

Cluster Validation

Ke Chen

Outline

- Motivation and Background
- Internal index
 - Motivation and general ideas
 - Variance-based internal indexes
 - Application: finding the “proper” cluster number
- External index
 - Motivation and general ideas
 - Rand Index
 - Application: evaluating performance with benchmarks
- Summary

Motivation and Background

- **Motivation**

Supervised classification

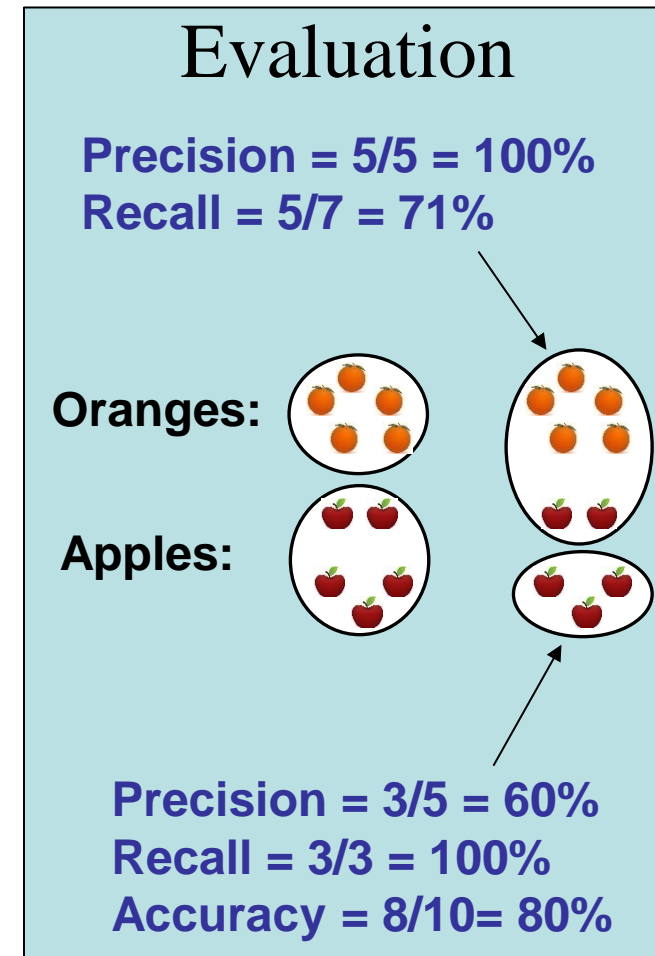
- Class labels known for ground truth
- Accuracy, precision, recall →

Clustering analysis

- No class labels
- Evaluation still demanded

Validation needs to

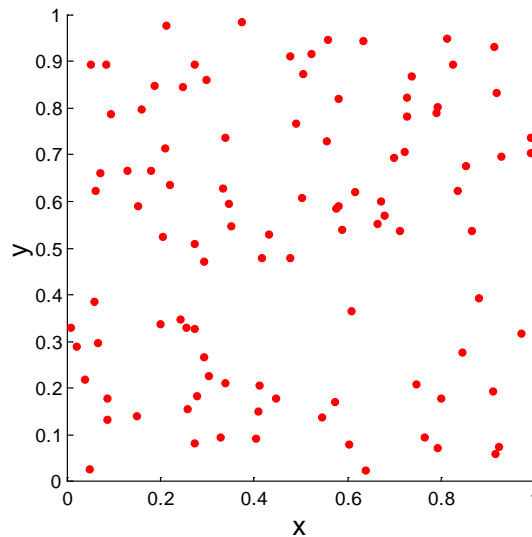
- Compare clustering algorithms
- Solve the number of clusters
- Avoid finding patterns in noise
- Find the “best” clusters from data



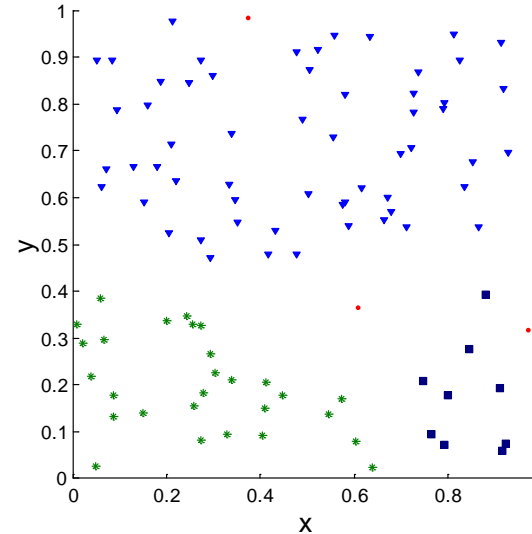
Motivation and Background

- Illustrative Example: which one is the “best”?

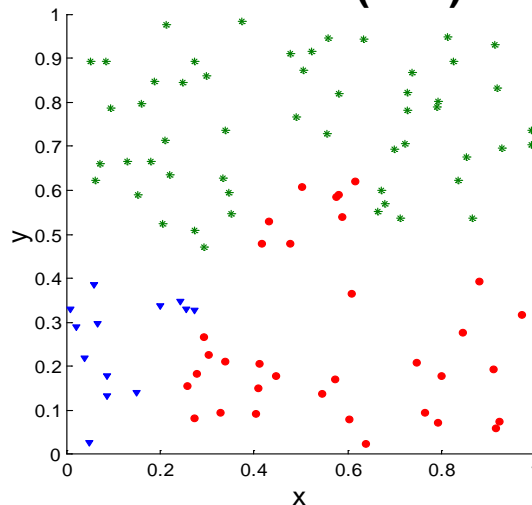
Data Set (Random Points)



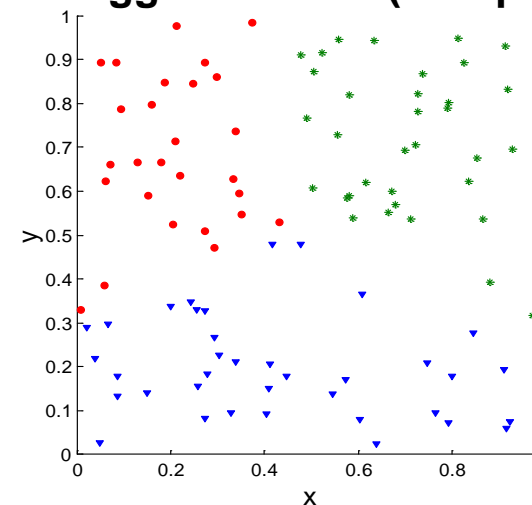
K-means ($K=3$)



K-means ($K=3$)

