The stability of Classifications of Binary Attribute Data

by

D. M. Jackson

Technical Report

70-65

Department of Computer Science
Cornell University

March 1970

Presented at the Classification Society meeting held at the Battelle Memorial Institute, Columbus, Ohio, April 7th.-8th., 1970

The stability of classifications of binary attribute data.

D. M. Jackson

The general intention of cluster analysis is to reduce detailed information about the objects or individuals of a population to generalisations about certain subsets of this population. This general statement of intent gives little guide for devising methods of constructing classifications, and indeed, taxonomists have approached the problem from a number of substantially different points of view, from the essentialism of Caesalpino, for example, to the neo-nominalism of the present day pheneticists and the evolutionism of the phylogeneticists. Classification techniques and, indeed, the foundations of these techniques still remain controversial as witness the discussions in recent taxonomic literature (see, for example, Colless (1)). The clustering algorithms which have been devised often lead to classifications at considerable variance with one another, even when applied to the same population and to the same set of date (Minkoff (2)). To a certain extent a classification is artificial since it depends, particularly in its form, on the purpose for which it was constructed and since many taxonomic concepts are difficult to define operationally. Classifications of the same data, but serving different purposes, will therefore justifiably differ from each other even in the case of allegedly natural classifications. However, if a classification is to represent in a tractable form, once this has been decided upon, the affinities which exist between subsets of the objects of the population, then it might reasonably be expected that

classifications devised by different methods should be in substantial agreement with one another. Stated negatively, there would be grounds for concern if the classifications had no, or at most few, points of agreement. Some variability, clearly, derives from the measure of affinity between pairs of objects. There is considerable flexibility in the way in which this might be carried out in practice, ranging from the determination by an appropriate formula from the character lists on the one hand (Sokal and Sneath (3)), to a reduction of the a priori "proximities" to conform to a specified accuracy with a generalised distance function on the other hand (Kruskal (4)). But the variability in the results from cluster analyses using the same set of affinities has led understandably to the questioning of the reliability and consistency of these techniques in practice, particularly when a second analysis has been performed with the intention of corroborating the findings of the first. Before accepting the results of a cluster analysis, it is therefore appropriate to determine whether or not the results satisfy a number of necessary conditions which characterise, in part, the relationship between an adequate classification and the original population. The "adequacy" of a classification in this sense does not invoke value judgments concerning the reliability or generality of the information inferable, for example, from a polythetic classification. Problems in taxonomy are analogous in this respect to those of numerical analysis, in which the numerical solution of a partial differential equation, for example, is acceptable as "adequate" only when solutions by different methods and different step

lengths agree to an appropriate number of significant figures, and when tolerable error bounds on the solution have been obtained. The purpose of this paper is to examine the conditions appropriate to the context of numerical taxonomy and to describe how, in certain cases, such conditions may be applied in practice.

The following is a set of necessary conditions which may be generally applied to classification procedures:

1) The classification must be well-defined. That is, application of the algorithm must supply a single result.

2) The classification must be stable. That is, the classification must not be grossly affected by small changes in the data.

3) The algorithm must be unaffected by a permutation of the names of the objects to be classified. That is, the algorithm must be independent of the labelling of the objects.

4) The algorithm must be independent of scale. That is, the algorithm must be unaffected by multiplication of the similarity matrix by a positive, non-zero constant.

These conditions are clearly not sufficient since they leave entirely unresolved the question of the adequacy of a particular configuration of classes in summarising the detailed inter-object affinities for particular purposes. The adequacy of the grouping as a classification (sensu strictu) of the objects of the population remains undecidable within this framework.

Independence of scale (condition (4)) is a natural requirement since the scale of similarity is essentially arbitrary. The stronger condition of invariance of classifications under a monotonic transformation of the similarity function has been suggested by Benzécri (5), with the result that a number of the similarity functions given by Sokal and Sneath (op. cit. pp. 129 - 130), namely those which are mutually monotonic (demonstrable algebraically), may be regarded as equivalent. A review of these methods is given by de la Véga (6).

Independence of labelling (condition (3)) requires that a classification should not be dependent on the order in which the objects are treated, and it is demonstrable only in the case in which all classes satisfying the classificatory criterion have been enumerated. Typically, clustering techniques consist of three distinct procedures: the <u>starting procedure</u>, dealing with the initiation of the search for a class; the <u>assignment procedure</u>, dealing with the allocation and reallocation of objects to putative classes; and the <u>stopping procedure</u>, dealing with the conditions under which the search for a class may be terminated. The stopping procedure establishes whether a set of objects satisfies a previously defined classificatory criterion, and if so terminates the search. The classificatory criterion is the condition which must be satisfied by the objects of the set as a whole before the set is admitted as a class of the classification. Typically, the criterion will be that a certain function is locally maximal or minimal for an acceptable class. In effect

