Choice of methods for automatic classification*

N. Jardine and R. Sibson

King's College Research Centre, Cambridge

We examine the controversial issue of criteria of adequacy for methods of automatic classification, and we suggest that the issue can be partially resolved by considering the various purposes for which scientists classify.

(Received May 1971)

1. The nature of the controversy

Methods of automatic classification have proliferated since the mid 1950's and are widely applied in the study of the differentiation and ecology of organisms and in the behavioural sciences. They have been applied also as heuristics in information storage and retrieval and pattern recognition. Comparative studies have shown that when different methods are applied to the same data there are often major discrepancies between the results obtained (see for example, Minkoff, 1965; Boyce, 1969; Wishart, 1969). This may be alarming to a scientist who hopes that the result obtained by a method of automatic classification will help him to explain the differentiation of the objects classified or will serve as a basis for prediction. In recent papers in this journal Jardine and Sibson (1968) and Williams, Lance, Dale, and Clifford (1971) have proposed criteria to guide scientists in their choice of methods of automatic classification. Unfortunately, the two sets of criteria are incompatible.

We outline first the nature of the methods which have led to controversy. The starting point for many methods of automatic classification is a description of the objects to be classified by each of a set of attributes. Some methods proceed directly from descriptions of objects to a system of clusters. Other methods proceed indirectly via a pairwise measure of similarity or dissimilarity on the set of objects. Amongst the possible end products of a method of automatic classification it is convenient to distinguish *simple clusterings* and *numerically stratified clusterings*. *A simple clustering* is a partition or covering of a set of objects; no cluster can include another. In *numerically stratified clusterings* (*NSC's*) clusters have associated numerical levels, the clusters at a given level being nested within clusters at higher levels. Hierarchic NSC's in which the clusters at each level are disjoint are often called *dendrograms*. The controversy has concerned methods which produce NSC's. The reason for concentrating on this kind of clustering is that it can convey more information about the data than can a simple clustering. Methods which generate simple clusterings have been discussed in detail by Lerman (1970).

Jardine and Sibson (1968) discussed methods which transform a *dissimilarity coefficient* (*DC*) on all pairs in a set of objects into an NSC. We showed that a dendrogram (hierarchic NSC) can be characterised mathematically as an *ultrametric* DC by writing down for each pair of objects the lowest level at which they appear in the same cluster. Cluster methods which generate dendrograms from a DC are thus characterised as transformations from DC's to ultrametric DC's. We suggested certain criteria of adequacy for such methods. The single-link cluster method proposed by Sneath (1957) is the only hierarchic cluster method which satisfies these criteria.

We showed also that it is possible to obtain more information about a DC than is given by the single-link method by using methods which lead to non-hierarchic NSC's in which overlap between clusters is allowed. Each of the sequence of generalisations of the single-link method which we suggested satisfies the proposed criteria of adequacy.

Williams *et al.* (1971) considered methods which transform a DC into a dendrogram and also methods which transform descriptions of objects directly into a dendrogram. They did not consider methods which lead to non-hierarchic clusterings. They offered detailed objections to three of our criteria of adequacy and suggested that three 'pragmatic' criteria should overrule them in many applications of automatic classification.

2. Purposes of classification

Criteria of adequacy for a method of automatic classification are intended to help scientists in the choice of methods appropriate for their purposes. The criteria of adequacy which we suggested are proposed for cases where the purpose of classification is to represent the mutual dissimilarities of objects in a way which may suggest or confirm hypotheses about the factors which cause, maintain, or influence the differentiation of objects (cf. Jardine, 1970). For example, suppose that a biologist wishes to know whether geographical isolation of populations of a species of butterfly from different islands has led to the evolution of distinct races or subspecies. One possible line of investigation is to apply a method of automatic classification to see whether the populations from each island form well-marked clusters.

In such applications it is appropriate that the classification should be determined by the structure of the DC; it is inappropriate that it should be determined by extrinsic constraints on the number, size, or some specific property of the classes sought. The scientist may hope to obtain a small number of well-marked clusters when this would confirm his hypothesis, but he should not use a method which tends to force objects into a few well-marked clusters regardless of the structure of the DC.

