Clustering: Unsupervised Learning

IF-3270 Pembelajaran Mesin

Teknik Informatika ITB

Modul 7: Clustering

05 Cluster Evaluation

IF3270 - Pembelajaran Mesin
(Machine Learning)

Fariska Z. Ruskanda, S.T., M.T.
(fariska@informatika.org)

KK IF -Teknik Informatika - STEI ITB

Outline

Cluster Validity

Internal Measures

External Measures

Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is
  - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    supervised clustering
    - Entropy, Purity, Precision, Recall, F-Measure, Rand Index, etc.
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    supervised clustering
    - Silhouette Coefficient, Dunn Index, Cophenetic Correlation Coefficient (CPCC), etc.
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster; compactness
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters; isolation
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$\boxed{WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2}$$

itung jarak titik dgn suatu objek pd cluster

  - Separation is measured by the between cluster sum of squares

$$\boxed{BSS = \sum_{i} |C_i| (m - m_i)^2}$$

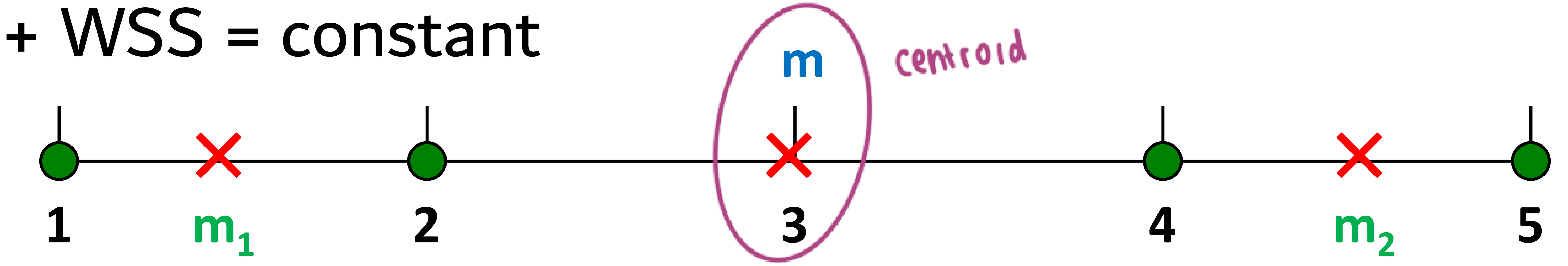
itung SSE antar cluster
(antara titik tengah dgn titik tengah pd cluster)

