displace the information statistic except in the special cases already mentioned.

3. Flexible. If used with a Euclidean measure, this will also be sensitive to skewness; but it has the advantage that the properties of the strategy are invariant under change of measure, and any measure can be used. Values associated with successive fusions lack, however, the statistical or geometrical interpretations that are available for the preceding strategies. The position of a group in the hierarchy will not be markedly size-dependent, nor will there be any tendency to produce 'nonconformist groups'; reallocation should be simple. If the arbitrariness of its fusion-measures is of no concern it appears to be the most attractive of the strategies; but more heuristic work is needed on the optimal value of \( \beta \), and it must be remembered that the optimal value may differ between applications.

4. Mean squared distance. A not very attractive strategy, and at present mainly used if, in the absence of reallocation, accuracy of group structure is of paramount importance, since its weak dilatation reduces the probability of misclassification. It is available for all measures, though its properties can only be examined rigorously in the Euclidean case. However, it seems likely that the flexible strategy with a suitably low value of \( \beta \), followed by reallocation if necessary, would always be preferable.

References


Controversy concerning the criteria for taxonometric strategies

W. T. Williams*, G. N. Lance†, M. B. Dale‡ and H. T. Clifford§

Recent growth of interest in numerical classificatory strategies has led to the search for criteria which such strategies should fulfill. Two largely incompatible schools of thought have now arisen, one in Cambridge the other in Australia: the article examines the differences in the basic premises of these schools with particular reference to linguistic differences and the wider implications of the two sets of criteria.

(Received October 1970)

1. Introduction

Researchers wishing to use numerical methods for the classification of their data quickly discover that there are at present at least two schools of thought concerning the criteria that taxonometric methods should be expected to meet. One of these schools, with which we are ourselves concerned, is of European descent but currently particularly vociferous in Australia; another is associated with the Computer Laboratory at Cambridge. In this issue are two papers: one (Williams, Clifford, and Lance, 1971) tacitly accepts the Australian strategies and examines their differential applicability, the other (Sibson, 1971) demonstrates that one of the Australian strategies (the 'flexible' system of Lance and Williams, 1967) fails to meet a particular criterion which the Cambridge school regards as essential. Both the schools claim considerable success over a wide range of applications; but unless a user can appreciate the differences in premises on which the controversy rests, he may well feel some misgivings in employing the methods of either school. This article is an attempt to examine objectively the two sets of premises and to highlight the differences in approach.

The examination falls conveniently into three parts. First, there are differences in linguistic usage: different terms may be used for the same concept, and the same term may be used by the two schools in different senses. Such differences do not reflect genuine differences in premises, but may appear to do so to a user unfamiliar with the field. Secondly, the Cambridge school has defined certain criteria which it considers should be met by all taxonometric methods, but which are not met by the methods of the Australian school. Lastly, the Australian school considers that there exists a further set of criteria which in certain circumstances should over-ride; we have perhaps

*CSIRO, Division of Tropical Pastures, St. Lucia, Queensland, Australia.
†CSIRO, Division of Computing Research, Canberra City, Australia.
‡CSIRO, Division of Plant Industry, Canberra City, Australia.
§Department of Botany, University of Queensland, St. Lucia, Queensland, Australia.
been at fault in that, although we have on occasion described these as properties of our methods, we have until now never formally enunciated them as criteria.

2. Linguistic differences
Four deserve particular mention:

1. Terms used for hierarchy-generation. Given a dissimilarity measure we (e.g. Lance and Williams, 1967) have described the process which builds this into an agglomerative hierarchy as a monotonic sorting strategy (we need not here consider non-monotonic strategies such as centroid and median); Burr (1970) calls the process a monotone fusion strategy; Jardine and Sibson (1968) call it an ultrametric transformation. These terms are all exactly equivalent; but the ‘ultrametric’ concept permits an elegant symbolic formulation of the process, so that this term has much to commend it. However, Jardine and Sibson completely ignore the existence of an important and widely-used set of strategies in which the succession of values which generate the hierarchy cannot be obtained from the primary dissimilarity matrix; in such cases all calculations are based on the raw data-matrix of individuals × attributes. All information-statistic strategies based on progressive minimisation of information gain are of this type.

2. Data. We always use this term for the ‘raw’ data, i.e. the matrix of individuals (OTU’s) specified by a set of attributes; but Jardine and Sibson (1968) appear to use it for the dissimilarity matrix obtained from this.

3. Agglomerative and divisive. Jardine (1970) regards the ultrametric transformation as the method, and ‘agglomerative’ and ‘divisive’ as merely defining algorithms which implement the method. Our difficulty in accepting this usage arises from the asymmetry of the system. Divisive strategies may be monothetic, but agglomerative strategies (except in a trivial sense) cannot be. The method of Edwards and Cavalli-Sforza (1965) and the ‘incremental sum of squares’ strategy of Burr (1970) are certainly divisive and agglomerative counterparts of the same method, but the scales on which they are computationally practicable differ by at least an order of magnitude; a similar discrepancy separates the divisive (Lance and Williams, 1968) and agglomerative (Williams, Lambert, and Lance, 1966) forms of information analysis. There is no known divisive counterpart of the ‘flexible’ strategy of Lance and Williams (1967), and no known agglomerative counterpart of the interesting 2-parameter model of Macnaughton-Smith (1965), at least in the case when simultaneous division of both OTU’s and attributes is required. In other words, the two approaches may imply different methods, or vice versa, so that we feel that to regard them simply as algorithms is misleading.