In other kinds of application it is appropriate to use methods which allow a classification to be determined both by the structure of the data and by extrinsic criteria. For example, a scientist may wish to represent his data as accurately as possible subject to a requirement that each class obtained be specified by some attribute or combination of attributes to facilitate subsequent indexing and identification. It is then appropriate to use one of the various methods which construct a classification by successively partitioning the set of objects on the basis of single attributes (see MacNaughton-Smith, 1965; Lance and Williams, 1968). Constraints on the number and size of classes are crucial in application of automatic classification to allocation problems. For example, suppose we wish to allocate activities to buildings, and to rooms within

*This article is a further contribution to the discussion promoted in Vol. 14, No. 2, pp. 156-165.

each building, in such a way as to minimise the cost of traffic of people and equipment between activities. Given as data estimates of the cost of traffic between activities, it may be useful to apply an automatic classification method which allows intial specification of the numbers and sizes of classes. Rocchio (1966) described a simple cluster method which seeks homogeneous clusters subject to constraints on the number and sizes of the clusters.

We discuss the two conflicting sets of criteria in the light of these remarks about the purposes of classifications.

3. Criteria for automatic classification

Williams *et al.* raised objection to three of the criteria of adequacy suggested by Jardine and Sibson (1968).

A. The transformation from a DC to a dendrogram (hierarchic NSC) should be well-defined;
B. The transformation from a DC to a dendrogram should be continuous;
C. If a DC is already ultrametric, it should be left unchanged.

It appears self-evident that criterion A is needed when a method is sought which represents the structure of a DC in a simplified form by a single clustering. There *are* circumstances under which the criterion is inappropriate. For example, we might wish to use a method which seeks a clustering which minimises some measure of badness-of-fit to the data, despite the fact that a unique best-fitting clustering does not exist for all data, so that the method is ill-defined. However, we would then be interested in all clusterings which achieved minima, and it would be misleading to attach much significance to one optimal clustering without comparing it with the others. In this respect such a method would be analogous to the multidimensional scaling method of Kruskal (1964).

Williams *et al.* (1971) defend the flexible strategy described in Lance and Williams (1967) against Sibson's (1970) demonstration that it is ill-defined when equal DC values occur on the grounds that this rarely happens. However, equal DC values arise in many practical applications when the DC is calculated from descriptions of objects by discrete-state attributes. They suggest an emended version of the flexible strategy. The emended version remains ill-defined under certain circumstances, but in practice these circumstances would rarely occur.

Criterion B ensures that small changes in DC values produce correspondingly small changes in the resultant classification. Williams *et al.* interpreted it as referring to changes in a small *number* of DC values which is incorrect. There are many ways in which such small changes in DC values can arise. For example, errors of measurement can occur in the description of individuals; and if populations rather than individuals are classified, sampling errors occur. This, or some related criterion, is needed whenever a scientist wishes to compare classifications derived from different information about a set of objects. If the criterion is violated he will be unable to tell whether differences between the classifications represent differences between the data sets or whether they are a byproduct of instability in the method of classification.

Williams *et al.* object to criterion C on the grounds that ultrametric DC's rarely arise. However, the output of any hierarchic cluster method is, as pointed out in Section 1, equivalent to an ultrametric DC. Clearly if a DC already represents a classification of the kind sought it is unreasonable to change it. The criterion is not, as Williams *et al.* suggest, equivalent to a requirement that the method used be single-link. Many hierarchic cluster methods satisfy it.

The only hierarchic cluster method which satisfies all the criteria suggested by Jardine and Sibson is single-link. This method is subject to the defect that it may produce highly inhomogeneous clusters when objects are connected together by chains of intermediates (see **Fig. 1**). This was called the chaining effect by Lance and Williams (1967). Cormack (1971) has suggested that this should be considered as an inherent defect of hierarchic clustering.

Williams *et al.* suggest that the three following pragmatic criteria may overide the criteria A, B and C listed above.

1. The grouping must be more intense than that implied by the original dissimilarity measures;
2. The grouping must be relatively insensitive to outlying values, due either to aberrant individuals, or to errors or inaccuracies in the data;
3. The ultrametric transformation should not be necessarily, or even usually, invariant over the entire population.

As stated, the first criterion appears inappropriate for applications of automatic classification in which a simplified representation of the data is required. Williams, Clifford and Lance (1971) defend the criterion on the grounds that a user often requires 'an intensely clustered, artificially sharpened analysis which will draw his attention to nodal situations existing in the data.' It is true that scientists often hope to find well-marked clusters, but this is no justification for using methods which find them even when they are not present.

