Computer programs for hierarchical polythetic classification ("similarity analyses")

By G. N. Lance and W. T. Williams*

It is demonstrated that hierarchical classifications have computational advantages over cluster analyses. Flexible programs are outlined providing two sorting strategies and four alternative similarity-coefficients. Preliminary results suggest that for qualitative data one of the strategies and two of the coefficients are superior to the remainder of the system; the further refinements desirable for a large-scale production program are discussed.

Introduction: properties of hierarchical systems

Methods of classifying groups of elements into sets can conveniently be divided into *hierarchical* and *non-hierarchical* methods. The non-hierarchical methods, which include the many variants of cluster analysis, are concerned to find groups whose members are, in accordance with some predetermined measure of "likeness", as like each other as possible. If the clustering coefficient used defines a simple mathematical model, it may then be possible to define an inter-group dissimilarity and thus embed the groups in a hierarchical structure; but the hierarchy so formed is essentially only a key, in that it is the best way of attaining predetermined groups, and not necessarily the most efficient means of successive subdivision. Moreover, in many forms of cluster analysis the coefficient used does not lend itself to a definition of inter-group dissimilarity, and no compatible hierarchy can be erected.

Hierarchical methods, on the other hand, seek to find the most efficient step at each stage in the progressive subdivision or synthesis of the population. They aim, in a sense which again must be predetermined, to find the best route from population to individuals; but this route may be found at some degree of sacrifice of the homogeneity of the groups through which the process passes. It is by no means certain that any method can be found which simultaneously maximizes hierarchical efficiency and cluster homogeneity; and certainly, no such method is known. Nearest to this ideal are those methods, such as that of Edwards and Cavalli-Sforza (1965), which examine all dichotomous choices; but since for n individuals this involves the examination of $(2^{n-1} - 1)$ possibilities at each division, such methods are normally regarded as computationally impracticable for large values of n. In practice, therefore, the user is required to decide whether he will optimize the clusters or the route.

All cluster-analysis methods necessarily involve three terms: a method of initiating clusters, a method of allocating new individuals to existing clusters, and a stopping-rule to determine when further allocation is unprofitable. The main disadvantage of such a method is that the allocation procedure involves repeated trial and error; this not only lengthens the computation time, but makes it almost impossible to estimate running time in advance with any degree of accuracy. On the other hand, the great advantage of such methods is that allocation is not irrevocable; there is in principle no objection to the later rejection from a cluster of an element previously included in it. In truly hierarchical methods this latter advantage is lost, as it is inherent in route-optimizing strategies that divisions or fusions are irrevocable. It is therefore important to establish whether hierarchical methods offer compensatory computational advantages.

The computational characteristics of hierarchical methods depend on whether these are *divisive*, using a strategy of progressive subdivision, or *agglomerative*, using progressive fusion. We will consider a population of n elements specified by s attributes, and define an "operation" as the calculation of a single likeness-coefficient. It is then true that the divisive monothetic method of "association analysis" (Lance and Williams, 1965) can be completely encompassed within $\frac{1}{2}s(s - 1)(t - 1)$ operations, where $t = 2^s$ or n, whichever is the smaller. However, it would be most unusual to continue the subdivision to its ultimate limits, and in any case an appreciable number of attributes will become indeterminate *en route*; this expression therefore represents a gross overestimate of the work required, and is virtually useless for estimating duration. The requirements of the polythetic divisive method of "dissimilarity analysis" (Macnaughton-Smith *et al.*, 1964) vary with the evenness of division that it produces. If an n-group divides into two groups of n' and $(n - n')$ members, the number of operations will have been $[(n'+1)(2n-1)-n]$. In the worst case, when $n' = n/2$, this reduces (if $n \geqslant 2$) to $3n^2/4$ approximately for a single division; in the best case, when $n' = 1$, it reduces to $(3n - 2)$. The time taken for analysis therefore depends on the course of the analysis, and cannot be estimated in advance.

