Agglomerative clustering algorithms

Procheta Sen

UNIVERSITY OF LIVERPOOL

Agglomerative (bottom-up) methods: main idea

- The individual objects are successively agglomerated into higher-level clusters.
- The main variation among the different methods is in the choice of objective function used to decide the merging of the clusters.

Agglomerative (bottom-up) methods: main idea

Input: dataset $\mathscr{D}$

1. **Initialise**: place every object in $\mathscr{D}$ in its own cluster
2. **Repeat**:
    1. Find ***closest*** pair of clusters i and j
    2. Merge clusters i and j
3. **Until** termination criterion
4. **Return** current clustering **or** hierarchy or clusterings

Agglomerative (bottom-up) methods: main idea

Input: dataset $\mathscr{D}$

1. **Initialise**: place every object in $\mathscr{D}$ in its own cluster
2. **Repeat**:
    1. Find ***closest*** pair of clusters i and j
    2. Merge clusters i and j
3. **Until** termination criterion
4. **Return** current clustering **or** hierarchy or clusterings

Need to specify measure of proximity between clusters

Agglomerative (bottom-up) methods: main idea

Input: dataset $\mathscr{D}$

1. **Initialise**: place every object in $\mathscr{D}$ in its own cluster
2. **Repeat**:
    1. Find ***closest*** pair of clusters i and j
    2. Merge clusters i and j
3. **Until** termination criterion
4. **Return** current clustering **or** hierarchy or clusterings

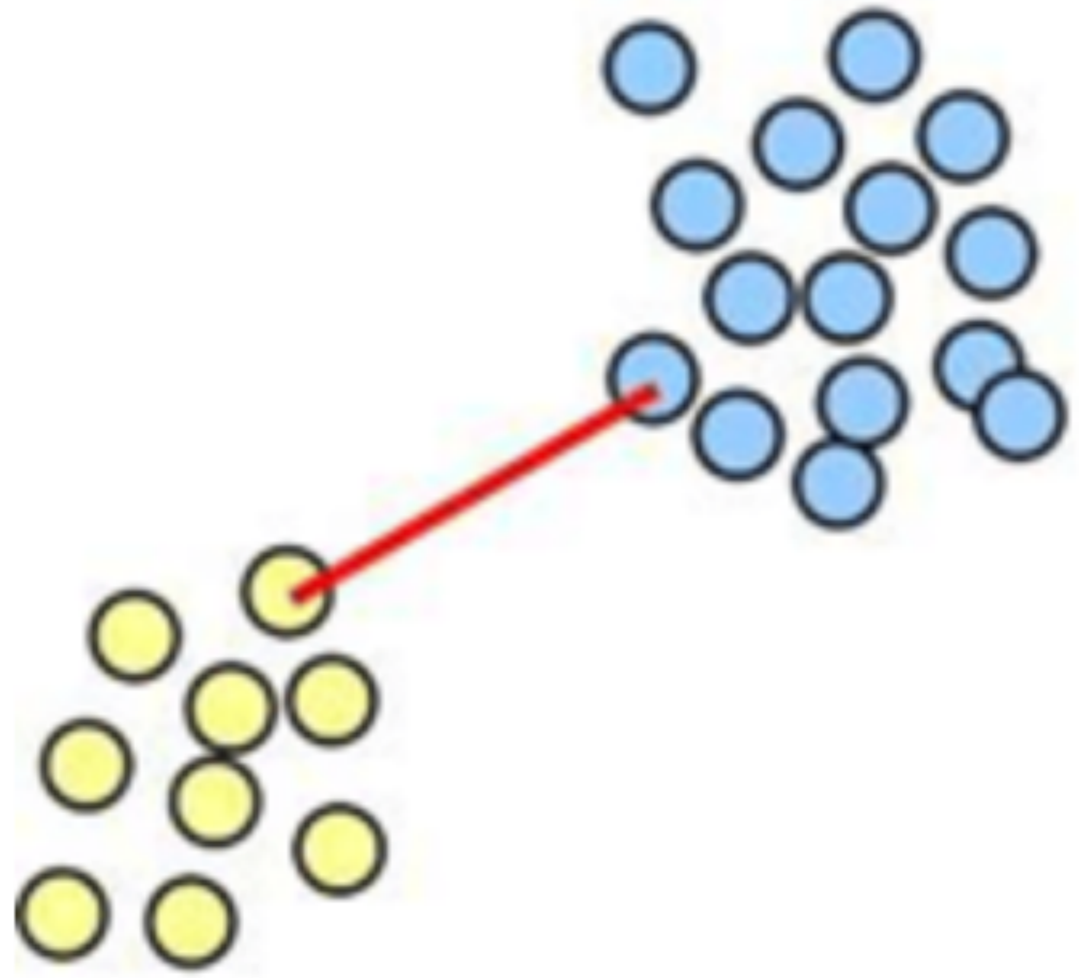
- distances between two merged clusters
- number of clusters