    - Where $|C_i|$ is the size of cluster i

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

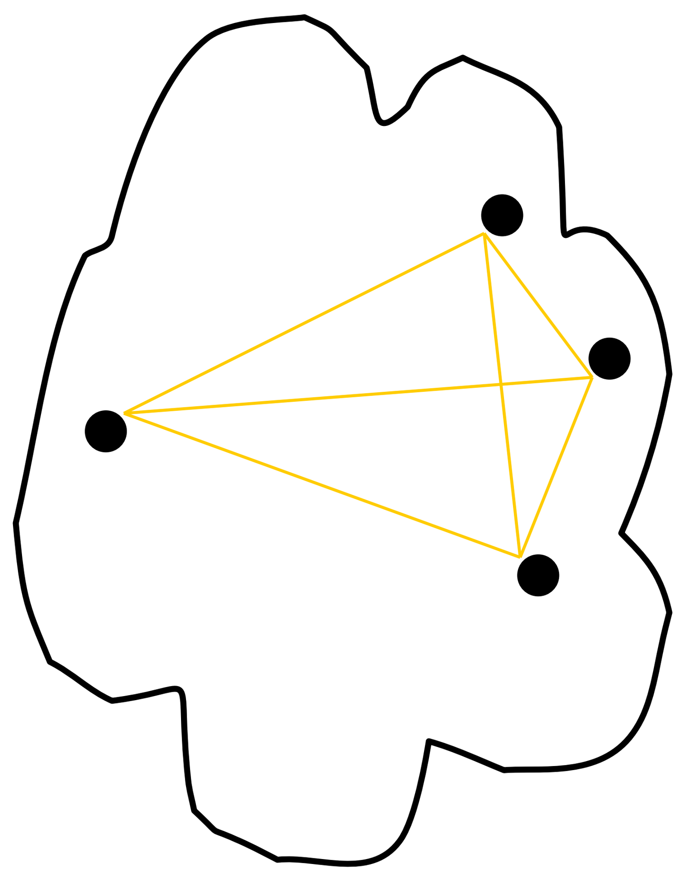
$$Total = 1 + 9 = 10$$

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

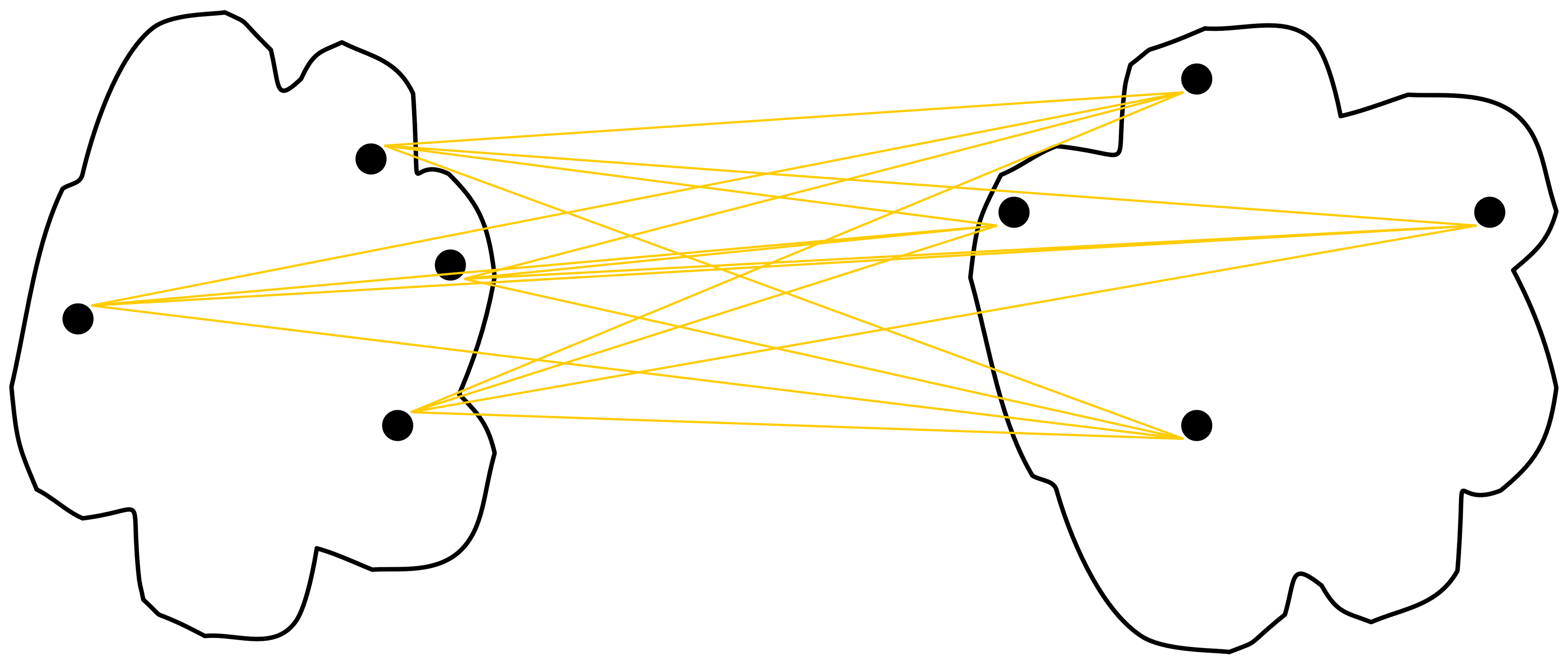
Internal Measures: Cohesion and Separation

