PBDB Intensive Summer Course 2007
Paleoecology Section – Thomas Olszewski

Cluster Analysis

Introduction

Purpose – classify either samples or species using *explicit* criteria (as distinct from "objective" criteria) into a smaller number of interpretable categories (i.e., *clusters*)

Dendrogram – a branching diagram that hierarchically nests objects into increasingly more inclusive groups; degree of similarity is depicted by length of branch; ordering axis is to prevent branches from crossing but is otherwise arbitrary

Divisive versus Agglomerative and Hierarchical versus Non-hierarchical Methods – The classical approach to clustering in paleoecology is agglomerative and hierarchical (historically, this reflects the development of these methods from the phenetic school of taxonomy); modern clustering techniques for very large data sets (e.g., genetic networks) are based on information theory and are typically neither hierarchical (i.e., cannot be depicted as dendrograms) nor agglomerative; rather, they are based on network linkage patterns

Steps in Clustering

1) SEARCH. Start with a similarity matrix (ALL clustering methods start at this point); find the cell with the highest similarity value and link that pair of objects (i.e., form a cluster); note that if there is more than one cell with equal similarity, link the first pair found (IMPORTANT: this means that the order of objects in the matrix can influence the outcome, particularly for large data sets with lots of redundancy!)
2) REDUCE. Recalculate similarities, treating the clusters as new objects; how clusters are treated differs among different algorithms
3) REPEAT until all objects are related to one another

Linking Algorithms

1) *Nearest Neighbor or Single Linkage Clustering*. Similarity between two clusters equals the maximum similarity between any two members of the clusters:

   $S_{(AB),C} = \max(S_{AB}, S_{AC})$ $S_{(AB),(CD)} = \max(S_{AC}, S_{AD}, S_{BC}, S_{CD})$

   Note. This algorithm tends to produce "chaining" (i.e., apparent addition of each object in a dendrogram, one by one); normally, this would be an indicator of gradient structure in the data, but single linkage clustering can produce it artifactually.

2) *Farthest Neighbor or Complete Linkage Clustering*. Similarity between two clusters equals the minimum similarity between any two members of the clusters:

   $S_{(AB),C} = \min(S_{AB}, S_{AC})$ $S_{(AB),(CD)} = \min(S_{AC}, S_{AD}, S_{BC}, S_{CD})$

   Note. Use this rule to recalculate similarity values in the matrix (in the REDUCING step), but continue to SEARCH for the greatest similarity values. This algorithm tends to produce very clear groups – i.e., by using minimum values, it tends to underestimate similarity between recognized clusters.

3) *Unweighted Group Average or Unweighted Pair Group Method with Arithmetic Averaging (UPGMA)*. Similarity between two clusters equals the mean similarity between all possible pair-group combinations:

$S_{(AB),C} = (S_{AC} + S_{BC})/2$
$S_{(AB),(CD)} = (S_{AC} + S_{AD} + S_{BC} + S_{BD})/4$
$S_{E,(C,(AB))} = (S_{AE} + S_{BE} + S_{CE})/3$

Note. The most commonly applied technique seen in paleoecology; degree of clustering is intermediate between single and complete linkage.

4) *Weighted Group Average or Weighted Pair Group Method with Arithmetic Averaging (WPGMA)*. Similarity between two clusters equals the mean similarity of previously existing clusters when they are grouped (average always involves only 2 terms and does not weight clusters by their size):

$S_{(AB),C} = (S_{AC} + S_{BC})/2$
$S_{(AB),(CD)} = [½(S_{AC} + S_{AD}) + ½(S_{BC} + S_{BD})]/2 = [S_{A,(CD)} + S_{B,(CD)}]/2$
$S_{E,(C,(AB))} = [½(S_{AE} + S_{BE}) + S_{CE})]/2 = (S_{E,(AB)} + S_{CE})/2$

Note. The first two cases are identical to UPGMA, but the third downweights the earlier cluster (AB).

5) *Centroid Clustering or Unweighted Pair Group Method with Centroid Averaging (UPGMC)*. Similarity between two clusters equals their similarities as composite objects (i.e., the sums of all their component samples):

$$x_{(AB),Ci} = \frac{x_{Ai} + x_{Bi} + x_{Ci}}{3}$$

Note. If collections composing a cluster are a cloud of points in a multivariate space, the cluster's similarity to others is represented by a point at the center of the cloud. Similarity should conform to *triangular inequality* to work properly (see note for WPGMC below).

