Rivista di Statistica, Vol. 18, N. 4 (Dicembre) 1985

STATISTICA COMPUTA-ZIONALE

Rubrica a cura di Natale Lauro

BASIC program for cluster analysis in numerical taxonomy

Francesco La Rosa, Dipartimento di Biologia Cellulare, Università degli Studi di Camerino.

Summary

This program IMCTAX, written in BASIC language for OLIVETTI E.S.E., applies the cluster analysis (UPGMA method) to similarity or dissimilarity matrixes derived from original data, by one of follow indexes: Ssm (Simple Matching Coefficient), So (Degree of Overlap between Superimposed Traces for Percent Series), D (Taxonomic Distance) and Sg (General Similarity Coefficient of Gower). The program IMCTAX could be utilized for many studies in numerical taxonomy.

1. Introduction

One problem among the greatest in taxonomic studies consists in the elaboration of a large number of parameters calculated with both quantitative and qualitative characters. The grouping by numerical methods of taxonomic units (OTU = Operational Taxonomic Unit) into taxa on the basis of their character states is called *numerical taxonomy* (Sneath, 1973). The estimation of resemblance between pair of OTUs is the most important and foundamental step in this procedure. The computation of a measure of resemblance (or distance) can be done in a variety of ways in relation to the type of studied characters; resemblance (or distance) coefficients are tabulated in matrix form with one coefficient for every pair of taxonomic entities (OTU). On this coefficient matrix you can apply the cluster analysis on the aim to define hierarchic groups utilized in the taxonomic studies.

In the commonest softwares, the used method for the cluster analysis consists in the *Single Linkage Clustering* applied to matrixes constructed by Bravais-Pearson correlation coefficients. Nevertheless some Authors don't always consider either possible or correct the application of «r» coefficient, and they think that the application of Single Linkage Method is not generalizable (Sneath, 1973).

For the reason given above, in this program — called IMCTAX — the method of *Average Linkage Clustering* is used and applied to matrixes constructed by several distance or similarity coefficients as: Ssm (Simple Matching Coefficient), So (Degree of Overlap between Superimposed Traces for Percent Series), D (Taxonomic Distance) and Sg (General Similarity Coefficient of Gower) (Bousfield, 1983; Dunn, 1982; Sneath, 1973).

2. Program characteristics and methods

The IMCTAX program has been written in BASIC language and adapted to Exended System Environment on OLIVETTI M40. The program is available in two versions: with utilization of external files and without them.

The coefficient used for the matrix calculation are:

1) *Ssm* = Simple Matching Coefficient for binary (+ or −) or dichotomous characters.

$$Ssm = (a + d) / (a + b + c + d)$$

from

		OTU_k	
		+	−
	+	a	b
OTU_j	−	c	d

2) *So* = Degree of Overlap between Superimposed Traces for Percent Series.

$$So = 100 - 1/n\Sigma_{i=1}^{n} |x_{ij} - x_{ik}|$$

where n is the number of percent series and x_{ij} and x_{ik} are the observations relative to j^{th} and k^{th} OTU respectively.

3) *D* = Taxonomic Distance, as euclidean distance, for quantitative characters.

$$D = [\Sigma_{i=1}^{n} (x_{ij} - x_{ik})^2]^{1/2}$$

where x_{ij} and x_{ik} are those in 2).

4) Sg = General Similarity Coefficient of Gower, for a mixture of binary characters, qualitative characters with more than two stages, and quantitative characters.

$$Sg = \Sigma_{i=1}^{n} S_{ijk} / \Sigma_{i=1}^{n} W_{ijk}$$

The weight W_{ijk} and S_{ijk} assume many values in relation to the used characters (Dunn, 1982).

The cluster grouping method is the *UPGMA* (Unweighted pairgroup method using arithmetic averages). The similarity $U_{l,jk}$ between every pair of OTUs is calculated as:

$$U_{l,jk} = (t_j / t_{j,k}) U_{l,j} + (t_k / t_{j,k}) U_{l,k}$$

where t is the number of OTUs of every cluster (Sneath, 1973).

The IMCTAX program provides control and correction procedures of data input.

The flowchart is reported in Fig. 1.