the stopping procedure examines all sets in an appropriately defined neighbourhood of the putative class to establish the existence of a local extremum. The assignment procedure regulates the admission or rejection of objects during the growth of a set until the criterion is satisfied. It is clear that if all classes satisfying the classificatory criterion are to be located, then all subsets of the objects of the population must be examined. Similarly, condition (1) is demonstrable only in the case in which all sets which satisfy the classificatory criterion have been extracted. Indeed, the failure of conditions (1) and (3) are related. Relabelling enables more classes to be found, and the examination of all permutations corresponds to inspecting all subsets of the objects to determine which subsets satisfy the classificatory criterion. Conditions (1) and (3) are susceptible to testing only when all subsets of objects of the population have been examined. For large populations, that is, populations for which examination of all subsets on a given computer configuration cannot be contemplated, these conditions cannot be applied. Accordingly, for such populations, the remaining condition, that of stability, is decisive.

A classification will be said to be stable if small changes in the material to be classified leads to small changes in the classification. This presupposes that the overall difference between two classifications, both of which may be incomplete in the sense described above, may be defined and quantified. A low

order technique for achieving this is discussed by Jackson (7). The need for stability arises when the reliability of the attribute descriptions of the objects is questionable. Errors in these descriptions may arise simply from a) faulty keypunching of the data. More seriously, errors may be caused by b) conflating "not applicable" with the state "absent", in the case of binary attributes, by c) the use of attributes which are stable in some objects of the population but variable in others, or by d) the faulty determination of homologues. The latter is of extreme importance in taxonomic studies. Some consideration has been given by Fisher and Rohlf (8) to the effect of the faulty determination of homologues on hierarchical classifications of a particular population. In each of these cases, a classification of the material must be obtained in spite of the errors it contains. Since the original material is not recoverable, the best which may be achieved is to ensure that the erroneous parts of the data do not exert a significant effect in the classification as a whole.

The case for stability may be argued on purely practical grounds. Suppose that a few additional observations are made on a population of objects which has already been classified. It would be expected that the classification constructed on the basis of the additional information would differ only slightly from the previous classification. Indeed, a large change in the classification would indicate extreme sensitivity to the

additional set of observations and would suggest, in the first instance, that insufficient observations had been gathered to provide adequate descriptions of the objects.

A connection may be seen between stability and continuity. Suppose that a classification is produced by a transformation of the data. Small changes in the data necessarily lead to small changes in the classification provided that the transformation is continuous. The classification is therefore stable under these circumstances and stability may be inferred by verifying the continuity of the transformation. In general, however, classification algorithms cannot be expressed in these terms. This is particularly true of classifications constructed from categorical data where the similarity function, although continuous, has a discrete domain and therefore a discrete range. It is also true of classifications in which class membership is not susceptible to quantification, for example, by a probabilistic weight, but in which class membership is a two-state relationship (present/absent).

In the cases in which a classification algorithm cannot be represented by a continuous transformation of the data, stability is not an analytic property of the algorithm. Rather, it is dependent both on the algorithm and the population which is to be classified. Stability must therefore be demonstrated each time a classification algorithm is applied to a set of data by

investigating the action of the algorithm on the particular set of data. Alternatively, stability must be ensured by appropriate modification of the algorithms currently in use. This possibility will not be examined, and the analysis will be applicable to phenetic clustering techniques utilising a matrix of inter-object affinities.

Let us consider those algorithms which may be expressed as hill-climbing algorithms. The assignment procedure involves the optimising of a given function g in a sequence of steps, each of involving the modification of the partition of the population achieved at the previous step. g is a function of this partition. Let g_k be the value of g after the k-th step. The k+1 th. partition is acceptable if g_{k+1} is smaller than g_k. Otherwise the k th. partition remains acceptable at the k+1 th. step. This is appropriate for minimising g. The stability condition requires that the reduction in g at the k+1 th. stage is attributable to a redistribution of the objects of the population rather than to the effect of errors in the data array. Now g_k is a function of the partition of the population at the kth. step. Let g'_k be the smallest value of g_k which may be obtained when the similarities between objects t_i and t_j are subjected to the appropriate fluctuations $\pm E_{ij}$, the latter being the expectation of the absolute error in the similarity between objects t_i and t_j on the assumption that errors occur <u>independently</u> and <u>equiprobably</u> with probability r. E will be called the

error matrix. Let e'_k be the decrease in g_k. Similarly, let g''_{k+1} be the largest value of g_{k+1} which may be obtained when the similarities between objects t_i and t_j are subjected to the appropriate fluctuations $\pm E_{ij}$: Let e''_{k+1} be the increase in g_{k+1}. From the stability condition, we require that g''_{k+1} is small than g'_k. Accordingly, the revised condition is that: $g_{k+1} - g_k < -(e'_k + e''_{k+1})$. This is the revised condition for minimization necessary to ensure the stability of the classification with respect to independent and equiprobable errors in the data, and is applied during the assignment procedure.