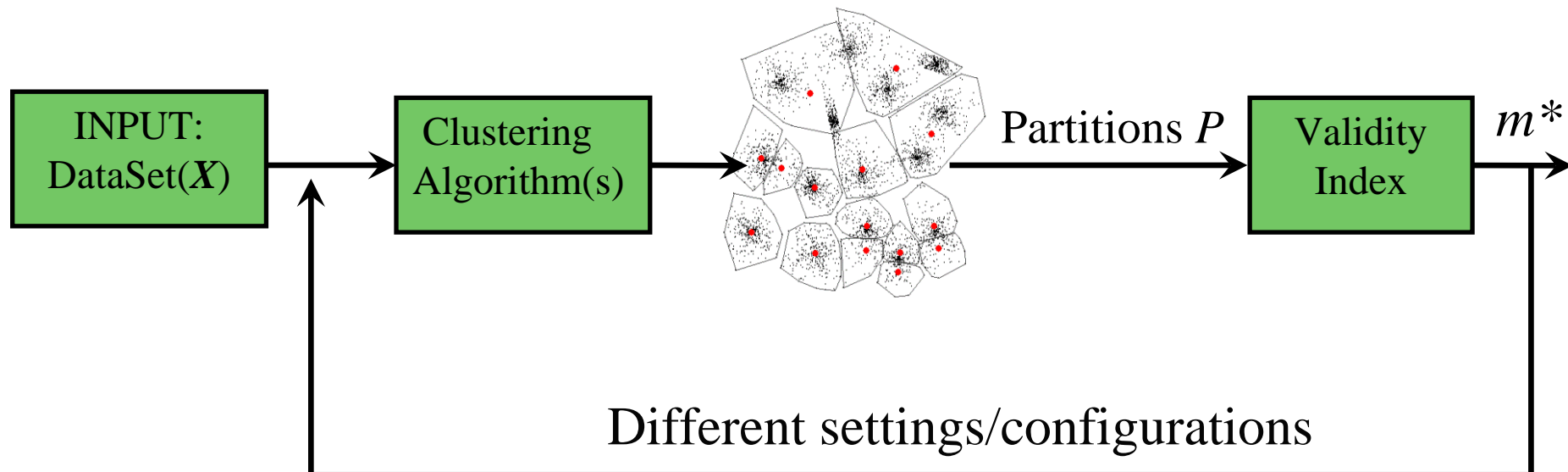


Agglomerative (Complete Link)



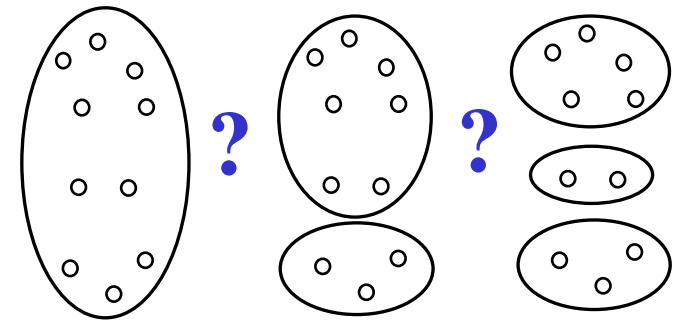
Motivation and Background

- *Cluster validation* refers to procedures that evaluate the results of clustering in a *quantitative* and *objective* fashion.
 - How to be “quantitative”: To employ the measures.
 - How to be “objective”: To validate the measures!

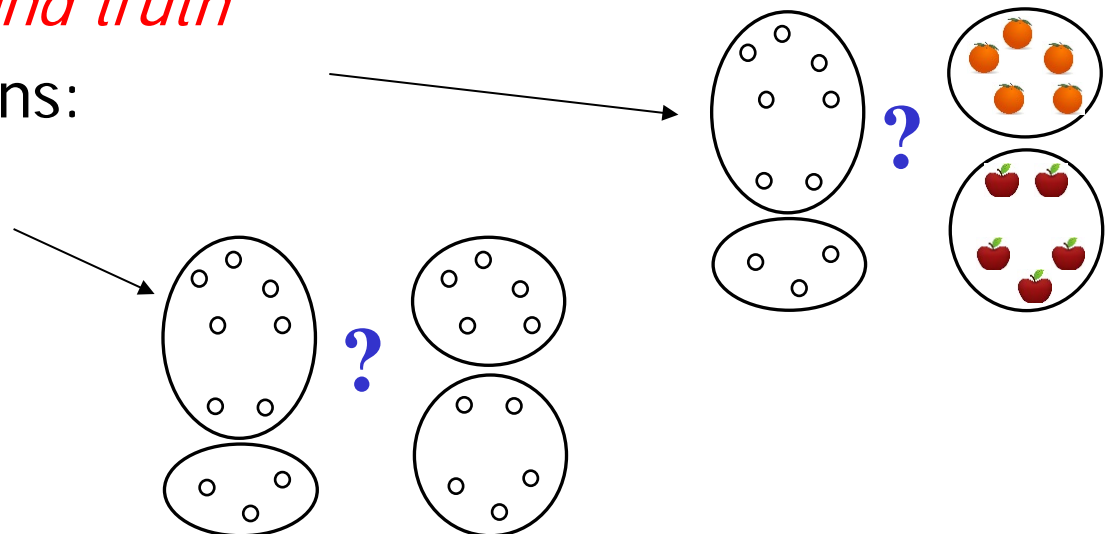


Motivation and Background

- Internal Criteria (Indexes)
 - Validate *without* external information
 - With different number of clusters
 - Solve the number of clusters



- External Criteria (Indexes)
 - Validate against *ground truth*
 - Compare two partitions: (how similar)

