CA Assignment - 2

COMP527-202425

MSc in Advanced Computer Science

University of Liverpool

Clustering Algorithms

Submission Date: [27th March, 2025]

- Jyothi Mani Ravi Sankar

Student ID - 201849581

4. Implement the Bisecting k-Means algorithm to compute a hierarchy of clustering's that refines the initial single cluster to 9 clusters. For each s from 1 to 9, extract from the hierarchy of clustering's the clustering with s clusters and compute the Silhouette coefficient for this clustering. Plot s in the horizontal axis and the Silhouette coefficient in the vertical axis in the same plot.

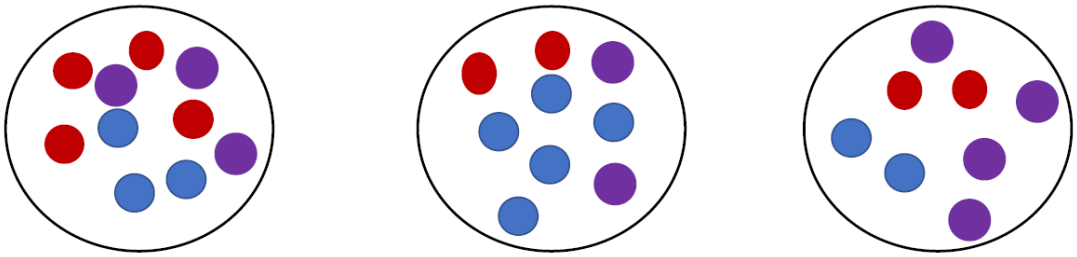

Figure 1: Outcome of a Clustering Algorithm

- **True Labels**:
  - Red points: Class 1
  - Blue points: Class 2
  - Purple points: Class 3
- **Predicted Labels**:
  - Left cluster: Cluster 1
  - Middle cluster: Cluster 2
  - Right cluster: Cluster 3

Step 1: Count the Points in Each Cluster

Based on the figure:

- **Cluster 1 (Left)**:
    - Red (Class 1): 4 points
    - Blue (Class 2): 3 points
    - Purple (Class 3): 3 points
    - Total: 10 points

- **Cluster 2 (Middle)**:
    - Red (Class 1): 2 points
    - Blue (Class 2): 5 points
    - Purple (Class 3): 2 points
    - Total: 9 points

- **Cluster 3 (Right)**:
    - Red (Class 1): 2 points
    - Blue (Class 2): 2 points
    - Purple (Class 3): 4 points
    - Total: 8 points

Total Points:

- Total = 10 + 9 + 8 = 27 points
- Per class:
    - Class 1 (Red): 4 + 2 + 2 = 8
    - Class 2 (Blue): 3 + 5 + 2 = 10
    - Class 3 (Purple): 3 + 2 + 4 = 9

Step 2: Construct the Confusion Matrix

The confusion matrix maps true labels (rows) to predicted labels (columns):

- Rows: True labels (Class 1: Red, Class 2: Blue, Class 3: Purple)
- Columns: Predicted labels (Cluster 1, Cluster 2, Cluster 3)

True Predicted	Cluster 1	Cluster 2	Cluster 3
Class 1 (Red)	4	2	2
Class 2 (Blue)	3	5	2
Class 3 (Purple)	3	2	4

$$\text{Confusion Matrix} = \begin{vmatrix} 4 & 2 & 2 \\ 3 & 5 & 2 \\ 3 & 2 & 4 \end{vmatrix}$$

Step 3: Map Clusters to True Classes

To compute Precision, Recall, and F-score, we map each predicted cluster to a true class based on the majority class in each cluster:

- Cluster 1: 4 red (majority) → Map to Class 1 (Red)
- Cluster 2: 5 blue (majority) → Map to Class 2 (Blue)
- Cluster 3: 4 purple (majority) → Map to Class 3 (Purple)

Step 4: Compute Precision, Recall, and F-score for Each Class

- **True Positives (TP)**: We predicted as positive and it is indeed positive
- **False Positives (FP)**: We predicted as positive but it turns out to be negative
- **False Negatives (FN)**: We predicted as negative but it turns out to be positive
- **True Negative (TN):** We predicted as negative and it is indeed negative

Precision = TP / TP + FP

Recall = TP / TP + FN

F – Score = 2 * Precision * Recall / Precision + Recall

Class 1 (Red, Cluster 1):

