Chapter DM:II (continued)

II. Cluster Analysis

- Cluster Analysis Basics
- Hierarchical Cluster Analysis
- Iterative Cluster Analysis
- Density-Based Cluster Analysis
- Cluster Evaluation
- Constrained Cluster Analysis

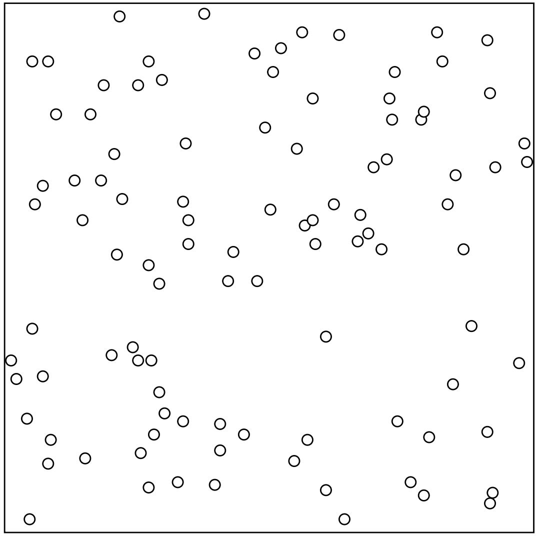
Cluster Evaluation

Overview

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

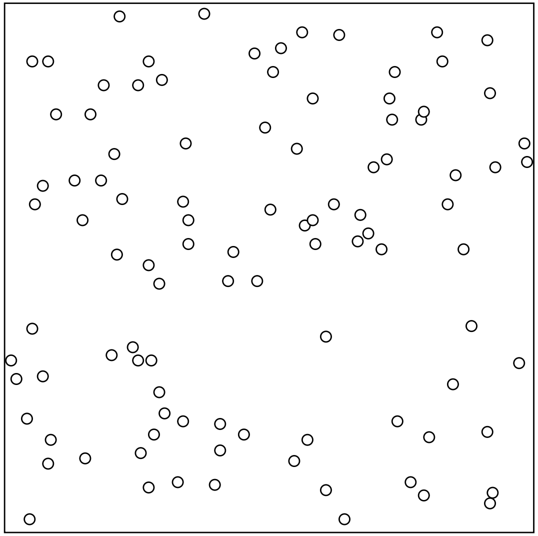
[Jain/Dubes 1990]

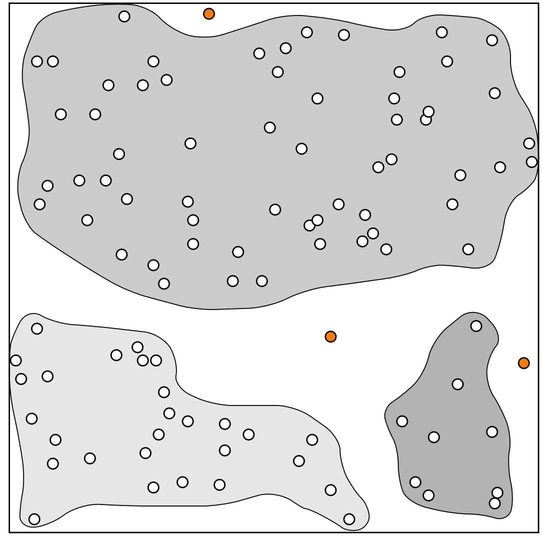
Cluster Evaluation [Tan/Steinbach/Kumar 2005]



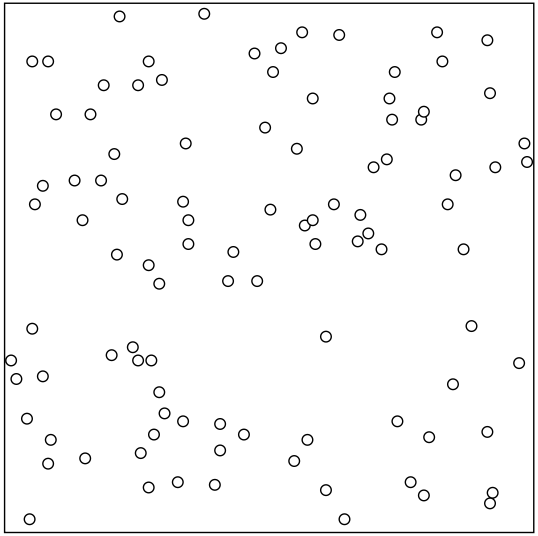
Random points

Cluster Evaluation [Tan/Steinbach/Kumar 2005]

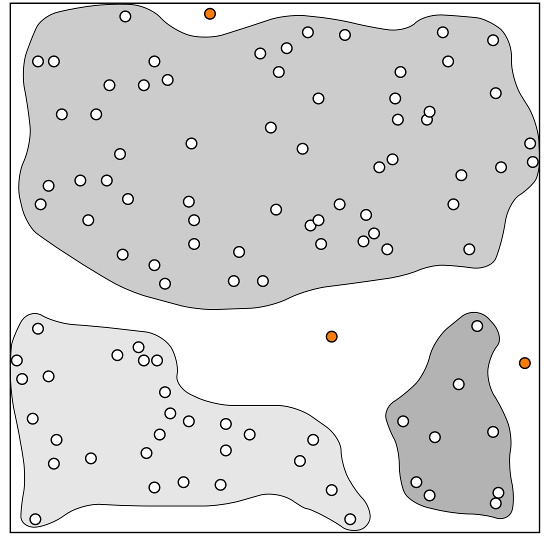

Random points


DBSCAN

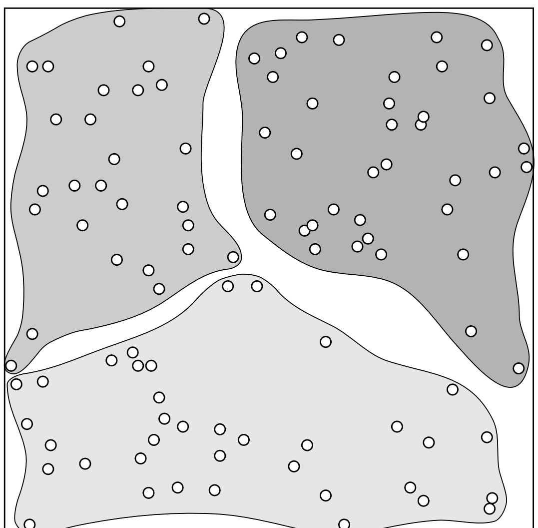
Cluster Evaluation [Tan/Steinbach/Kumar 2005]



Random points

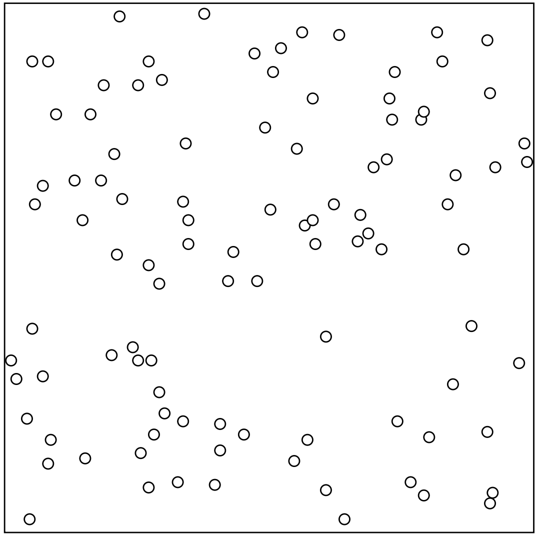


DBSCAN

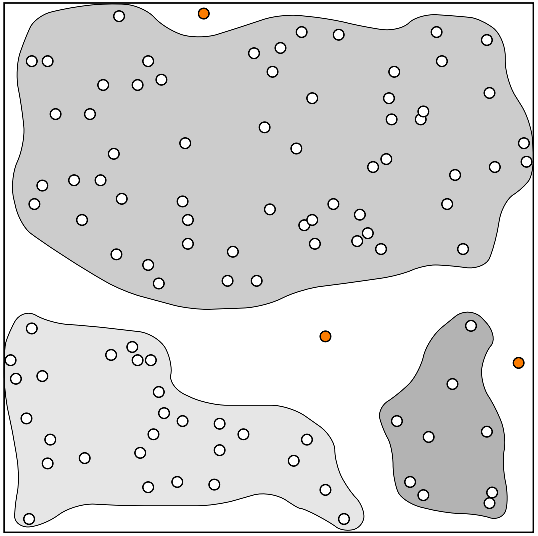


k-means

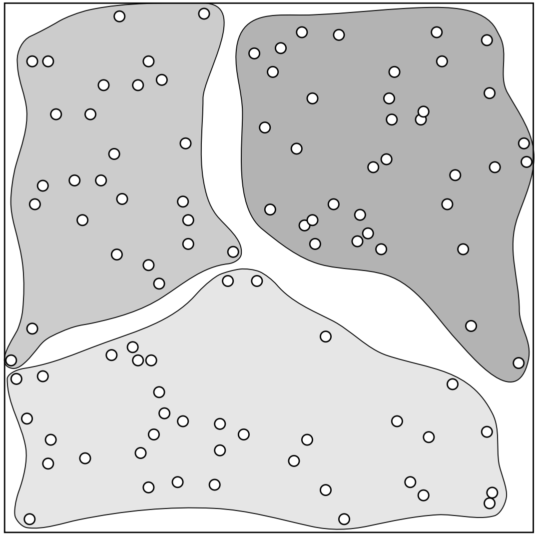
Cluster Evaluation [Tan/Steinbach/Kumar 2005]



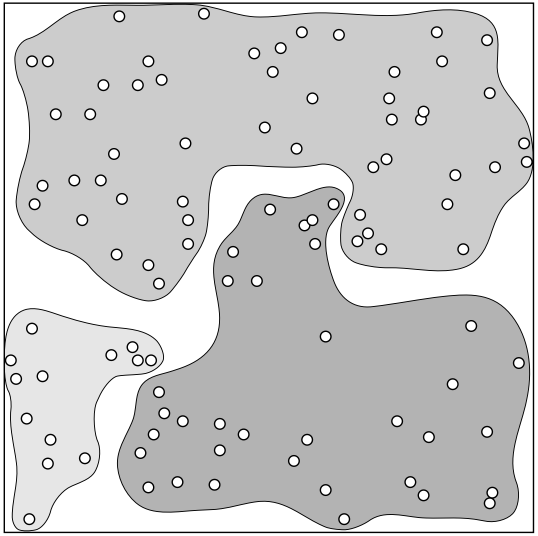
Random points



DBSCAN



k-means



Complete link

Cluster Evaluation

Overview

Cluster evaluation can address different issues:

- Provide evidence whether data contains non-random structures.
- Relate found structures in the data to externally provided class information.
- Rank alternative clusterings with regard to their quality.
- Determine the ideal number of clusters.
- Provide information to choose a suited clustering approach.

Cluster Evaluation

Overview

Cluster evaluation can address different issues:

- Provide evidence whether data contains non-random structures.
- Relate found structures in the data to externally provided class information.
- Rank alternative clusterings with regard to their quality.
- Determine the ideal number of clusters.
- Provide information to choose a suited clustering approach.

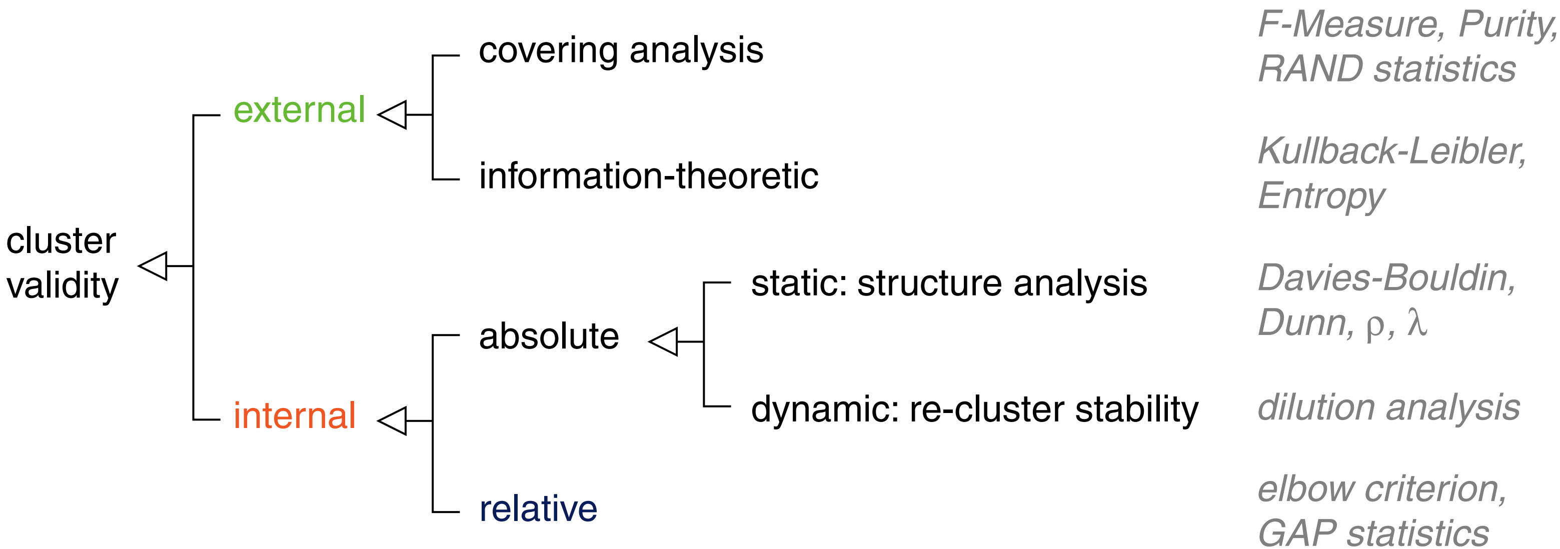
(1) External validity measures:
Analyze how close is a clustering to an (external) reference.

(2) Internal validity measures:
Analyze intrinsic characteristics of a clustering.

(3) Relative validity measures:
Analyze the sensitivity (of internal measures) during clustering generation.

Cluster Evaluation

Overview



(1) External validity measures:

Analyze how close is a clustering to an (external) reference.

(2) Internal validity measures:

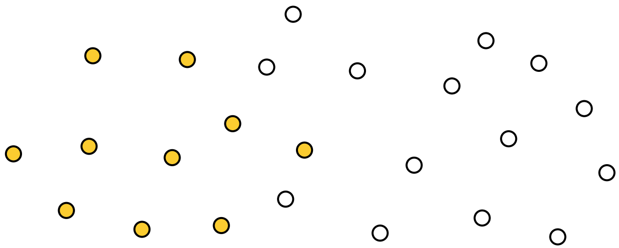
Analyze intrinsic characteristics of a clustering.

(3) Relative validity measures:

Analyze the sensitivity (of internal measures) during clustering generation.