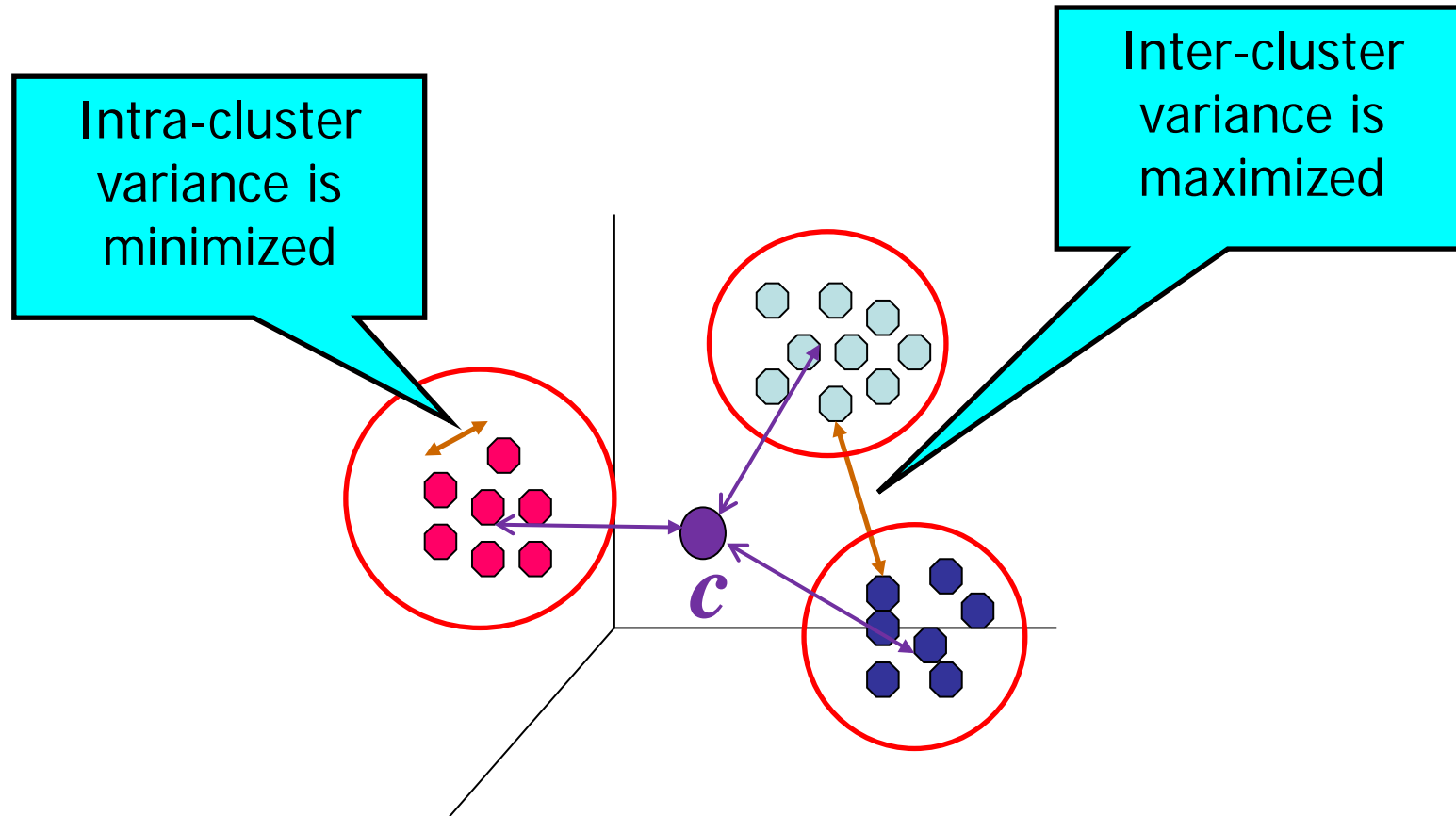


Internal Index

- Ground truth is *unavailable* but unsupervised validation must be done with “*common sense*” or “*a priori knowledge*”.
- There are a variety of internal indexes:
 - Variances-based methods
 - Rate-distortion methods
 - Davies-Bouldin index (DBI)
 - Bayesian Information Criterion (BIC)
 - Silhouette Coefficient
 - Minimum description principle (MDL)
 - Stochastic complexity (SC)
 - Modified Huber’s Γ (MH Γ) index

Internal Index

- Variances-based methods
 - Minimise within cluster variance (SSW)
 - Maximise between cluster variance (SSB)



Internal Index

- Variances-based methods (cont.)

Assume an algorithm leads to a partition of m clusters where cluster i has n_i data points and \mathbf{c}_i is its centroid. $d(.,.)$ is a distance used in this algorithm.

- Within cluster variance (SSW)

$$SSW(m) = \sum_{i=1}^m \sum_{j=1}^{n_i} d^2(\mathbf{x}_{ij}, \mathbf{c}_i)$$

- Between cluster variance (SSB)

$$SSB(m) = \sum_{i=1}^m n_i d^2(\mathbf{c}_i, \mathbf{c})$$

where \mathbf{c} is the mean (centroid) of the whole data set.

Internal Index

- Variance based F-ratio index
 - Measures ratio of between-cluster variance against the within-cluster variance (original F-test)
 - **F-ratio index** (W-B index) for a partition of m clusters

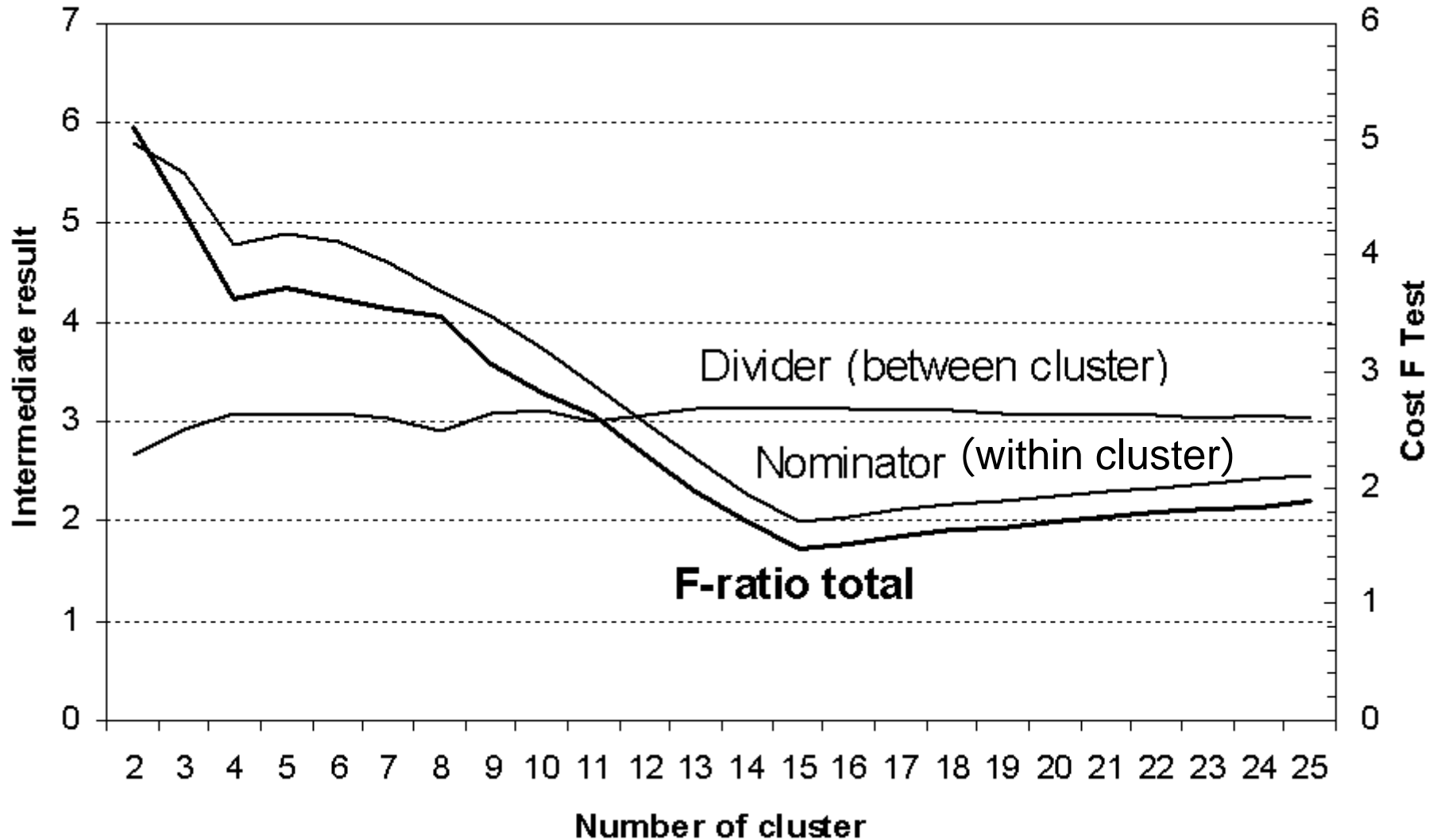
$$F(m) = \frac{mSSW(m)}{SSB(m)} = \frac{m \sum_{i=1}^m \sum_{j=1}^{n_i} d^2(\mathbf{x}_{ij}, \mathbf{c}_i)}{\sum_{i=1}^m n_i d^2(\mathbf{c}_i, \mathbf{c})}$$

