Pages 125-132 in C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock & Y. Baba (eds.), *Data Science, Classification, and Related Methods.* Springer, Tokyo. 1998.

Explanatory Variables in Classifications and the Detection of the Optimum Number of Clusters

János Podani

Department of Plant Taxonomy and Ecology
Loránd Eötvös University, Ludovika tér 2
H-1083 Budapest, Hungary
Fax: +36 1 1338 764. Email: PODANI@LUDENS.ELTE.HU

Summary: An ordinal approach to the *a posteriori* evaluation of the explanatory power of variables in classifications is proposed. The contribution of each variable is assessed in a way fully compatible with the distance or dissimilarity function used in the clustering process. Then, a simple ranking-based measure is applied to express the relative agreement or disagreement of variables with a given partition. This measure treats all variables equally, no matter how influential they were when the classification was actually created. The sum of measures for all variables reflects their overall agreement and can be used to select an optimal partition from a hierarchical classification.

1. Introduction

An integral part of the interpretation of clustering results is to evaluate how the individual variables explain the classes. Finding an order of importance of variables for an existing classification is often called *a posteriori* **feature selection** (cf. Dale et al. 1986), as opposed to *a priori* feature selection, when the variables are ranked before the analysis starts (e.g., Orlóci 1973, Stephenson and Cook 1980), and to *forward* selection, in which evaluation of variables is part of the algorithm (e.g., Jancey and Wells 1987, Fowlkes et al. 1988).

Attention in *a posteriori* feature selection may be focused on two fundamental aspects of classification: cluster **cohesion** and **separation** (*sensu* Gordon 1981). The analysis can be restricted to either of these aspects (e.g., to contributions to within-cluster sum of squares only). Alternatively, the effect of variables on the distinction between clusters as well as on the internal "homogeneity" of clusters is simultaneously incorporated in the study, even if the clustering method did not actually consider both. A simple possibility which comes to the mind first is to compute for each variable the ratio of within-group and between-group sum of squares as an index of explanatory power.

It is emphasized, however, that there is no point in examining cluster cohesion and separation in terms of sum of squares when, say, the starting matrix contained chord distances or percentage dissimilarity values and the algorithm was single or complete linkage sorting. In other words, the evaluation procedure has to be **compatible** with the distance coefficient used in creating the classification. Since in many fields of science, e.g., in biological taxonomy, relatively few classifications are based on sum of squares or variance, and often the clustering models are not even Euclidean, a more generally applicable, yet flexible, criterion is required. Godehardt's (1990) multigraph approach, in which each variable is treated independently, seems to satisfy this requirement.

The third point emphasized here is that the importance of variables may be judged in two

ways. The more obvious one is the measurement of the **absolute effect** of each variable upon the creation of clusters. For example, in case of Euclidean distance and centroid clustering, we can examine how far apart the cluster centroids are for each variable, and then order the variables on this basis. This ordering will emphasize variables that dominated the classification process, and may neglect others that are equally if not more interesting for the *a posteriori* interpretation of clusters. In fact, any variable supporting the given partition may prove useful in subsequent descriptions, no matter how small this support is in absolute terms. Thus, an alternative procedure free from the implicit variable weighting, i.e., measurement of the **relative importance** of the variable may prove useful. Lance and Williams (1977) are early proponents of this approach, by suggesting taking the ratio of between-cluster sum of squares to the total for each continuous variable or to compute Cramer's index (see also Anderberg 1973) for each binary or multistate variable. These criteria are thus only data-type-dependent and do not consider the manner in which the dissimilarities were calculated. The procedure described in this paper releases implicit weighting by introducing an **ordinal measure of the explanatory power** of variables in non-hierarchical classifications. This measure also satisfies the requirement of being compatible with the dissimilarities used and relies equally on both cluster separation and cohesion.

I will also provide an alternative approach to the familiar problem of detecting the optimum number of clusters. The sum of measures of explanatory power for all variables will be defined as an **overall measure of the agreement** (in a sense: consensus) among the variables regarding the partition of m objects into t clusters. Plotting the sum over a reasonable range of t values provides a graphical means to find the optimum, if any. This approach, as will be seen, is radically different from most of the methods reviewed and compared by Milligan and Cooper (1985).