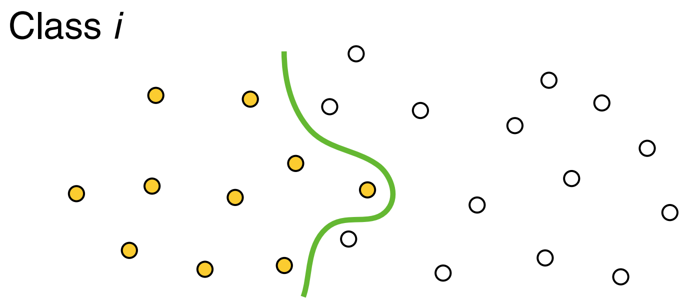
Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



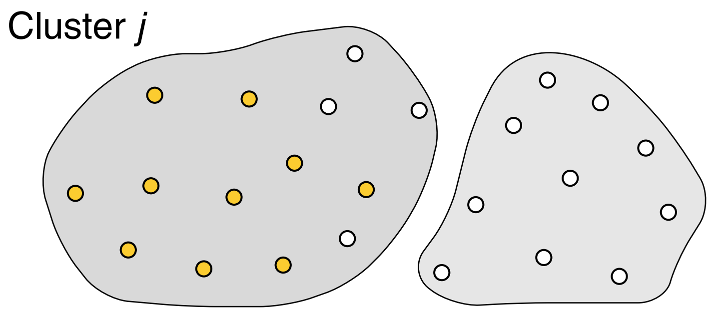
Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



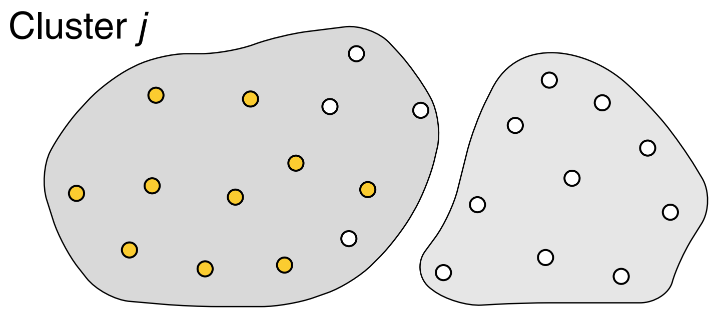
Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)

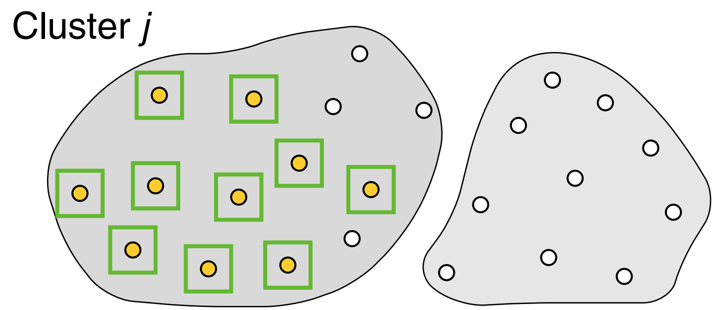


(node-based analysis)

		Truth	
		P	N
Hypothesis	P		
	N		

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)

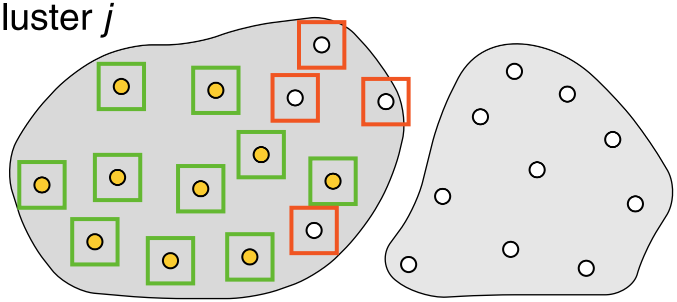


(node-based analysis)

		Truth	
		P	N
Hypothesis	P	TP (a)	
	N		

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)

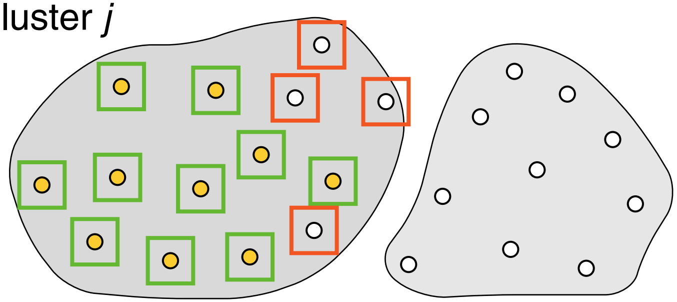


(node-based analysis)

		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N		

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)

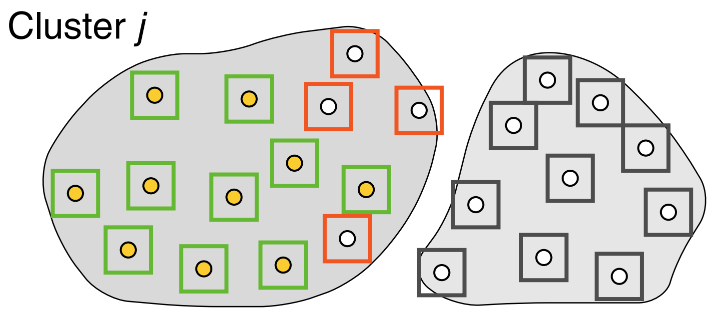


(node-based analysis)

		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)

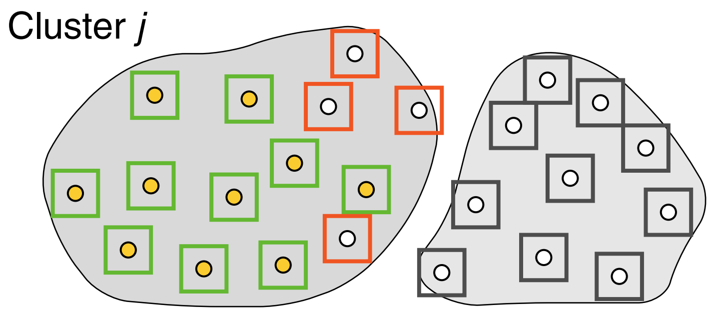


(node-based analysis)

		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



(node-based analysis)

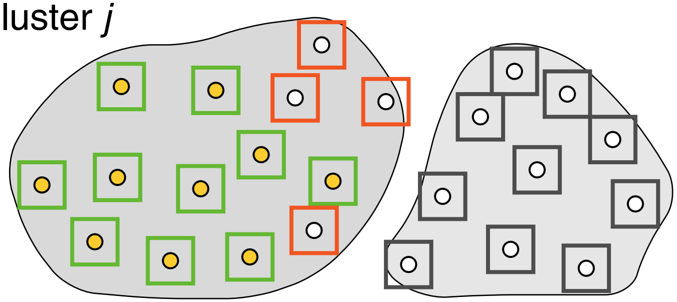
		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Precision: $\dfrac{a}{a+b}$

Recall: $\dfrac{a}{a+c}$

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



(node-based analysis)

		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Precision: $\frac{a}{a+b}$

Recall: $\frac{a}{a+c}$

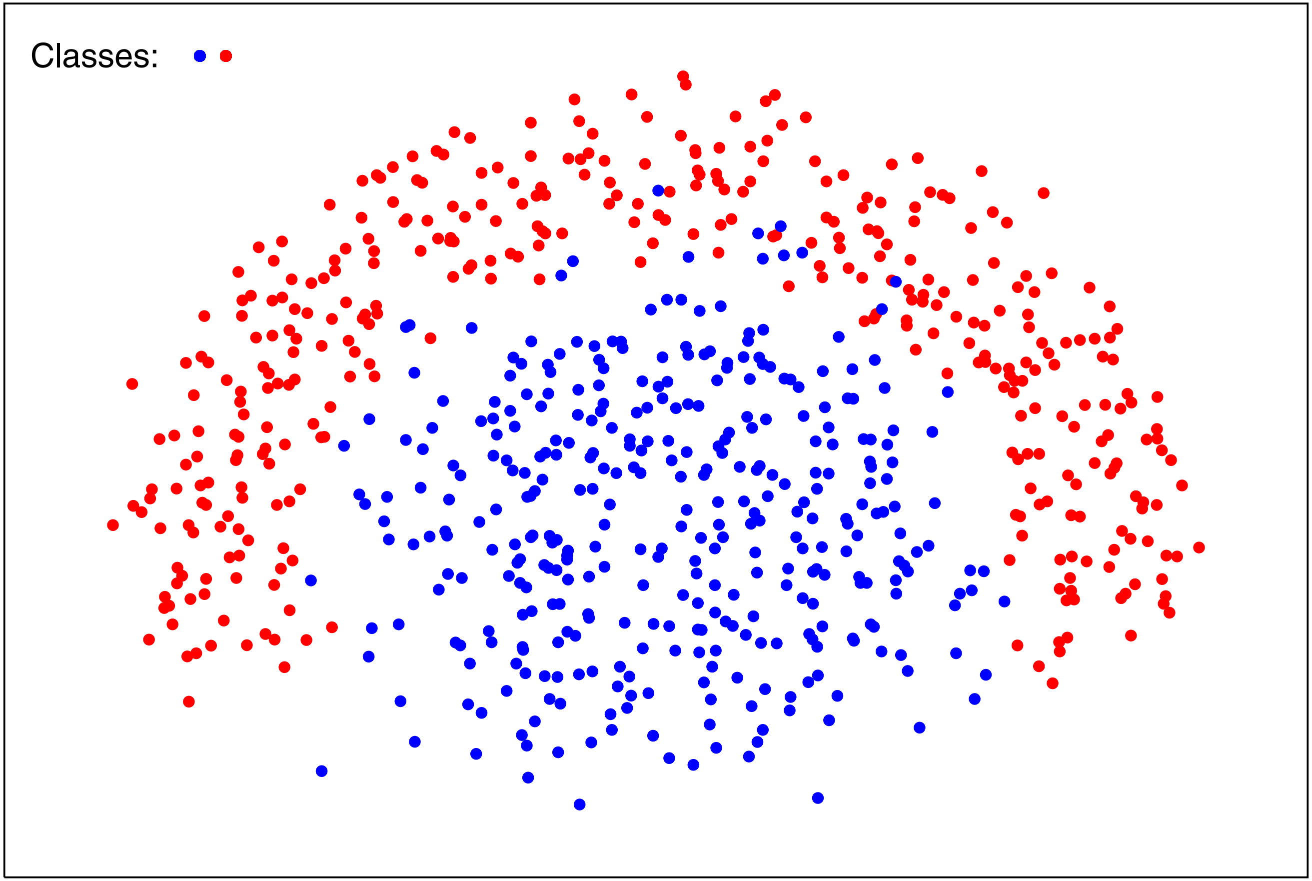
F-measure:

$$F_\alpha = \frac{1+\alpha}{\frac{1}{\textit{precision}} + \frac{\alpha}{\textit{recall}}}$$

$\alpha = 1$	harmonic mean
$\alpha \in (0;1)$	favor precision over recall
$\alpha > 1$	favor recall over precision

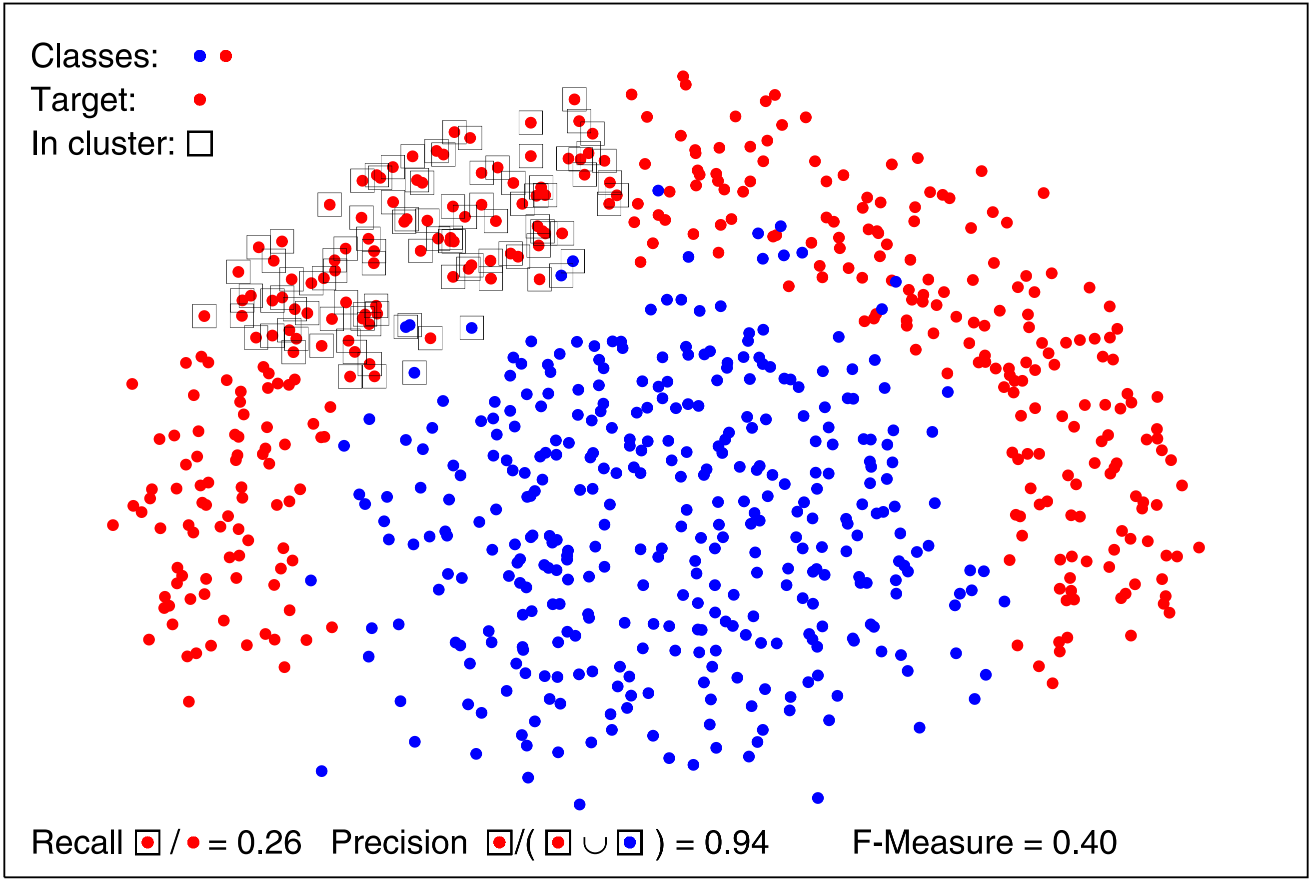
Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Target Class)



Cluster Evaluation

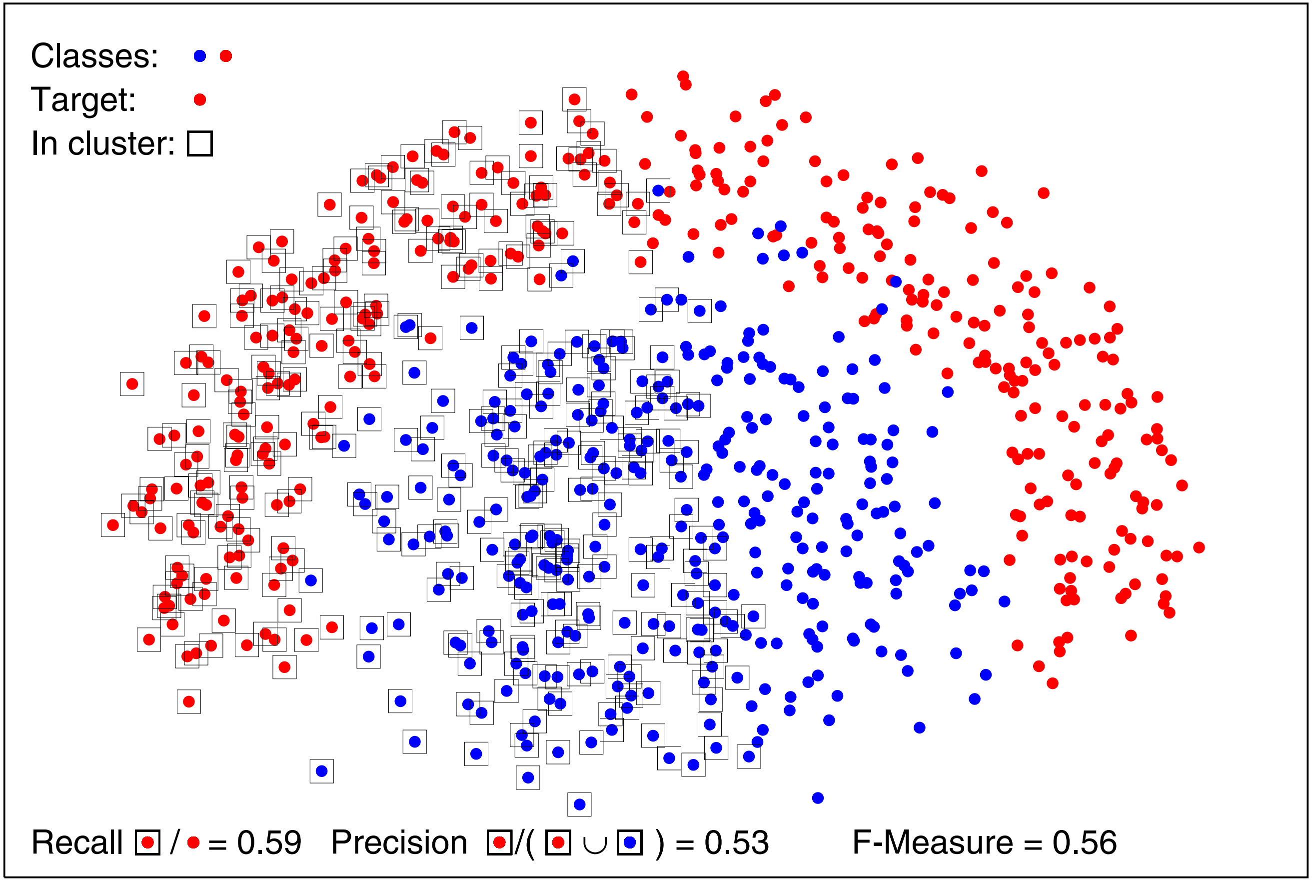
(1) External Validity Measures: F-Measure (for a Target Class)



High precision, low recall $\Rightarrow$ low F-measure.