Agglomerative methods have strikingly different properties. The two major fusion-strategies with which we shall be concerned in this communication—"nearest-neighbour" and "centroid"—require in each case a number of operations for completion which depends

* *C.S.I.R.O. Computing Research Section, Canberra, A.C.T., Australia.*

only on n; this is $\frac{1}{2}n(n-1)$ for nearest-neighbour and $(n-1)^2$ for centroid. The computational requirements can therefore easily be calculated in advance with reasonable precision.

We are not here concerned with specific user requirements, which may themselves generate a preference for hierarchical or non-hierarchical methods, but to suggest an improved strategy whereby the undoubted computational advantages of hierarchical agglomerative methods can be more fully exploited than has been the case hitherto. These methods ("similarity analyses") are relatively old, stemming at least from Kulczynski (1928), and their many variants are discussed in Sokal and Sneath (1963); but no comparative study of a critically-chosen set of variants on the same set of data has previously been attempted. We shall outline the strategy of a set of flexible programs, for the Control Data Corporation 3600 computer, intended for such a methodological survey, and report preliminary results that suggest strongly that two of the variants are of greater power than the remainder. The existence of the programs has been briefly announced in a communication (Williams and Lance, 1965) dealing primarily with problems of inference; detailed specifications can be obtained from G.N.L.

The new computer programs: facilities

1. *Data.* Qualitative (i.e. binary) or quantitative.

2. *Sorting strategies.* Two are provided:

   (i) *Nearest neighbour.* Similarity coefficients are calculated between all pairs of individuals, tagged and ordered. Individuals are progressively fused by reference to the ordered list so obtained; the distance between two groups is thus defined as the distance between their nearest neighbouring individuals. Since the structure of the groups is not itself used in subsequent calculation, the information level never rises above that of individual comparisons, and the strategy is theoretically lacking in power.

   (ii) *Centroid.* The process begins as before by the calculation of all inter-individual similarities; but on fusion the fused individuals are replaced by a new synthetic individual representing the sum by attributes of the constituent individuals (or mean sum, according to the coefficient in use). New similarities are calculated between the new individual and all others (original or synthetic) that remain in the analysis.

3. *Coefficients.* Four are provided:

   (i) *Correlation coefficient.* For qualitative data and nearest-neighbour sorting this is calculated from a 2 × 2 table as the Pearson ϕ-coefficient; for numerical data (and therefore for centroid sorting) it is the conventional product-moment coefficient. The coefficient is undefined for an individual with zero variance (normally a qualitatively-specified individual lacking or possessing all attributes); in the CDC 3600 program, relationships with such individuals are allocated the impossible coefficient of $-2\cdot0$ and these individuals thereby segregated from the analysis.

   (ii) *Squared Euclidean distance.* Qualitative data are accommodated by taking the jth co-ordinate for an individual as 1 if he possess the jth of the attributes, and 0 if he lacks it; in the usual (a, b, c, d) symbolism of a 2 × 2 table the squared distance between two such individuals reduces to $(b + c)$. Provision is also made for the prior standardization by attributes to zero mean and unit variance, or for reading in an external vector of attribute weighting coefficients.

   (iii) *Non-metric coefficient.* We imply by this the coefficient, for binary data, $(b + c)/(2a + b + c)$. It is the complement of a coefficient apparently first used by Czekanowski (1913), which is monotonic with the coefficient $a/(a + b + c)$, used by Sneath in his earlier work to avoid counting double-negative matches. Its quantitative form between two vectors (x_{1j}) and (x_{2j}) is $(\Sigma|x_{1j} - x_{2j}|)/\Sigma(x_{1j} + x_{2j})$, and has been used in ecological work. The coefficient is undefined if both individuals being compared are everywhere zero; and since it is desirable that such individuals should be grouped together, the coefficient is put equal to zero if $(2a + b + c)$ or $\Sigma(x_{1j} + x_{2j})$ is zero.