Part of the meaning of this criterion may be that the groups formed by a classification method should satisfy a homogeneity criterion. The single-link method may yield clusters which are highly inhomogeneous when objects are linked by chains of intermediates. It is not surprising that attempts to generate clusters which are more homogeneous than those obtained by single-link leads to arbitrariness in a cluster method. Thus, in Fig. 1 it can be seen that division of the larger single-link cluster into two disjoint homogeneous clusters involves arbitrary allocation of the intermediate object.

The second criterion does not conflict with our criteria. The requirement that the grouping be insensitive to aberrant individuals is satisfied by the single-link method. An individual which is highly dissimilar from all others appears as a single-element cluster at a relatively high level and cannot affect the

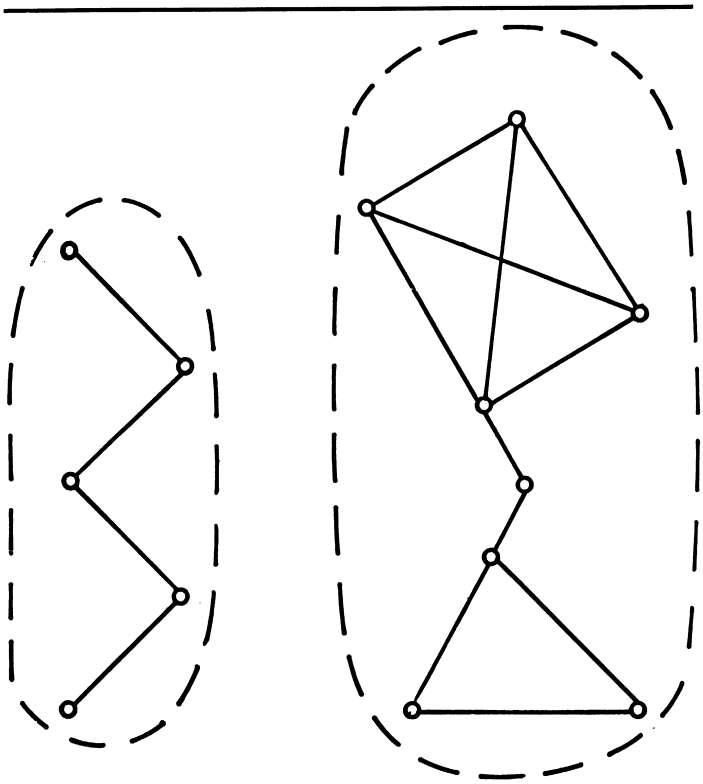

Fig. 1. Single-link clusters. Vertices represent objects and edges join pairs of vertices which represent objects with DC values less than or equal to some threshold.

ranges of clusters below that level. The requirement that the grouping be relatively insensitive to errors or inaccuracies in the data is identical with our criterion B.

Williams *et al.* support the third criterion by an example from biological classification.

> If a sample of grass genera including members of what are currently recognised as pooid, panicoid, festucoid 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.

If the bamboos are more heterogeneous they will appear as a cluster in the dendrogram obtained by single-link at a higher level than clusters which correspond to the other groups of grasses. But this indicates only that when a scientist interprets the results obtained by single-link, the clusters which he finds of interest may not all lie at a single level. A method which concealed the heterogeneity of the bamboos would be seriously misleading.

Each of the two pragmatic criteria (1 and 3) which conflict with our criteria of adequacy for a hierarchic cluster method implies that clusters should satisfy a homogeneity criterion. In Jardine and Sibson (1968) we showed that there need be no conflict between a cluster homogeneity criterion and our criteria of adequacy for representation of a DC when non-hierarchic cluster methods are used. When hierarchic cluster methods are used the conflict cannot be avoided.

As we mentioned earlier there are many applications of automatic classification where it is appropriate to allow representation of data to be subject to extrinsic criteria. The requirement of Williams *et al.* (1971) that clusters in a hierarchic clustering be highly homogeneous is one such external constraint, and the flexible method of Lance and Williams (1967) may, for certain choices of its parameters, be appropriate in such cases, as may the various monothetic divisive methods. However, even when a hierarchy of highly homogeneous clusters is wanted there is a strong case for generating them in such a way that the investigator has some control over the arbitrary choices involved. For example, single-link may be applied to a DC and a selection then made of the clusters which satisfy a homogeneity criterion. The ball-cluster definition of Jardine (1969) yields all single-link clusters which are neighbourhoods of each of their members, and the cluster method of van Rijsbergen (1970) yields all single-link clusters which satisfy an even stronger homogeneity criterion. This approach will often fail to generate any clusters from a DC. The following approaches are less stringent. A hierarchy of clusters more homogeneous than can be obtained by single-link may be obtained by generating first a non-hierarchic clustering, and then reallocating objects which lie in the overlaps of clusters. Another technique which can be used to generate more homogeneous clusters is the calculation of secondary DC's. From a DC a secondary DC is derived by ranking from each object the remaining objects in order of increasing dissimilarity, and calculating a rank-order correlation measure between the rankings. As this process is iterated it generally happens that the clusters obtained from the DC's by single-link become progressively more homogeneous, although no convergence proof is known. Related methods for 'emphasising' clusters have been described by Bonner (1964) and Vaswani (1968).