Cluster Evaluation

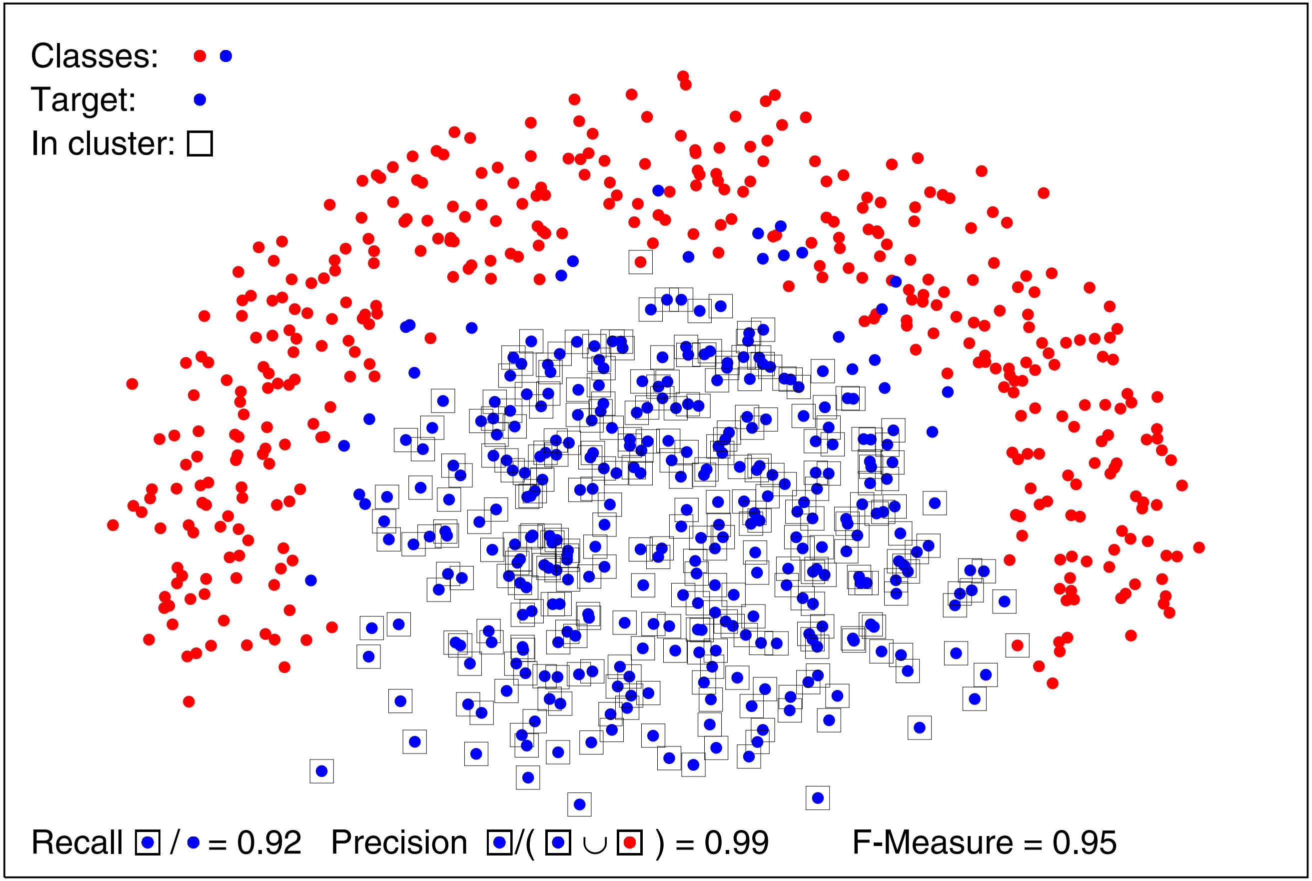
(1) External Validity Measures: F-Measure (for a Target Class)



Low precision, low recall $\Rightarrow$ low F-measure.

Cluster Evaluation

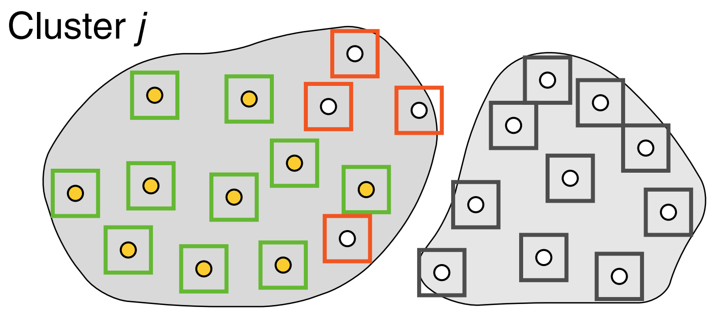
(1) External Validity Measures: F-Measure (for a Target Class)



High precision, high recall $\Rightarrow$ high F-measure.

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Clustering)



(node-based analysis)

- Clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ and classification $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ of D.

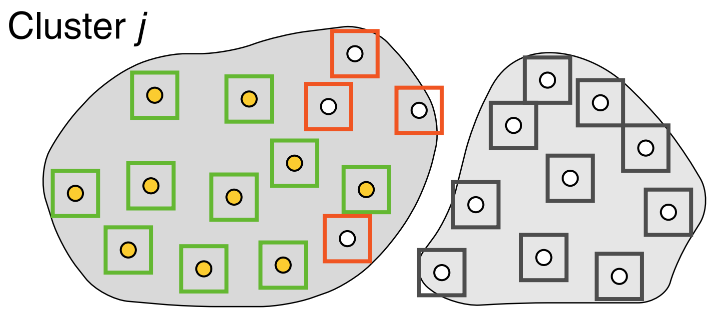
- $F_{i,j}$ is the F-measure of a cluster j computed *with respect to a class* i.

  Precision of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_j|$ (here: $Prec_{i,j} = 0.71$)

  Recall of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_i^*|$ (here: $Rec_{i,j} = 1.0$)

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Clustering)



(node-based analysis)

- Clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ and classification $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ of D.
- $F_{i,j}$ is the F-measure of a cluster j computed *with respect to a class* i.

  Precision of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_j|$ (here: $\mathit{Prec}_{i,j} = 0.71$)

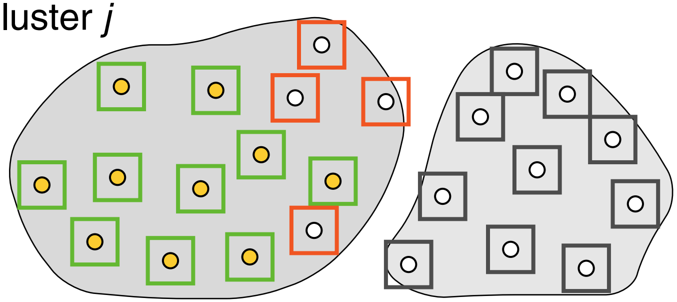
  Recall of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_i^*|$ (here: $\mathit{Rec}_{i,j} = 1.0$)

➜ Maximize micro-averaged F-measure for $\langle D, \mathcal{C}, \mathcal{C}^* \rangle$:

$$F = \sum_{i=1}^{l} \frac{|C_i^*|}{|D|} \cdot \max_{j=1,\ldots,k} \{F_{i,j}\}$$

Cluster Evaluation

(1) External Validity Measures: F-Measure (for a Clustering)



(node-based analysis)

- Clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ and classification $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ of D.
- $F_{i,j}$ is the F-measure of a cluster j computed *with respect to a class* i.

  Precision of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_j|$ (here: $\mathit{Prec}_{i,j} = 0.71$)

  Recall of cluster j with respect to class i is $|C_j \cap C_i^*|/|C_i^*|$ (here: $\mathit{Rec}_{i,j} = 1.0$)

➜ Maximize macro-averaged F-measure for $\langle D, \mathcal{C}, \mathcal{C}^* \rangle$:

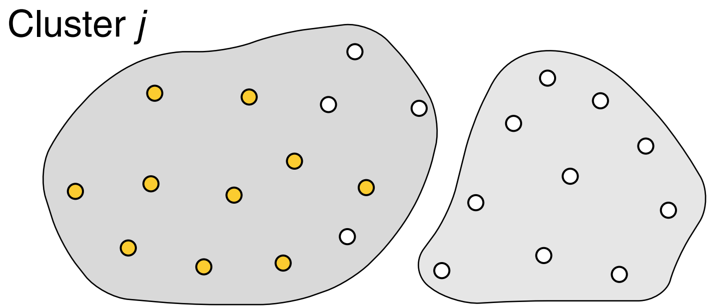
$$F = \frac{1}{l} \sum_{i=1}^{l} \max_{j=1,\ldots,k} \{F_{i,j}\}$$

Remarks:

- Micro averaging treats objects equally, whereas macro averaging treats classes equally.

Cluster Evaluation

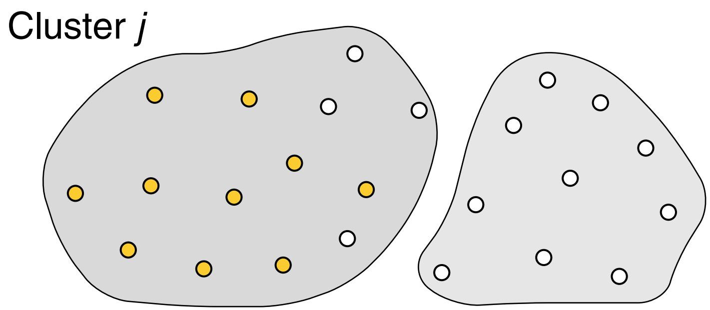
(1) External Validity Measures: Entropy



(node-based analysis)

Cluster Evaluation

(1) External Validity Measures: Entropy

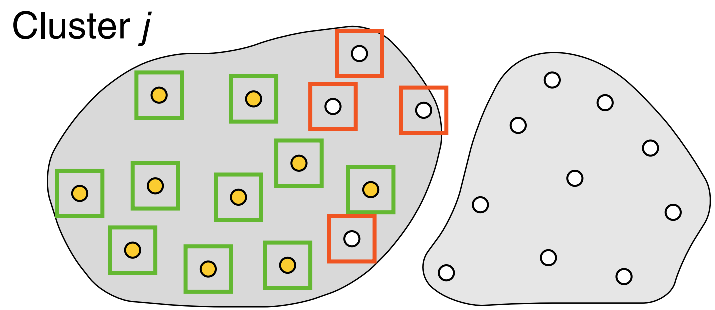


(node-based analysis)

- A cluster C acts as information source $\mathcal{L}$.
  $\mathcal{L}$ emits cluster labels $L_1, \ldots, L_l$ with probabilities $P(L_1), \ldots, P(L_l)$.

Cluster Evaluation

(1) External Validity Measures: Entropy



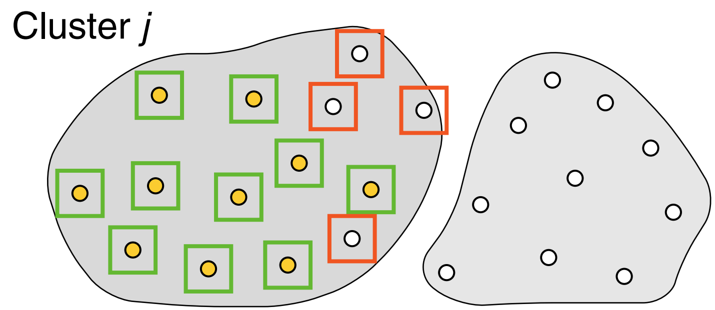
(node-based analysis)

- A cluster C acts as information source $\mathcal{L}$.
  $\mathcal{L}$ emits cluster labels $L_1, \ldots, L_l$ with probabilities $P(L_1), \ldots, P(L_l)$.

  $L_1 = \square$, $L_2 = \square$, $\hat{P}(\square) = 10/14$, $\hat{P}(\square) = 4/14$

Cluster Evaluation

(1) External Validity Measures: Entropy



(node-based analysis)

- A cluster C acts as information source $\mathcal{L}$.
  $\mathcal{L}$ emits cluster labels $L_1, \ldots, L_l$ with probabilities $P(L_1), \ldots, P(L_l)$.

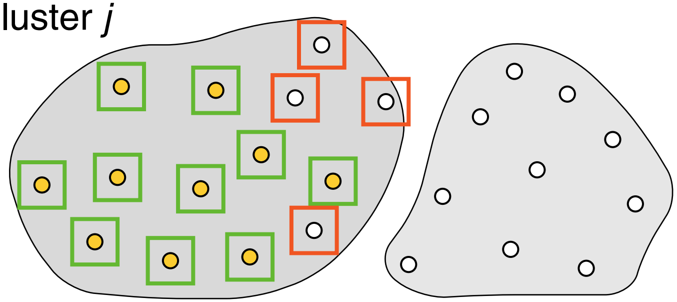
  $L_1 = \square$, $L_2 = \square$, $\hat{P}(\square) = 10/14$, $\hat{P}(\square) = 4/14$

- Entropy of $\mathcal{L}$: $H(\mathcal{L}) = -\sum_{i=1}^{l} P(L_i) \cdot \log_2(P(L_i))$

  Entropy of C_j wrt. $\mathcal{C}^*$: $H(C_j) = -\sum_{C_j \cap C_i^* \neq \emptyset} |C_j \cap C_i^*|/|C_j| \cdot \log_2(|C_j \cap C_i^*|/|C_j|)$

Cluster Evaluation

(1) External Validity Measures: Entropy



(node-based analysis)

- A cluster C acts as information source $\mathcal{L}$.
  $\mathcal{L}$ emits cluster labels $L_1, \ldots, L_l$ with probabilities $P(L_1), \ldots, P(L_l)$.

  $L_1 = \square$, $\quad L_2 = \square$, $\qquad \hat{P}(\square) = 10/14, \quad \hat{P}(\square) = 4/14$

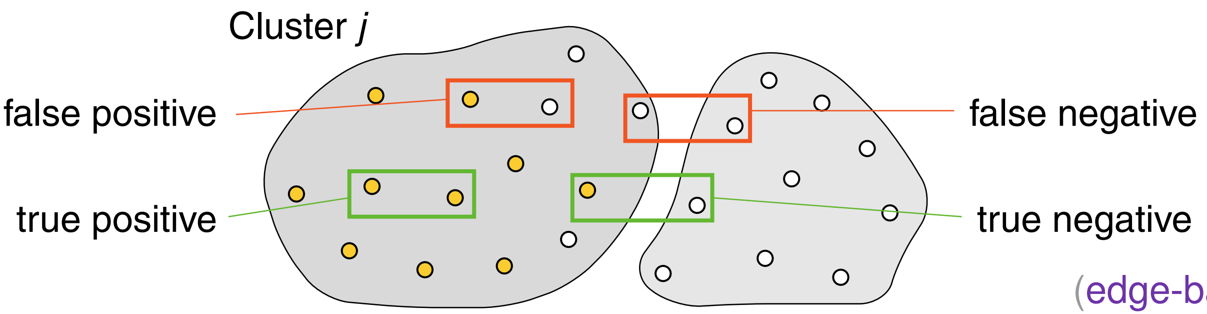
- Entropy of $\mathcal{L}$: $\qquad H(\mathcal{L}) = -\sum_{i=1}^{l} P(L_i) \cdot \log_2(P(L_i))$

  Entropy of C_j wrt. $\mathcal{C}^*$: $\quad H(C_j) = -\sum_{C_j \cap C_i^* \neq \emptyset} |C_j \cap C_i^*|/|C_j| \cdot \log_2(|C_j \cap C_i^*|/|C_j|)$

➜ Minimize entropy of $\mathcal{C}$ wrt. $\mathcal{C}^*$: $\quad H(\mathcal{C}) = \sum_{C_j \in \mathcal{C}} |C_j|/|D| \cdot H(C_j)$

Cluster Evaluation

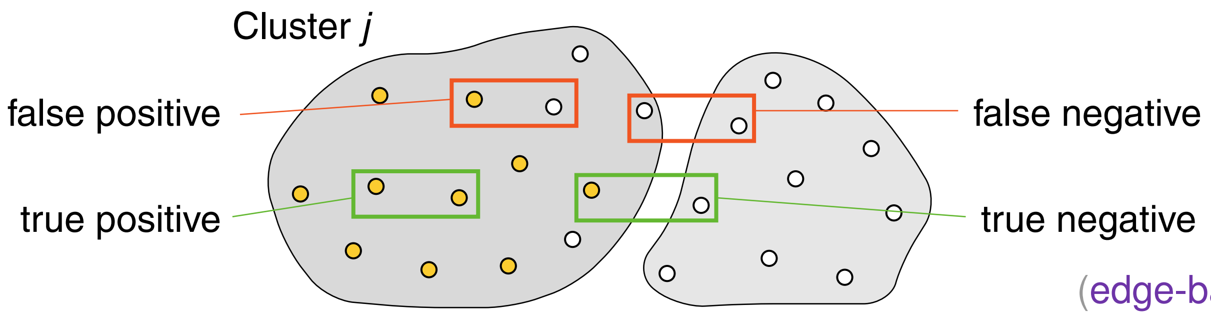
(1) External Validity Measures: Rand, Jaccard



