A Re-examination of Four Classificatory Fusion Strategies

D. J. ABEL*

CSIRO Division of Computing Research, Davies Laboratory, PMB Aitkenvale, Queensland 4814

W. T. WILLIAMS

Australian Institute of Marine Science, PMB 3, Townsville, Queensland 4810

Space-distortion has been recognised as an important factor in selection of a classificatory fusion strategy but has not been considered in depth. This paper suggests that two separable phenomena are present. Analysis shows that the group-average strategy is free from both forms and that the flexible strategy is not group-size-dependent. While use of the centroid and incremental sum of squares strategies has been regarded as dubious with other than distance measures and squared Euclidean distance, it is argued that both can be viewed as formulated algebraically and then retain their established space-distortion properties.

1. INTRODUCTION

The user of numerical methods of classification has at his disposal a diverse set of techniques, with selection of the most appropriate technique for a particular problem calling for an intelligent matching of the characteristics of his data set and the style of classification desired to the properties of the chosen technique. The restrictions on choice imposed by the characteristics of the data set are the more obvious. For example, it is common in ecological studies to represent sites only by the absences and presences of a set of species, so that the use of specialist dissimilarity measures and techniques able to accommodate such measures is demanded. The style of the classification is more subtle. For the moment we note that some techniques can accentuate the differences between individuals and groups and inherently tend to lead to classification with homogeneous groups and groups of outliers, while others tend to yield diffuse groups. A user might see accentuation of differences as desirable in some instances and not in others: to a certain extent the choice will reflect the user's philosophical approach to classification as well as the perceived requirements of the task in hand. We suggest, however, that a knowledge of this aspect of a technique's performance is necessary when rationally choosing a technique from the several available.

This paper considers agglomerative techniques and gives reasons for choosing between alternatives. A technique is agglomerative if it begins with a set of isolated elements and progressively fuses these into groups of increasing size until a required number of groups is reached. The nominal objective (the global objective) is minimisation of some quantity related to a certain number of groups and the allocation of elements to those groups. It is more usual, however, to proceed by minimising a local criterion for each fusion in turn, so that the algorithm is sub-obtimal or non-exact in the terminology of Muller-Merbach.[11] For example, under the within-group sum of squares method of Ward,[16] the global objective is minimisation of the within-group sum of squares, while the local objective is to minimise the increase in the sum of squares as the result of the next fusion to be performed. Strategies within the agglomerative approach are defined by the various local criteria adopted. In this paper we are particularly concerned with the centroid, group average, incremental sum of squares and the flexible strategies which are the more widely used.

Two aspects are considered. The first is space distortion, identified by Lance and Williams[8] as an aspect of behaviour considerably influencing the characteristics of the classifications generated. The term is based on a model of inter-element dissimilarity measures as defining a space of known properties. As fusion proceeds it need not follow that the inter-group measures preserve the properties of the space. If a strategy does preserve the properties, it is termed 'space-conserving'. If, on the other hand, the group–element or group–group dissimilarities require a view of the space as being contracted or dilated in the immediate vicinity of the groups, then the strategy is defined as 'space distorting', or more specifically as 'space contracting' or 'space dilating'. Space-dilating strategies were noted as tending to lead to 'intense' groups where all elements in a group were closely similar to other members of the group, while space-contracting strategies tend to form diffuse groups which could be relatively heterogeneous. In subsequent investigations,[18] space distortion was considered to be an effect of group size only, so that 'space distortion' has been usually accepted as synonymous with size distortion. We shall consider space distortion more closely and shall identify another effect termed 'range distortion'. Analysis will also more clearly identify the relative propensity of the strategies to be influenced by space distortion.