   (iv) *Information statistic.* The statistic we have used is derived as follows. Let a system be capable of existing in any one of a number of discrete states, and let the probability of the ith state be p_i; then Shannon (1948) has shown that an appropriate measure of the entropy of the system is given by

   $$H = -\sum_i p_i \log p_i.$$

   If a system (such as a single qualitative attribute) has only two states with probabilities p and $(1 - p)$, this reduces to

   $$H = -[p \log p + (1 - p) \log (1 - p)].$$

   Let there be a group of n individuals specified by the presence or absence of s qualitative attributes, and let the probability of the presence of the jth attribute be p_j; then the mean entropy of the system will be given by

   $$H = -\sum_{j=1}^{s} [p_j \log p_j + (1 - p_j) \log (1 - p_j)].$$

   If there are a_j individuals possessing the jth attribute, the best available estimate of p_j is a_j/n; and we may define an "information content" of the entire system, I, and write $I = nH$. Making the necessary substitutions, we have

$$I = sn \log n - \sum_{j=1}^{s} [a_j \log a_j + (n - a_j) \log (n - a_j)].$$

The criterion is the increase in I on fusion (ΔI or I-gain), which is to be minimum. The base of the logarithms is at arbitrary choice, and we have made use of the tables of $n \log n$ to base e in Kullback (1959). The relationship between functions of this type and χ^2 is established in Kullback, and their relationship with likelihood functions has been exploited by Macnaughton-Smith (1965).

The coefficient is not defined for continuously-varying numerical data; and it cannot be usefully employed with nearest-neighbour sorting, since for a pair of individuals it reduces to $2(b+c) \log 2$, i.e. to a constant multiple of the squared Euclidean distance.

4. *Organization*

Three separate programs are provided: QUALNEAR (qualitative data, nearest-neighbour sort), CENTROID (qualitative or quantitative data, centroid sort) and NUMENEAR (quantitative data, nearest-neighbour sort).

Results

1. *Criteria for assessment*

An ideal strategy would fulfil three criteria:

(1) The value of the similarity-coefficient should change monotonically with successive fusions.

(2) The process should, so far as the data permit, fuse the population into clearly-separated groups, and not continually add single individuals (for genuine "chain" data, other methods are more appropriate).

(3) The coefficient should define an objective level below which details of individual fusions may be disregarded as without interest.

Concerning Criterion (1), nearest-neighbour sorting is monotonic by definition. In centroid sorting, the information statistic (I as distinct from ΔI), being completely additive, is also necessarily monotonic. Consideration of the geometrical models involved will show that Euclidean distances and the correlation coefficient must be liable to occasional failure of monotonicity; theoretical information concerning the non-metric coefficient and ΔI is lacking. Concerning Criterion (2), only in the case of the centroid information statistic is there a known theoretical reason why the criterion should be fulfilled. For, since $\{(n + 1) \log (n + 1) - n \log n\}$ increases with n, the disturbance caused by the addition of an aberrant individual will increase with group size. It follows that the analysis will tend to delay the fusion of large groups, or the addition of outlying members to existing groups, until relatively late in the analysis. Similarly, only the information statistic is known to fulfil Criterion (3); for $2\Delta I$ is substantially distributed as a χ^2 with as many degrees of freedom as there are attributes (for original references and restrictions see Kullback, 1959), and may be used as a significance test. Since it is minimized on fusion, it represents a conservative, and not a random, estimate. However, the conservatism will tend to reduce the final degree of subdivision of the population, and will thus minimize the chance of retaining unprofitably fine divisions; the objection we have by implication raised against conservative χ^2 estimates in divisive systems (Lance and Williams, 1965), which have the opposite effect, does not here apply.

2. *Empirical tests*

These have so far been confined to qualitative data in plant ecology (Dr. J. M. Lambert) and taxonomy (Mr. L. Watson). Monotonicity failure in centroid analyses has been found to be rather common with Euclidean distances, somewhat less common with the correlation coefficient, and relatively infrequent with the non-metric coefficient. In fourteen different analyses there has been no monotonicity failure of ΔI; this suggests to us that this coefficient may be necessarily monotonic, but we have not been able to obtain a formal proof.