(edge-based analysis)

Cluster Evaluation

(1) External Validity Measures: Rand, Jaccard



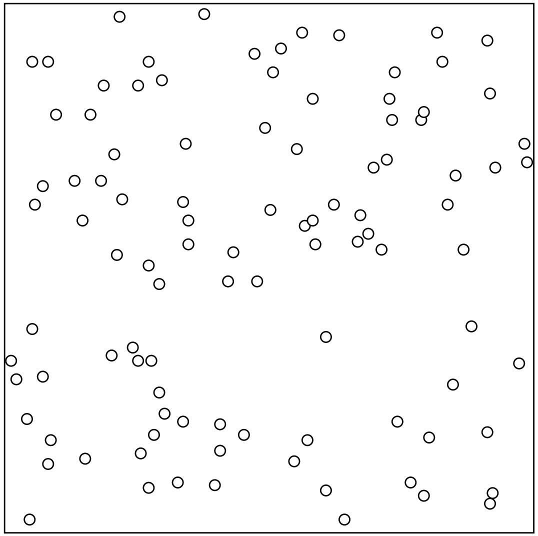
(edge-based analysis)

- $$R(\mathcal{C}) = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} = \frac{|TP| + |TN|}{n(n-1)/2},$$ with $n = |D|$

- $$J(\mathcal{C}) = \frac{|TP|}{|TP| + |FP| + |FN|}$$

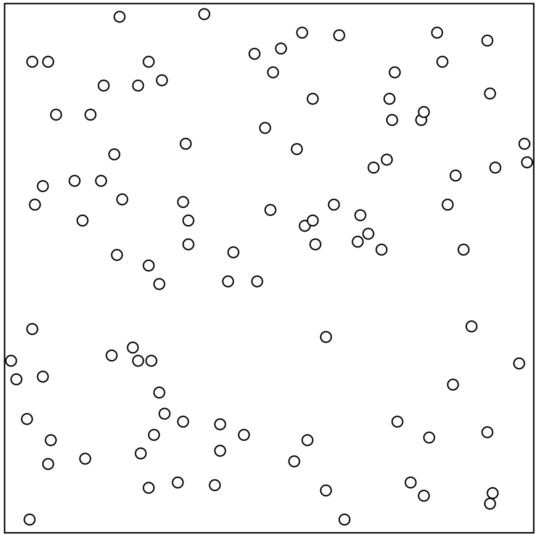
Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



Cluster Evaluation

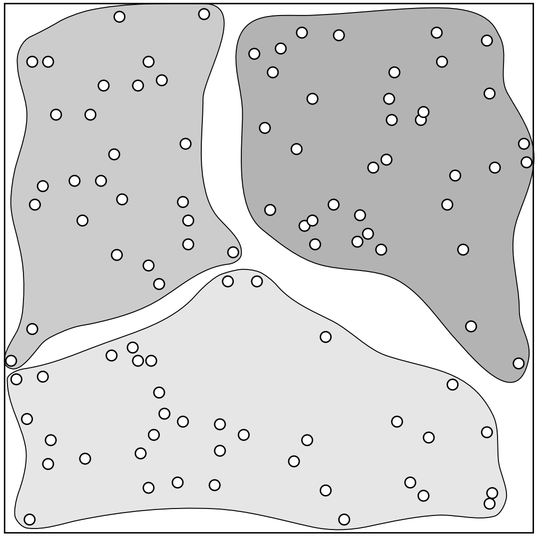
(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



$$\begin{pmatrix} 1.0 & 0.2 & 0.1 & 0.3 & \dots & 0.1 & 0.0 \\ - & 1.0 & 0.1 & 0.0 & \dots & 0.0 & 0.2 \\ & & & & \vdots & & \\ - & - & - & - & - & 1.0 & 0.6 \\ - & - & - & - & - & - & 1.0 \end{pmatrix}$$

Cluster Evaluation

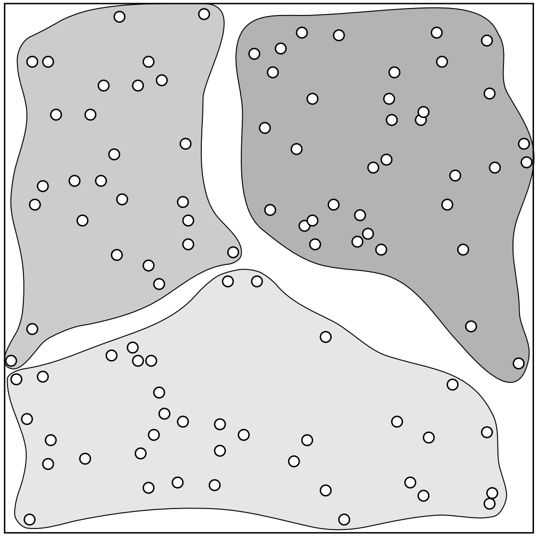
(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



$$\begin{pmatrix} 1.0 & 0.2 & 0.1 & 0.3 & \dots & 0.1 & 0.0 \\ - & 1.0 & 0.1 & 0.0 & \dots & 0.0 & 0.2 \\ & & & & \vdots & & \\ - & - & - & - & - & 1.0 & 0.6 \\ - & - & - & - & - & - & 1.0 \end{pmatrix}$$

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]

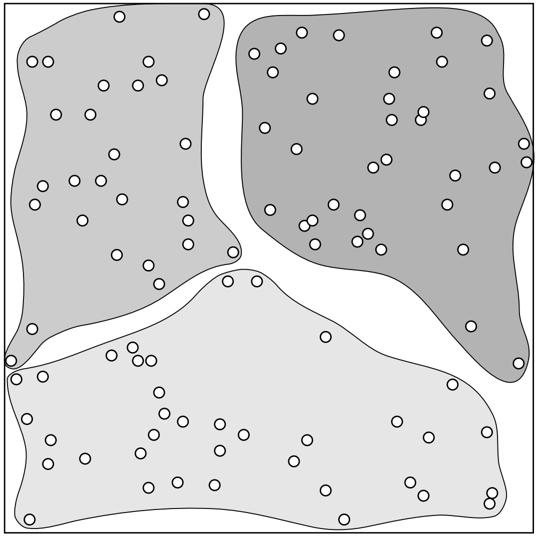


$$\begin{pmatrix} 1.0 & 0.2 & 0.1 & 0.3 & \ldots & 0.1 & 0.0 \\ - & 1.0 & 0.1 & 0.0 & \ldots & 0.0 & 0.2 \\ & & & & \vdots & & \\ - & - & - & - & - & 1.0 & 0.6 \\ - & - & - & - & - & - & 1.0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 & \ldots & 0 & 0 \\ - & 1 & 0 & 0 & \ldots & 0 & 1 \\ & & & & \vdots & & \\ - & - & - & - & - & 1 & 1 \\ - & - & - & - & - & - & 1 \end{pmatrix}$$

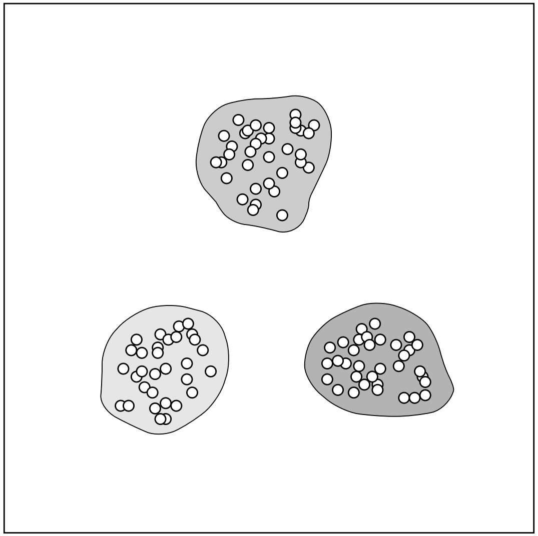
- Construct occurrence matrix based on cluster analysis.
- Compare similarity matrix to occurrence matrix: correlation τ

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



k-means
$\tau = 0.58$



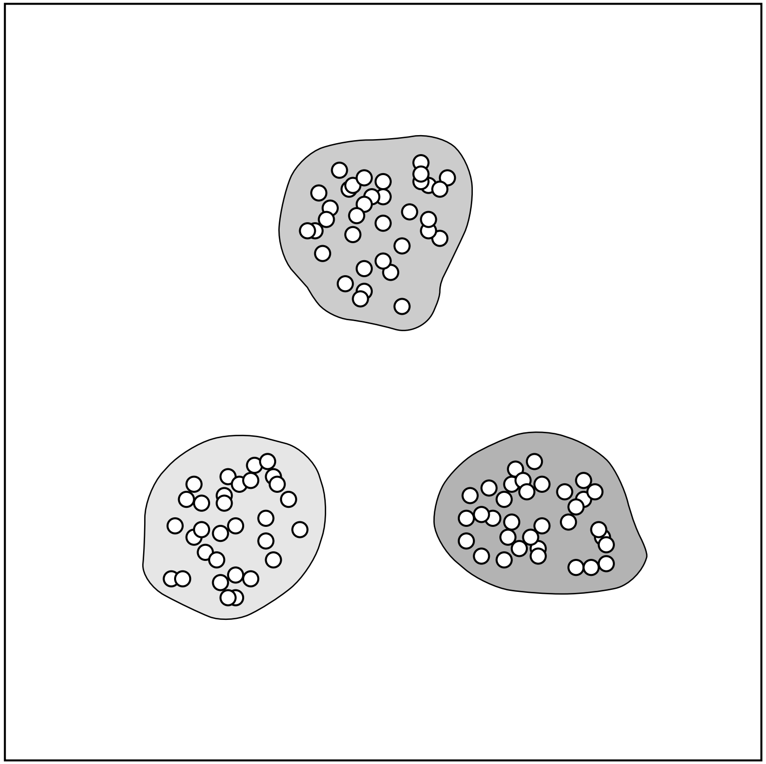
k-means
$\tau = 0.92$

$$\begin{pmatrix} 1.0 & 0.2 & 0.1 & 0.3 & \dots & 0.1 & 0.0 \\ - & 1.0 & 0.1 & 0.0 & \dots & 0.0 & 0.2 \\ & & & & \vdots & & \\ - & - & - & - & - & 1.0 & 0.6 \\ - & - & - & - & - & - & 1.0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 & \dots & 0 & 0 \\ - & 1 & 0 & 0 & \dots & 0 & 1 \\ & & & & \vdots & & \\ - & - & - & - & - & 1 & 1 \\ - & - & - & - & - & - & 1 \end{pmatrix}$$

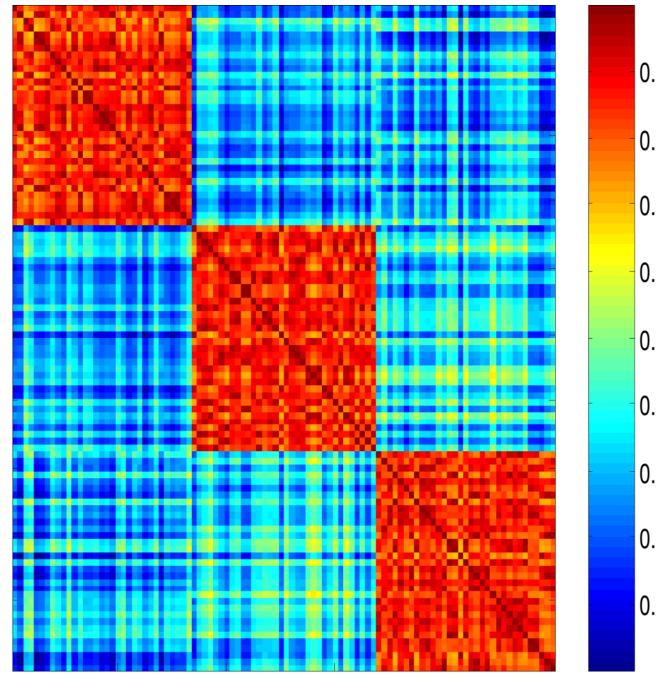
- Construct occurrence matrix based on cluster analysis.
- Compare similarity matrix to occurrence matrix: correlation τ

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



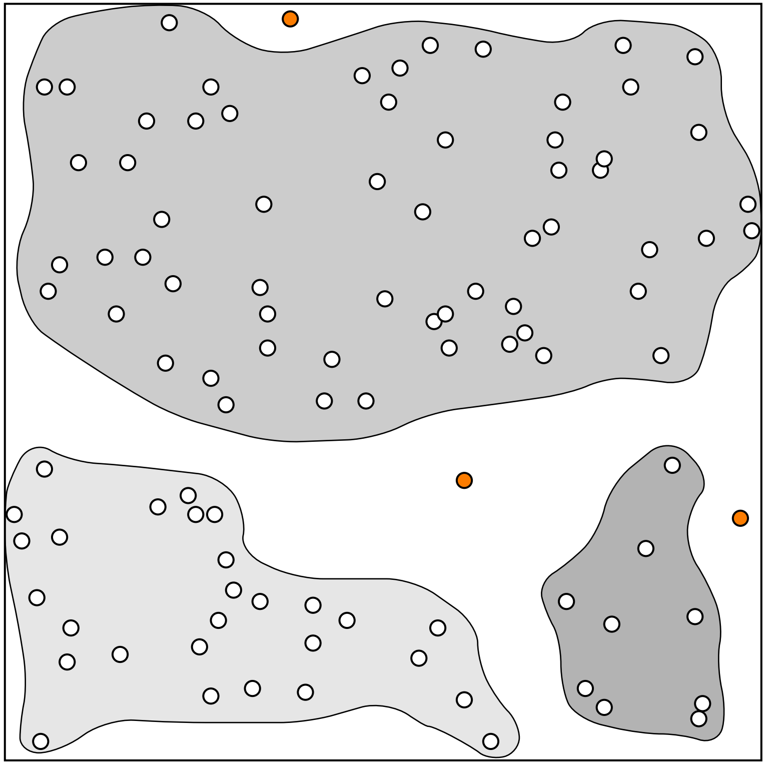
k-means at structured data.



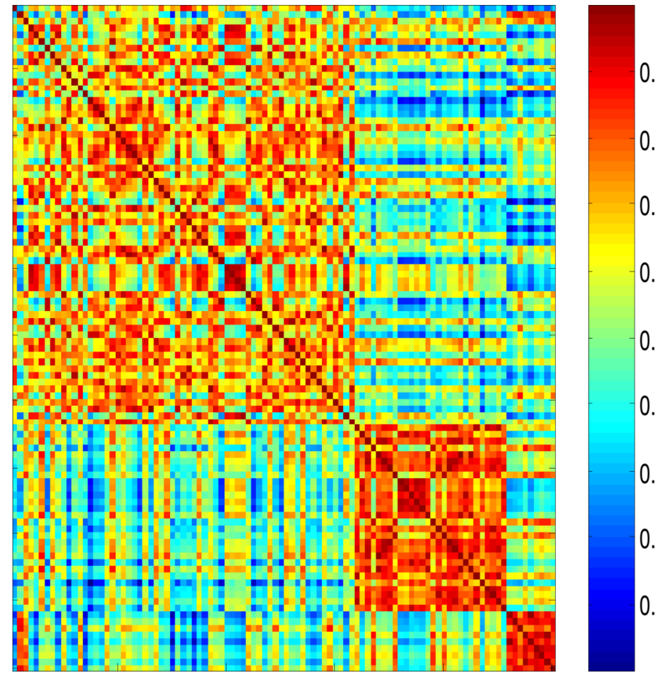
Similarity matrix sorted by cluster label.

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



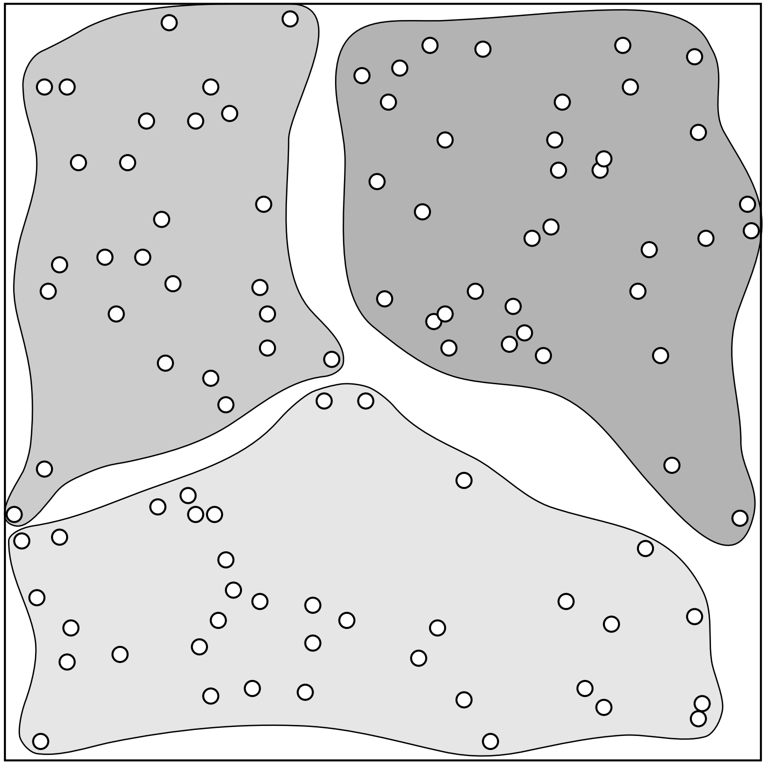
DBSCAN at random data.