The second aspect to be considered is the range of dissimilarity measures able to be appropriately treated by the combinatorial forms of the strategies, i.e. the measures compatible with the strategies. Combinatorial forms for (amongst others) the centroid and group average strategies were advanced by Lance and Williams,[9] who were also led to propose the flexible strategy with its purely algebraic formulation. Under a combinatorial form of a strategy, a dissimilarity matrix is initially defined for all pairs of elements and is maintained for all pairs of outstanding elements and groups. At each step, the groups or elements fused are chosen as those with the smallest dissimilarity, and the measures for the new group and the remaining elements or groups defined from group sizes and the measures in force immediately before the fusion. This is considerably more efficient than earlier

[*] To whom correspondence should be addressed.

approaches where evaluation of the criterion was performed by reference to the attribute values for elements and usually demands less storage, as the raw data can be discarded when fusion commences. Lance and Williams[9] showed the group average and centroid strategies to be compatible with Euclidean distance measures and with the squared Euclidean distance (D^2). They also suggested that the group average strategy was compatible with all measures 'providing the concept of an average measure is acceptable'. Wishart[17] and Burr[2] independently provided a combinatorial formulation of the incremental sum of squares or Ward's method,[16] considering D^2 only.

It is apparent, however, that there is a number of measures other than Euclidean distance and D^2 in widespread use, such as the metric Canberra[9] and Gower[5] measures and the quasi-metric Bray–Curtis[1] and Jaccard[6] measures. More generally, it is worthwhile to consider if dissimilarity measures not possessing the four properties required for a measure to be a metric or equivalently a distance measure[14] can be accommodated within the combinatorial formulations of strategies. Such measures are termed non-metrics. There is some disagreement in the literature on this point. Sneath and Sokal[14] and Ross[13] see such usage as dubious, while Clifford and Stephenson[3] state it is permissible. Burr[2] notes the use of the ISS strategy with measures other than D^2 as not specifically addressed. There is, however, some empirical evidence that use of the ISS strategy with non-metrics does produce readily interpreted classifications.[12, 15] We shall argue that the combinatorial formulations of the common strategies can be viewed as purely algebraic statements of strategies and so can be applied with non-metrics even if a geometric interpretation is not available. We also note that the space-distortion properties remain the same as shown for the distance measures.

2. COMBINATORIAL FORMULATIONS OF STRATEGIES

We suppose that two elements or groups i (with n_i elements) and j (with n_j elements) have just fused to form a composite group k (with $n_k = n_i + n_j$) elements. Consider a further group h (with n_h elements); it is then required to calculate the measure d_{hk} from the already-known values of d_{ij}, d_{hi} and d_{hj}. Lance and Williams[8] define the following general relation for combinatorial strategies:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \, | d_{hi} - d_{hj} | \qquad (1)$$

where α_i, α_j, β and γ are functional parameters defining the particular strategy in use. The parameter γ is used only for the extreme strategies of 'nearest and furthest neighbour', and will not be discussed further. The remaining parameters are functions of n_h, n_i and n_j (for all strategies except the flexible) or constants, once β is specified in the flexible. The four strategies to be considered are implemented by parameter definitions as follows.

Centroid:

$$\alpha_i = n_i/n_k; \quad \alpha_j = n_j/n_k; \quad \beta = -n_i n_j/n_k^2 \qquad (2)$$

Group average:

$$\alpha_i = n_i/n_k; \quad \alpha_j = n_j/n_k; \quad \beta = 0 \qquad (3)$$

Incremental Sum of Square (ISS):

$$\begin{aligned} \alpha_i &= (n_h + n_i)/(n_h + n_k); \\ \alpha_j &= (n_h + n_j)/(n_h + h_k); \\ \beta &= -n_h/(n_h + n_k) \end{aligned} \qquad (4)$$

Flexible:

$$\alpha_i = \tfrac{1}{2}(1 - \beta); \quad \alpha_j = \tfrac{1}{2}(1 - \beta); \quad \beta < 1 \qquad (5)$$

It will be convenient to define two further quantities Δ_i and Δ_j where $\Delta_i = d_{hi} - d_{ij}$ and $\Delta_j = d_{hj} - d_{ij}$. We note that a desirable property of a strategy is that at each fusion the value of the local criterion is not less than that at any preceding fusion. Strategies with this property are said to be monotone; those for which this is not invariably true are said to exhibit reversals. Lance and Williams[8] show that the condition for monotonicity is simply $\alpha_i + \alpha_j + \beta \geqq 1$; from this it follows that the centroid strategy is subject to reversals, whereas the rest are monotone.