Full details of the analyses will be submitted for publication in appropriate user journals, but an indication of comparative clarity of classification may be obtained from **Fig. 1** (an ecological population of 20 individuals specified by 76 qualitative attributes). Fig. 1(*a*) shows the form of hierarchy resulting from nearest-neighbour sorting of Euclidean distance: there is virtually no grouping, the individuals being added in succession. All nearest-neighbour results are of this type. Fig. 1(*b*) shows the corresponding centroid analysis, which is little improved. However, the centroid sorting shows improved grouping with the correlation coefficient, further improvement with the non-metric coefficient, and (Fig. 1(*c*)) strikingly clear grouping with the information statistic (ΔI), in which the major groups can be emphasized by re-plotting as I instead of ΔI (Fig. 1(*d*)).

Since the information statistic appears to be immune from monotonicity failure, groups clearly, and defines an objective level of profitable subdivision, it appears to be ideal for qualitative data; but the ability to group is not won without price. If a population contains subpopulations with substantially homogeneous "cores" but with which are associated occasional peripheral members, this strategy is liable to sweep up all peripheral members into a single group of non-conforming individuals, irrespective of their intrinsic affinities. In ecology, where the interest lies primarily in the central pattern, this behaviour is acceptable and even advantageous: it is not acceptable in taxonomy, where every individual must be accounted for in the best possible way. Our tentative conclusions are, therefore, (*a*) that in any project where general patterns are required, centroid information-statistic strategy is indicated, and