4. Conclusion

We conclude that controversy about criteria of adequacy for methods of automatic classification can be partially resolved by considering the purposes for which scientists classify. The criteria of adequacy proposed by Jardine and Sibson (1968) are appropriate when the aim of classification is to simplify data in ways which may suggest or confirm hypotheses. In other classificatory problems it may be appropriate to use methods of classification which impose constraints on the homogeneity of clusters, as suggested by Williams *et al.* (1971), or which impose constraints on the number or sizes of the clusters. Such external constraints are incompatible with the criteria of adequacy for simplified representation of the dissimilarities of objects by a hierarchic clustering. If a scientist wishes both to obtain homogeneous clusters and to obtain a simplified representation he should use a non-hierarchic cluster method. If he definitely requires a hierarchic system of clusters he must make a choice between adequacy of representation and homogeneity of clusters by considering carefully the purpose for which he is classifying.

References

BONNER, R. E. (1964). On some clustering techniques, *IBM J. Res. Dev.*, Vol. 8, p. 22.

BOYCE, A. J. (1966). Mapping diversity: a comparative study of some numerical methods, In A. J. Cole (ed.) *Numerical Taxonomy*, Academic Press, London and New York, p. 1.

JARDINE, N. (1969). Towards a general theory of clustering (abstract of paper), *Biometrics*, Vol. 25, p. 609.

JARDINE, N. (1970). Algorithms, methods, and models in the simplification of complex data, *The Computer Journal*, Vol. 13, p. 116.

JARDINE, N., and SIBSON, R. (1968). The construction of hierarchic and non-hierarchic classifications, *The Computer Journal*, Vol. 11, p. 177.

KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: a numerical method, *Psychometrika*, Vol. 29, p. 115.

LANCE, G. N., and WILLIAMS, W. T. (1967). A general theory of classificatory sorting strategies. 1. Hierarchical systems, *The Computer Journal*, Vol. 9, p. 373.

LANCE, G. N., and WILLIAMS, W. T. (1968). Note on a new information-statistic classificatory program, *The Computer Journal*, Vol. 11, p. 195.

LERMAN, I. C. (1970). *Les bases de la classification automatique*, Gauthier-Villars, Paris.

MACNAUGHTON-SMITH, P. (1965). Some statistical and other techniques for classifying individuals, H.M.S.O. London.

MINKOFF, E. C. (1965). The effects on classification of slight alterations in numerical technique, *Syst. Zool.*, Vol. 14, p. 196.

ROCCHIO, J. J. (1966). Document retrieval systems—optimisation and evaluation, Report ISR 10 to National Science Foundation, Harvard Computer Lab., Chap. 4.

SIBSON, R. (1971). Some observations on a paper by Lance and Williams, *The Computer Journal*, Vol. 14, p. 156.

SNEATH, P. H. A. (1957). The application of computers to taxonomy, *J. Gen. Microbiol.*, Vol. 17, p. 201.

VASWANI, P. K. T. (1968). A technique for cluster emphasis and its application to automatic indexing. IFIP Congress, Edinburgh, 1968. Booklet G, p. 1. North Holland Publishing Company, Amsterdam.

WILLIAMS, W. T., CLIFFORD, H. T., and LANCE, G. N. (1971). Group size dependence: a rationale for choice between numerical classifications, *The Computer Journal*, Vol. 14, p. 157.

WILLIAMS, W. T., LANCE, G. N., DALE, M. B., and CLIFFORD, H. T. (1971). Controversy concerning the criteria for taxonometric strategies, *The Computer Journal*, Vol. 14, p. 162.

WISHART, D. (1969). FORTRAN II programs for 8 methods of cluster analysis (CLUSTAN I), Computer Contribution No. 38 to State Geological Survey, University of Kansas.

VAN RIJSBERGEN, C. J. (1970). A clustering algorithm, *The Computer Journal*, Vol. 13, p. 113.