2. Variable contributions

The procedure starts with evaluating the contribution of each variable to the distances or dissimilarities between objects. To ensure compatibility, the determination of this contribution must be specific to the distance or dissimilarity function used. As an example, the total contribution of variable i to all the $z=m(m-1)/2$ values in the lower semimatrix of $\mathbf{D}^2$ containing the squared Euclidean distances for m objects is computed as

$$\Phi_i = \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} g_{ijk} ,$$

where $g_{ijk} = (x_{ij} - x_{ik})^2$ is the contribution of variable i to d^2_{jk} and is written as an element of matrix $\mathbf{G}_i$. The contributions are strictly additive, the matrix of squared distances is therefore reproduced as

$$\mathbf{D}^2 = \sum_{i=1}^{n} \mathbf{G}_i ,$$

with n as the number of variables. Formulae for computing contributions have been derived and are presented without proofs for 15 other distance and dissimilarity measures (Table 1). The measures themselves are not shown here, because most of them are well-known from the clustering literature. A full list is found in Podani (1994), although the

Tab. 1: Contribution of variable i to d_{jk} for several, well-known dissimilarity and distance functions. For presence/absence coefficients we assume that x_{ij}=1 for presence) and x_{ij}=0 for absence. Contributions are ranked in ascending order for most measures, except for those marked with an *, for which ranking is the reverse (see text).

Euclidean distance	$(x_{ij} - x_{ik})^2$
Manhattan distance, 1-simple match. coeff.	$\lvert x_{ij} - x_{ik} \rvert$
Penrose SIZE	$x_{ij} - x_{ik}$
Chord distance *	$\dfrac{x_{ij}\, x_{ik}}{\sqrt{\sum_{h=1}^{n} x_{hj}^2 \sum_{h=1}^{n} x_{hk}^2}}$
Canberra metric	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\lvert x_{ij} \rvert + \lvert x_{ik} \rvert}$
Percentage difference, 1-Sorensen	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} x_{hj} + x_{hk}}$
1-Ruzicka, 1-Jaccard	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} \max[x_{hj}, x_{hk}]}$
1-Similarity ratio *	$\dfrac{x_{ij}\, x_{ik}}{\sum_h x_{hj}^2 + \sum_h x_{hk}^2 + \sum_h x_{hj} x_{hk}}$
1-Russell - Rao	$(1 - x_{ij} x_{ik})/n$
1-Rogers - Tanimoto	$\dfrac{2 \lvert x_{ij} - x_{ik} \rvert}{n + \sum_{h=1}^{n} 2 \lvert x_{hj} - x_{hk} \rvert}$
1-Sokal - Sneath	$\dfrac{2 \lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} \max[x_{hj}, x_{hk}] + \lvert x_{hj} - x_{hk} \rvert}$
1-Anderberg 1 *	$\dfrac{x_{ij} x_{ik}}{\sum_h x_{hj}} \cdot \dfrac{x_{ij} x_{ik}}{\sum_h x_{hk}} \cdot \dfrac{(1-x_{ij})(1-x_{ik})}{n - \sum_h x_{hj}} \cdot \dfrac{(1-x_{ij})(1-x_{ik})}{n - \sum_h x_{hk}}$
1-Kulczynski *	$\dfrac{\min[x_{ij}, x_{ik}]}{\sum_h x_{hj}} + \dfrac{\min[x_{ij}, x_{ik}]}{\sum_h x_{hk}}$

reader may also consult Anderberg (1973), Sneath and Sokal (1973) and Orlóci (1978). Note that some indices known generally as similarity functions are expressed as complements.

3. A new measure of explanatory power

The second part of the analysis involves determining the the **rank order** of the g_{ijk} scores for each variable i for a given partition P_t of m objects into t clusters, each with m_s objects, s=1,...,t. For most coefficients of distance, e.g., Manhattan and Euclidean distances, it is reasonable to state that a variable completely explains P_t if **all of its within-cluster contributions are smaller than the between-cluster contributions**, and have therefore the smallest ranks. (For functions with an asterisk in Table 1 the situation is the reverse, however. In these cases, ranking is done in descending order to keep the generality of the statement that within-cluster contributions should be ranked first in the optimal case. In the sequel, we assume the previous type to simplify discussion.) Consequently, we have the **minimum sum of ranks of within-cluster contributions** of any variable, denoted by $R_{(t)min}$. This quantity is obtained as

$$R_{(t)min} = (q^2+q)/2 \quad \text{where } q = \sum_{s=1}^{t} \binom{m_s}{2}.$$

Let, further, $R_{(i,t)obs}$ be an **observed sum of ranks of within-cluster contributions** for variable i. Clearly, $R_{(t)min} \leq R_{(i,t)obs}$. Ties in the rank order can be resolved randomly which has negligible effects for large values of z. The sum of ranks for between-cluster contributions will not be used, because it conveys no extra information.