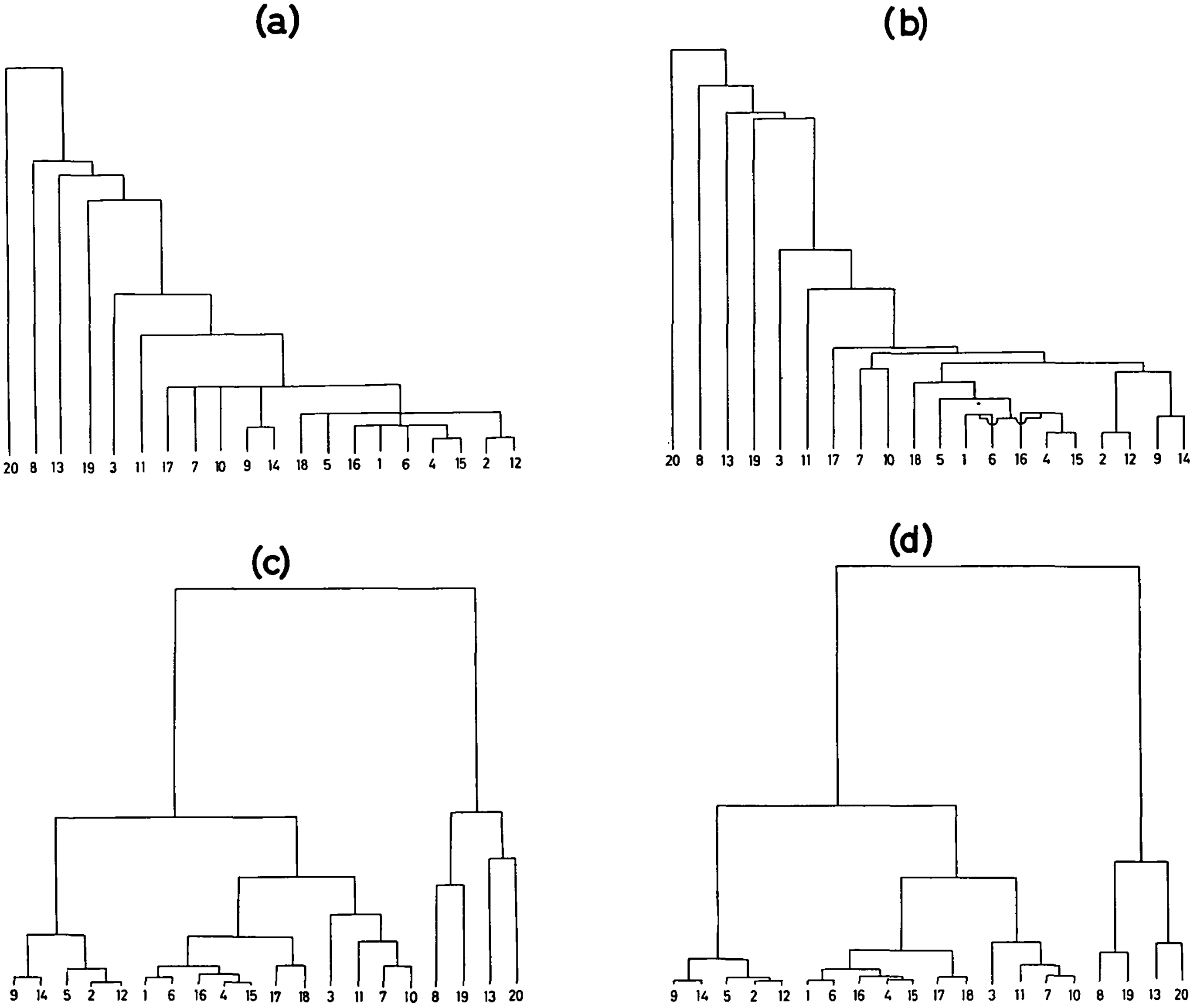

Fig. 1.—Comparison of hierarchies: (*a*) Euclidean distance, nearest-neighbour sort; (*b*) Euclidean distance, centroid sort; (*c*) information statistic (ΔI), centroid sort; (*d*) as (*c*) but plotted as total information (I).

(*b*) although in strictly taxonomic situations the information statistic may be invaluable for indicating the likely level of profitable subdivision, it must be checked by reference to another strategy—our present experience suggests centroid sorting with the non-metric coefficient, because of its relatively good grouping and monotonicity.

Requirements for production programs

We conclude from the foregoing that production programs are desirable using centroid sorting and both the information and "non-metric" statistics. Consideration must be given to the practicability of incorporating two facilities which are lacking in our pilot program—provision for dealing with missing or inapplicable values, and provision for resolving ambiguities encountered early in the analysis.

1. *Missing or inapplicable values*

(i) *Information statistic*

Let a group of n individuals contain, so far as the jth qualitative attribute is concerned, a_j individuals for which the answer is known to be yes ($=1$), b_j for which it is known to be no ($=0$), and c_j for which the answer is unknown, or to which the question is inapplicable. The information content is now associated with two degrees of freedom and can be partitioned as follows:

(for known/unknown)

$$I' = n \log n - (a_j + b_j) \log (a_j + b_j) - c_j \log c_j$$

(for yes/no if known)

$$I'' = (a_j + b_j) \log (a_j + b_j) - a_j \log a_j - b_j \log b_j$$

If both were of interest, the two totals could be accumulated separately; this is the exact information counterpart of the sum-of-squares partition suggested by Williams and Dale (1962) for the corresponding problem in fully-quantitative data. For most taxonomic purposes only I'' is of interest, and it is this that should be accumulated towards the I total for the group. A vector of $(a_j + b_j)$ values for each attribute must be stored, so that each individual or group is in effect held twice.

(ii) *Non-metric coefficient*

The problem here is not the calculation of the coefficient, but the value to be taken, in the formation of the "centroid" itself, for the proportion of the group containing the attribute. Using the symbols of the previous paragraph, is the proportion to be taken as a_j/n or $a_j/(a_j + b_j)$? The former is computationally simpler, and meets the objection that an attribute known for only a small number of members should not weigh equally with one known throughout; but it is formally equivalent to taking the value as zero if unknown, so that the "unknown" values are taken as being "known/no," which is clearly undesirable. Our tentative recommendation is therefore to use $a_j/(a_j + b_j)$; it will again be necessary to hold, for each group, a vector of $(a_j + b_j)$ values.

