Clustering evaluation

Procheta Sen

UNIVERSITY OF LIVERPOOL

Clustering quality evaluation

1. Extrinsic methods (supervised methods)
    - ground truth labels are used
    - assign a score to a clustering given the ground truth labels

2. Intrinsic methods (unsupervised methods)
    - only the partition of the objects into clusters is used
    - examine how well the clusters are separated and how compact the clusters are

Extrinsic methods: evaluate clustering as classification

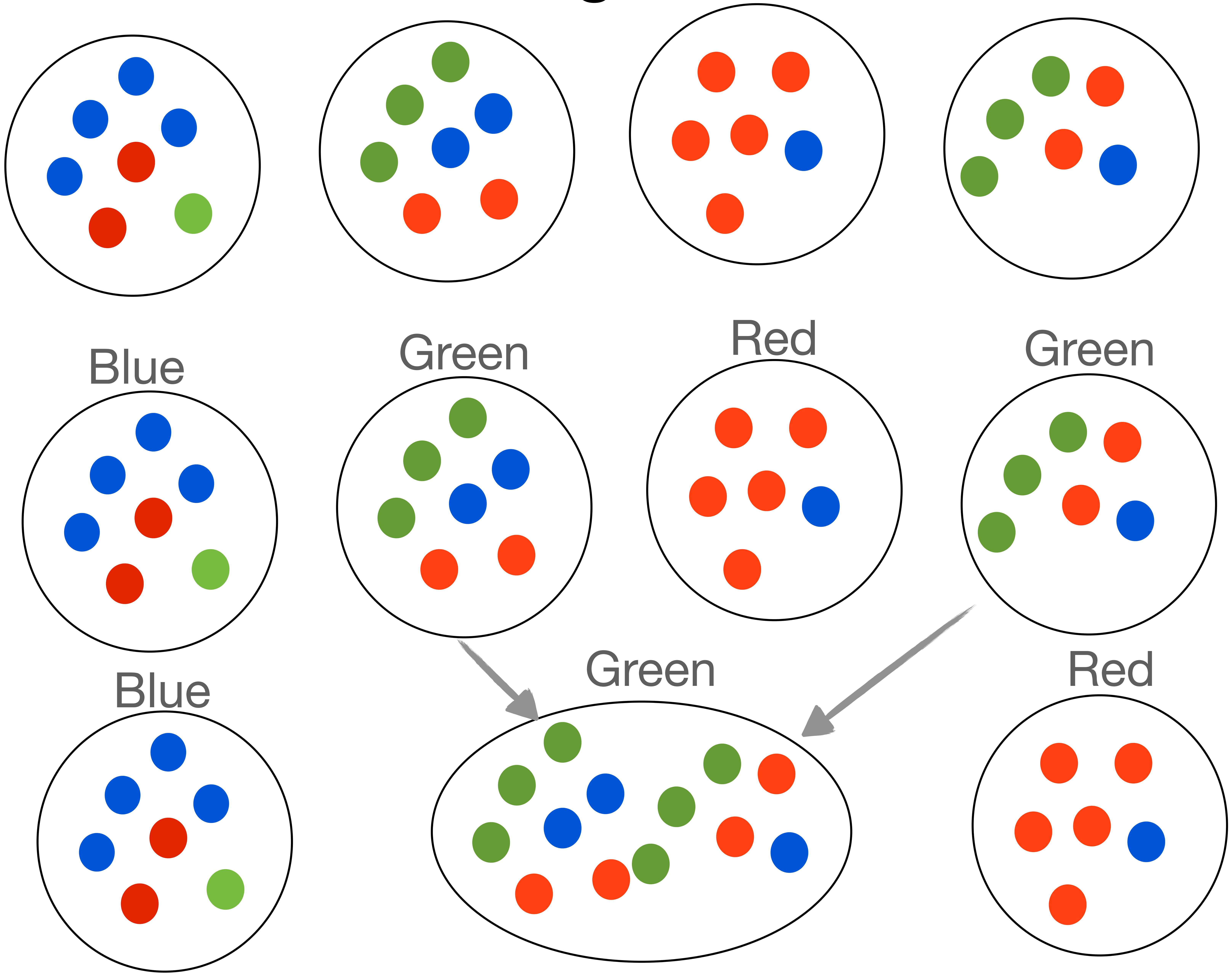
1. Assign each cluster the label that appears most in that cluster
2. Merge clusters with the same label
3. Measure Precision, Recall, and F-measure for each label type
4. Compute the macro-average, i.e. the average over all label types (classes), of
    - Precision
    - Recall
    - F-score

Extrinsic methods: evaluate clustering as classification

Clustering outcome

Assign each cluster the label that appears most in that cluster

Merge clusters with the same label



Extrinsic methods: B-CUBED Measure

- Proposed in (Bagga B. Baldwin = B^3)
  - A. **B**agga and **B**. **B**aldwin. *Entity-based cross document coreference resolution using the vector space model*, In Proc. of 36th COLING-ACL, pages 79--85, 1998.
- We would like to evaluate clustering without labelling any clusters.

$$\text{precision}(x) = \frac{\text{No. of items in C(x) with A(x)}}{\text{No. of items in C(x)}}$$

$$\text{recall}(x) = \frac{\text{No. of items in C(x) with A(x)}}{\text{Total no. of items with A(x)}}$$

A(x): label of x

C(x): The ID of the cluster that x belongs to

Extrinsic methods: B-CUBED Measure

- Compute the average over all the items (instances) that appear in all clusters (N)

$$\text{Precision} = \frac{1}{N} \sum_{p \in DataSet} \text{Precision}(p)$$

$$\text{Recall} = \frac{1}{N} \sum_{p \in DataSet} \text{Recall}(p)$$

$$F{-}\text{Score} = \frac{1}{N} \sum_{p \in DataSet} \text{F}(p)$$

Intrinsic methods: Silhouette Coefficient

1. Let $C_1, C_2, \ldots, C_k$ be the clusters
2. For object x (assume $x \in C_i$):
   - $a(x)$: The mean distance between x and all other points in the cluster of x.
   - $b(x)$: The mean distance between x and all other points in the *next nearest cluster*.

$$a(x) = \frac{1}{|C_i| - 1} \sum_{y \in C_i, y \neq x} d(x, y) \qquad b(x) = \min_{j=1,\ldots,k, j \neq i} \frac{1}{|C_j|} \sum_{y \in C_j} d(x, y)$$

$a(x)$ is a measure of how dissimilar x is to its own cluster, a small value means it is well matched.

$b(x)$ is a measure of how badly x is matched to its neighbouring cluster.

Intrinsic methods: Silhouette Coefficient

3. **Silhouette Coefficient** of x

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}, \text{ if } |C_i| > 1$$

$$s(x) = 0, \text{ if } |C_i| = 1$$

$s(x)$ close to 1 means that the data is appropriately clustered

$s(x)$ close to -1 means that it would be more appropriate if x was clustered in its neighbouring cluster

$s(x)$ close to 0 means that x is on the border of two natural clusters

4. **Silhouette Coefficient** of the dataset is the average of Silhouette Coefficients of all objects in the dataset

measure of how appropriately the data have been clustered

Intrinsic methods: Silhouette Coefficient