where \mathbf{x}_{ij} is the j th data point in cluster \mathbf{c}_i

n_i is the number of data points in cluster \mathbf{c}_i

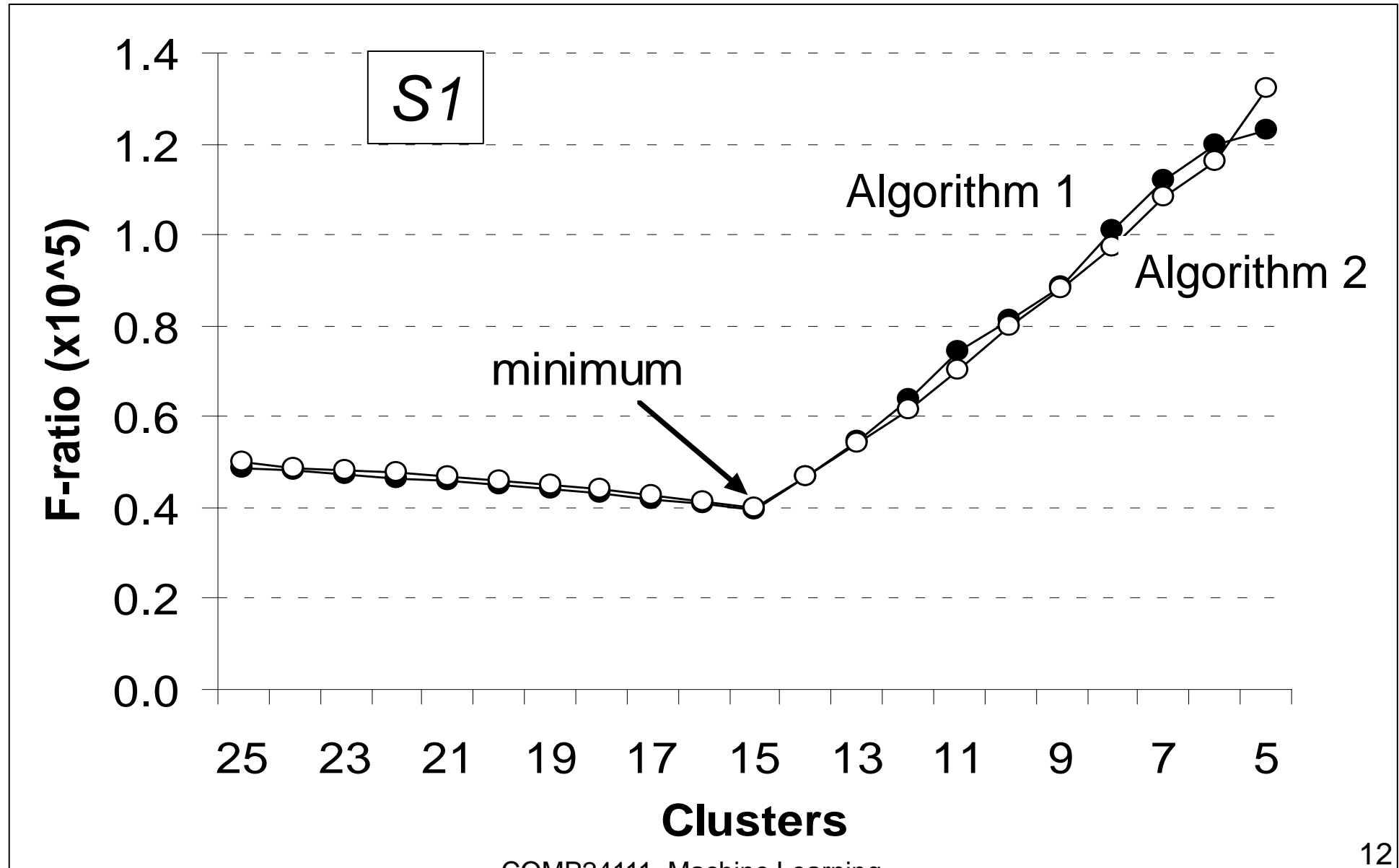
Internal Index

- Variance based F-ratio index (cont.)