Let $R_{(t)exp}$ denote the **random expectation** for the null situation, i.e., when the variable does not make distinction as to whether a contribution is within- or between clusters (indifferent variable). In other words, contributions are arranged at random with the expected sum of within-cluster ranks given by

$$R_{(t)exp} = p\,(z^2+z)/2$$

where $p=q/z$ is the probability that a randomly chosen value in the rank order is a within-cluster contribution. The expectation will be used below as a reference basis for constructing the formula.

Then, the **explanatory power** of the variable is defined as the complement of the deviation of the actual sum of ranks from the minimum as divided by the deviation of the expectation from the minimum:

$$r_{(i,t)} = 1.0 - \frac{R(i,t)obs - R(t)min}{R(t)exp - R(t)min}$$

$r_{(i,t)}$ values close to 1.0 indicate high explanatory power, values around zero reflect indifference, whereas negative values correspond to a situation when variable i is contradictory with P_t. (I deliberately avoid using the term discriminatory power, because it is usually coined with variance-related concepts as in discriminant analysis.) The variables may be ordered based on their r scores, to facilitate interpretation of clusters and to detect variables which happen to be indifferent or even contradictory with the given partition.

4. Detecting the optimum number of clusters

The coefficient of explanatory power can be used in turn to determine the optimum number of clusters in hierarchical classifications. The intuitive basis for this is that the more variables support a given partition the more acceptable it is. The criterion to be used is defined as the sum of coefficients of explanatory power,

$$\sigma_t = \sum_{i=1}^{n} r_{(i,t)}$$

which will be called the **coefficent of cluster separation**. The upper bound of this coefficient is n, reached in the unanimous situation with all variables fully explaining the partition. The σ_t coefficient is computed for each level of interest in the hierarchy and the results are plotted against t. For data sets with group structure, the curve shows a peaked effect allowing to detect the number of clusters at which the **majority of variables support the same clustering** in terms of their ranked within-cluster contribution scores. For very many clusters, each with 1-2 objects only, the increase of σ_t is a necessity, but such trivial clusters attract no interest anyway (these clusters are usually excluded from such studies, cf. Milligan and Cooper 1985). Absence of clear-cut peaks is indicative of either strong disagreements between variables as to the "optimum" value of t, or complete lack of group structure, so the $r_{(i,t)}$ values must be inspected.

Computer program SYN-TAX 5.02 (Podani 1994) designed for classification purposes includes an option for computing the explanatory power of variables and the coefficient of cluster separation at several levels in a dendrogram and for plotting the graph automatically (available for PCs and Macintosh computers).

5. Example

The method will be demonstrated by an actual example coming from community ecology. A total of 80 vegetational plots (objects) represent a sample of dolomite grasslands of Sas-hill, Budapest, Hungary (for more details, see Podani 1985). The plots have been described in terms of percentage cover scores of 123 vascular plant species. For the purpose of illustration, the matrix of Euclidean distances of objects was subjected to complete linkage clustering (Fig. 1). The explanatory power of variables and the coefficient of cluster separation were computed for the top ten cut levels in the dendrogram, i.e., for t=2 to 11. The plot of cluster separation against the number of clusters (Fig. 2) indicates high agreement of variables for two and three clusters, with the maximum at t=3. When t is raised from 3 to 4, the coefficient drops by more than 50% and for more groups it remains about the same. The analysis thus suggests that the given classification is best supported by the species at the 3-cluster level. It is therefore worthwhile to examine the rank order of variables based on their explanatory power values for t=3 (Tab. 2). To save space, the table lists only the first ten and the last ten species from the rank order. Those at the beginning of the list are the best indicators of difference between closed (the two smaller groups) and open grasslands (the large group), whereas species with negative scores counter-support this classification because they tend to differentiate the large group even further. These species have been widely used as discriminatory species to subdivide the relatively open communities. The analysis revealed, however, that the majority of species are contradictory with this, showing that the classical syntaxonomic classification was subjectively based on a narrow subset of species.

