L. Orloci, C. R. Rao, and W. M. Stiteler, (eds.),
Multivariate Methods in Ecological Work, pp. 203-209.

GENERALIZED STRATEGY FOR HOMOGENEITY-OPTIMIZING HIERARCHICAL CLASSIFICATORY METHODS

JÁNOS PODANI

Research Institute for Botany
Hungarian Academy of Sciences
2163 Vacratot, Hungary

SUMMARY. A new scheme, similar to the route-optimizing strategy of Lance and Williams (1966), is proposed for homogeneity-optimizing sorting procedures. Cluster homogeneity is defined in three ways and three algorithms compatible with the scheme are briefly discussed.

KEY WORDS. classification, sorting, strategies, clusters, homogeneity

1. INTRODUCTION

Some cluster analytical procedures used frequently in mathematical ecology and numerical taxonomy start with a resemblance matrix between entities and do not require the initial data. These methods have many computational advantages. The well-known hierarchical and agglomerative algorithms of this type have been called 'combinatorial' by Lance and Williams (1967) and reviewed by Cormack (1971). A basic problem of these strategies is the definition of inter-cluster similarity, distance, or dissimilarity. Lance and Williams (1966) gave a recurrence formula to compute the dissimilarity between group z_h and group z_{ij} obtained by the fusion of groups z_i and z_j :

$$d_{h(ij)} = \alpha_i d_{hi} + \alpha_j D_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \ . \qquad (1)$$

The values of parameters α, β, and γ are determined by the

nature of the strategy used (see Cormack, 1971). Algorithms satisfying relation (1), however, optimize the route by which the clusters are formed, such as in single linkage and complete linkage, the sum of squares agglomeration and the centroid sorting method. Internal structure or some kind of homogeneity of clusters is taken into consideration by the sum of squares agglomeration method only (see Ward, 1963; Orloci, 1967). This strategy, however, minimizes the increase of the within-group sum of squares so that the homogeneity of the new clusters is not necessarily optimal. Contrary to these route-optimizing strategies, it may be desired to optimize the homogeneity of the new clusters.

In the classification of plant and animal individuals, communities or other entities, the primary aim is to produce groupings whose homogeneity is as high as possible. Hierarchies obtained by even exact and well-defined route-optimizing procedures may be of secondary significance. It is reasonable to make a further distinction among the classificatory techniques. The family of hierarchical and agglomerative methods can be divided into two groups: the route-optimizing (called r-hierarchical) and the homogeneity-optimizing (called h-hierarchical) procedures. It will be shown that some of the h-hierarchical strategies are 'combinatorial.' This rather ambiguous term is used and accepted in this paper for lack of a better terminology.

The concept of homogeneity may of course be defined in a number of different ways. I shall use three definitions to illustrate my general classification scheme for h-hierarchical and combinatorial procedures. It is worth mentioning that most fruitful information-theroretical definitions are not compatible with any combinatorial model.

2. A NEW GENERAL SCHEME AND ITS APPLICATION

2.1 The Basic Equation. Let us assume that in the course of the computations we have already three clusters denoted by z_h, z_i, and z_j with the number of elements respectively n_h, n_i, and n_j. Let w_h, w_i, and w_j denote the homogeneity or heterogeneity of the clusters and let w_{hi} denote the homogeneity of group z_{hi} obtained by the fusion of z_h and z_i (see Sections 2.2, 2.3, 2.4 for definitions of cluster homogeneity). Thus we have the following semi-matrix,

$$\underset{\sim}{W} = \begin{bmatrix} w_h & w_{hi} & w_{hj} \\ & w_i & w_{ij} \\ & & w_j \end{bmatrix} ,$$

and vector

$$\underset{\sim}{N} = [n_h, n_i, n_j] .$$

In the case of homogeneity, let w_{ij} be the greatest value in the upper triangular portion of $\underset{\sim}{W}$. Then we amalgamate groups z_i and z_j to form a new group z_{ij}. After this we can compute w_{hij} from the pre-existing homogeneity measures in $\underset{\sim}{W}$ and values of $\underset{\sim}{N}$ using the following formula,

$$w_{hij} = \alpha_i w_{hi} + \alpha_j w_{hj} + \beta w_{ij} + \gamma_h w_h + \gamma_i w_i + \gamma_j w_j . \qquad (2)$$

If heterogeneity measures are given, the same relation holds. The values of the parameters for three h-hierarchical strategies may be found in Table 1.

Computations by the h-hierarchical and combinatorial clustering methods are based on the values of inter-entity matrix $\underset{\sim}{W}$, after calculation of which the original data need not be retained in the memory of the computer. Application of equation (2) differs from that of equation (1) since the values of the diagonal of $\underset{\sim}{W}$ are of importance. Strategies compatible with equation (2) are given below.