di dlm   di luar

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



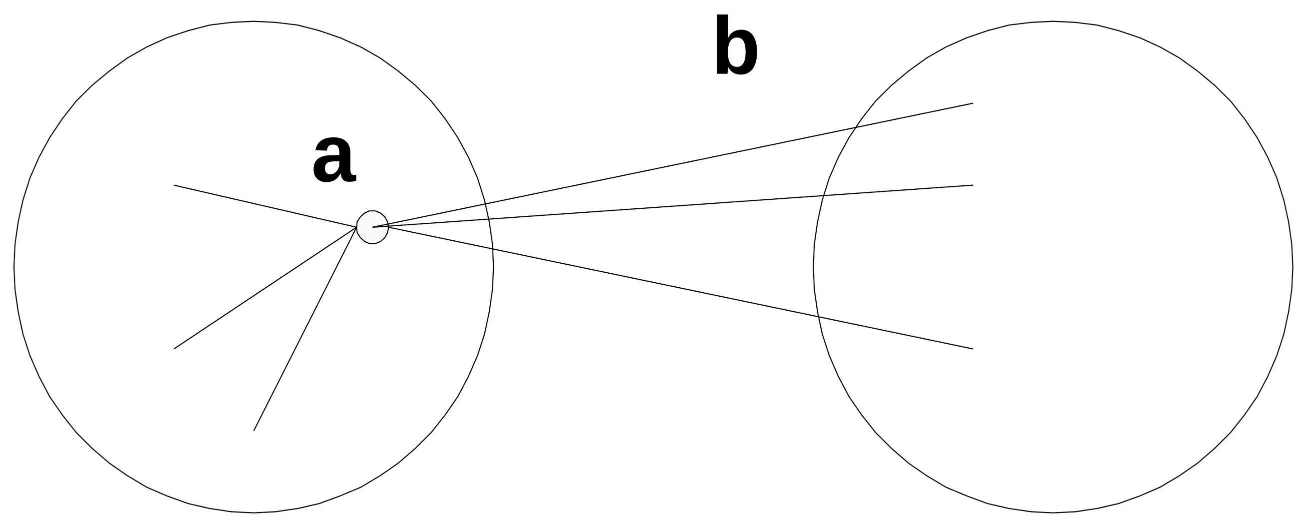
separation

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings.
- For an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by

$$\boxed{s = \frac{b - a}{max(a, b)}}$$



  - The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.
  - Scores around zero indicate overlapping clusters.

- Can calculate the Average Silhouette width for a cluster or a clustering

https://arxiv.org/pdf/1905.05667

External Measures: Purity

- Purity: ratio between the dominant class in the cluster and the size of the cluster.

$$\boxed{Purity(w_i) = \frac{1}{n_i}\max_j(n_{ij}) \quad j \in C}$$

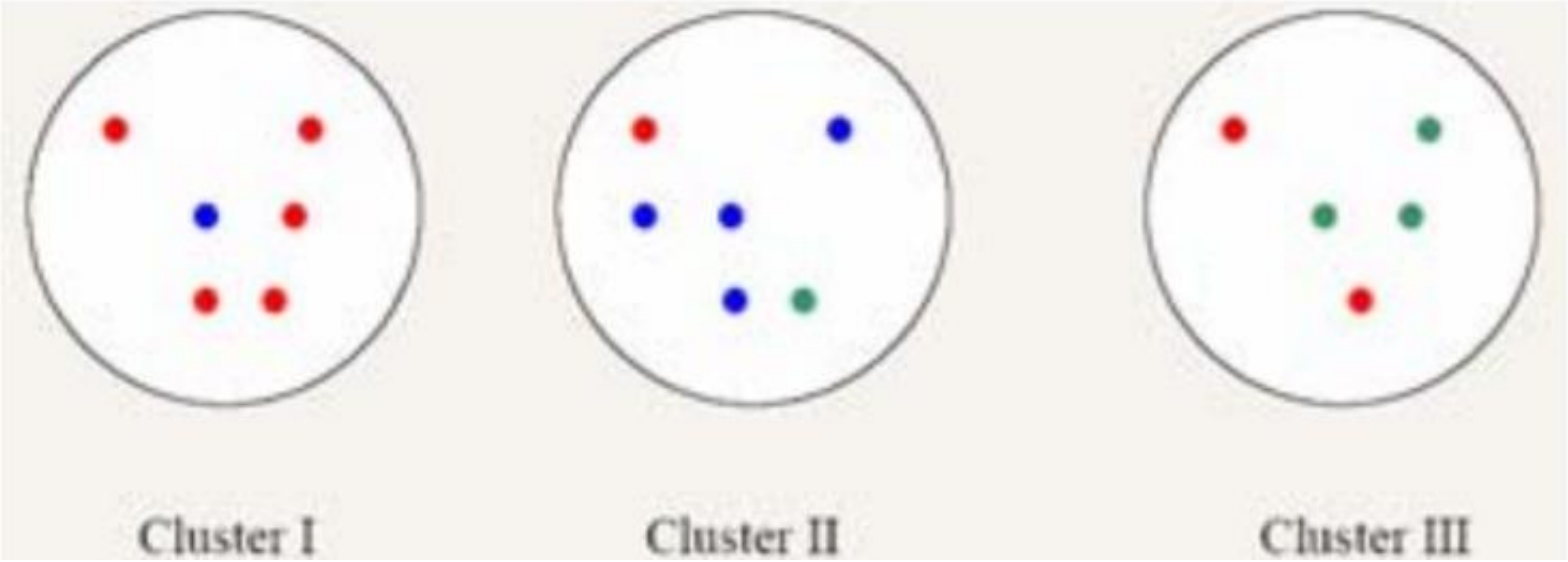
- Assume documents with C gold standard classes, while our clustering algorithm produce K clusters, w1, w2, … wk with ni members.

https://www.cs.cmu.edu/~epxing/Class/10701-08s/Lecture/lecture15-Clustering.pdf

External Measures: Purity

- Example:

merah biru hijau



m b h

- Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6
- Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6
- Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

https://www.cs.cmu.edu/~epxing/Class/10701-08s/Lecture/lecture15-Clustering.pdf

External Measures: Entropy

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

https://www.cs.purdue.edu/homes/clifton/cs590d/Cluster.ppt

Terima Kasih