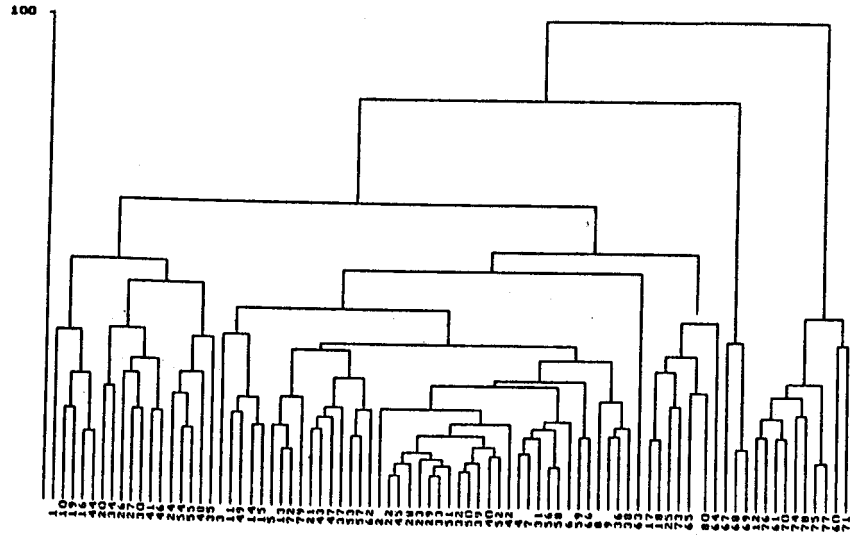

Fig. 1: Dendrogram showing complete linkage clustering of 80 vegetational plots, based on the Euclidean distances among objects using percentage cover scores of 123 species.

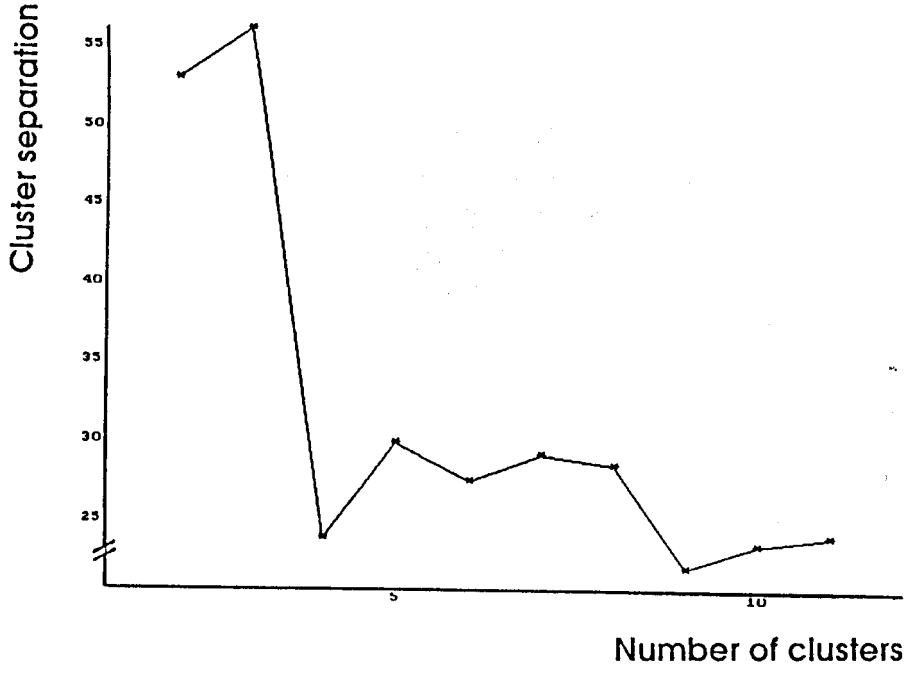

Fig. 2: Plot showing the relationship between the number of clusters and the coefficient of cluster separation for the top ten partitions obtained from the dendrogram of Fig. 1.

Tab. 2: The first ten and the last ten species in the rank order of variables for the three-cluster partition obtained from the dendrogram in Fig. 1.

Rank	Species	r_i	Rank	Species	r_i
1	Cytisus hirsutus	.865	114	Dianthus serotinus	.057
2	Festuca sulcata	.835	115	Andropogon ischaemum	.050
3	Bupleurum falcatum	.803	116	Helianthemum canum	.047
4	Pimpinella saxifraga	.771	117	Thymus praecox	.024
5	Asyneuma canescens	.763	118	Sanguisorba minor	.005
6	Veronica spicata	.758	119	Stipa eriocaulis	-.004
7	Polygonatum odoratum	.757	120	Festuca pallens	-.051
8	Campanula sibirica	.725	121	Carex liparocarpos	-.053
9	Carlina intermedia	.719	122	Chrysopogon gryllus	-.162
10	Adonis vernalis	.717	123	Seseli leucospermum	-.177

6. Discussion

The measure of explanatory power proposed in this paper is a non-metric criterion because actual differences are irrelevant: it is the rank order of contributions that matters. Thus, even if a variable had negligible effects on the distances (because of lack of commensurability, for example), it may turn out to be a good explanatory variable afterwards. Also, ranking variables based on the r values reveals an aspect rarely emphasized: the identification of variables that do not agree with the partition. Finding these variables may lead to revisions of former classifications. The possibility is also raised here that after the removal of these variables a repeated classification based on the reduced set of variables may provide a more noise-free classification. This is certainly an aspect which merits future investigations.