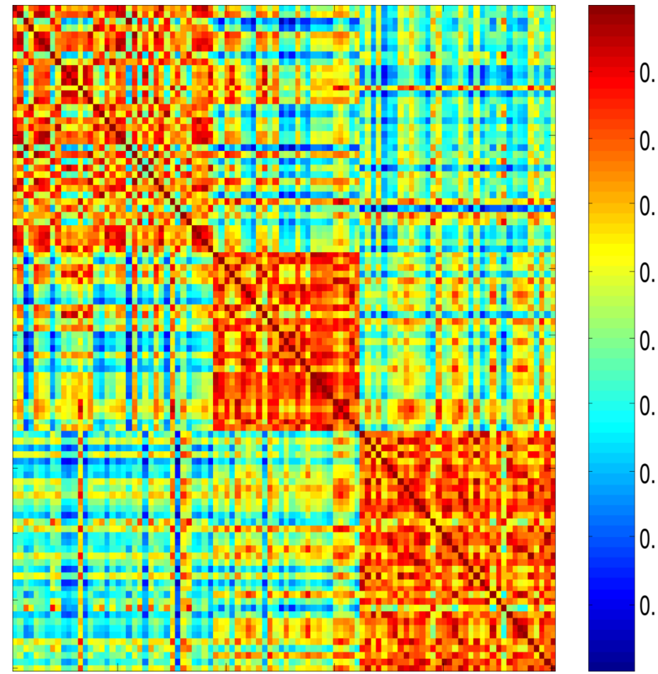
Similarity matrix sorted by cluster label.

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



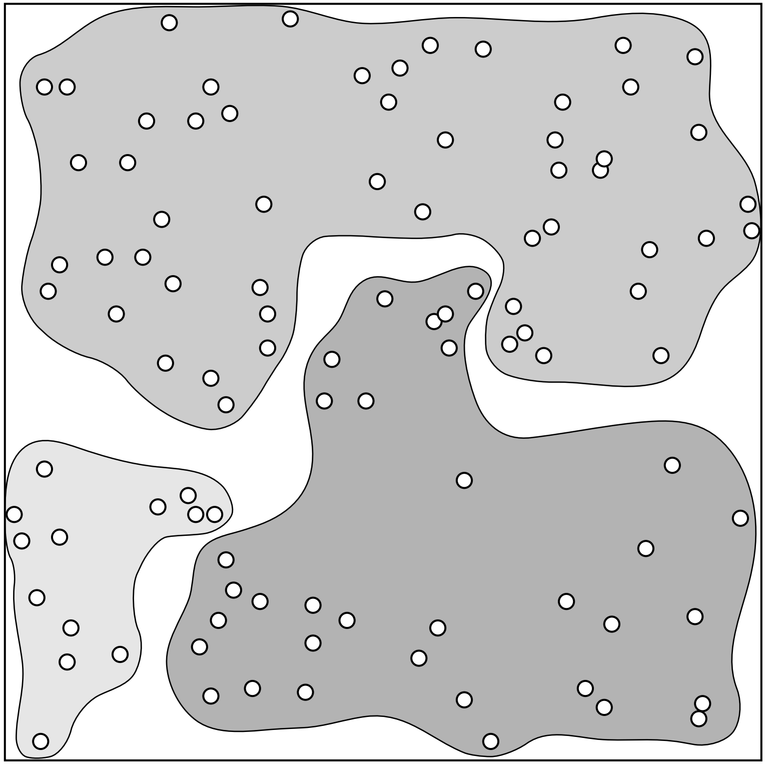
k-means at random data.



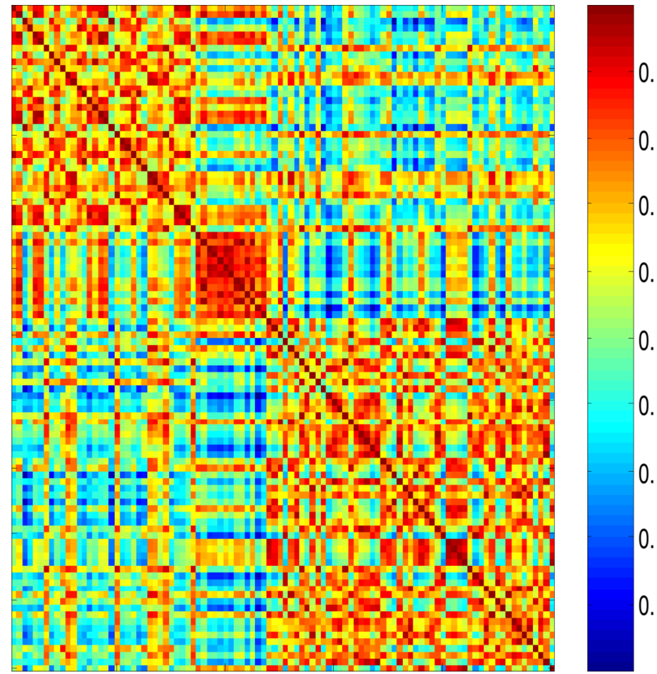
Similarity matrix sorted by cluster label.

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



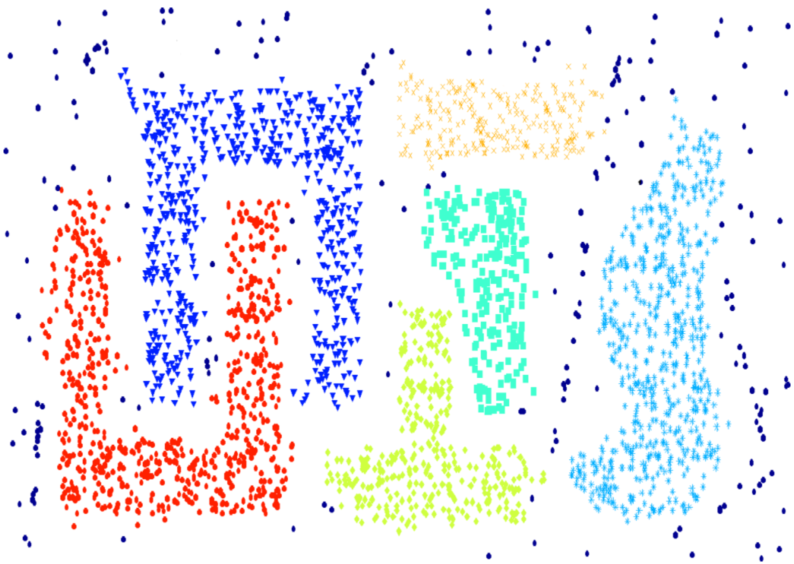
Complete link at random data.



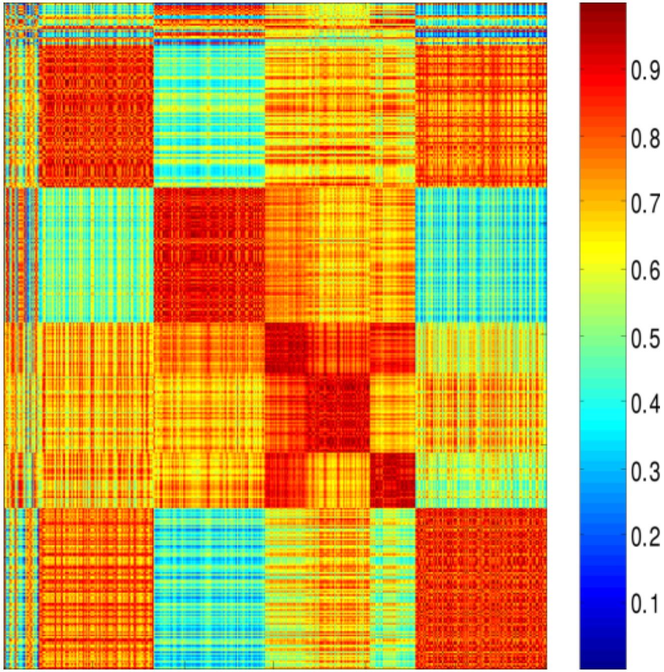
Similarity matrix sorted by cluster label.

Cluster Evaluation

(2) Internal Validity Measures: Edge Correlation [Tan/Steinbach/Kumar 2005]



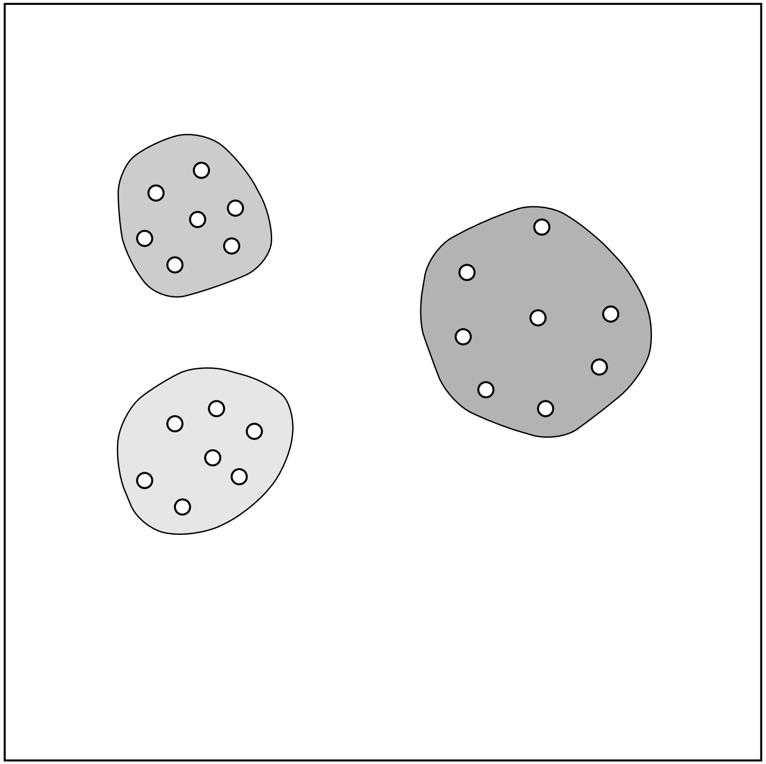
DBSCAN at structured data.



Similarity matrix sorted by cluster label.

Cluster Evaluation

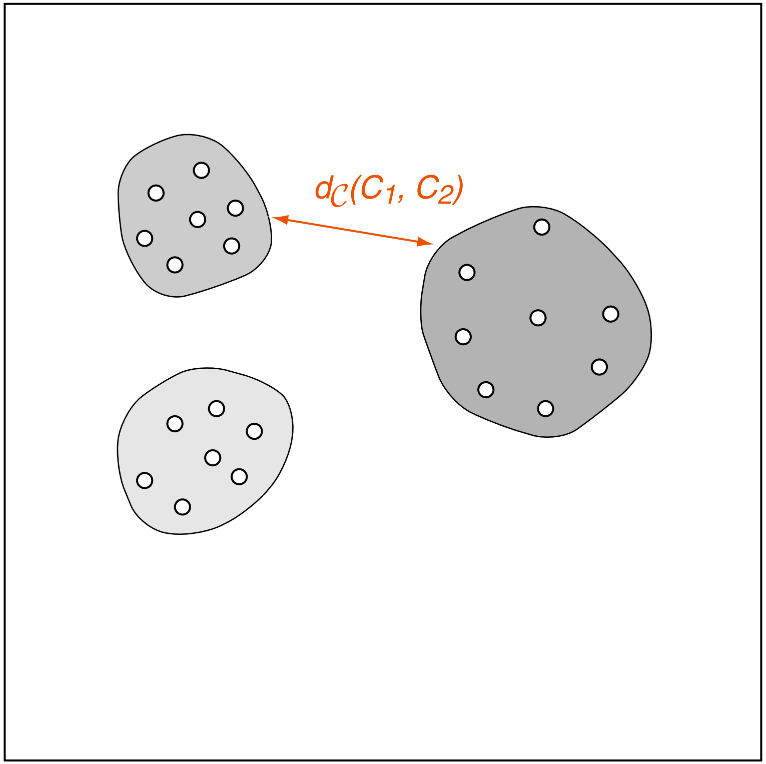
(2) Internal Validity Measures: Structural Analysis



- Distance between two clusters, $d_{\mathcal{C}}(C_1, C_2)$.
- Diameter of a cluster, $\Delta(C)$.
- Scatter within a cluster, $\sigma^2(C)$, RSS.

Cluster Evaluation

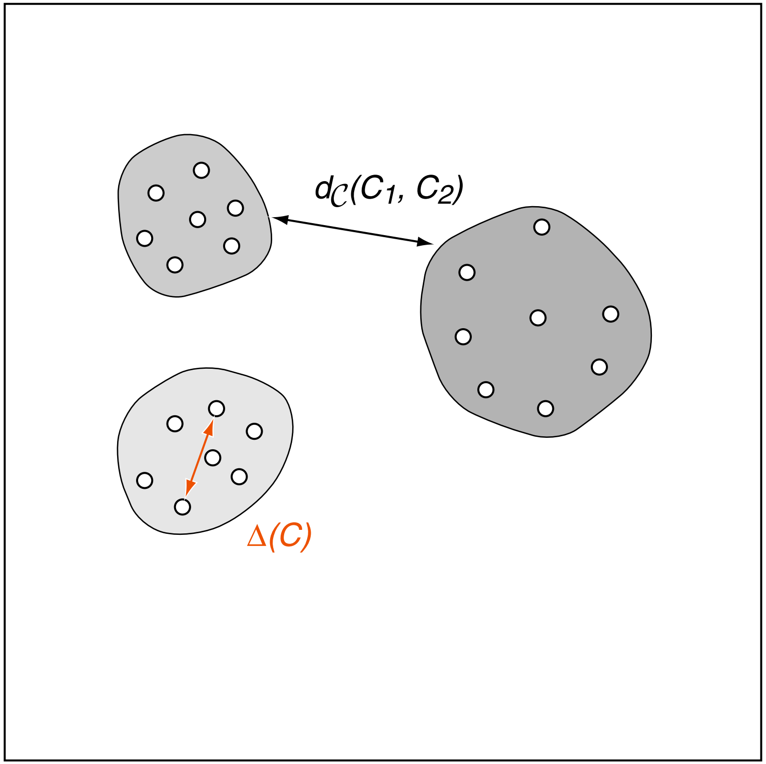
(2) Internal Validity Measures: Structural Analysis



- Distance between two clusters, $d_{\mathcal{C}}(C_1, C_2)$.
- Diameter of a cluster, $\Delta(C)$.
- Scatter within a cluster, $\sigma^2(C)$, RSS.

Cluster Evaluation

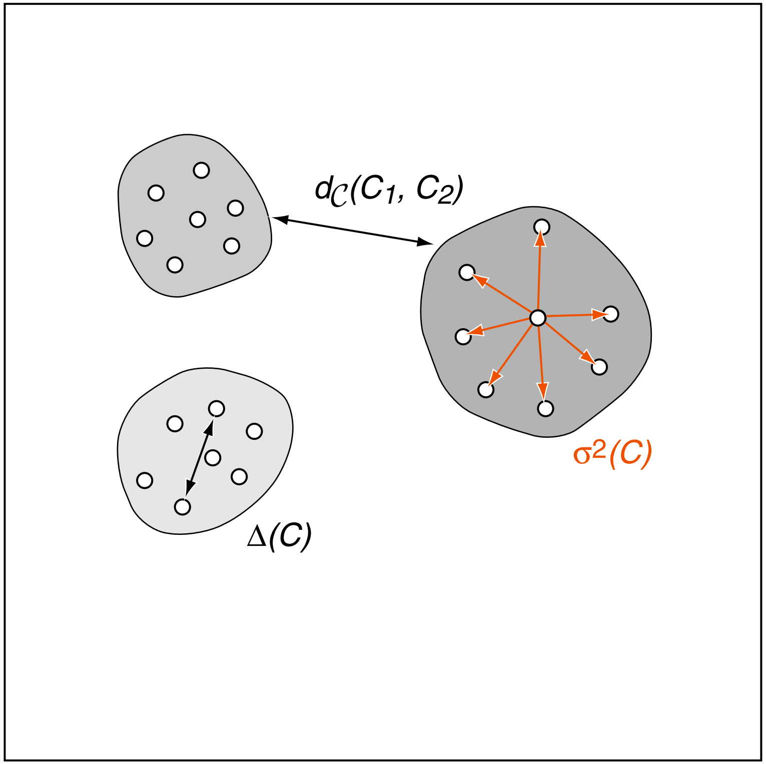
(2) Internal Validity Measures: Structural Analysis



- Distance between two clusters, $d_{\mathcal{C}}(C_1, C_2)$.
- Diameter of a cluster, $\Delta(C)$.
- Scatter within a cluster, $\sigma^2(C)$, RSS.