Equations (2) to (5) can be written in functional form, treating the various n and d as parameters. We have

$$d_{hk} = F(n_k, s, n_h, d_{ij}, d_{hi}, d_{hj}) \qquad (6)$$

where $s = n_i/n_j$. The advantage of this functional form is simply that it allows the use of parametric analysis to examine the effects of the various constituents of a group's definition.

3. A RE-EXAMINATION OF SPACE DISTORTION

Studies to date[2, 3, 8, 18] have tacitly assumed that space-distortion is a single phenomenon with 'group-size dependence' synonymous with space distortion. The parameterized form of the general combinatorial formulation of the strategies suggests, however, that there is another phenomenon present. While the two may be confounded in practice, a recognition of them as separable provides a clearer assessment of the influences on the classifications yielded from the various common agglomerative strategies given particular characteristics of the presented set of elements. We shall term these phenomena size distortion and range distortion.

Some discussion of 'space distortion' is desirable to elucidate the concepts involved. The geometric basis for such strategies as the centroid and group average encourages a view of a group as a point or element in attribute space, defined by the group's centroid and mean over attributes respectively. A single-element representation of a group under the ISS strategy is less obvious, but there remains a view of a group defined as a region in space enclosing the group's members. This model can give a misleading impression of the dissimilarity of a group and another element in some cases. Consider, for example, the fusion under the ISS strategy of eleven elements, with the first to tenth elements identical. After five of the identical elements have fused, the eleventh element is more dissimilar to the group than any of its constituent elements, i.e. it appears to have receded from the region defined by the group. It appears even more dissimilar to the group once it includes all ten elements. To reconcile the view of a group as a region in attribute space and the actual dissimilarities it is necessary to introduce distortion of the space in the vicinity of the region occupied by the group. We observe in passing that, for the centroid, group average and ISS strategies, space distortion is not an artefact of the combinatorial formulations, as these have been shown to be exactly equivalent in their computed dissimilarities to evaluations referencing the individual elements.[8, 17, 2]

That is, the dissimilarity between a group and some reference element is not simply a function of the relative placement of the reference element and the region occupied by the group but also reflects other group parameters. Size distortion is attributable to the number of elements within the group, while range distortion is attributable to the dispersion of elements within the group's region.

3.1 Size Distortion

This is the classical case as introduced by Lance and Williams.[8] A strategy is said to exhibit size distortion if the dissimilarity measure for two groups is sensitive to the group sizes, i.e. n_h and n_k. Applying the functional form of (6), a strategy is formally defined as size-distorting if

$$F(bn_k, s, n_h, d_{ij}, d_{hi}, d_{hj}) \neq F(n_k, s, n_h, d_{ij}, d_{hi}, d_{hj})$$

where b is greater than unity and for some fixed n_h, s, d_{ij}, d_{hi} and d_{hj}. The various types of size distortion can be defined in the same manner. For example, a strategy is size-dilating if

$$F(bn_k, s, n_h, d_{ij}, d_{hi}, d_{hj}) > F(n_{k,} s, n_h, d_{ij}, d_{hi}, d_{hj}).$$

3.2 Range-distortion

We prefer this term to the perhaps more general, but cumbersome, 'heterogeneity-distortion'. We define range-distortion as occurring when a strategy is sensitive to the range of elements within the region of attribute space occupied by their group. The only parameter available to use within the general combinatorial formula is d_{ij}. Formally, a strategy is said to exhibit range-distortion if

$$F(n_k, s, n_h, d_{hi}, d_{hj}, bd_{ij}) \neq F(n_k, s, n_h, d_{hi}, d_{hj}, d_{ij})$$

where b is greater than unity and d_{ij} is non-zero.

The situation has been confused by the demonstration in Ref. 18 that the difference between the square of the distance between the centroids for two groups and the value for D^2 is equal to the sum of the variances of the two groups. We submit that this is not space distortion, but rather shows an identity relating two local objective functions, for the group average and centroid strategies. Indeed, by the test advanced for range distortion the group average strategy must be range-conserving, as its combinatorial formulation contains no term in d_{ij}.

4. ANALYSIS

(i) *Size distortion*

We adopt what is essentially a form of sensitivity analysis. To determine whether a strategy is size-distorting or size-conserving, we examine the change in d_{hk} with increasing n_k, holding other parameters constant. If that change increases with n_k, then the strategy is size-dilating; if the change is zero, the strategy is size-conserving. These criteria differ somewhat from those of Ref. 18 in not considering cumulative displacement.