The coefficient of cluster separation, being based on ranked contributions, is considerably different from the currently known indices of optimum number of clusters as reviewed by Milligan and Cooper (1985). In addition to the ranking technique, the most substantial difference is that whereas the other methods are less dependent on the number of variables (so that they can be best demonstrated with a two-dimensional example), the present technique is more meaningful when there are quite a few variables. Therefore, it may perform very poorly in a low dimensional situation if compared to the other methods, and evaluation of the method proposed here along the lines of Milligan and Cooper's study would be irrelevant. The only exception seems to be the Ratkowsky and Lance (1978) criterion, which involves computation of the ratio used by Lance and Williams (1977) for each variable, and takes the average over variables. It is perhaps an explanation for the fairly poor performance of this measure in the two-dimensional case of Milligan and Cooper's study, though Ratkowsky and Lance reported high success, usually with many dimensions. It is also noted here that whereas almost all methods provide the same result after rigid rotation of the axes (rotation invariance) the Ratkowsky and Lance criterion and the one suggested in this paper are exceptions.

As with other formulae for detecting the optimum number of clusters in hierarchical classifications, the possibility to incorporate the measure directly as a clustering criterion may be examined. The coefficient of cluster separation is computationally very

demanding, however. (The actual example presented in this paper took ten hours on a PC 486.) Building clusters based on a global nonmetric criterion similar to σ_t will provide a clustering procedure completely compatible with the optimality measure.

Acknowledgements:

The author expresses his sincerest thanks for receiving an OMFB Travel Grant (No. MEC 96-0176) and an OTKA Travel Grant. (No. U21456) to participate at IFCS'96, Kobe, where this contribution was presented. This study was funded by the OTKA Hungarian National Research Grant No. T19364. I am grateful to A. D. Gordon (University of St. Andrews, U.K.) for his comments on the manuscript, and to M. B. Dale (CSIRO, Australia) and Sz. Bokros (ELTE, Budapest) for discussions.

References:

Anderberg, M. R. (1973): *Cluster Analysis for Applications*. Academic, New York.

Dale, M. B., Beatrice, M., Venanzoni, R. and Ferrari, C. (1986): A comparison of some methods of selecting species in vegetation analysis. *Coenoses*, **1**, 35-52.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1988): Variable selection in clustering. *Journal of Classification*, **5**, 205-228.

Godehardt, E. (1990): *Graphs as Structural Models: The Application of Graphs and Multigraphs in Cluster Analysis* (2nd ed.). Vieweg & Sohn, Braunschweig.

Gordon, A. D. (1981): *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, London.

Jancey, R. C. and Wells, T. C. (1987): Locality theory: the phenomenon and its significance. *Coenoses*, **2**, 31-37.

Lance, G. N. and Williams, W. T. (1977): Attribute contributions to a classification. *Australian Computer Journal*, **9**, 128-129.

Milligan, G. W. and Cooper, M. C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159-179.

Orlóci, L. (1973): Ranking characters by a dispersion criterion. *Nature*, **244**, 371-373.

Orlóci, L. (1978): *Multivariate Analysis in Vegetation Research*. Junk, The Hague.

Podani, J. (1985): Syntaxonomic congruence in a small-scale vegetation survey. *Abstracta Botanica*, **9**, 99-128.

Podani, J. (1994): *Multivariate Data Analysis in Ecology and Systematics*. SPB Publishing, The Hague.

Ratkowsky, D. A. and Lance, G. N. (1978): A criterion for determining the number of groups in a classification. *Australian Computer Journal*, **10**, 115-117.

Sneath, P.H.A. and Sokal, R. R. (1973): *Numerical Taxonomy*. Freeman, San Francisco.

Stephenson, W. and Cook, S. D. (1980): Elimination of species before cluster analysis. *Australian Journal of Ecology*, **5**, 263-273.