Cluster Evaluation

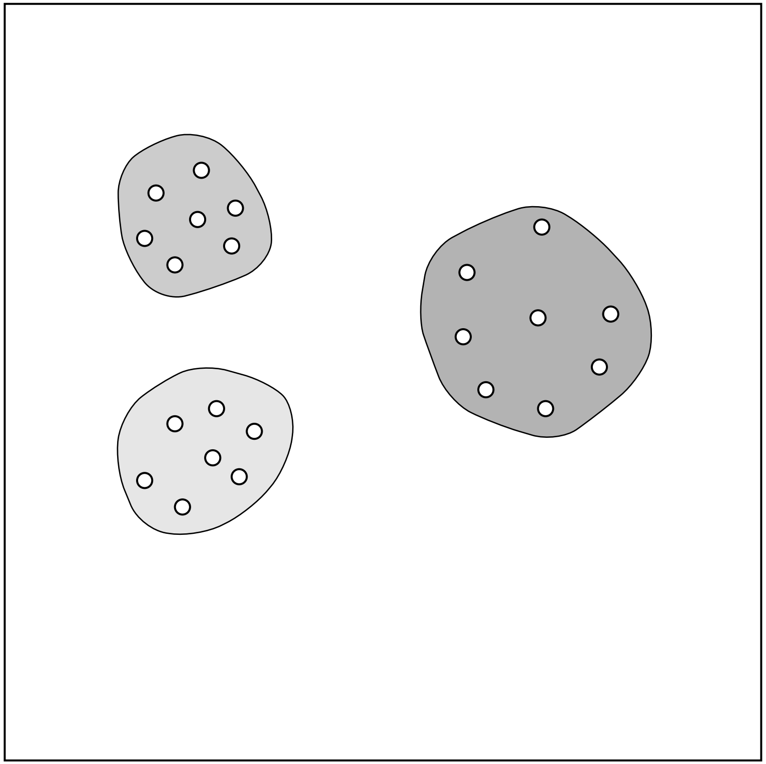
(2) Internal Validity Measures: Structural Analysis



- Distance between two clusters, $d_{\mathcal{C}}(C_1, C_2)$.
- Diameter of a cluster, $\Delta(C)$.
- Scatter within a cluster, $\sigma^2(C)$, RSS.

Cluster Evaluation

(2) Internal Validity Measures: Dunn Index

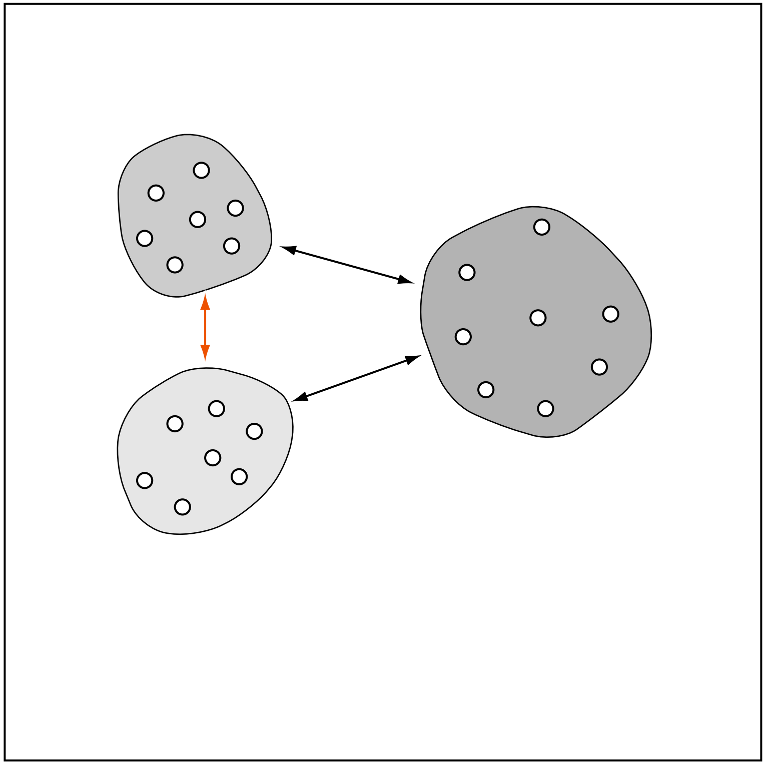


$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}},$$

$$I(\mathcal{C}) \rightarrow \max$$

Cluster Evaluation

(2) Internal Validity Measures: Dunn Index



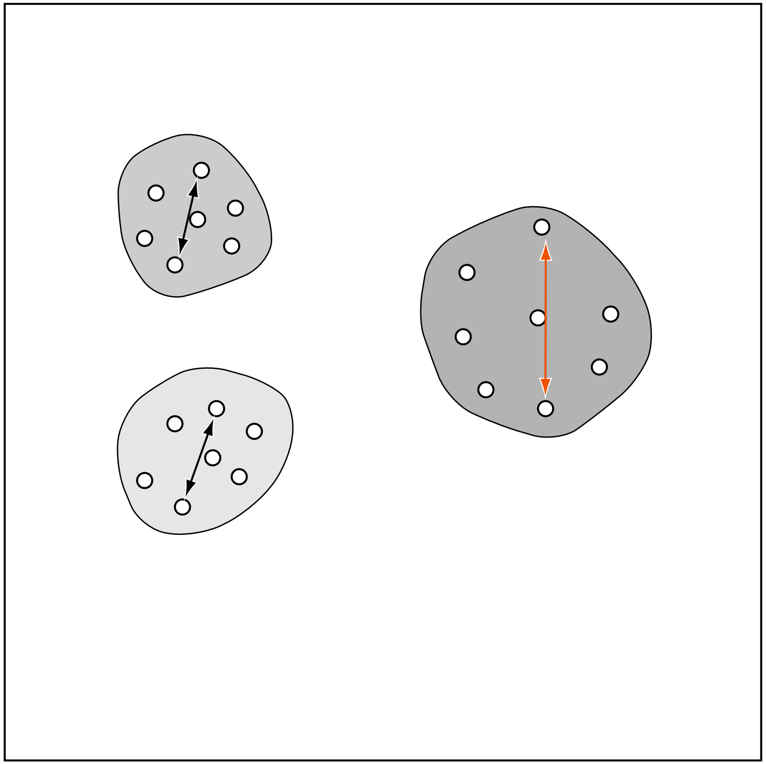
Cluster distance

$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}},$$

$$I(\mathcal{C}) \to \max$$

Cluster Evaluation

(2) Internal Validity Measures: Dunn Index



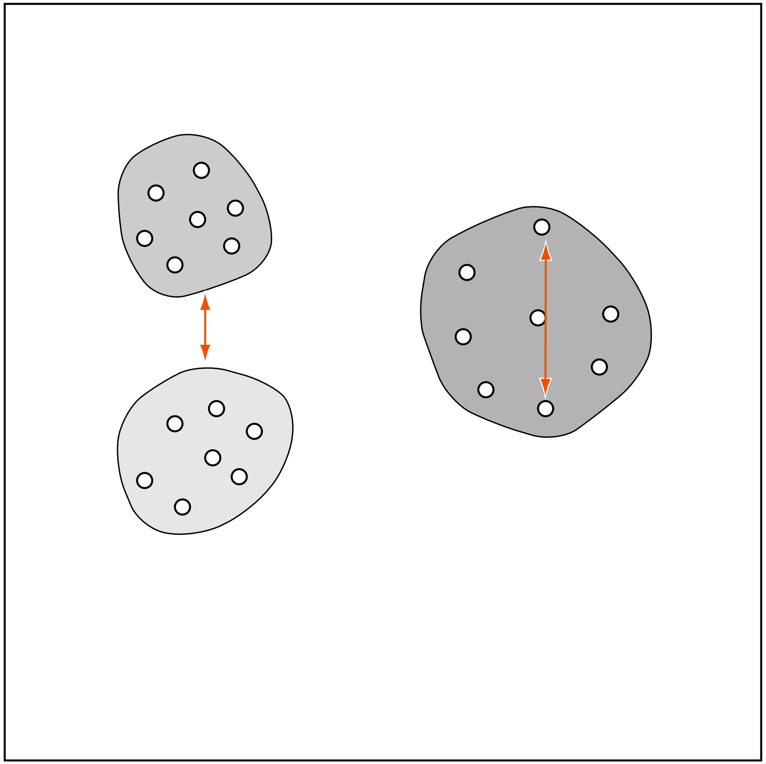
$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}},$$

$$I(\mathcal{C}) \rightarrow \max$$

Cluster diameter

Cluster Evaluation

(2) Internal Validity Measures: Dunn Index



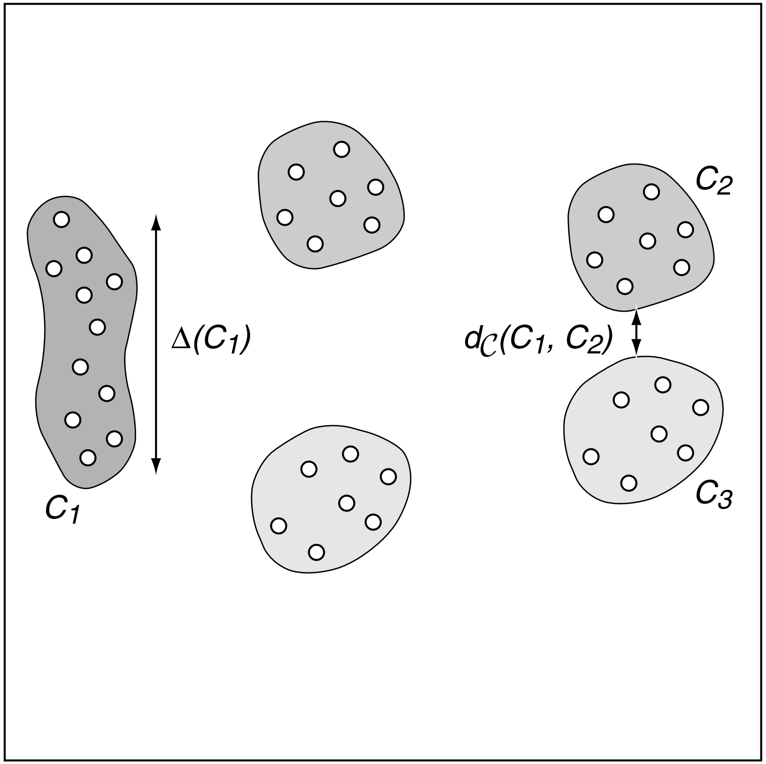
$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}},$$

$$I(\mathcal{C}) \rightarrow \max$$

- Dunn is susceptible to noise.
- Dunn is biased towards the worst substructure in a clustering (cf. the $\min$).

Cluster Evaluation

(2) Internal Validity Measures: Dunn Index



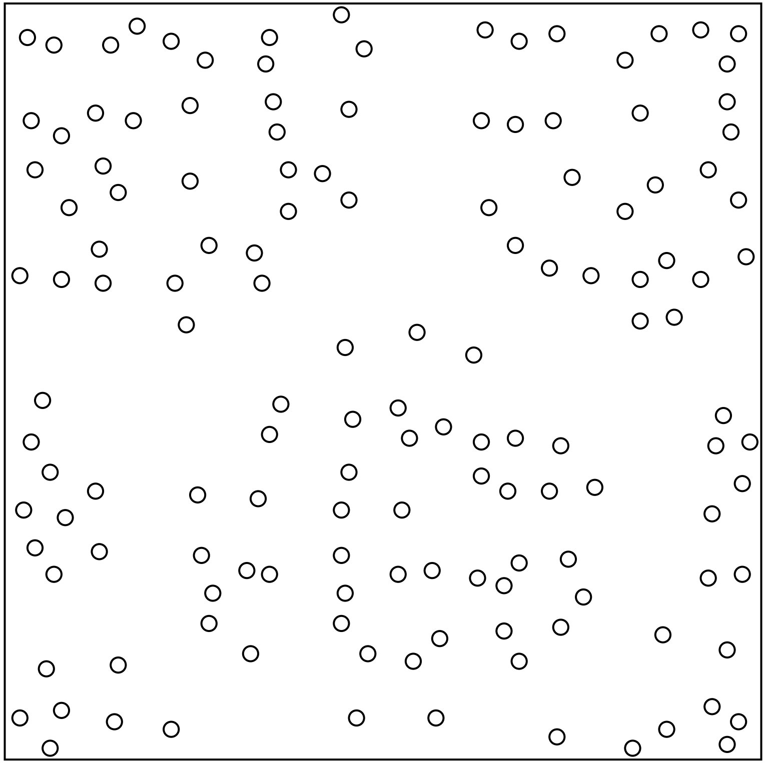
$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}},$$

$$I(\mathcal{C}) \rightarrow \max$$

- Dunn is susceptible to noise.
- Dunn is biased towards the worst substructure in a clustering (cf. the $\min$).
- Dunn value too low since distances and diameters are not put into relation.

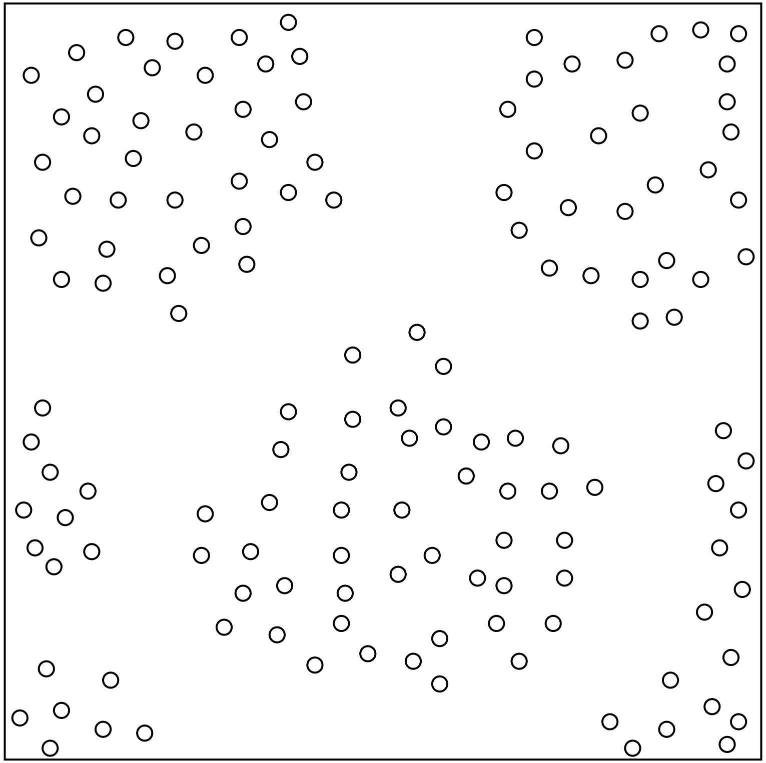
Cluster Evaluation

(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Cluster Evaluation

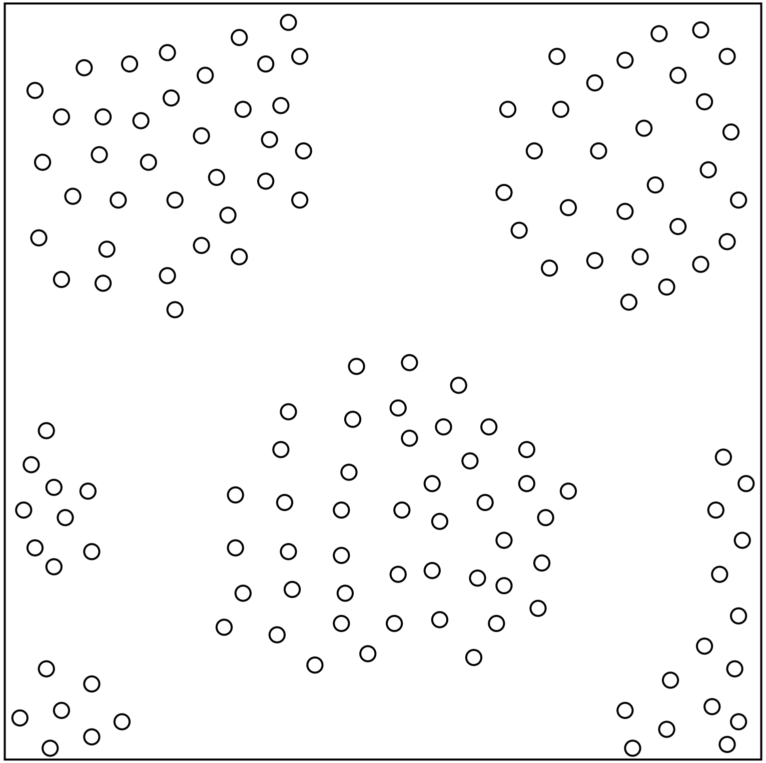
(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Different models (feature sets) yield different similarity graphs.

