Notes 06-6: Terminology

For numeric attributes, clusters can be described by several characteristic values. Assume a cluster K_b consisting of n m-dimensional points $\{(p_{11}, p_{12}, \ldots, p_{1m}), (p_{21}, p_{22}, \ldots, p_{2m}), \ldots (p_{n1}, p_{n2}, \ldots, p_{nm})\}$.

The *centroid*, C_a, of a cluster K_a is the "middle" point of the cluster (it need not be an actual point in the cluster) and is described by $C_a = (p_1, p_2, \ldots, p_m)$, where p_u, the u-th attribute of the centroid, is given by

$$p_u = \frac{\sum_{i=1}^{n} p_{iu}}{n}$$

The *radius*, R_a, of a cluster K_a is the square root of the average mean squared distance from all points in the cluster to the centroid, and is given by

$$R_a = \sqrt{\frac{\sum_{i=1}^{n} \sum_{u=1}^{m} |p_{iu} - p_u|^2}{n}}$$

The diameter, $Diameter_a$, of cluster K_a is the square root of the average mean squared distance between all pairs of points in the cluster, and is given by

$$Diameter_a = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{u=1}^{m} |p_{iu} - p_{ju}|^2}{n(n-1)}}$$

Many clustering algorithms require that the *distance between clusters* be determined (as opposed to the *distance between objects*) to identify when two clusters are of sufficient similarity to be linked together (i.e., amalgamated).

The *single linkage* (or *nearest neighbor*) method links clusters when the distance between the two closest objects in the different clusters is below some threshold.

The *complete linkage* (or *furthest neighbor*) method links clusters when the distance between the two furthest objects in the different clusters is below some threshold.

The *pair-group average* method links clusters when the average distance between all pairs of objects in the different clusters is below some threshold.

The *pair-group centroid* method links clusters when the distance between centroids is below some threshold.

The *pair-group medoid* method links clusters when the distance between medoids is below some threshold.