- **TP**: 4
- **FP**: 3 (blue) + 3 (purple) = 6
- **FN**: 2 (Cluster 2) + 2 (Cluster 3) = 4
- **Precision** = 4 /4 + 6 = 4/10 = 0.4
- **Recall** = 4 / 4 + 4 = 4 / 8 = 0.5
- **F-Score** = 2 * (0.4 * 0.5 / 0.4 + 0.5) ≈ 0.4444

Class 2 (Blue, Cluster 2):

- **TP**: 5
- **FP**: 2 (red) + 2 (purple) = 4
- **FN**: 3 (Cluster 1) + 2 (Cluster 3) = 5
- **Precision**: $\frac{5}{5+4} = \frac{5}{9} \approx 0.55565$
- **Recall**: $\frac{5}{5+5} = \frac{5}{10} = 0.5$
- **F-score**: $2 \times \frac{0.5556 \times 0.5}{0.5556+0.5} = 2 \times \frac{0.2778}{1.0556} \approx 2 \times 0.2632 = 0.52632$

Class 3 (Purple, Cluster 3):

- **TP**: 4
- **FP**: 2 (red) + 2 (blue) = 4
- **FN**: 3 (Cluster 1) + 2 (Cluster 2) = 5
- **Precision**: $\frac{4}{4+4} = \frac{4}{8} = 0.5$
- **Recall**: $\frac{4}{4+5} = \frac{4}{9} \approx 0.44444$
- **F-score**: $2 \times \frac{0.5 \times 0.4444}{0.5+0.4444} = 2 \times \frac{0.2222}{0.9444} \approx 2 \times 0.2353 = 0.4706$

Step 5: Compute Macro-Averaged Metrics

- **Macro-averaged Precision:** $\frac{0.4+0.5556+0.5}{3} = \frac{1.4556}{3} \approx 0.4852$
- **Macro-averaged Recall:** $\frac{0.5+0.5+0.4444}{3} = \frac{1.4444}{3} \approx 0.4815$
- **Macro-averaged F-score:** $\frac{0.4444+0.5263+0.4706}{3} = \frac{1.4413}{3} \approx 0.4804$

Final Answer

$$\text{Confusion Matrix} = \begin{vmatrix} 4 & 2 & 2 \\ 3 & 5 & 2 \\ 3 & 2 & 4 \end{vmatrix}$$

Macro-averaged Precision: 0.4852

Macro-averaged Recall: 0.4815

Macro-averaged F-score: 0.4804

6. For the same clusters as in Figure 1, compute B-CUBED Precision, Recall, and F-score.

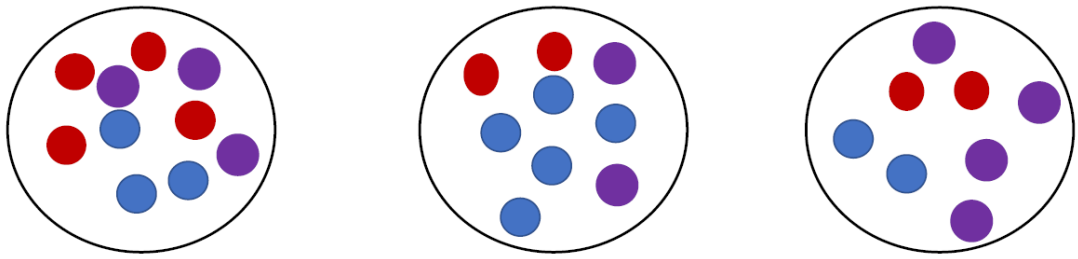

Figure 1: Outcome of a Clustering Algorithm

- **True Labels**:
    - Red points: Class 1
    - Blue points: Class 2
    - Purple points: Class 3

- **Predicted Labels**:
    - Left cluster: Cluster 1
    - Middle cluster: Cluster 2
    - Right cluster: Cluster 3

Step 1: Count the Points in Each Cluster

Based on the figure:

- **Cluster 1 (Left)**:
    - Red (Class 1): 4 points
    - Blue (Class 2): 3 points
    - Purple (Class 3): 3 points
    - Total: 10 points

- **Cluster 2 (Middle)**:
    - Red (Class 1): 2 points
    - Blue (Class 2): 5 points
    - Purple (Class 3): 2 points
    - Total: 9 points