Cluster Evaluation

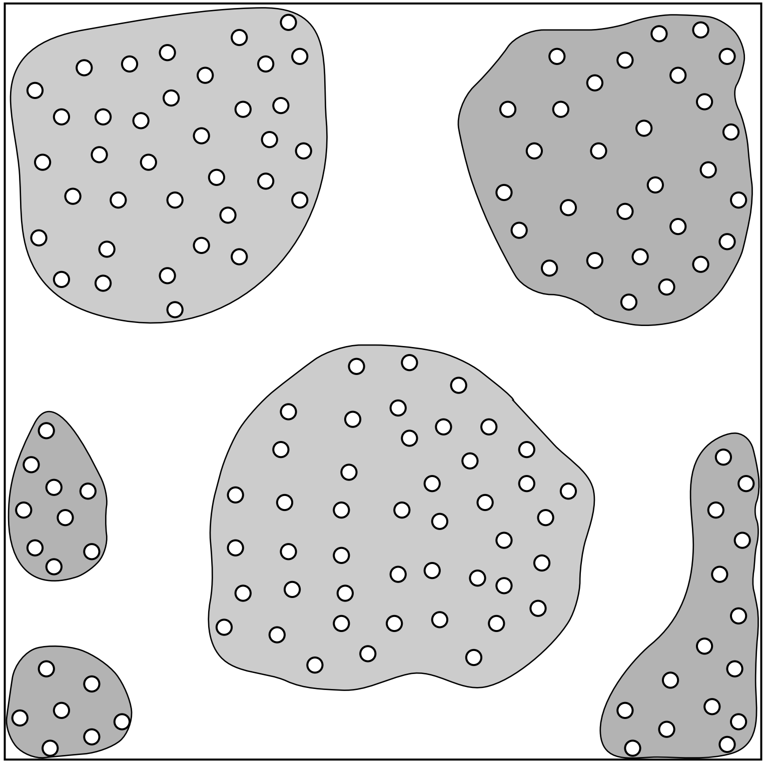
(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Different models (feature sets) yield different similarity graphs.

Cluster Evaluation

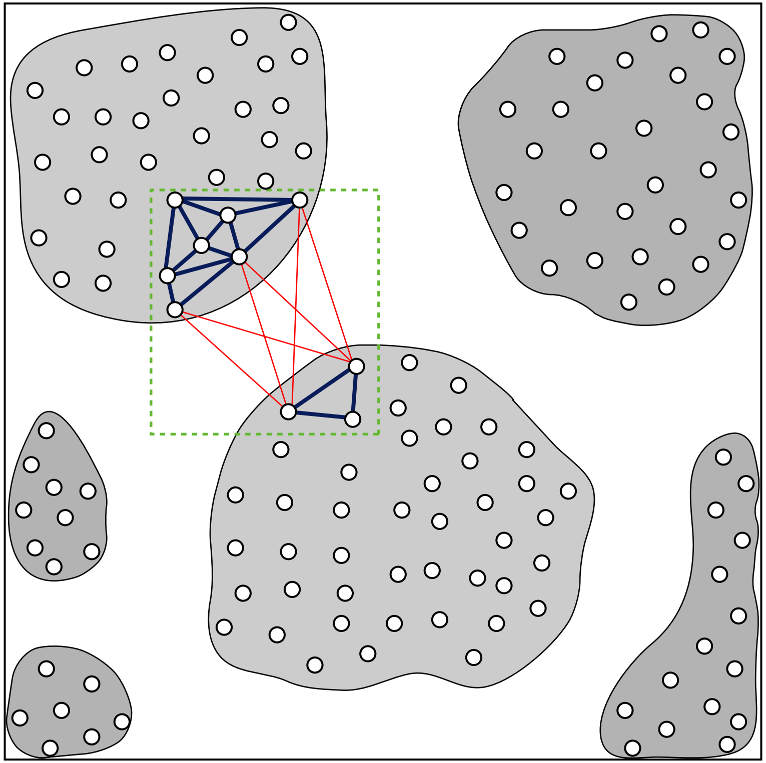
(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Compare (for alternative clusterings) the similarity density within the clusters to the average similarity of the entire graph.

Cluster Evaluation

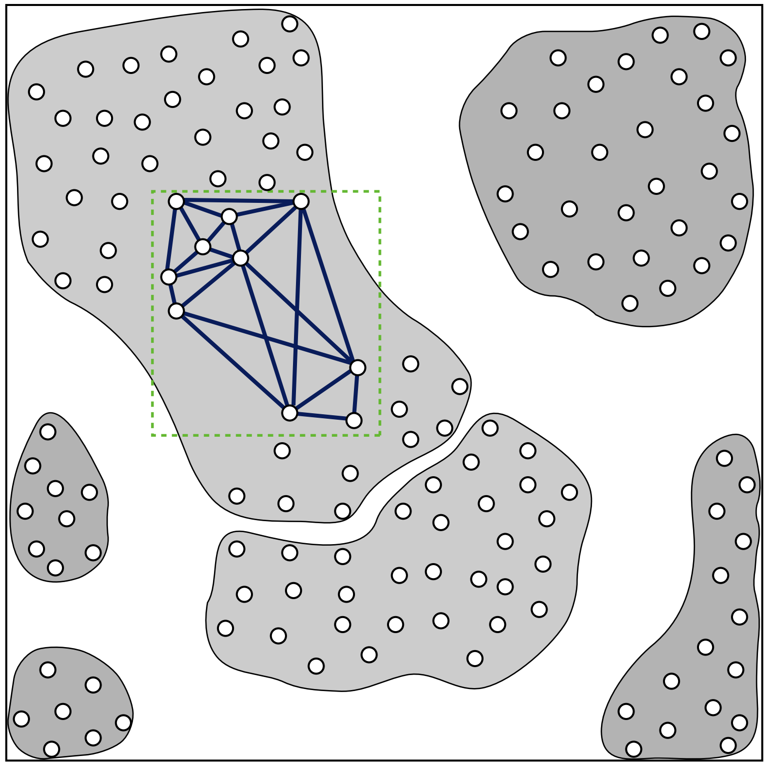
(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Compare (for alternative clusterings) the similarity density within the clusters to the average similarity of the entire graph.

Cluster Evaluation

(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]



Compare (for alternative clusterings) the similarity density within the clusters to the average similarity of the entire graph.

Cluster Evaluation

(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]

Graph $G = \langle V, E \rangle$:

- G is called sparse if $|E| = O(|V|)$, G is called dense if $|E| = O(|V|^2)$

➜ the density θ computes from the equation $|E| = |V|^\theta$

Cluster Evaluation

(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]

Graph $G = \langle V, E \rangle$:

- G is called sparse if $|E| = O(|V|)$, G is called dense if $|E| = O(|V|^2)$

➜ the density θ computes from the equation $|E| = |V|^\theta$

Similarity graph $G = \langle V, E, w \rangle$:

- $|E| \sim w(G) = \sum_{e \in E} w(e)$

➜ the density θ computes from the equation $w(G) = |V|^\theta$

Cluster Evaluation

(2) Internal Validity Measures: Expected Density ρ [Stein/Meyer zu Eissen 2007]

Graph $G = \langle V, E \rangle$:

- G is called sparse if $|E| = O(|V|)$, G is called dense if $|E| = O(|V|^2)$

➜ the density θ computes from the equation $|E| = |V|^\theta$

Similarity graph $G = \langle V, E, w \rangle$:

- $|E| \sim w(G) = \sum_{e \in E} w(e)$

➜ the density θ computes from the equation $w(G) = |V|^\theta$

Cluster C_i induces subgraph G_i :

➜ the expected density ρ relates the density of G_i to the density average in G

$$\rho(G_i) = \frac{w(G_i)}{|V_i|^\theta}$$

Cluster Evaluation

(3) Relative Validity Measures: Elbow Criterion

1. Hyperparameter alternatives of a clustering algorithm: $\pi_1, \ldots, \pi_m$
   - number of centroids for k-means
   - stopping level for hierarchical algorithms
   - neighborhood size for DBSCAN

2. *Set* of clusterings $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_{\pi_1}, \ldots, \mathcal{C}_{\pi_m}\}$ associated with $\pi_1, \ldots, \pi_m$.

3. Points of an error curve $\{\big(\pi_i,\, e(\mathcal{C}_{\pi_i})\big) \mid i = 1, \ldots, m\}$.

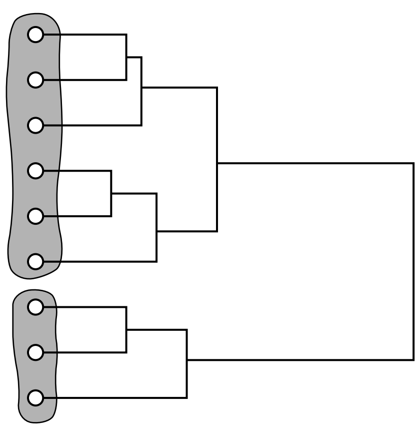
Cluster Evaluation

(3) Relative Validity Measures: Elbow Criterion

1. Hyperparameter alternatives of a clustering algorithm: $\pi_1, \ldots, \pi_m$
   - number of centroids for k-means
   - stopping level for hierarchical algorithms
   - neighborhood size for DBSCAN

2. *Set* of clusterings $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_{\pi_1}, \ldots, \mathcal{C}_{\pi_m}\}$ associated with $\pi_1, \ldots, \pi_m$.

3. Points of an error curve $\{\big(\pi_i,\, e(\mathcal{C}_{\pi_i})\big) \mid i = 1, \ldots, m\}$.



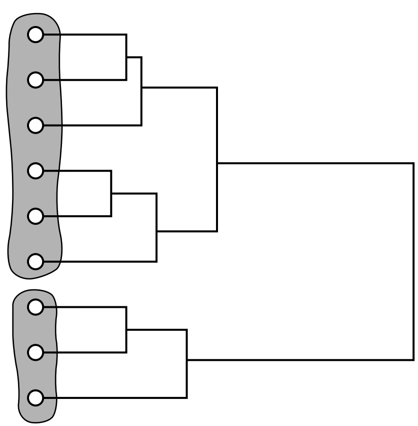
Cluster Evaluation

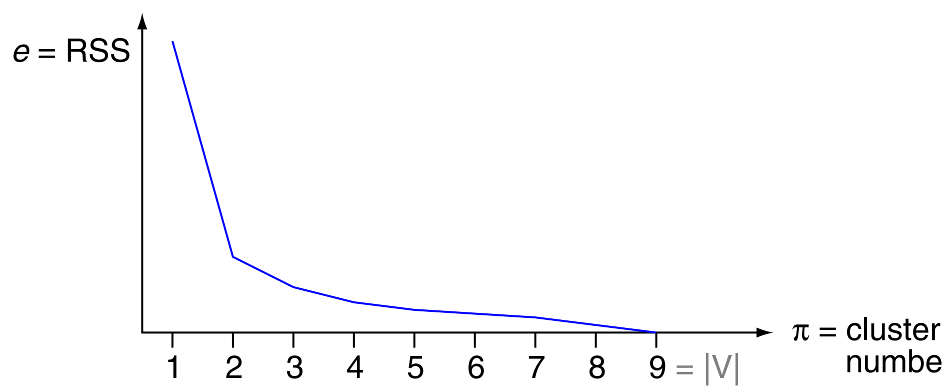
(3) Relative Validity Measures: Elbow Criterion

1. Hyperparameter alternatives of a clustering algorithm: $\pi_1, \ldots, \pi_m$
   - number of centroids for k-means
   - stopping level for hierarchical algorithms
   - neighborhood size for DBSCAN

2. *Set* of clusterings $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_{\pi_1}, \ldots, \mathcal{C}_{\pi_m}\}$ associated with $\pi_1, \ldots, \pi_m$.

3. Points of an error curve $\{\big(\pi_i,\, e(\mathcal{C}_{\pi_i})\big) \mid i = 1, \ldots, m\}$.





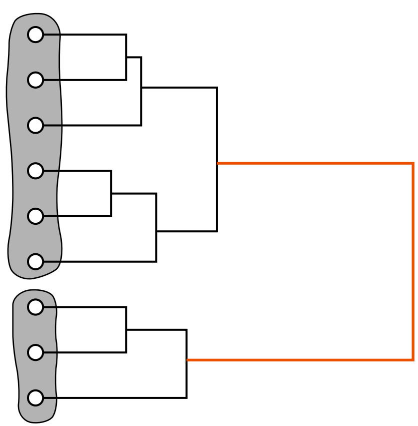
Cluster Evaluation

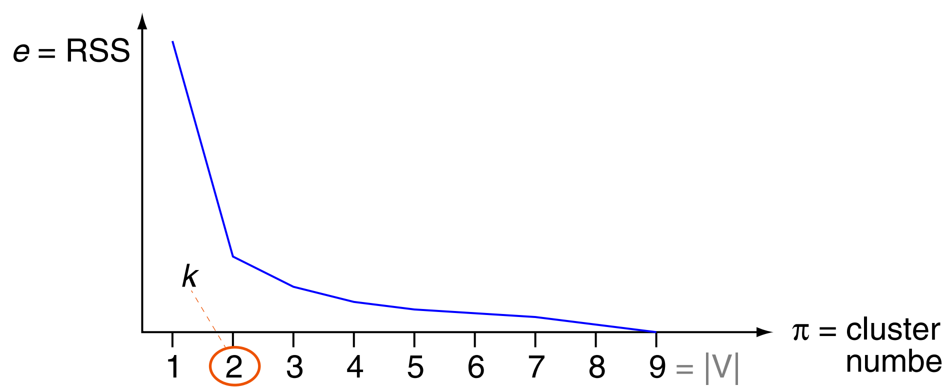
(3) Relative Validity Measures: Elbow Criterion

1. Hyperparameter alternatives of a clustering algorithm: $\pi_1, \ldots, \pi_m$
   - number of centroids for k-means
   - stopping level for hierarchical algorithms
   - neighborhood size for DBSCAN

2. *Set* of clusterings $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_{\pi_1}, \ldots, \mathcal{C}_{\pi_m}\}$ associated with $\pi_1, \ldots, \pi_m$.