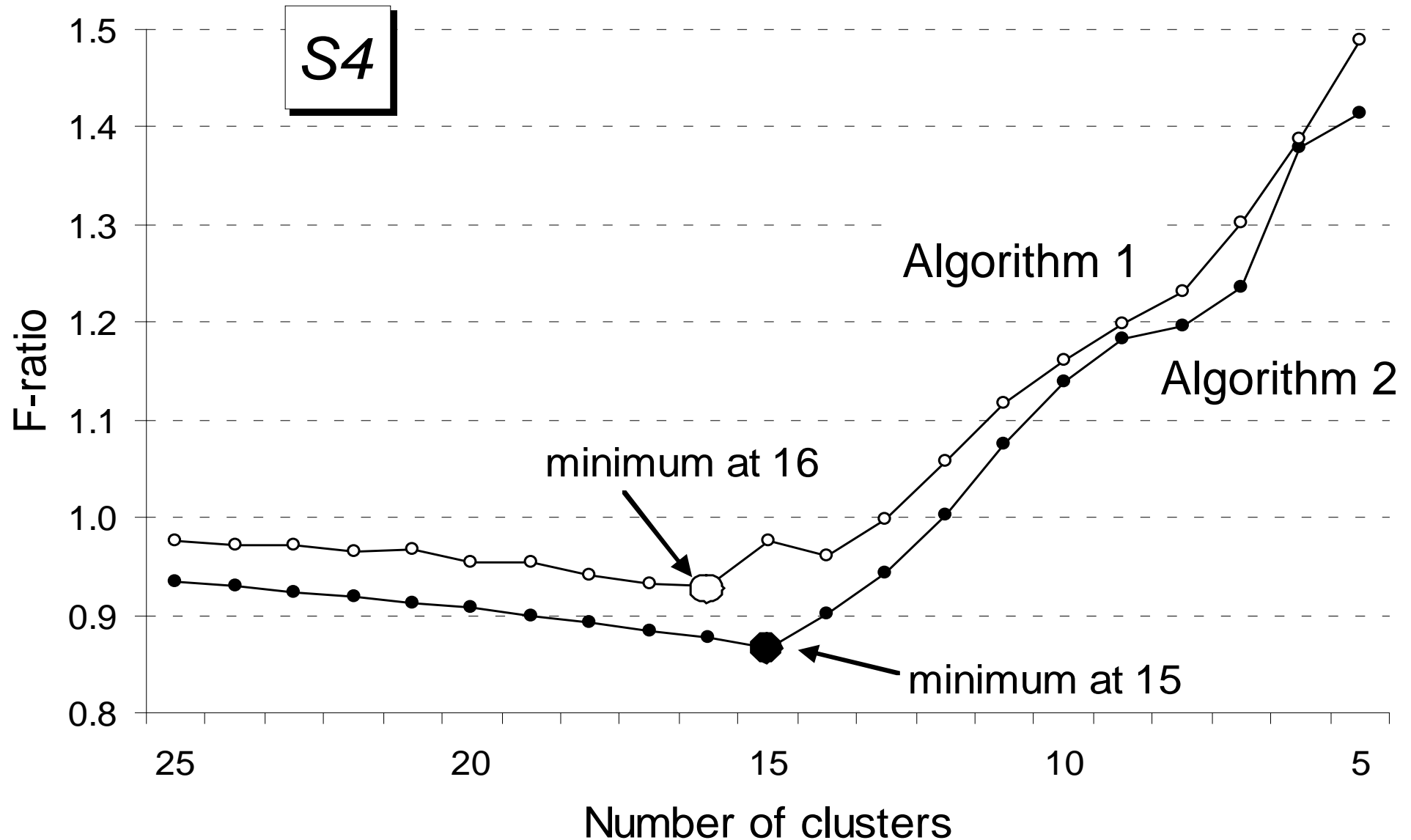
Internal Index

- Application: finding the “proper” cluster number



Internal Index

- Application: finding the “proper” cluster number



External Index

- “Ground truth” is *available* but an clustering algorithm *doesn't* use such information during unsupervised learning.
- There are a variety of internal indexes:
 - Rand Index
 - Adjusted Rand Index
 - Pair counting index
 - Information theoretic index
 - Set matching index
 - DVI index
 - Normalised mutual information (NMI) index

External Index

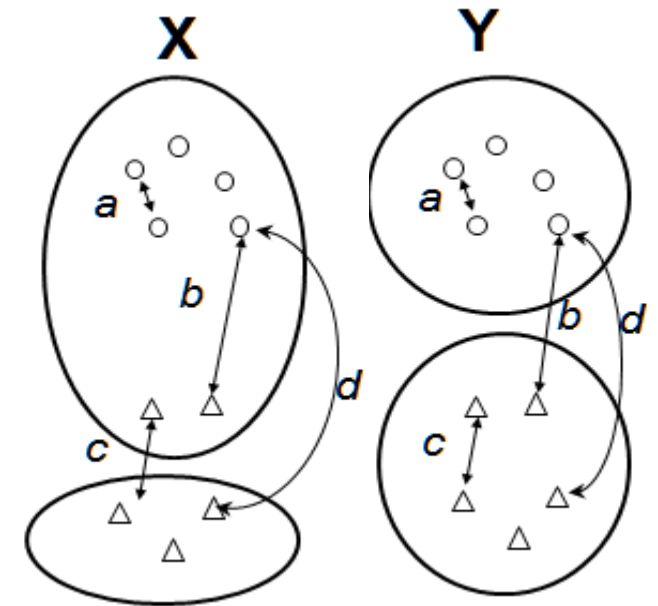
- Main issues

- If the “ground truth” is known, the validity of a clustering can be verified by comparing the class or clustering labels.
- However, this is much more complicated than in supervised classification (where labels used in training)
 - The cluster IDs in a partition resulting from clustering have been assigned *arbitrarily* due to unsupervised learning – **permutation**.
 - The number of clusters may be different from the number of classes (the “ground truth”) – **inconsistence**.
 - The most important problem in external indexes would be how to find all possible correspondences between the “ground truth” and a partition (or two candidate partitions in the case of comparison).

External Index

- Rand Index
 - This is the first external index proposed by Rand (1971) to address the “correspondence” problem.
 - Basic idea: considering all pairs in the data set by looking into both **agreement** and **disagreement** against the “ground truth”
 - The index defined as $RI(X, Y) = (a + d) / (a + b + c + d)$

X/Y	Y: Pairs in the same class	Y: Pairs in different classes
X: Pairs in the same cluster	<i>a</i>	<i>b</i>
X: Pairs in different clusters	<i>c</i>	<i>d</i>



External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

- To calculate a , b , c and d , we have to list all possible pairs (excluding any data point to itself)

[1, 2], [1, 3], [1, 4], [1, 5]

[2, 3], [2, 4], [2, 5]

[3, 4], [3, 5]

[4, 5]

External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Initialisation: $a, b, c, d \leftarrow 0$

For data point pair [1, 2], we have

X: [1, 2] assigned to (i, ii) \rightarrow in different clusters

Y: [1, 2] labelled by (q, p) \rightarrow in different classes

Thus, $d \leftarrow d + 1 = 1$

External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Current Status: $a=0$, $b=0$, $c=0$, $d=1$

