Cluster Analysis

Sayantee Jana

Why cluster analysis ?

- Market Segmentation
- Market Structure Analysis
- Balanced Portfolios
- Industry Analysis

Hierarchical Agglomerative Clustering Algorithm

1. Start with n clusters (each record = cluster).

2. The two closest records are merged into one cluster.

3. At every step, the two clusters with the smallest distance are merged. This

means that either single records are added to existing clusters or two existing

clusters are combined.

Limitations of Hierarchical Clustering

1. Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow.

2. The hierarchical algorithm makes only one pass through the data. This means that records that are allocated incorrectly early in the process can-not be reallocated subsequently.

3. Low stability: Reordering data or dropping a few records can lead to a different solution.

Limitations contd ...

4. With respect to the choice of distance between clusters, single and complete linkage are robust to changes in the distance metric (e.g., Euclidean, statistical distance) as long as the relative ordering is kept. In contrast, average linkage is more influenced by the choice of distance metric, and might lead to completely different clusters when the metric is changed.

5. Sensitive to outliers.

K-MEANS CLUSTERING ALGORITHM

1. Start with k initial clusters (user chooses k).

2. At every step, each record is reassigned to the cluster with the “closest” centroid.

3. Recompute the centroids of clusters that lost or gained a record, and repeat Step 2.

4. Stop at convergence or when moving any more records between clusters increases cluster dispersion.