Measure of proximity between clusters: single-linkage

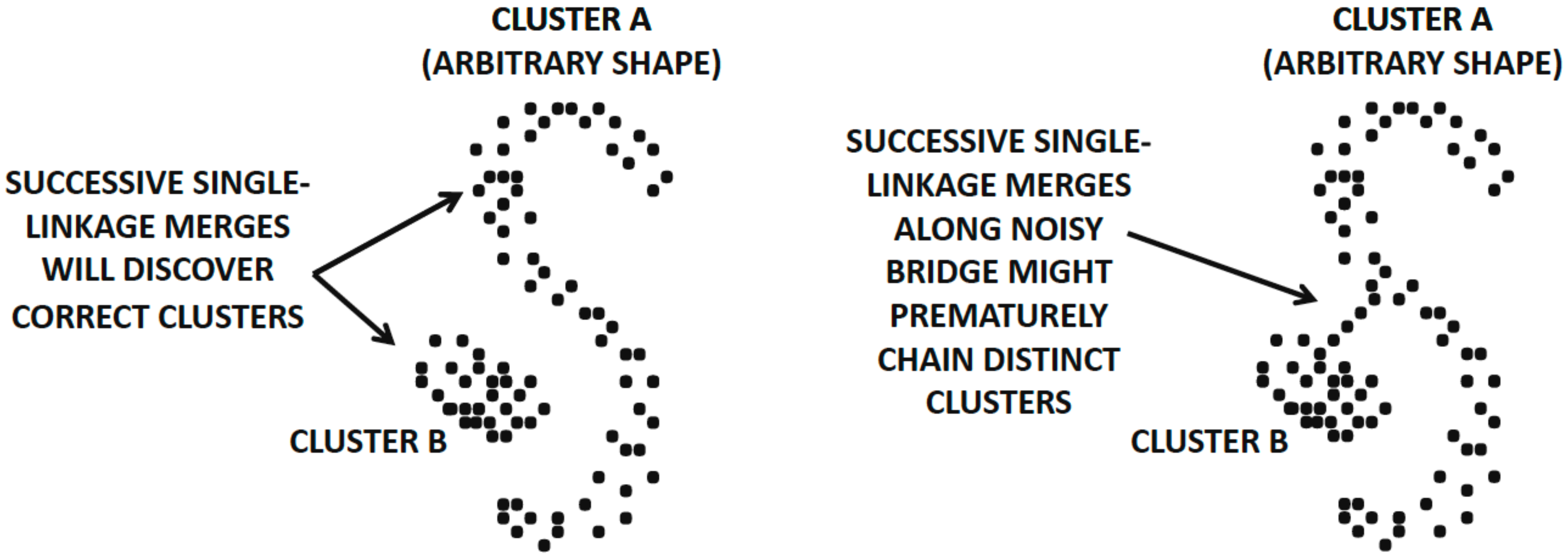
Let P and Q be two clusters. Assume we have a distance function $d(\cdot, \cdot)$ for objects.

Best (single) linkage: the distance between P and Q is the <u>minimum</u> distance between a pair of objects one of which is in P and the other in Q.

$$dist(P, Q) = \min_{\overline{X} \in P, \overline{Y} \in Q} d(\overline{X}, \overline{Y})$$



