On the determination of optimal levels in phytosociological classification*,**,***

J. Popma1, L. Mucina2, O. van Tongeren1 & E. van der Maarel1
1Division of Geobotany, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
2Permanent address: Department of Geobotany, Institute of Experimental Biology and Ecology of S.A.S., Sienkiewiczova 1, 814 34 Bratislava, Czechoslovakia
3Institute of Ecological Botany, University of Uppsala, 75122 Uppsala, Sweden

Keywords: Arrhenatheretum, Classification, Clustering, Mesobromion, Optimal level, Phytosociological table, Spartina communities, Stopping rule

Abstract

A method is introduced to compare results of a clustering technique at different levels of abstraction, or of different clustering techniques. The method emphasizes within cluster homogeneity as well as discontinuities between clusters. It has been derived from Hogeweg’s method with some important changes. First each cluster is handled separately to determine the ratio between homogeneity and similarity to the nearest neighbour cluster. For a given clustering a weighted average value is computed over all clusters. This average value is standardized using an expected average value for a cluster configuration with the same number of clusters having the same sizes. A low level of the ratio between expected and observed values is supposed to indicate an optimal clustering. A derivation of the criterion is given and results from three sets of data with different properties are evaluated.

Introduction

In various numerical clustering techniques, notably association analysis (Williams & Lambert, 1959) and information analysis (Williams, Lambert & Lance, 1966) stopping rules were introduced (Lambert & Williams, 1966; Lance & Williams, 1965). The function of a stopping rule in a divisive technique is to terminate the division process at a certain level if the clusters obtained are sufficiently homogeneous, and/or sufficiently distinct from each other. A well-known stopping rule is the chi square statistic. No further divisions are performed when the chi square value for species associations drops below a selected (significance) level.

In agglomerative strategies stopping rules are seldom used. Here the direct phytosociological or ecological interpretability of the clusters seems to be decisive. Still, optimality criteria or stopping rules would be useful.

There is a clear need to interpret the classification hierarchy in terms of an optimal classification regardless the hierarchical position of the clusters. Various approaches, both formalized and unformalized are available. We do not intend to review them all but refer to the review by Pielou (1977). We may distinguish three types of criteria to obtain a final result:
(1) A priori fixed number of clusters; (2) Ecological interpretation of different results; (3) Pattern of the within and/or between cluster similarities.

The a priori choice of an optimal number of clusters is largely arbitrary, although some relation may exist with the number of relevés treated in the clustering. According to M. B. Dale (pers. comm.)
the intuitive optimal number of clusters is often in the order of the square root of the number of individuals to be classified. This could merely reflect a more or less fixed redundance in (subjective or representative) sampling after reconnaissance in the area under investigation has revealed an idea of the presumable number of types. Van der Maarel (1966), using an information clustering, found an optimum number of types at a level where the average information content of the clusters was ca. 50% of the total information of the relevé set. Optimality in number of clusters could have some relevance in practice, e.g. vegetation mapping.

Optimality in terms of a satisfactory ecological interpretation (cf. Clifford, 1976) should not be regarded as a separate criterion. It is rather a check of any optimality criterion based on the intrinsic structure of the data.

Obviously, optimality in the patterns of within/between cluster homogeneity is the essential feature of any clustering to be tested. This approach is called a 'probabilistic' one (Pielou, 1977), because statistical tests are used. In many cases however, the tests applied are only pseudo- or quasi-statistical. Since phytosociological samples are usually preferential (Orlóci, 1978) and not random, optimal levels cannot be determined in any statistical way. Still it may be useful to find optimal levels in representative sampling based clusterings. In the following survey we will not consider statistical significance.

There are three possible ways of using homogeneity (heterogeneity) measures: (1) within-group homogeneity only, (2) between-group heterogeneity only, and (3) a combination of both. Within-group homogeneity will usually, though not necessarily, depend on the clustering technique used (e.g. complete versus single linkage clustering). With a sum of squares method (Ward, 1963; Orlóci, 1967) within-group dispersion provides a measure of heterogeneity (Pielou, 1977). Our preference is to use a combined measure.

Another choice concerns the way the dendrogram (in case of hierarchical clustering) is interpreted. First, it is possible to consider each branch of a dichotomy separately. Hill (1980a; 1980b) used the criterion of Ratkowsky and Lance (1978) which is based on the ratio of between-group sum of squares and total sum of squares for each pair of clusters obtained from any division. As Hill (1980b) himself remarked, a disadvantage of this method is that small clusters with a low number of attributes may easily remain separated and therefore the method results in the determination of a maximum number of clusters rather than an optimal one. Second, it is possible to compare clusters having the same dichotomy rank in the hierarchy, but not necessarily the same level of homogeneity. In these ways the cutting lines in the dendrogram can run through different levels. An example is the method based upon information statistic of Lausi & Feoli (1979; Feoli & Lausi 1980). Third, clusters may be compared at a given level of homogeneity or cluster-similarity. In this case the cutting lines, 'phenon lines' (sensu Sneath & Sokal, 1973), are straight. The use of phenon lines in space dilating clustering strategies (cf. Lance & Williams, 1965) was criticized by Clifford & Williams (1973) and Hill (1980a, b). The main objection here concerns the dependence of the clustering technique involved on the number of objects in a given cluster. We consider this irrelevant, because in evaluating a certain level, one can apply weighting to equalize the sizes. Furthermore it seems correct to choose an optimal clustering from a given set of solutions irrespective of the technique used. Drawbacks of the clustering technique are of no importance for the practical evaluation of its results. If a clustering technique fails in elucidating data structure, an optimal solution resulting from this technique will be inferior to another solution resulting from a more suitable technique.

A frequently used phenon line approach is Orlóci's (1967) 'classification efficiency', using $Q_w$ (within), $Q_b$ (between) and $Q_t$ (total sum of squares), with three options: $Q_w/Q_t$ (Orlóci, 1967), $Q_b/Q_t$ Stanek, 1973; Orlóci, 1978) and $Q_b/Q_w$ (Orlóci & Stanek, 1980). The combination of within- and between-cluster homogeneity measures occurs also in the approach of Hogeweg (1976). The special feature of this method is that within-cluster homogeneity ('ultrametric distance' sensu Hogeweg, 1976) is compared to the mean of the cluster homogeneities of each cluster to its nearest cluster. The criterion reads:

$$H = \frac{\sum I_i}{\sum U_i}$$

where $I_i$ is the between cluster homogeneity between cluster $i$ and the cluster to which it is joined first and $U_i$ is the within cluster homogeneity of