4. Cluster. The formidable linguistic difficulties which surround this word have been examined in detail by Gasking (1960); it will be clear from his account that any universally applicable definition will be so permissive as to be of dubious value for numerical applications. We have always confined the term ‘cluster analysis’ to those strategies which optimise the structure of groups of individuals by making them as internally homogeneous as possible; in such cases there may be no definition of boundaries within the group, or of the relationships of groups to the entire population. In contrast, to us a hierarchical strategy optimises a route between individuals and population; in so doing it incidentally passes through ‘groups’, which we have attempted to avoid calling ‘clusters’. No precise definition of the word is given by the Cambridge school and their use of the term is not clear to us; but it is obviously different from ours. There are now so many variant uses of the word that it would probably be better if we all avoided it completely.

3. The Cambridge criteria
Seven criteria, A to G, are erected by Jardine and Sibson (1968); those at issue are A, B, C and (though these are not raised as criticisms by them) F and G.

(A) The transformation is to be well-defined; i.e. a unique result should be obtained from a given dissimilarity matrix. It is this criterion which Sibson (1971) discusses in detail. Our strategies will fail this test only when there are ambiguities (i.e. equal dissimilarities) which are arbitrarily resolved. We believe that in practice true ambiguities are rare; but the problem exists and requires further consideration. In ‘non-combinatorial’ strategies (i.e. strategies in which all calculations are based on the raw data-matrix) the problem can be overcome by permitting multiple fusions, or resolved by weighting attributes by information imported from the population parameters (Williams and Dale, 1965); but neither of these solutions is available for the ‘flexible’ strategy of Lance and Williams (1967). We assume that the criterion cannot be interpreted as requiring that the order in which ambiguous fusions are taken is to be without effect; for we believe that this would be equivalent to a requirement that the strategy should be single-link, so that the demonstration that only single-link strategies fulfil it would be tautologous. In any strategy based on group-definitions an ambiguity must be resolved by reference to information outside the ambiguity under consideration; in the case of a ‘flexible’ sorting of a given dissimilarity-matrix the appropriate information consists of the relationships with all other entities (elements or groups) currently in the matrix. We suppose that to represent such an entity (\(a_{ij} = a_{ji} = 0\)), and we suppose we have encountered a particular ambiguity \(a_{qi} = a_{qg}\). The overall relationships between entities \(q, r\) and \(s\) and the remainder of the matrix can be defined as \(\sum_{j \in r} \sum_{m \in s} A_{mj} = \sum_{j \in q} \sum_{m \in s} A'_{mj}\), \(\sum_{j \in r} \sum_{m \in s} A_{mj} = \sum_{j \in q} \sum_{m \in s} A'_{mj}\), \(\sum_{j \in r} \sum_{m \in s} A_{mj} = \sum_{j \in q} \sum_{m \in s} A'_{mj}\). Then if \(|A_{qj} - A_{qj}^*| < |A_{qj} - A_{qj}^*|\), \(q\) is fused with \(r\). The technique is available for all strategies, including the non-combinatorial information-statistic systems; and since our present arbitrary resolution appears to be causing concern, we shall incorporate this procedure in our programs.

(B) The transformation is to be continuous; i.e. small changes in data (sensu Jardine and Sibson) should produce only small changes in the dendrogram. If the criterion is intended to be applied to small changes—i.e. to one or two entries—in the dissimilarity matrix, it appears to us to be artificial; apart from the trivial case of an error in calculation, we do not see how the situation could arise. If the criterion were to be applied to raw data, it would be unrealistic; an alteration of a single attribute-value for a single individual will affect every dissimilarity measure involving that individual, and may well make appreciable changes in the dendrogram. This is not unreasonable, since in effect the attribute-set has been modified; similarly, the addition of a new individual alters the context of the pre-existing individuals.

(C) If the dissimilarity measure is already ultrametric it is to remain unchanged by the transformation. We know of no measure which is ultrametric over its whole range; but in any case this again seems to us to be equivalent to a requirement that the strategy shall be single-link, so that the relevant part of the later demonstration is tautologous. It is perhaps of interest to note that, if the dissimilarity measure is a metric it will remain so under flexible sorting only so long as \(\beta = 0\). Since \(\beta\) is invariably negative in practice, this means that the metric property will be lost; but we fail to understand why this should be important or even relevant.

(F) The transformation is to commute with any permutation
of individuals. This requirement is not, in fact, independent of the ambiguity problem; if there are ambiguities, and if (as currently in our programs) these are resolved by the order in which they are encountered within the computer, the criterion will not be met. We note in passing that we have recently (Dale, Lance, and Albrecht, 1970) needed to devise a strategy in which this criterion is deliberately violated, to meet the needs of stratigraphers and social geographers who require that fusions be constrained to topographically adjacent sites.