The derivation of the error matrix E, for a general statistical function and a general similarity function (of the association type) has been discussed by Jackson (9) and will not be given here. The problem is related to a class of "Matching Problems" described by David and Barton (10). It may be shown that the expectation E(A,B) of the absolute error in the similarity f between two objects A and B with N attributes (binary) when each of the attributes is susceptible to error equiprobably and independently with probability r, is given, in the case of the Jaccard, Russel and Rao, Kulczynski and Dice similarity functions, as follows:

$$E(A,B) = \prod_j n_j! \cdot \sum_{\underline{m}} |f(\underline{n}) - f(\underline{m})| \cdot h(\underline{m},\underline{n})$$

where $h(\underline{m},\underline{n})$ =

$$(1 - r)^{2N} \cdot \sum^{*} \frac{2^{m_2 - I(2,2)}}{\prod_{i,j} I(i,j)} \cdot (1 + t^2)^{I(2,2)} \cdot t^{N+I(3,1)+I(1,3)}$$

$$- \sum_{i}' I(i,i)$$

and where:

1) $\sum^{*}$ denotes summation over all tables I such that:

2) $\sum_{j} I(i,j) = m_i$; $\sum_{i} I(i,j) = n_j$; $\sum_{j} m_i = \sum_{i} n_j = N$

3) I(i,j) is integer and non-negative

4) $\underline{n} = (n_1, n_2, n_3)$; $\underline{m} = (m_1, m_2, m_3)$; $t = r/(1-r)$

5) n_1 = no. of attributes which have state 0 in A and 0 in B

n_2 = no. of attributes which have state 0 in A and 1 in B or state 1 in A and 0 in B

n_3 = no. of attributes which have state 1 in A and 1 in B

The error matrix is constructed by determining E(A,B) for all relevant (AB), or more efficiently, for all (n_2, n_3) which occur in the pairs of objects (A,B).

It has already been remarked that the same polynomial $h(\underline{m},\underline{n})$ is applicable to each of the four well known similarity functions of Jaccard, Russell and Rao, Kulczynski and Dice. Providing a technique may be found for computing $h(m,n)$ rapidly, considerable generality has been achieved. Since taxonomic studies frequently involve using a number of such functions exploratively, this generality is valuable in practice and will be retained in techniques for evolving the matrix E. The computation of $h(\underline{m},\underline{n})$, however, is not straightforward for it involves summation over all three-by-three integer, non-negative tables, I, with row and column sum constraints. Some results have been obtained by O'Neil (11), for different row-sum and column-sum constraints, in the case where the tables are of larger dimension and have zero or one entries. Once all feasible tables have been constructed, the value of $h(\underline{m},\underline{n})$ may be computed by substituting a suitable numerical value for r. However, the number of feasible tables is large and not all tables will result in substantial contributions to E(A,B). Provided the polynomial expansion of $h(\underline{m},\underline{n})$ in terms of t may be truncated after a specified number of terms with tolerable error, and provided those tables which contribute to the retained terms may be isolated, then some economy may be achieved. A number of results have already been obtained which may be useful for this purpose. The probability of error, r, may be set to any realistic value, for example, 0.05 representing a 5% error rate over the population as a whole. It is suggested that the use of the error matrix E in the fashion

described above will result in greater stability in classification. Classifications so constructed will be insensitive to errors occuring independently and equiprobably at this rate.

References

1. Colless, D. H. The phylogenetic fallacy. Syst. Zool., 16, 1967, pp. 289 - 295.

2. Minkoff, E. C. The effects on classification of slight alterations in numerical technique. Syst. Zool., 14, 1965, pp. 196 - 213.

3. Sokal, R. R. and Sneath, P. H. A. Principles of Numerical Taxonomy. Freeman. 1963.

4. Kruskal, J. B. Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis. Psychometrika 29, 1964, pp. 1 - 27.

5. Benzécri, J. P. Sur les algorithmes de classification. 1965. (texte multigraphie), 8pp.

6. de la Véga, W. F. Techniques de classification automatique utilisant un indice de ressemblance. Revue francaise de Sociologie, VIII(4), 1967, pp. 506 - 520.

7. Jackson, D. M. Comparison of classifications. In: Numerical Taxonomy, (A. J. Cole ed.), Academic Press, 1969, 324pp.

8. Fisher, D. R. and Rohlf, F. J. Robustness of numerical taxonomic methods and errors in homology. Syst. Zool., 18, 1969, pp. 33 - 36.

9. Jackson, D. M. An error analysis for functions of qualitative attributes with application to information retrieval. Proceedings of the Third International Symposium on Computer and Information Science, Florida, 1969 (in press).

10. David, F. N. and Barton, D. E. Combinatorial Chance. Hafner, 1962.

11. O'Neil, P. E. Asymptotics and random matrices with row-sum and column-sum restrictions. Bulletin of the Americal Mathematical Society, 75(6), 1969, pp. 1276 - 1282.