- **Cluster 3 (Right)**:
    - Red (Class 1): 2 points
    - Blue (Class 2): 2 points
    - Purple (Class 3): 4 points

- Total: 8 points

Total Points:

- Total = 10 + 9 + 8 = 27 points
- Per class:
  - Class 1 (Red): 4 + 2 + 2 = 8
  - Class 2 (Blue): 3 + 5 + 2 = 10
  - Class 3 (Purple): 3 + 2 + 4 = 9

Step 2: Compute B-CUBED Metrics for Each Point

For each point, compute Precision and Recall:

$$\text{Precision}(x) = \frac{\text{Number of points in } x\text{'s cluster with } x\text{'s true label}}{\text{Total number of points in } x\text{'s cluster}}$$

$$\text{Recall}(x) = \frac{\text{Number of points in } x\text{'s cluster with } x\text{'s true label}}{\text{Total number of points with } x\text{'s true label}}$$

Cluster 1 (10 points):

- 4 points (Class 1, Red):

  Precision = 4 / 10 = 0.4

  Recall = 4 / 8 = 0.5

- 3 points (Class 2, Blue):

  Precision = 3 / 10 = 0.3

  Recall = 3 / 10 = 0.3

- 3 points (Class 3, Purple):
  Precision = 3 / 10 = 0.3
  Recall = 3 / 9 = 0.3333

Cluster 2 (9 points):

- 2 points (Class 1, Red):
  Precision = 2 / 9 = 0.2222
  Recall = 2 / 8 = 0.25
- 5 points (Class 2, Blue):
  Precision = 5 / 9 = 0.5556
  Recall = 5 / 10 = 0.5
- 2 points (Class 3, Purple):
  Precision = 2 / 9 = 0.2222
  Recall = 2 / 9 = 0.2222

Cluster 3 (8 points):

- 2 points (Class 1, Red):
  Precision = 2 / 8 = 0.25
  Recall = 2 / 8 = 0.25
- 2 points (Class 2, Blue):
  Precision = 2 / 8 = 0.25
  Recall = 2 / 10 = 0.2
- 4 points (Class 3, Purple):
  Precision = 4 / 8 = 0.5
  Recall = 4 / 9 = 0.4444

Step 3: Compute B-CUBED Precision and Recall

B-CUBED Precision is the average precision across all individual points.

Cluster 1 (10 points):

- $(4 \times 0.4) + (3 \times 0.3) + (3 \times 0.3)$

  $= 1.6 + 0.9 + 0.9$

  $= 3.4$

Cluster 2 (9 points):

- $(2 \times 0.2222) + (5 \times 0.5556) + (2 \times 0.2222)$

  $= 0.4444 + 2.778 + 0.4444$

  $= 3.6668$

Cluster 3 (8 points):

- $(2 \times 0.25) + (2 \times 0.25) + (4 \times 0.5)$

  $= 0.5 + 0.5 + 2.0$

  $= 3.0$

Total Precision:

$3.4 + 3.6668 + 3.0 = 10.0668$

B-CUBED Precision:

$10.0668 / 27 \approx 0.3728$

B-CUBED Recall is the average recall across all individual points.

Cluster 1 (10 points):

- $(4 \times 0.5) + (3 \times 0.3) + (3 \times 0.3333)$

  $= 2.0 + 0.9 + 0.9999$

  $= 3.8999$

Cluster 2 (9 points):

- $(2 \times 0.25) + (5 \times 0.5) + (2 \times 0.2222)$

  $= 0.5 + 2.5 + 0.4444$

  $= 3.4444$

Cluster 3 (8 points):

- $(2 \times 0.25) + (2 \times 0.2) + (4 \times 0.4444)$

  $= 0.5 + 0.4 + 1.7776$

  $= 2.6776$

Total Recall:

$3.8999 + 3.4444 + 2.6776 = 10.0219$

B-CUBED Recall:

$10.0219 / 27 \approx 0.3712$

Step 4: Compute B-CUBED F-score

$$\textbf{B-CUBED F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\textbf{B-CUBED F-score} = \frac{2 \times (0.3728 \times 0.3712)}{0.3728 + 0.3712}$$

$$= \frac{2 \times 0.1383}{0.7440}$$

$$\approx \frac{0.2766}{0.7440}$$

$$\approx 0.3718$$

B-CUBED Precision: 0.3728

B-CUBED Recall: 0.3712

B-CUBED F-score: 0.3718