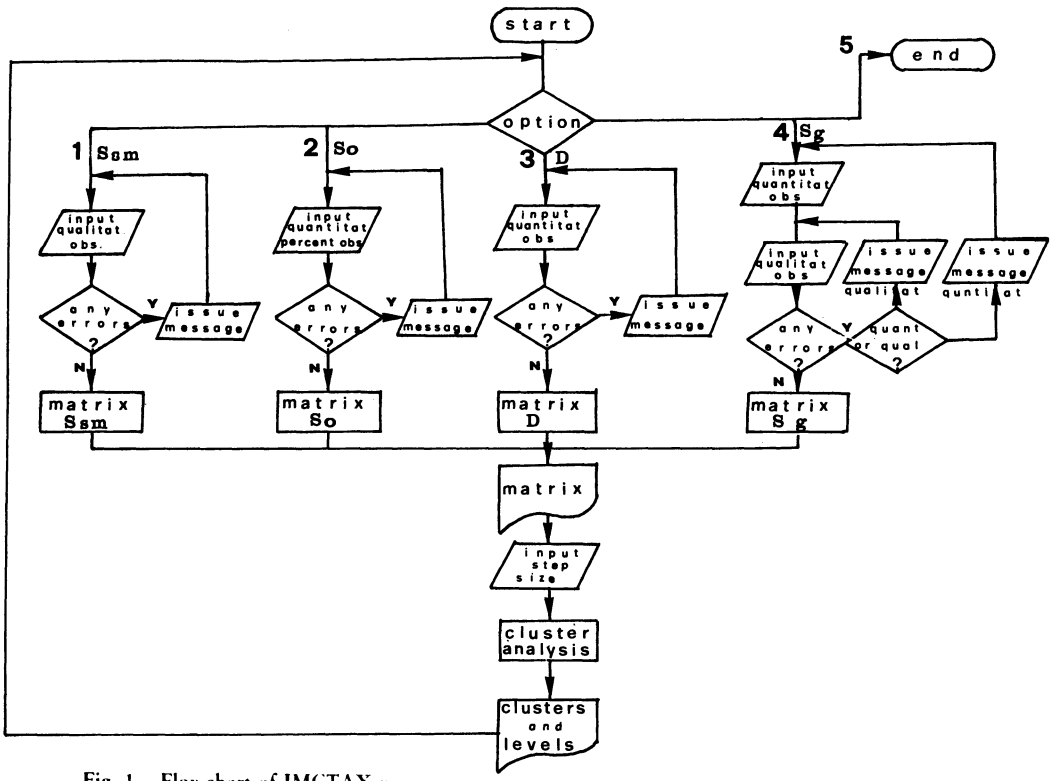

Fig. 1 – Flox-chart of IMCTAX program.

The IMCTAX program has been written in BASIC language on OLIVETTI M40 (E.S.E.) and then it is related to its hardware. The utilization of flowchart reported in Fig. 1 could avoid the restrictions to utilize the software in different computers. The listing program, on the other hand, is available and can be obtained from the Author; the language can be traslated, with necessary modification, for the requirements of other kind of computers.

3. Example of application

Some phenotipic properties of four hypothetical bacterial clones have been reported in Tab. 1 (Dunn, 1982).

Tab 1

	Presence of spores	*Mean cell diameter (μm)*	*GC in DNA*	*Colony morphology*	*Colony colour*
Clone 1	+	*1.0*	*57*	*Smooth*	*Brown*
Clone 2	+	*0.7*	*70*	*Rough*	*White*
Clone 3	−	*0.5*	*30*	*Mucoid*	*Yellow*
Clone 4	−	*0.7*	*40*	*Smooth*	*Yellow*

The first character — presence of spores — is binary; the second and the third ones — mean cell diameter and per cent of guanine + cytosine in DNA — are quantitative; the fourth and the fifth ones are qualitative three-stages.

The choice must fall on «General Coefficient of Gower», which is an index for a mixture of several characters.

The program asks the option:

The matrix (max 50 OTU's) is defined as:
≠ 1 Ssm similarity matrix
≠ 2 So similarity matrix
≠ 3 D dissimilarity matrix
≠ 4 Sg similarity matrix
≠ 5 END

and the input must be: 4.
After *ENTER THE NUMBER OF OTU's?*
and the input must be: 4.
The number of OTU's is 4
After *ENTER THE NUMBER OF QUANTITATIVE OBSERVATIONS?*
and the input must be: 2.

The number of quantitative observations is 2.

			OBSERVATION ≠ 1
1	*.7*	*.5*	*.7*
			OBSERVATION ≠ 2
57	*70*	*30*	*40*

Now the program asks:
ENTER THE NUMBER OF QUALITATIVE OBSERVATIONS?
and the input must be: 3.

The number of qualitative observations is 3.

The fourth and the fifth characters must be input only by the initial letter as it follows: Smooth = *S*, Rough = *R*, etc.

			OBSERVATION ≠ 1
+	+	−	−
			OBSERVATION ≠ 2
S	*R*	*M*	*S*
			OBSERVATION ≠ 3
B	*W*	*Y*	*Y*

Now the program asks:

≠ 1 Cluster analysis

≠ 2 Correction of data

If all inputs are right type # 1, otherwise type # 2 and the program asks which characters must be corrected.

For the cluster analysis the program writes the matrix of coefficient as it follows:

MATRIX OF COEFFICIENTS

OTU ≠ 1
Sg 1 2 .415
Sg 1 3 .065
Sg 1 4 .395

OTU ≠ 2
Sg 2 3 .12
Sg 2 4 .25

OTU ≠ 3
Sg 3 4 .5875

After the program asks the value of step size for the clustering:
ENTER STEP SIZE?
and the input, e.g., could be: .01

Step size is: .01

Now the program prints the clusters and the levels:

CLUSTER 3 4
LEVEL: .59

CLUSTER 1 2
LEVEL: .42

CLUSTER 1 2 3 4
LEVEL: .22

The program goes to the first request of index option. The choice of other indexes involves similar procedure both in input and in output.

References

Bousfield I.J., Smith G.L., Dando T.R. and Hobbs G.: Numerical Analysis of Total Fatty Acid Profiles in the Identification of Coryneform, Nocardioform and Some Other Bacteria, *J. Gen. Microbiol.* ***129***, 1983, 375-394.

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

Sneath P.H.A. and Sokal R.R.: *Numerical taxonomy. The principles and practice of numerical classification*, W.H. Freeman and Co., S. Francisco, 1973.

Sommario

Il programma IMCTAX, scritto in linguaggio BASIC applica l'analisi dei clusters con il metodo UPGMA, a matrici di similarità, o dissimilarità, costituite a partire dalle osservazioni originali, con uno dei seguenti indici: Ssm (simple matching coefficient), So (degree of overlap between superimposed traces for percent series), D (taxonomic distance) e Sg (general similarity coefficient of Gower).

Il programma, di semplice utilizzazione, trova utile applicazione in molti studi di tassonomia numerica.