For data point pair [1, 3], we have

X: [1, 3] assigned to (i, ii) \rightarrow in different clusters

Y: [1, 3] labelled by (q, q) \rightarrow in the same class

Thus, $c \leftarrow c + 1 = 1$

External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Current Status: $a=0$, $b=0$, $c=1$, $d=1$

For data point pair [1, 4], we have

X: [1, 4] assigned to (i, i) \rightarrow in the same cluster

Y: [1, 4] labelled by (q, p) \rightarrow in different classes

Thus, $b \leftarrow b + 1 = 1$

External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Current Status: $a=0$, $b=1$, $c=1$, $d=1$

For data point pair [1, 5], we have

X: [1, 5] assigned to (i, i) \rightarrow in the same cluster

Y: [1, 5] labelled by (q, q) \rightarrow in the same class

Thus, $a \leftarrow a + 1 = 1$

External Index

- Rand Index (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Current Status: $a=1$, $b=1$, $c=1$, $d=1$

On-class Exercise: continuing until have the final value of a , b , c and d .

External Index

- Rand Index: Contingency Table
 - In general, a , b , c and d are calculated from a contingency table

N	.	
.	$n_{..}$	

 $=$

n_{11}	n_{12}	\dots	n_{1l}	$n_{1.}$
n_{21}	n_{22}	\dots	n_{2l}	$n_{2.}$
\vdots	\vdots	\ddots	\vdots	\vdots
n_{k1}	n_{k2}	\dots	n_{kl}	$n_{k.}$
$n_{.1}$	$n_{.2}$	\dots	$n_{.l}$	$n_{..}$

Assume there are k clusters in \mathbf{X} and l classes in \mathbf{Y}

n_{ij} : the number of points in both cluster i and class j

External Index

- Rand Index: Contingency Table (cont.)
 - Example: for a 5-point data set, we have

	1	2	3	4	5
X	i	ii	ii	i	i
Y	q	p	q	p	q

Contingency table

	p	q	
i	1	2	3
ii	1	1	2
	2	3	5

External Index

- Rand Index: measure the number of pairs in Same cluster/class in X and Y

$$a = \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^l n_{ij} (n_{ij} - 1)$$

Same cluster in X / different classes in Y

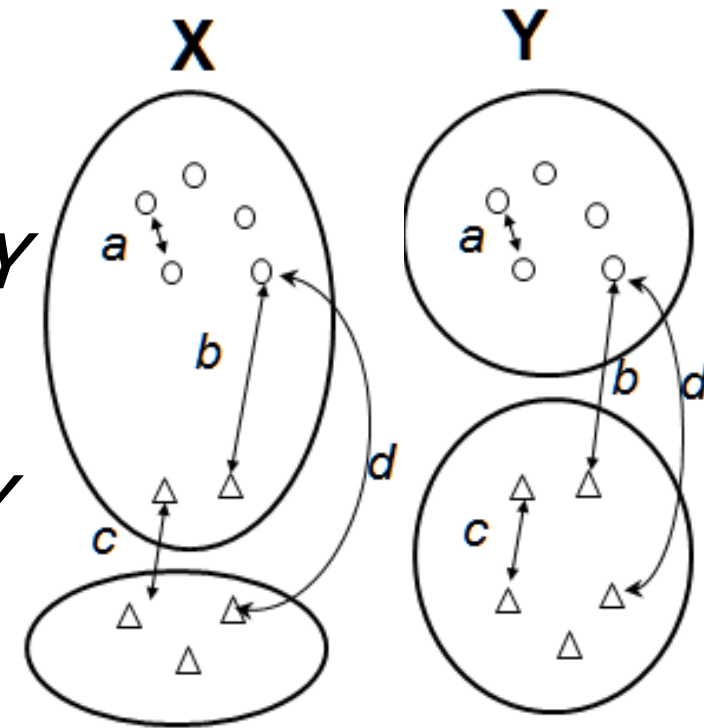
$$b = \frac{1}{2} \left(\sum_{j=1}^l n_{.j}^2 - \sum_{i=1}^k \sum_{j=1}^l n_{ij}^2 \right)$$

Different clusters in X / same class in Y

$$c = \frac{1}{2} \left(\sum_{i=1}^k n_{i.}^2 - \sum_{i=1}^k \sum_{j=1}^l n_{ij}^2 \right)$$

Different clusters/classes in X and Y

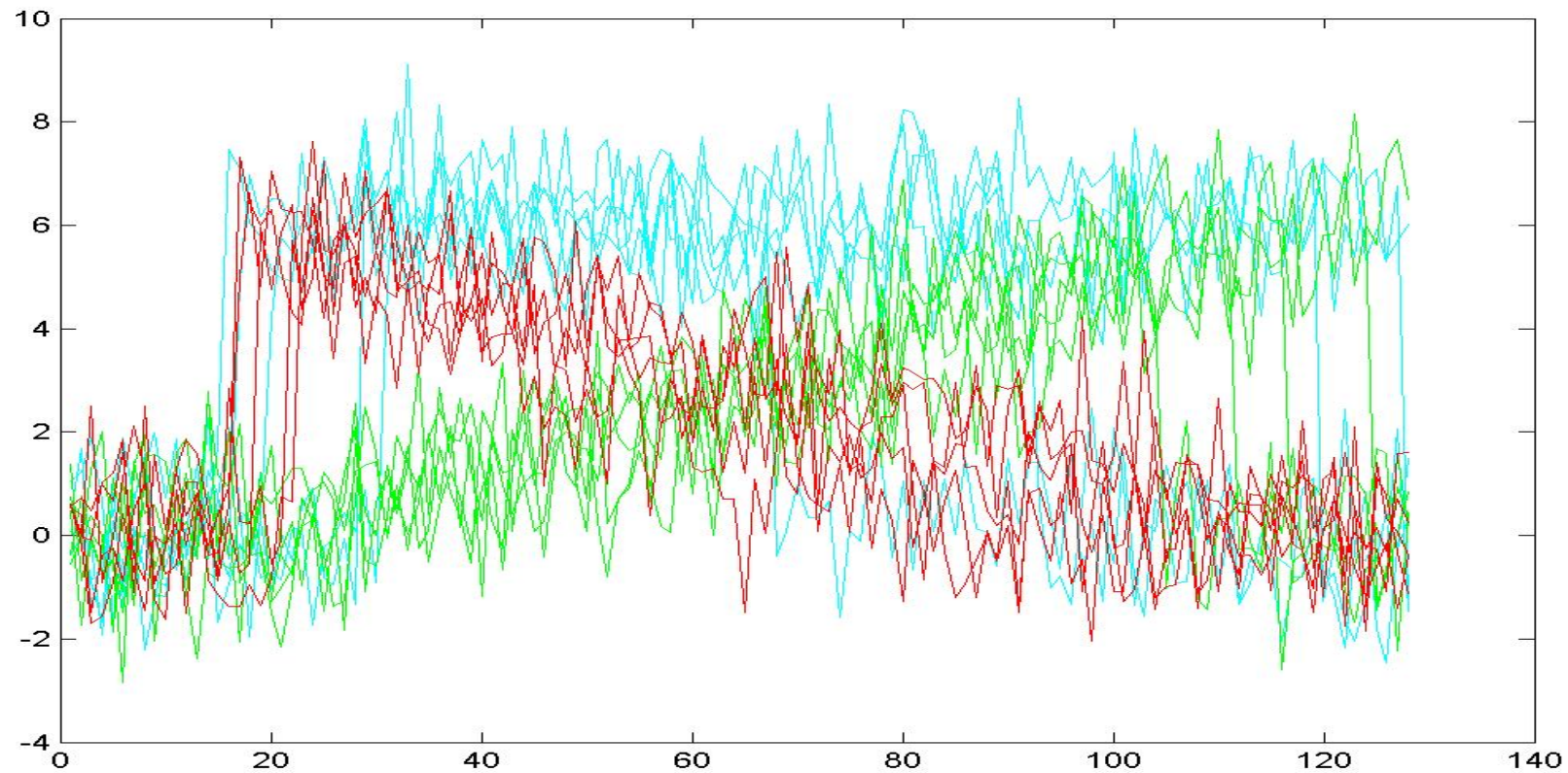
$$d = \frac{1}{2} \left(N^2 + \sum_{i=1}^k \sum_{j=1}^l n_{ij}^2 - \left(\sum_{i=1}^k n_{i.}^2 + \sum_{j=1}^l n_{.j}^2 \right) \right)$$