3. Points of an error curve $\{\big(\pi_i,\, e(\mathcal{C}_{\pi_i})\big) \mid i = 1, \ldots, m\}$.

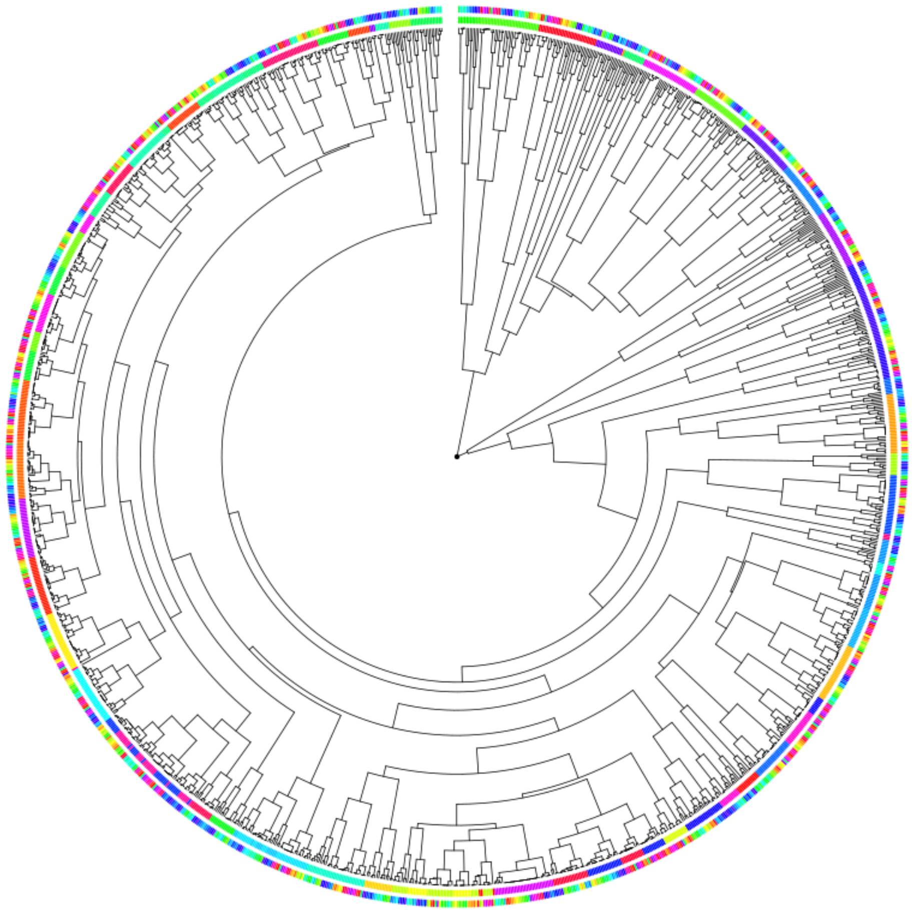




4. Find the point that maximizes error reduction with regard to its successor.

Cluster Evaluation

(3) Relative Validity Measures: Elbow Criterion



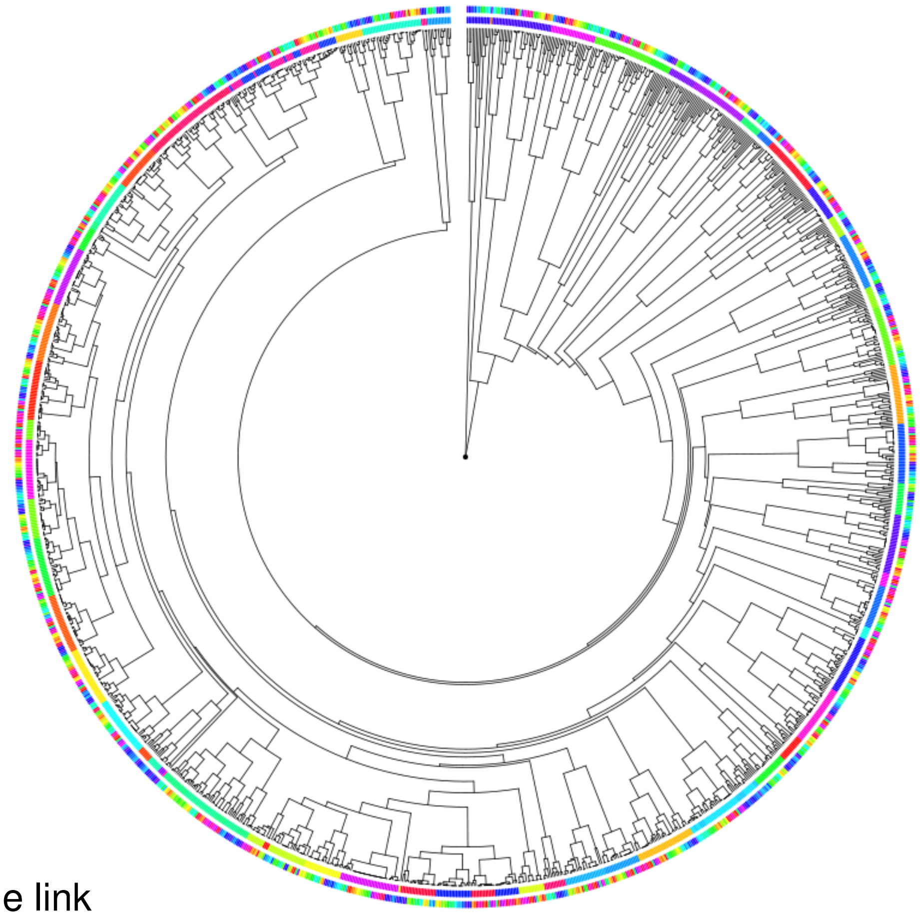
$d_{\mathcal{C}}$: Hamming distance
Merging: complete link

http://cs.jhu.edu/~razvanm/fs-expedition/2.6.x.html

Relations between 1377 file systems for Linux Kernel 2.6.0. [Razvan Musaloiu 2009]

Cluster Evaluation

(3) Relative Validity Measures: Elbow Criterion



$d_{\mathcal{C}}$: Hamming distance
Merging: group average link

http://cs.jhu.edu/~razvanm/fs-expedition/2.6.x.html

Relations between 1377 file systems for Linux Kernel 2.6.0. [Razvan Musaloiu 2009]

Cluster Evaluation

Correlation between External and Internal Measures

In the wild, we are not given a reference classification.

➜ An external evaluation is not possible.

(though many papers report on such experiments)

➜ Resort to an internal evaluation.

(connectivity, squared error sums, distance-diameter heuristics, etc.)

Cluster Evaluation

Correlation between External and Internal Measures

In the wild, we are not given a reference classification.

➜ An external evaluation is not possible.
(though many papers report on such experiments)

➜ Resort to an internal evaluation.
(connectivity, squared error sums, distance-diameter heuristics, etc.)

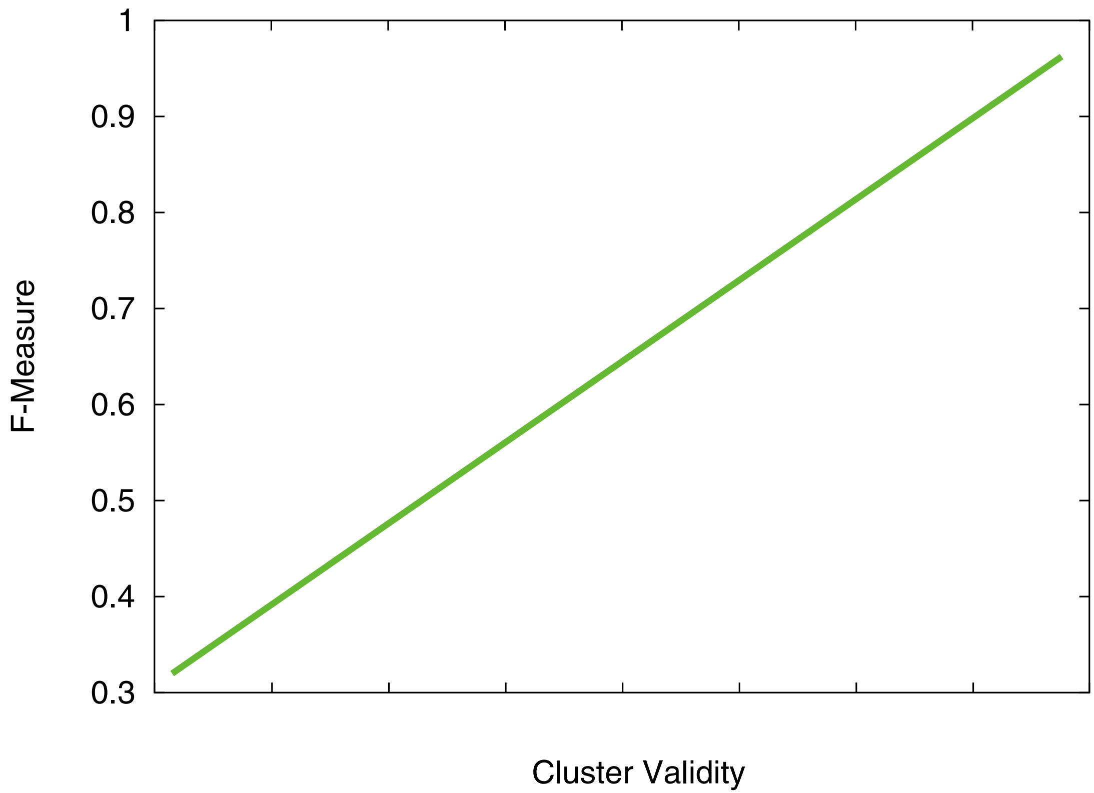
"To which extent can an internal evaluation ϕ be used to predict for a clustering its distance from the best reference classification—say, to predict the F-measure?"

$$\underset{\phi}{\text{argmax}} \, \{\tau\langle X, Y\rangle \mid x = F(\mathcal{C}),\ y = \phi(\mathcal{C}),\ \mathcal{C} \in \boldsymbol{\mathcal{C}}\}$$

[Stein/Meyer zu Eissen 2007]

Cluster Evaluation

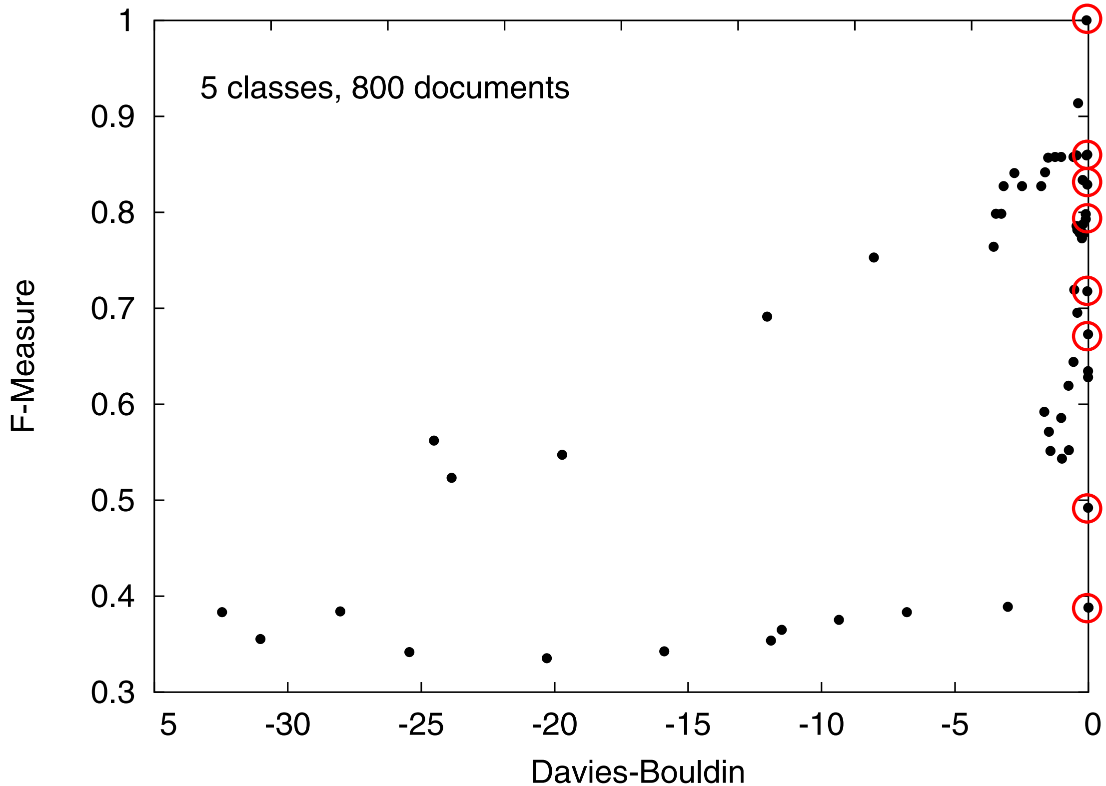
Correlation between External and Internal Measures



Perfect correlation (desired).

Cluster Evaluation

Correlation between External and Internal Measures

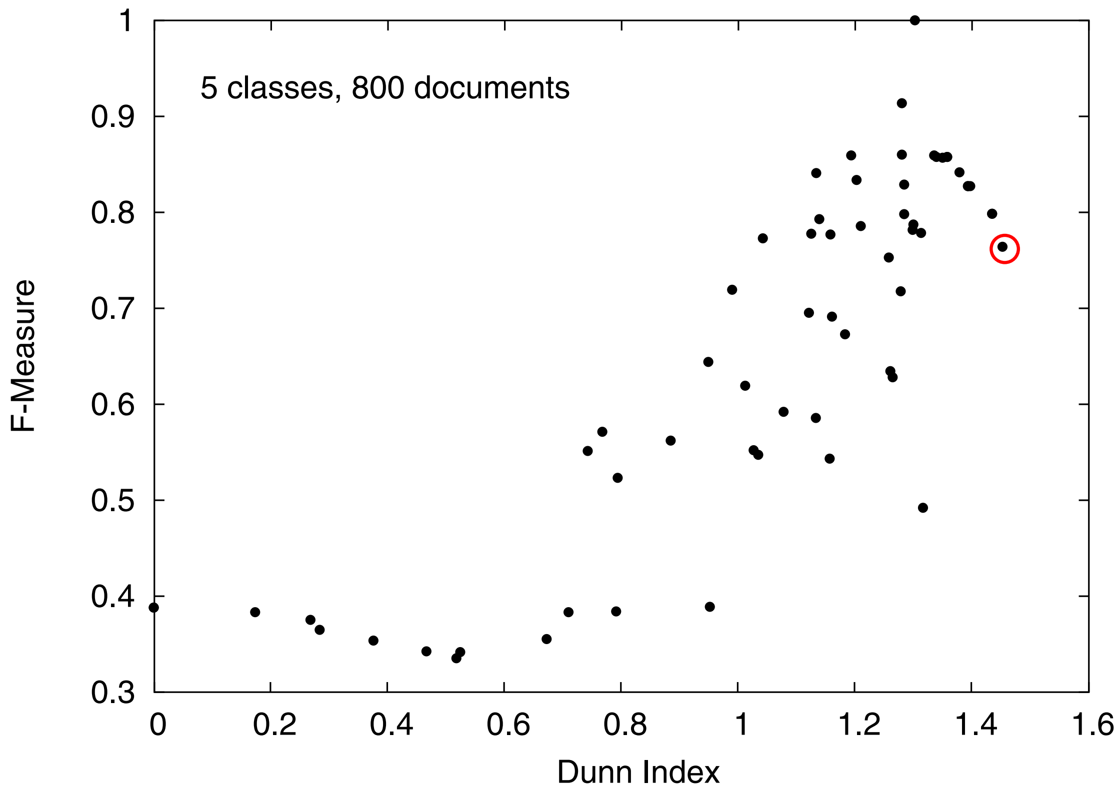


Davies-Bouldin: $$\frac{1}{k} \cdot \sum_{i=1}^{k} \max_j \frac{s(C_i) + s(C_j)}{d_{\mathcal{C}}(C_i, C_j)}$$

Prefers spherical clusters.

Cluster Evaluation

Correlation between External and Internal Measures

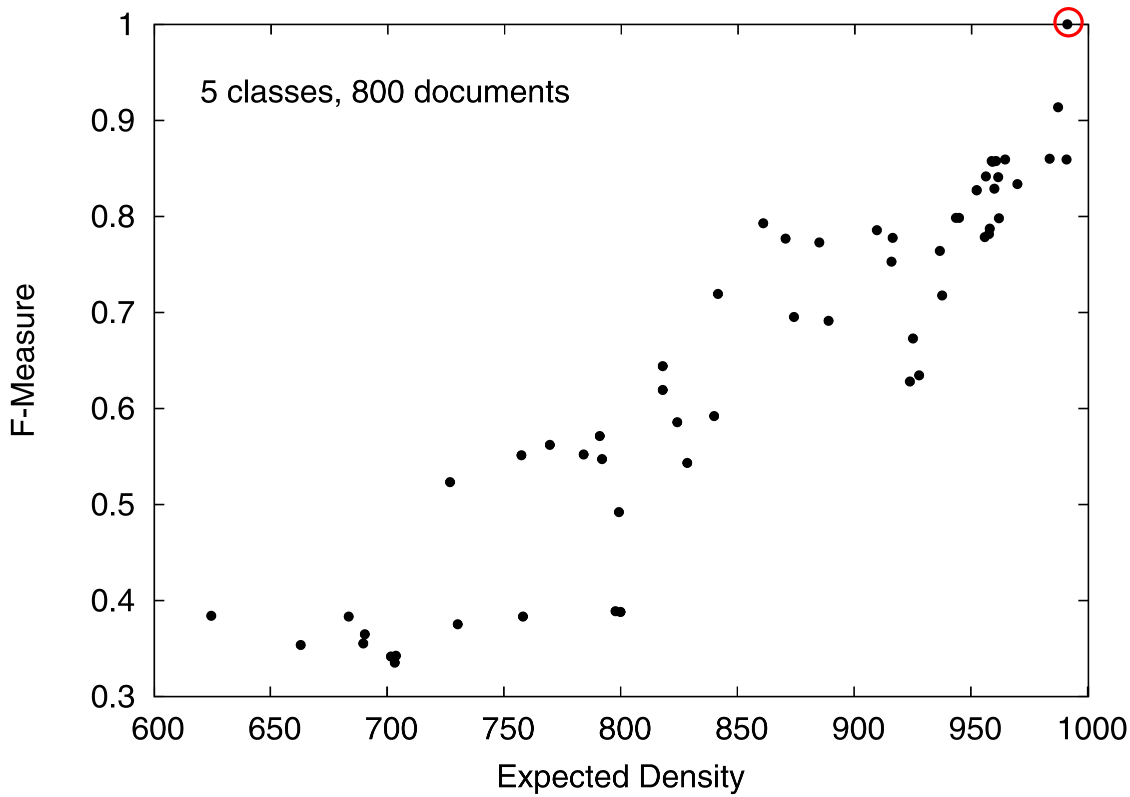


Dunn Index: $$\frac{\min_{i \neq j}\{d_{\mathcal{C}}(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}}$$

Maximizes dilatation = inter/intra-cluster-diameter.

Cluster Evaluation

Correlation between External and Internal Measures



Expected Density: $$\bar{\rho} = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}$$

Independent of cluster forms and sizes.