can be readily extended to situation where the O.T.Us are described by a set of variables which include binary, qualitative and quantitative measures. In such a case, instead of using dissimilarity viz., Euclidean distance, a similarity coefficient proposed by Gower (1971) can be employed. The character states of an unclassified O.T.U. with variable characters can be simulated by generating random numbers from various probability distributions. For instance a continuous variable can be simulated by drawing random numbers from a normal distribution. The assignment criterion would be the maximum average similarity of the unclassified O.T.U. with a certain group. The details of this algorithm would be described elsewhere.

References

Davidson, A.C., D.V. Hinkley and E. Schechtman. 1986. Effi cient bootstrap simulation. *Biometrika.*, 73: 555-566.

Dunn, G. and B.S. Everitt. 1982. *An introduction to mathematical taxonomy.* Cambridge Univ. Press, Cambridge.

Efron, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika.*, 68: 589-599.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. *Conference Series in applied mathematics, Report 38. Society for Industrial and Applied Mathematics, Phila delphia.*

Ganesalingam, S. 1989. Classification and mixture approaches to clustering via maximum likelihood. *Appl. Statist.*, 38: 455-466.

Gleason, J.R. 1988. Algorithms for balanced bootstrap simulations. *Amer. Statist.*, 42: 263-266.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics.*, 27: 857-872.

Hussain, S.Z., S.S. Shaukat and G.M. Sinnreich. 1992. A chemosystematic study of the genus *Thymus,* section *Pseudothymbra,* Manuscript.

O'Donnell, A.G., H.J.H. FacFie and J.R. Norris. 1980. An investi gation of the relationship between *Bacillus cereus, Bacillus thuringiensis* and *Bacillus mycoides* using pyrolisis gas-liquid chromatography. *J. Gen. Microbiol.*, 119: 109-194.

Orloci, L. and N.C. Kenkel. 1985. *Introduction to data analysis: with examples from population and community ecology.* ICPH, Fair land, Maryland.

Pankhurst, R.J. 1975. *Biological identification with computers.* Academic Press, London.

Payne, R.W. and D.A. Preece. 1980. Identification keys and diagnostic tables-a review. *J. Royl. Statist. Soc., Ser.* A 143: 253-282.

Sneath, P.H.A. and R.R. Sokal. 1973. *Numerical Taxonomy.* W.H. Freeman, San Francisco.

Wilcox, W.R., S.P. Lapage and B. Holmes. 1980. A review of numerical methods in bacterial identification. *Antonie van Leeuwenhoek,* 46: 233-299.

(Received for Publication 3 February 1992)

APPENDIX

Computer program

```
10      REM PROGRAM: ASSIGNMENT1
20      REM PROBABILITY FILE CONTAINS THE
30      REM PROBABILITY OF PRESENCE.
40      INPUT "SPECIFY GROUP DATA FILE ";F$
50      INPUT "SPECIFY PROBABILITY FILE ";G$
60      INPUT "NUMBER OF ROWS (VARIABLES) ";T
70      INPUT "# OF INDIV. IN THE GROUP ";K
80      INPUT "# OF RESAMPLES ";B
90      DIM X(T), P(T), Y(T,K), P1(T,K)
100     DIM S(K)
110     PRINT CHR$ (4);"OPEN ";F$
120     PRINT CHR$ (4);"READ ";F$
130     FOR I = 1 TO T: FOR J = 1 TO K
140     INPUT A: Y(I,J) = A
150     NEXT J,I
160     PRINT CHR$ (4);"CLOSE ";F$
170     PRINT CHR$ (4);"OPEN";G$
180     PRINT CHR$ (4);"READ";G$
190     FOR i = 1 TO T: FOR J = 1 TO K
200     NEXT J,I
220     PRINT CHR$ (4);"CLOSE";G$
230     PRINT "PROVIDE THE UNKNOWN VECTOR "
240     FOR i = 1 TO T
250     PRINT "VARIABLE # ";I
260     INPUT X (I)
270     PRINT "BROBABILITY OF X(";I;")"
280     INPUT P(I)
290     NEXT I
300     PRINT
310     REM CREATE PROBABILISTIC VECTOR
320     FOR I = 1 TO T
330     IF P(I) = 1 OR P(I) = O THEN 370
340     R = RND (1)
350     IF R < = P(I) THEN X(I) = 1: GOTO 370
360     X(I) = 0
370     NEXT I
380     PRINT
390     PRINT "SIMULATED VECTOR "
400     FOR I = 1 TO T: PRINT X(I);" ";: NEXT I
410     PRINT
```

```
420     FOR I = 1 TO T
430     FOR J = 1 TO K
440     IF P1(I,J) = 1 OR P1(I,J) = 0 THEN 480
450     R2 = RND (1)
460     IF R2 < = P1(I,J) THEN Y(I,J) = 1: GOTO 480
470     Y(I,J) = 0
480     NEXT J
490     NEXT I
500     PRINT : PRINT "GROUP MATRIX"
510     FOR J = 1 TO K
520     FOR I = 1 TO T
530     PRINT Y(I.J); " ";
540     NEXT I: PRINT : NEXT J
550     PRINT
560     TT = 0
570     FOR I = 1 TO K
580     R3 = INT (K * RND (1)) + 1
590     S(I) = R3
600     NEXT I
610     FOR L = 1 TO K
620     J = S(L)
630     D1 = 0
640     FOR I = 1 TO T
650     D1 = D1 + (Y(I,J) - X(I)) ^ 2
660     NEXT I
670     D2 = SQR (D1)
680     TT = TT + D2
690     NEXT L
700     E1 = TT / K
710     PRINT "AVERAGE DISTANCE = ";E1
720     T1 = T1 + E1
730     T2 = T2 + E1 ^ 2
740     C = C + 1
750     IF C < B THEN 310
760     M = T1 / B
770     PRINT  CHR$ (4); "PR#1"
780     PRINT
790     PRINT "GRAND MEAN DISTANCE=";M
800     V = ((T2 - T1 ^ 2 / B) / B)
810     PRINT 'VARIANCE=";V
820     PRINT CHR$ (4); "PR#O"
```