Clustering Algorithms

I. Nearest neighbor (Single linkage):

The distance between two clusters, D(A,B) where A is made up of entities a_i and B of entities b_i is:

$D(A,B) = \min (D(a_i,b_i))$

2. Farthest neighbor (Complete linkage):

$D(A,B) = \max (D(a_i,b_i))$

3. Centroid linkage

$D(A,B) = D(\text{centroid}(A),\text{centroid}(B))$

In a cluster of m entities with S species, the centroid $C = (c_1, c_2, ..., c_S)$ where:

$$c_1 = \frac{1}{m}(x_{11} + x_{12} + \ldots + x_{1m}) = \frac{1}{m}\sum_{j=1}^{m} x_{1j}$$

$$c_2 = \frac{1}{m}(x_{21} + x_{22} + \ldots + x_{2m}) = \frac{1}{m}\sum_{j=1}^{m} x_{2j}$$

$$\vdots$$

$$c_s = \frac{1}{m}(x_{11} + x_{12} + \ldots + x_{Sm}) = \frac{1}{m}\sum_{j=1}^{m} x_{Sj}$$

where x_{ij} = value for ith species and the jth sample.

D^2 is often more convenient measure of distance in centroid clustering where:

$$D^2(C(A),C(B)) = \sum_{i=1}^{S} (c_{iA} - c_{iB})^2$$

This formula requires going back to the original data matrix for each cluster step. Calculation formula is also available which allows recalculation of new distances from existing distance matrix:

$$D^2((A,B),(E)) = \frac{m}{m+n}D^2(A,E) + \frac{n}{m+n}D^2(B,E) - \frac{mn}{m+n}D^2(A,B)$$

where m=# of samples in A and n=# of samples in B.This is called unweighted centroid method or just centroid method.

This formula weights the contribution of each previous group to the new group in proportion to the number of points it contains. To give equal weight to each of the previous groups in deciding the centroid of the new group:

$$D^2((A-B),E) = \frac{1}{2}D^2(A,E) + \frac{1}{2}D^2(B,E) - \frac{1}{4}D^2(A,B)$$

This is called weighted centroid method or median method.

4. Average Distance Linkage

-uses average of interpoint distances:

D(A,B) = avg($D(a_i,b_j)$)

$$D(A,B) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} D(a_i,b_j)$$

Computational formula for this is:

$$D((A-B),E) = \frac{m}{m+n}D(A,E) + \frac{n}{m+n}D(B,E)$$

This is called Group mean or Unweighted Pair Group method. If A and B are given equal weights we get Weighted Pair Group Method:

$$D((A-B),E) = \frac{1}{2}D(A,E) + \frac{1}{2}D(B,E)$$