6) *Median Clustering or Weighted Pair Group Method with Centroid Averaging (WPGMC).* Similarity between two clusters equals their similarities as composite objects, but determining their composite composition using only the last two objects (samples or clusters) to be joined in each cluster:

$$x_{(AB),Ci} = \frac{\frac{(x_{Ai+} + x_{Bi})}{2} + x_{Ci}}{2} = \frac{(x_{Ai+} + x_{Bi})}{4} + \frac{x_{Ci}}{2}$$

Note. This method downweights the first objects to be joined. With median or centroid clustering, if the similarity metric does not conform to the triangular inequality, it is possible to get non-monotonic clustering – i.e., two large clusters may be more similar to each other than components within either cluster.

7) *Ward's Method or Minimum Variance Clustering or Orloci's Sum of Squares*. Similarity is calculated as in UPGMA or UPGMC but clusters are created differently – in this algorithm, objects are grouped that minimize the variance of the similarities between all member objects within a cluster:

3 clusters: ABC, DE, F; which two should be clustered next? Pick the two that produce the minimum variance ($\overline{S}$ = mean of all S's in that cluster):

$$V_{(ABC)(DE)} = \frac{(S_{AD} - \overline{S})^2 + (S_{AE} - \overline{S})^2 + (S_{BD} - \overline{S})^2 + (S_{BE} - \overline{S})^2 + (S_{CD} - \overline{S})^2 + (S_{CE} - \overline{S})^2}{6}$$

$$V_{(ABC)F} = \frac{(S_{AF} - \overline{S})^2 + (S_{BF} - \overline{S})^2 + (S_{CF} - \overline{S})^2}{3}$$

$$V_{(DE)F} = \frac{(S_{DF} - \overline{S})^2 + (S_{EF} - \overline{S})^2}{2}$$

Note. This method also requires a metric that adheres to the triangular inequality. This method has seen relatively little application in paleoecology.

Interpretation of Common Dendrogram Patterns

Chaining – if not artifactual, suggests presence of a continuous gradient; can reflect either the algorithm used or a gradient in sampling intensity

Defining clusters

1) "phenon" level
2) relative homogeneity
3) relative stem length
4) visually evident breaks

Two-way Cluster Analysis

Use ordering axis of Q-mode (samples) and R-mode (taxa) analyses of the same data to reorganize the rows and columns of the data matrix – a very effective way of understanding which taxa are related to which cluster of samples. Remember that the purpose of the ordering axis in cluster analysis is to keep branches from crossing, not to show a statistically justified gradient, so links can be "reflected" and "spun". In this sense, two-way cluster analysis is arbitrary, but it is nevertheless a very effective way of relating the R and Q modes and incorporating data into a figure.

Cophenetic or Matrix Correlation

Cluster analysis inherently discards some information from the original similarity matrix, which lead to the question of how well a dendrogram represents the original matrix?

Each pair of objects has a similarity value in the original similarity matrix and a depicted similarity in the dendrogram (the linkage level of the pair). For each pair, these values can be cross-plotted, with dendrogram similarity on one axis and measured similarity on the other. The degree of correlation between the two (quantified using r) is a measure of how well the clustered pattern retained the underlying information.

r_{CS} shows that cluster analysis generally is good at linking very similar objects but loses its ability to accurately depict patterns at lower levels of similarity

Limitations of Cluster Analysis

1) imposes hierarchical structure on data, whether real or not
2) it does not depict data with multiple, independent underlying controls well
3) because these are based on algorithms rather than formal mathematics, solutions can be non-unique

References for Cluster Analysis

Girvan, M. and Newman, M.E.J., 2002, Community structure in social and biological networks. Proceedings of the National Academy of Sciences, v. 99, p. 7821-7826.

Hopcroft, J., Khan, O., Kulis, B., and Selman, B., 2004, Tracking evolving communities in large linked networks. Proceedings of the National Academy of Sciences, v. 101, p. 5249-5253.

Morris, S.A. and Yen, G.G., 2004, Crossmaps: visualization of overlapping relationships in collections of journal papers. Proceedings of the National Academy of Sciences, v. 101, p. 5291-5296. (An independent re-invention of 2-way cluster analysis.)

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T., 2005, Uncovering the overlapping community structure of complex networks in nature and society. Nature, v. 435, p. 814-818.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., 2004, Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, v. 101, p. 2658-2663.

Slonim, N., Atwal, G.S., Tkačik, G., and Bialek, W., 2005, Information-based clustering. Proceedings of the National Academy of Sciences, v. 102, p. 18297-18302.

Sneath, P.H.A. and Sokal, R.R., 1973, Numerical Taxonomy: The Principles and Practice of Numerical Classification. W.H. Freeman and Co., 573 p. (A classic – the basis for virtually all commonly used phenetic and paleoecological clustering methods.)