The change in dissimilarity measure can be developed as the difference in dissimilarity measures to a group h from a group k (defined as previously) and from a group K, where K has been formed from two groups I and J identical to i and j respectively in all parameters except group size. As the ratio n_i/n_j has been discounted as irrelevant to size dilatation, we also require that n_i/n_j and n_I/n_J are equal. The group parameters for k and I are then related by:

$$d_{hi} = d_{hI}; \quad d_{hj} = d_{hJ}; \quad d_{ij} = d_{IJ}$$
$$n_I/n_i = n_J/n_j = n_K/n_k = s$$

It will also be convenient to write $r = n_i/n_k$ and $t = n_h/n_k$. (We note in passing that, at least for some strategies, a group such as I can be formed by fusing s groups identical with i; but the analysis does not assume that all members of such a group are identical.)

Presentation of the sensitivity analyses will be simplified by introducing a third subscript to identify the strategy being considered. We use 1 to denote the centroid strategy, 2 for the group average, 3 for the ISS and 4 for the flexible. Thus d_{hkl} denotes the dissimilarity of h and k under the centroid strategy.

It is trivial to show that

$$d_{hK1} - d_{hk1} = d_{hK2} - d_{hk2} = d_{hK4} - d_{hk4} = 0$$

It follows that the centroid, group average and flexible strategies are free from size distortion. However,

$$d_{hK3} - d_{hk3} = \frac{t(s-1)}{(t+1)(t+s)}\{(1-r)\Delta_i + r\Delta_j\}$$

This expression is always positive if $s > 1$; the ISS strategy is therefore always size-dilating.

(ii) *Range-distortion*

A similar form of analysis can be applied to assess the range-distorting properties of the four strategies. We now take I, J and K defined by

$$d_{hI} = d_{hi}; \quad d_{hJ} = d_{hj}; \quad d_{IJ} > d_{ij}$$
$$n_I = n_i; \quad n_J = n_j; \quad n_K = n_k$$

The differences in distance measures then reduce to:

$$d_{hK1} - d_{hk1} = n_i n_j (d_{IJ} - d_{ij})/n_k^2$$
$$d_{hK2} - d_{hk2} = 0$$
$$d_{hK3} - d_{hk3} = -n_h(d_{IJ} - d_{ij})/(n_h + n_k)$$
$$d_{hK4} - d_{hk4} = \beta(d_{IJ} - d_{ij})$$

We observe that the group average strategy is perfectly range-conserving. The centroid and ISS strategies are range-contracting, while the flexible strategy can be range-contracting, -conserving or -dilating depending on the value of β. For the usual value of $\beta = -0.25$, strategy 4 is range-contracting.

(iii) *Comparison of strategies*

The six possible comparisons are summarised below:

$$d_{hk2} - d_{hk1} = n_i n_j d_{ij}/n_k^2$$

$$d_{hk3} - d_{hk1} = n_h(n_j \Delta_i + n_i \Delta_j)/n_k(n_h + n_k) + n_i n_j d_{ij}/n_k^2$$

$$d_{hk3} - d_{hk2} = n_h(n_j \Delta_i + n_i + \Delta_j)/n_k(n_h + n_k)$$

$$\begin{aligned} d_{hk4}-d_{hk1} &= \Delta_i\{\tfrac{1}{2}(1-\beta)-n_i/n_k\} \\ &\quad +\Delta_j\{\tfrac{1}{2}(1-\beta)-n_j/n_k\}+n_i n_j d_{ij}/n_k^2 \\ d_{hk4}-d_{hk2} &= \Delta_i\{\tfrac{1}{2}(1-\beta)-n_i/n_k\}+\Delta_j\{\tfrac{1}{2}(1-\beta)-n_j/n_k\} \\ d_{hk4}-d_{hk3} &= \Delta_i\left\{\tfrac{1}{2}(1-\beta)-\frac{n_h+n_i}{n_h+n_k}\right\} \\ &\quad +\Delta_j\left\{\tfrac{1}{2}(1-\beta)-\frac{n_h+n_j}{n_h+n_k}\right\} \end{aligned}$$

Considering first the centroid, group average and ISS strategies, we note that, trivially, d_{hk1} is always less than d_{hk2}, which in turn is always less than d_{hk3}. The difference between strategies of the centroid and group average resides entirely in the term in d_{ij}. We can therefore suggest that, in place of the usual Euclidean geometric explanation, the loss of monotonicity of the centroid strategy is due to its range-contracting properties.