2. *Ambiguities*

These are of two types: independent ambiguities (AB : CD), which may be taken in any order, and linked ambiguities (AB : AC), which in theory may be taken in any order only if they are zero. The basic problem thus concerns non-zero linked ambiguities when these are next in order for fusion. Such an ambiguity can only be resolved by the importation of further information; and the only source of such information within the data is the entire fused population, even though this may obscure local concentrations of interest. For the Euclidean case it has been suggested (Williams *et al.*, 1964) that the population values of $\sum_{k \neq j} \chi^2{}_{jk}$ should be used as attribute-weighting coefficients; for the information statistic the obvious analogy is to calculate the contribution (I_j) of each attribute to the I of the whole population, and use this as a weight. The weighted I would then be calculated as

$$I_w = \sum_j \; [\{n \log n - a_j \log a_j - (n - a_j) \log (n - a_j)\} I_j]$$

making due allowance for unknown values as necessary. (Such a value would only be used for discrimination, and not accumulated as the I value for the group under study.)

No comparable strategy is available for the non-metric coefficient, since this is not additive over attributes. It could be made so by replacing

$$(\Sigma|x_{1j}-x_{2j}|)/\Sigma(x_{1j}+x_{2j}) \text{ by } \Sigma[(|x_{1j}-x_{2j}|)/(x_{1j}+x_{2j})],$$

but we have some empirical evidence that this seriously impairs the otherwise good monotonic properties of the coefficient. Fortunately, ambiguities are usually somewhat less common than with the information statistic.

It is clear that (*a*) a substantially rigorous solution is available only for the information statistic, which is itself in any case liable to produce some small degree of misclassification of aberrant individuals, and (*b*) the inclusion of such a procedure would appreciably complicate the program, since it would involve listing and examining all ambiguities, followed by further computation for discrimination. Our tentative conclusion is that the additional precision obtained by formal resolution of ambiguities would not justify the additional computation, and we recommend that ambiguities be taken in any convenient order—normally, the first encountered by the sorting strategy in use.

We express our indebtedness to Mr. P. Macnaughton-Smith of the Home Office Research Unit, who suggested both the computational strategy for centroid analysis and the use of the information statistic.

References

Czekanowski, J. (1913). *Zarys metod statystycznyck*, Warsaw.

Edwards, A. W. F., and Cavalli-Sforza, L. (1965). "A method for cluster analysis," *Biometrics*, Vol. 21, p. 362.

Kulczynski, S. (1928). "Die Pflanzenassoziationen der Pieninen," *Bull. int. Acad. pol. Sci., Ser. B*, Suppl. 2, p. 57.

Kullback, S. (1959). *Information theory and statistics*, Wiley: New York (Chapman & Hall: London).

Lance, G. N., and Williams, W. T. (1965). "Computer programs for monothetic classification ('Association analysis')," *Comp. J.*, Vol. 8, p. 246.

Macnaughton-Smith, P. (1965). *Some statistical and other numerical techniques for classifying individuals*, H.M.S.O. Home Office Research Unit Report, No. 6.

Macnaughton-Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. (1964). "Dissimilarity analysis: a new technique of hierarchical subdivision," *Nature*, Vol. 202, p. 1034.

Shannon, C. E. (1948). "A mathematical theory of communication," *Bell System Tech. J.*, Vol. 27, p. 379 and p. 623.

Sokal, R. R., and Sneath, P. H. A. (1963). *Principles of numerical taxonomy*, W. H. Freeman: San Francisco and London.

Williams, W. T., and Dale, M. B. (1962). "Partition correlation matrices for heterogeneous quantitative data," *Nature*, Vol. 196, p. 602.

Williams, W. T., Dale, M. B., and Macnaughton-Smith, P. (1964). "An objective method of weighting in similarity analysis," *Nature*, Vol. 201, p. 426.

Williams, W. T., and Lance, G. N. (1965). "Logic of computer-based intrinsic classifications," *Nature*, Vol. 207, p. 159.