2.2 Optimization of Dispersion within New Clusters. This strategy is the h-hierarchical version of the sum of squares agglomeration method (Ward, 1963; Orloci, 1967; Wishart, 1969). The sum of squared distance from the centroid within cluster z_h is the measure of z_h's heterogeneity and is denoted by q_h. This quantity can be calculated from the distances between entities,

TABLE 1: Parameters for three homogeneity-optimizing strategies. ($n. = n_h+n_i+n_j$)

Name	α_i	α_j	β	γ_h	γ_i	γ_j
Edge-density	$\frac{(n_h+n_i)(n_h+n_i-1)}{n.^2-n.}$	$\frac{(n_h+n_j)(n_h+n_j-1)}{n.^2-n.}$	$\frac{(n_i+n_j)(n_i+n_j-1)}{n.^2-n.}$	$-\frac{n_h^2-n_h}{n.^2-n.}$	$-\frac{n_i^2-n_i}{n.^2-n.}$	$-\frac{n_j^2-n_j}{n.^2-n.}$
Dispersion	$\frac{n_h+n_i}{n.}$	$\frac{n_h+n_j}{n.}$	$\frac{n_i+n_j}{n.}$	$-\frac{n_h}{n.}$	$-\frac{n_i}{n.}$	$-\frac{n_j}{n.}$
Average Dispersion	$\frac{(n_h+n_i)^2}{n.^2}$	$\frac{(n_h+n_j)^2}{n.^2}$	$\frac{(n_i+n_j)^2}{n.^2}$	$-\frac{n_h^2}{n.^2}$	$-\frac{n_i^2}{n.^2}$	$-\frac{n_j^2}{n.^2}$

$$q_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{n_h} d_{ij}^2}{2n_h} , \qquad (3)$$

where d_{ij} denotes the distance between entities e_i and e_j. The analysis starts with matrix $\underset{\sim}{Q} \equiv \{q_{ij}\}$, in which

$$q_{ij} = d_{ij}^2/2 . \qquad (4)$$

2.3 Optimization of Average Dispersion within New Clusters. This is an improved version of the previous procedure. Let us assume that $q_h = q_i$ such that $n_h > n_i$. Cluster z_h is obviously more compact, therefore less heterogeneous than cluster z_i, thus measuring cluster heterogeneity by the average dispersion seems to be reasonable.

An element of the starting matrix is

$$q_{ij} = d_{ij}^2/4 . \qquad (5)$$

Distance between entities may be defined by the Euclidean distance, such as the chord distance (Orloci, 1967), or other standardized measures in both dispersion-minimizing strategies. These methods are equally applicable to binary and quantitative data.

2.4 Optimization of Edge Density in Subgraphs Representing New Clusters. This strategy (Podani, 1978) is based on graph theoretical considerations and is applicable to binary data only. Let $\underset{\sim}{E} \equiv \{e_i\}$ be the set of entities to be classified, $\underset{\sim}{A} \equiv \{a_k\}$ be the set of attributes describing e_i, and m be the number of attributes. Let, further, $\underset{\sim}{R} \equiv \{r_k\}$ be the set of symmetric relations between entities such that relation $e_i r_k e_j$ holds if e_i agrees with e_j with respect to attribute a_k (joint presence or joint absence). Thus we have an undirected graph $\underset{\sim}{G}$ in which vertex g_i represents e_i and the edges symbolize the existing relations between entities. In this way the maximum number of edges connecting any two vertices is m.

The homogeneity of cluster z_h represented by subgraph

$\underset{\sim}{G}_h$ may be measured by the edge-density of $\underset{\sim}{G}_h$. This quantity can be calculated according to ψ_h,

$$\psi_h = 2 \frac{\text{number of edges in } \underset{\sim}{G}_h}{m\, n_h\, (n_h-1)} \tag{6}$$

The edge-density of $\underset{\sim}{G}_h$ may also be determined using the following formula:

$$\psi_h = 1 + \frac{2n_h}{m(n_h-1)} \sum_k \hat{p}_k(\hat{p}_k-1) , \tag{7}$$

where $\hat{p}_k$ is the estimated probability or relative frequency of the presence of attribute a_k in z_h, $0 \leq \psi_h \leq 1$. If $\psi_h=1$ then the homogeneity of z_h is maximal. The minimum value of ψ_h is, however, greatly affected by n_h such that $\min \psi_h=0$ if and only if $n_h=2$. If $n_h > 2$, there will be necessary joint presences and absences in z_h, therefore the edge-density of $\underset{\sim}{G}_h$ must be greater than zero. The possible minimum of ψ_h can be calculated using the following formulae:

$$\min \psi_h = \frac{\frac{n_h}{2} - 1}{n_h - 1} \tag{8}$$

for even values of n_h, and

$$\min \psi_h = \frac{\frac{n_h}{2} - 1 + \frac{1}{2n_h}}{n_h - 1} \tag{9}$$

for odd values of n_h. This property may or may not be considered in the construction of a sorting algorithm but the strategy is combinatorial in the latter case.

The cluster analysis starts with a similarity matrix $\underset{\sim}{S}$

computed based on the coefficient of Sokal and Michener (1958) given by

$$S_{ij} = (a+d)/(a+b+c+d) , \qquad (10)$$

where the symbols are those regularly used in 2×2 contingency tables. Index (10) is the special case of expression (6) for $n_h=2$.

ACKNOWLEDGEMENTS

The author is grateful to P. Juhász-Nagy and Z. Szőcs for their helpful suggestions.

REFERENCES

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A*, 134, 321-353.

Lance, G. N. and Williams, W. T. (1966). A generalized sorting strategy for computer classifications. *Nature*, 212, 218.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal*, 9, 373-380.

Orloci, L. (1967). An agglomerative method for classification of plant communities. *Journal of Ecology*, 55, 193-205.

Podani, J. (1978). *Hierarchical classificatory methods for the analysis of binary ecological data.* Ph.D. thesis, Eötvös University, Budapest.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

Wishart, D. (1969). An algorithm for hierarchical classifications. *Biometrics*, 25, 165-170.

[*Received June* 1978. *Revised January* 1979]