No simple universal statement can be made for d_{hk4}. The relative performance of this strategy is clearly a function of the value of β adopted for the classification, and of the values of n_i, n_j and n_h in force at a particular fusion. Some observations on particular cases, however, suggest that under certain configurations of data, and for part of the fusion cycle, the flexible strategy can perform similarly to the centroid or group average. For example, it will perform similarly to group average 2 when $\beta = 0$; for inspection of $(d_{hk4}-d_{hk2})$ shows that the difference between strategies will then be small if (n_i-n_j) and $(\Delta_i-\Delta_j)$ are both small. Particular interest resides in the comparison between ISS and flexible with $\beta = -0.25$. If we write $\Delta_i = \Delta_j = \Delta$, the quantity $(d_{hk4}-d_{hk3})$ then reduces to $\Delta\{0.25-n_h/(n_h+n_k)\}$, which vanishes when $n_k = 3n_h$. Flexible, with $\beta = -0.25$, thus becomes more space-dilating when $n_k > 3n_h$.

5. DISCUSSION

We return to the two aspects of selection of a strategy from the four considered to match a user's requirements for an analysis.

5.1 Compatibility of measures and strategies

As noted earlier, the literature has largely tended to argue that strategies should be applied only where the dissimilarity measure used is compatible in the sense that 'measures calculated later in the analysis are exactly of the same kind as the initial inter-element measures; they have the same dimensions (if any), are subject to the same constraints, and can be illustrated by an exactly comparable model'.[8] Incompatible strategies are then seen as undesirable through difficulties in interpretation of group–element or group–group measures. The flexible strategy is an anomaly: while inspired by the general form of the other combinatorial strategies, there is no explicit statement of a geometric model. This has not precluded its use, and it has been recommended where space contraction or dilatation is required when a user has non-metric measures other than D^2.

Clearly the combinatorial formulations of the other strategies can be viewed in the same sense as algebraic formulations of strategies with the fixed values of α_i, α_j and β of the flexible strategy replaced by terms in the n_i, n_j and n_h applying at the time of the fusion. We observe that the analyses presented here have not made use of the properties of metrics or D^2, so that the space-distortion characteristics would remain applicable for all dissimilarity measures. The ISS strategy would then provide more predictable performance when size dilatation is desired than would the flexible strategy with $\beta = -0.25$. Similarly, if range-contraction is desired, then the centroid strategy appears more desirable than the flexible strategy.

Such a usage with measures other than metrics or D^2 does require the loss of a local criterion able to be interpreted geometrically, with the formulation now able to be viewed only as a mechanism to define a between-group or group–element dissimilarity measure from the group's two component group parameters. Depending on the philosophical attitude of the pattern analyst, there is a requirement either to judge the form of the derived measure as meaningful in some way or to treat the classification with caution and to refer to external data to assess its merits. Minimally, however, use of the strategies with other than the measures considered in their original formulation can be seen as a device to apply the concepts of the flexible strategy while avoiding the inherent problems in striking arbitrary weights on the constitutent dissimilarity measures.

5.2 Space-distortion as a guide to strategy selection

The centroid strategy appears to be mainly of historical interest. It was at one time regarded as the only extant space-conserving strategy although it has been found here to be subject to range distortion. However, it is now seldom used because of the inconvenience of its not infrequent reversals and possibly as it has been treated as incompatible with dissimilarity measures other than Euclidean distance or D^2.[8, 13, 14]

The group-average strategy has been surmised to be 'substantially' space-conserving although it has been taken as sensitive to the within-group variances.[18] While it has been stigmatised as a 'not very attractive strategy',[18] in our recent experience it is being increasingly used by biologists who wish to avoid space dilatation. Lance and Williams[8] suggest that it can appropriately be used for all dissimilarity measures, 'providing the concept of an average measure is acceptable'. The analysis here confirms it as both size- and range-conserving.