(G) An excised portion of a dendrogram should be invariant when rerun after excision. The symbolic representation of Jardine and Sibson is not applicable to all strategies, but the general requirement remains clear. They state that ‘most known methods’ satisfy this criterion; but in fact it is not satisfied by most of our strategies. We regard this as the single most cogent criticism of our methods; the phenomenon is known to be connected with the group-size dependence of our more intensely-grouping strategies, and the paper elsewhere in this issue (Williams, Clifford, and Lance, 1971) reports a preliminary investigation into the nature and extent of the phenomenon.

4. The Australian criteria

Jardine (1970) makes an interesting (and we believe invalid) distinction between two uses of these methods. If the problem for solution has been stated explicitly, he suggests that any algorithm is acceptable ‘if it works well’; but if the need is to simplify data in a way which will suggest fruitful hypotheses, he implies that the method used must meet criteria such as those discussed above. We do not dispute his statement concerning the need for ‘knowledge of the mathematical properties of methods’; but we believe that this applies equally to both his uses. His contention appears to be that the pragmatic approach is justified only for an explicitly-stated problem, and with this we strongly disagree.

Three of us are professionally responsible for carrying out numerical classifications for users with a very wide range of data types and consequently we are biased towards the pragmatic approach. Most of our cases involve data-simplification rather than the solution of explicit problems; and it has been our experience that in the great majority of cases our strategies provide our users with the type of lead they require, whereas single-link strategies do not. We incline to the view that the highly non-linear nature of our strategies corresponds rather closely with human intuitive mental processes; but the strategies have also been shown to meet objective external criteria (such as the requirement, in Clifford, Williams, and Lance, 1969, that grass genera known to hybridise should fall into the same group). It is therefore clearly our duty to attempt to enumerate the essentially pragmatic criteria which we believe may in practice over-ride the mathematical criteria of the Cambridge school. We distinguish three such criteria:

1. The grouping must be more intense than that implied by the original dissimilarity measures. This is to ensure that the analysis produces a small number of clearly-defined groups, even though the raw data may be more nearly continuous. Such a result is essential if the user is, as frequently in ecology, more interested in the boundaries between the groups than in the structure of the groups themselves; but it appears also greatly to facilitate the erection of hypotheses, particularly hypotheses concerning relationships with external information. Once a hypothesis has been erected a user can, and frequently does, return to the individual dissimilarity measures to make minor corrections.

2. The grouping must be relatively insensitive to outlying values, due either to aberrant individuals, or to errors or inaccuracies in the data. This—like the ‘common-factor’ requirement in factor analysis—is to ensure that a broad overall picture is obtained. Both nearest- and furthest-neighbour strategies fail to meet this requirement, since outliers are inherently likely to be peripheral in their groups. The only single-link strategy known to us which is likely to meet this requirement is that of Wishart (1968), which uses the $k$th nearest neighbour; but we have no direct experience of its efficacy.

3. The ultrametric transformation should not be necessarily, or even usually, invariant over the entire population. This is probably the most fundamental difference between the two schools, and we give a simple example from our own experience. If a sample of grass genera including members of what are currently recognised as pooid, panicoid, festucoal and bambusoid grasses is classified by any strategy which seeks to preserve the basic dissimilarity measures, the bambusoid group will be fragmented into separate genera before the other three tribes have separated from each other. To a worker unfamiliar with the grasses this is not very helpful. The more intense strategies such as the agglomerative information-statistic system divide the set into four groups, one of which is manifestly more heterogeneous than the others. A coarser metric is in fact required for the bamboo than for the rest of the grasses. Most divisive programs possess this characteristic. For example, the monothetic divisive systems based on $\chi^2$ or an information statistic define a disjoint metric space (Williams and Dale, 1965) which can be metrised within, but not between groups. The information-statistic program of Wallace and Boulton (1970) calculates anew, after each division, the sub-population parameters from which the message-lengths will be calculated. Any such system can be regarded as a ‘piecewise’ simplification process, where within groups a reasonably simple model fits well, although this model may have different parameters in different groups. The process is adaptive, in that any decision to fuse, or divide, affects subsequent decisions; that is not the case in single-link systems, but it again appears to us both intuitively and from user experience, to correspond rather closely with human reasoning processes.

5. Conclusions

We are not here primarily concerned either to attack the views of the Cambridge school, or to defend ourselves against attack from them. We are, however, concerned to demonstrate to users that we have here an example of the distinction, well known in philosophy, between the study of the problem of choice, and the study of the criteria by means of which choices are resolved. The two schools have simply erected different systems of criteria, both internally consistent but mutually incompatible. We do not believe that either should be regarded as ‘right’ or ‘wrong’. Nor do we think it likely that either will ever completely displace the other; as in similar situations in the past, the two sets can exist side-by-side, and continued use of both approaches may eventually clarify the conditions under which one set or the other is the more profitable. Moreover, we cannot ignore the possibility that still other approaches may ultimately prove more profitable than either.

References