Single-linkage clustering



(a) Good case with no noise

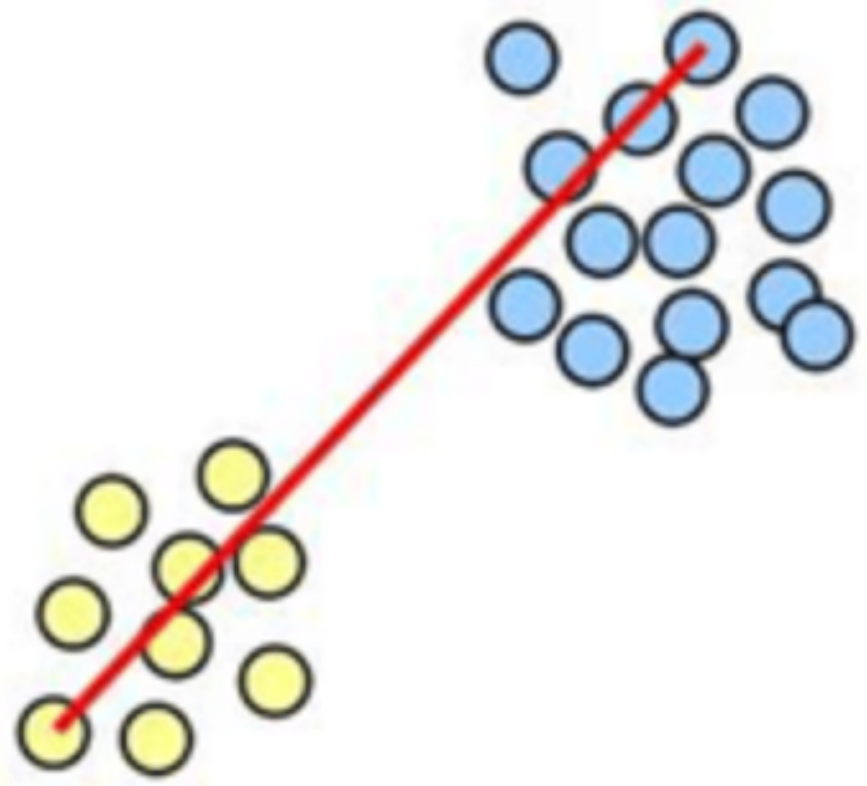
(b) Bad case with noise

Measure of proximity between clusters: complete-linkage

Let P and Q be two clusters. Assume we have a distance function $d(\cdot, \cdot)$ for objects.

Worst (complete) linkage: the distance between P and Q is the maximum distance between a pair of objects one of which is in P and the other in Q.

$$dist(P, Q) = \max_{\overline{X} \in P, \overline{Y} \in Q} d(\overline{X}, \overline{Y})$$

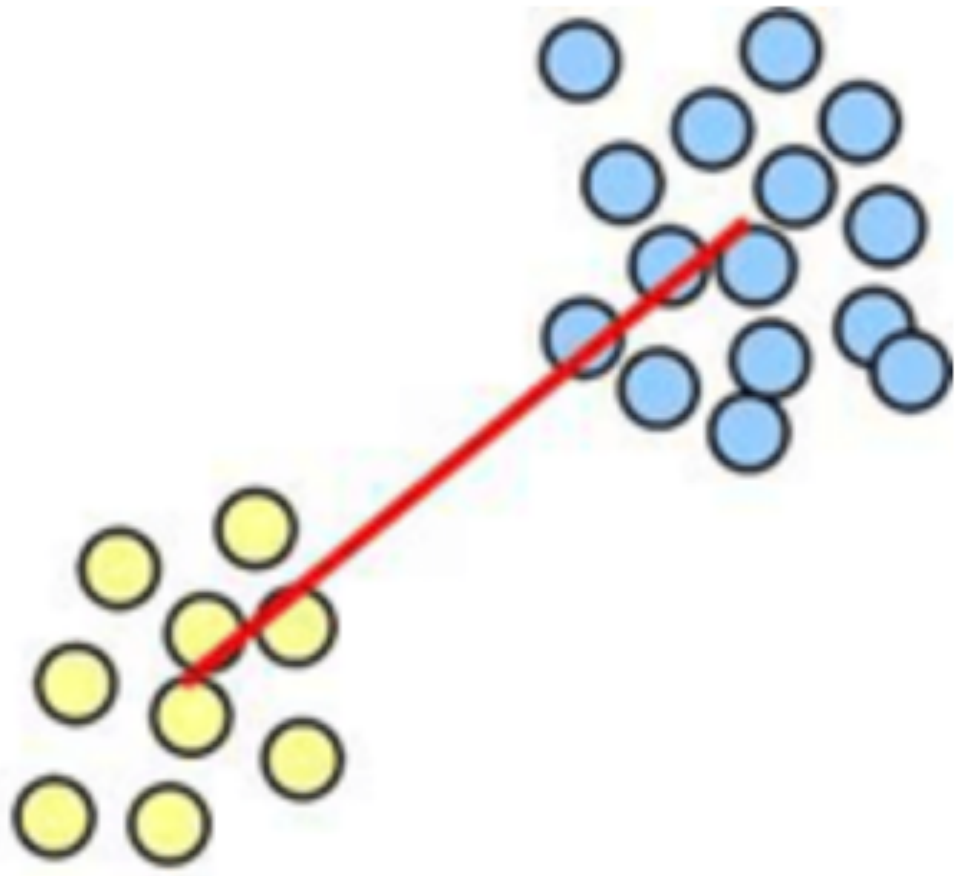


Implicitly attempts to minimize the maximum diameter of a cluster (i.e. as the largest distance between any pair of points in the cluster).

Measure of proximity between clusters: group-average linkage

Let P and Q be two clusters with p and q objects respectively. Assume we have a distance function $d(\cdot, \cdot)$ for objects.

Group-average linkage: the distance between P and Q is the average distance between all pairs of objects one of which is in P and the other in Q.



$$dist(P, Q) = \frac{1}{p \cdot q} \sum_{\overline{X} \in P, \overline{Y} \in Q} d(\overline{X}, \overline{Y})$$