The incremental sum-of-squares strategy is probably now the most widely used strategy when some degree of dilatation is demanded. The analyses of this paper confirm its size dilatation while showing it to be range-contracting.

The flexible strategy should now be viewed as mainly of theoretical interest. When β is positive, it is space-contracting; but on the rare occasions when biologists have required such a strategy they have normally preferred to use the older 'nearest-neighbour' strategy, which has well-known mathematical advantages.[7] When β is zero it is space-conserving, but though preferred for this purpose by Williams *et al.*[18] it has not ousted the group average strategy. When β is negative it is space-dilating; and its main interest in the past has rested in the fact that it has been considered the only appropriate strategy when a user required to use a space-dilating strategy and a non-metric dissimilarity measure.

It can of course be argued that space dilatation is inherently advantageous in tending to yield homogeneous and sharply defined groups, with the groups of outliers usually able to be readily recognised as such. We submit, however, that it is the user's prerogative to select space dilatation or conservation as most appropriate to his purposes or even his general philosophy. Our investigation should be interpreted as providing a sounder basis for identifying the most appropriate strategy once the user has made his choice.

In summary, two strategies have been shown to be of outstanding value. The group average strategy, in being free from all forms of space distortion, is clearly the ideal space-conserving strategy and is preferable to the less predictable flexible strategy with β of 0. The incremental sum-of-squares strategy remains a dilating strategy and again is to be preferred to the flexible strategy with negative β.

REFERENCES

1. J. R. Bray and J. T. Curtis, An ordination of upland forest communities, *Ecological Monographs* **27**, 325–349 (1957).
2. E. J. Burr, Cluster sorting with mixed character types. II. Fusion strategies, *Australian Computer Journal* **2**, 98–103 (1970).
3. H. T. Clifford and W. Stephenson, *An Introduction to Numerical Classification*. Academic Press, New York (1975).
4. H. T. Clifford and W. T. Williams, Classificatory dendrograms and their interpretation, *Australian Journal of Botany* **21**, 151–162 (1973).
5. J. C. Gower, A general coefficient of similarity and some of its properties, *Biometrics* **27**, 857–871 (1971).
6. P. Jaccard, Nouvelles recherches sur la distribution florale, *Bulletin de la Société vaudoise des sciences naturelles* **44**, 223–270 (1908).
7. N. Jardine and R. Sibson, The construction of hierarchic and non-hierarchic classifications, *The Computer Journal* **11**, 1177–1184 (1958).
8. G. N. Lance and W. T. Williams, A general theory of classificatory sorting strategies. I. Hierarchical systems, *The Computer Journal* **9**, 373–380 (1967*a*).
9. G. N. Lance and W. T. Williams, Mixed-data classificatory programs. I. Agglomerative systems, *Australian computer Journal* **1**, 15–20 (1967*b*).
10. R. Mojena, Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal* **20**, 359–363 (1977).
11. H. Muller-Merbach, Modelling techniques for combinatorial problems. In *Combinatorial Programming: Methods and Applications*, edited B. Roy. Proceedings of the NATO Advanced Study Institute (1975).
12. N. Revelante, W. T. Williams and J. S. Bunt, Temporal and spatial distributions of diatoms, dinoflagellates and trichodesium in the waters of the Great Barrier Reef, *Journal of Experimental Marine Biology and Ecology* **63**, 27–45 (1983).
13. D. Ross, *TAXON Users' Manual*. Canberra: CSIRO Division of computing Research, edition P3 (1982).
14. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francisco (1973).
15. J. A. Taylor, Merino sheep and the intra-paddock pattern of herbaceous species on the Northern Tablelands of New South Wales, Australia, Ph.D. Thesis, University of New England (1980).
16. J. H. Ward, Hierarchical grouping to optimise an objective function, *Journal of the American Statistical Association*, **58**, 236–244 (1963).
17. D. Wishart, An algorithm for hierarchical classifications, *Biometrics* **25**, 165–170 (1969).
18. W. T. Williams, H. T. Clifford and G. N. Lance, Group-size dependence: a rationale for choice between numerical classifications, *The Computer Journal* **14**, 157–161 (1971).