GENERALIZED STRATEGY FOR HOMOGENEITY-OPTIMIZING HIERARCHICAL CLASSIFICATORY METHODS

JÁNOS PODANI

Research Institute for Botany
Hungarian Academy of Sciences
2163 Veszprém, 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 $z_i$ obtained by the fusion of groups $z_i$ and $z_j$:

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

The values of parameters $\alpha$, $\beta$, and $\gamma$ 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 $\omega_h$, $\omega_i$, and $\omega_j$ denote the homogeneity or heterogeneity of the clusters and let $\omega_{hi}$ 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,
GENERALIZED CLUSTERING STRATEGY

\[ W = \begin{bmatrix} w_h & w_{hi} & w_{hj} \\ w_i & w_{ij} \\ w_j \end{bmatrix} \]

and vector

\[ N = [n_h, n_i, n_j] \]

In the case of homogeneity, let \( w_{ij} \) be the greatest value in the upper triangular portion of \( W \). 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 \( W \) and values of \( N \) using the following formula,

\[ w_{hij} = \alpha_i w_{hi} + \alpha_j w_{hj} + \beta w_{ij} + \gamma_i w_i + \gamma_j w_j \]

(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 \( W \), 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 \( W \) are of importance. Strategies compatible with equation (2) are given below.

2.2 Optimisation 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>
<thead>
<tr>
<th>Name</th>
<th>$a_1$</th>
<th>$a_2$ - $n$</th>
<th>$n_2$ - $n$</th>
<th>$n_2$ - $n/m$</th>
<th>$n_2$ - $n/m$</th>
<th>$n_2$ - $n/m$</th>
<th>$n_2$ - $n/m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge-density</td>
<td>$n_2$ - $n/m$</td>
<td>$(n_2/m(n_2/m - 1))$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
</tr>
<tr>
<td>Dispersion</td>
<td>$n_2$ - $n/m$</td>
<td>$(n_2/m)^2$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
</tr>
<tr>
<td>Average dispersion</td>
<td>$n_2$ - $n/m$</td>
<td>$(n_2/m)^2$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
<td>$n_2$ - $n/m$</td>
</tr>
</tbody>
</table>
\[ q_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{n_h} d_{ij}^2}{2n_h} , \]  

where \( d_{ij} \) denotes the distance between entities \( e_i \) and \( e_j \).

The analysis starts with matrix \( Q \equiv \{q_{ij}\} \), in which

\[ q_{ij} = \frac{d_{ij}^2}{2} . \]

2.3 Optimization of Average Dispersion within New Clusters.
This is an improved version of the previous procedure. Let us assume that \( q_h = q_1 \) such that \( n_h > n_1 \). Cluster \( z_h \) is obviously more compact, therefore less heterogeneous than cluster \( z_1 \), thus measuring cluster heterogeneity by the average dispersion seems to be reasonable.

An element of the starting matrix is

\[ q_{ij} = \frac{d_{ij}^2}{4} . \]

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 \( \mathcal{A} \equiv \{a_k\} \) be the set of attributes describing \( e_i \), and \( m \) be the number of attributes. Let, further, \( \mathcal{R} \equiv \{r_k\} \) 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 \( \mathcal{G} \) in which vertex \( v_i \) 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
$G_h$ may be measured by the edge-density of $G_h$. This quantity can be calculated according to $\psi_h$,

$$\psi_h = \frac{\text{number of edges in } G_h}{m n_h (n_h - 1)}$$

The edge-density of $G_h$ may also be determined using the following formula:

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

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 $\psi_h$ 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 $G_h$ must be greater than zero. The possible minimum of $\psi_h$ can be calculated using the following formulae:

$$\min \psi_h = \frac{n_h - 1}{n_h - 1}$$

for even values of $n_h$, and

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

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

$$ S_{ij} = \frac{a+d}{(a+b)(c+d)} $$

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

ACKNOWLEDGEMENTS

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

REFERENCES


[Received June 1978. Revised January 1979]