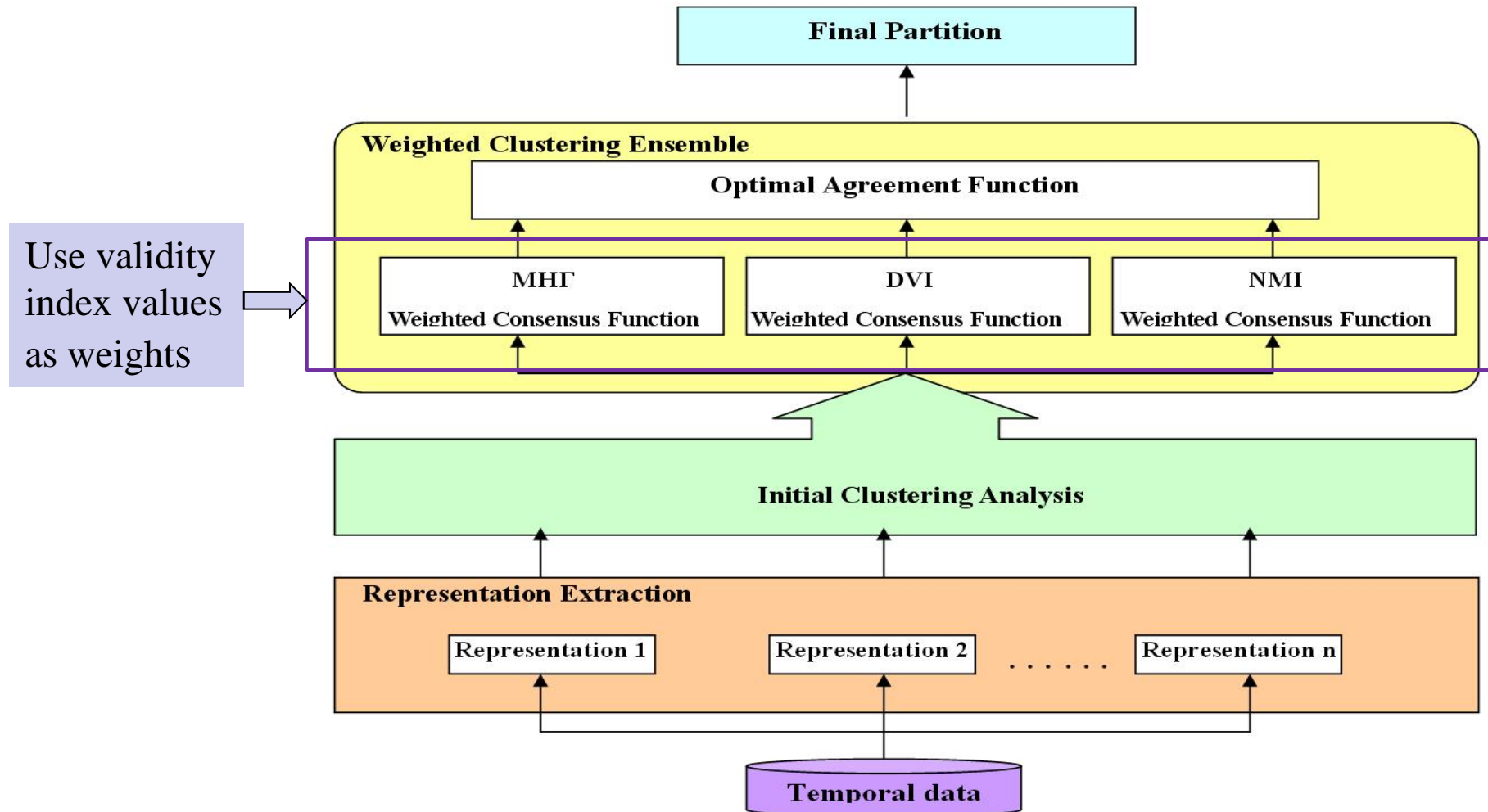
Ex. 4

External Index

- Application: Performance evaluation
 - Time series clustering



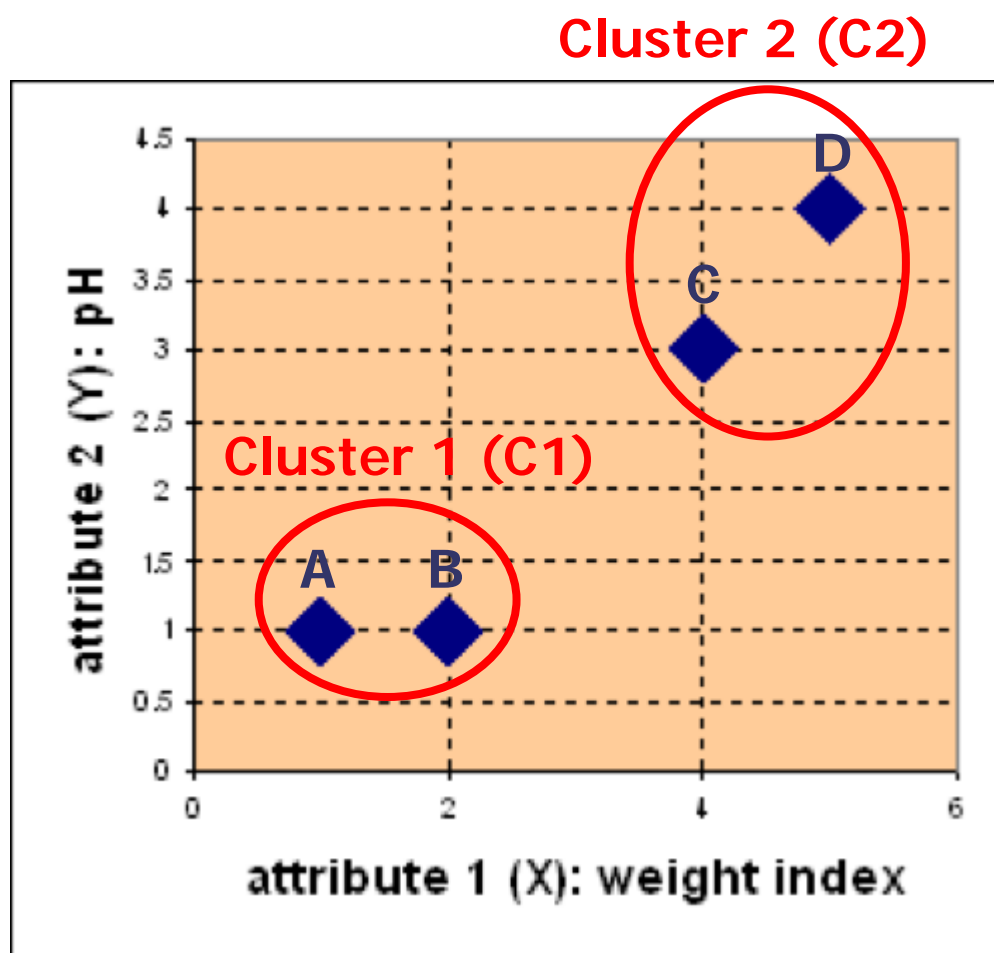
Weighted Clustering Ensemble



Yun Yang and Ke Chen, "Temporal data clustering via weighted clustering ensemble with different representations," *IEEE Transactions on Knowledge and Data Engineering* **23**(2), pp. 307-320, 2011.

Initial Clustering Analysis

□ Binary Indicator Matrix vs. "Distance" Matrix



Binary Indicator
Matrix

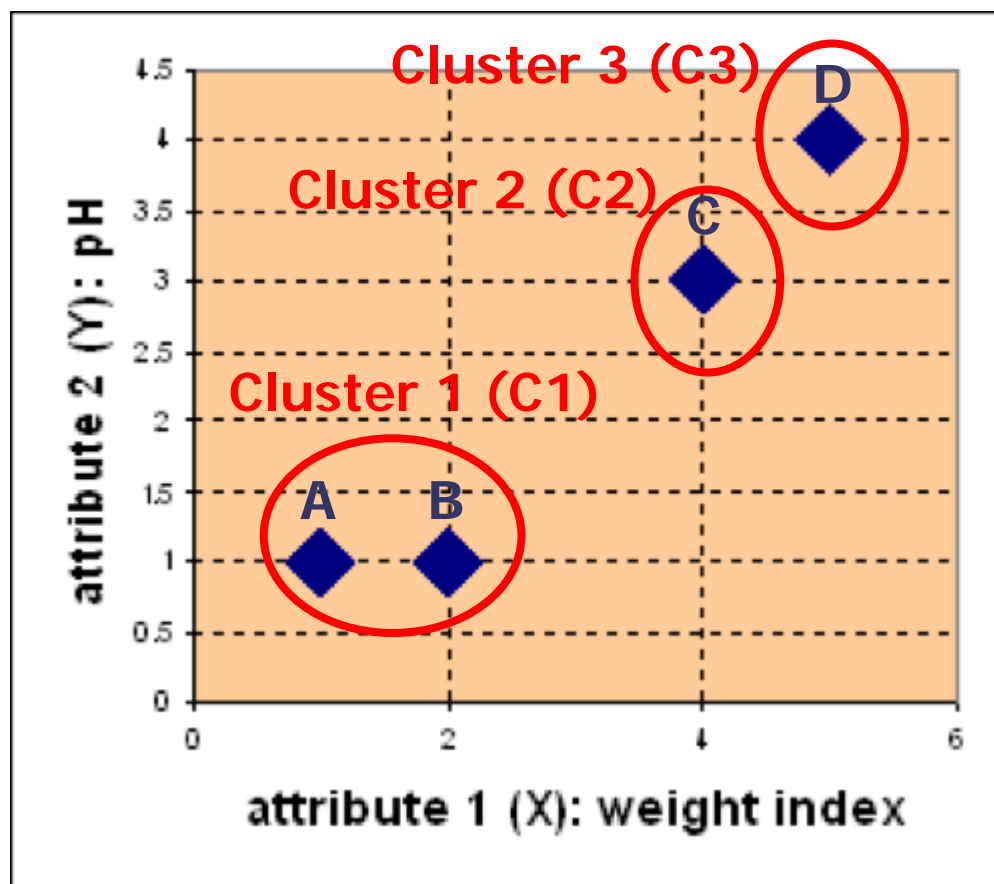
$$H_1 = \begin{matrix} & \begin{matrix} C1 & C2 \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \end{matrix}$$

"distance" Matrix

$$D_1 = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \end{matrix}$$

Initial Clustering Analysis

□ Binary Indicator Matrix vs. "distance" Matrix



Binary Indicator
Matrix

$$H_2 = \begin{matrix} & \begin{matrix} C1 & C2 & C3 \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix}$$



"distance Matrix"

$$D_2 = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{matrix}$$

Initial Clustering Analysis

□ Ensembled "distance" Matrix (evidence-accumulation)

$$D_1 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \quad D_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$D_E = \frac{1}{3}[(w_{M1} + w_{D1} + w_{N1})D_1 + (w_{M2} + w_{D2} + w_{N2})D_2] = \begin{bmatrix} 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 \\ 2 & 2 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix}$$

when $w_{M1} = w_{M2} = 1$, $w_{D1} = w_{D2} = 1$, $w_{N1} = w_{N2} = 1$.

External Index

- Application: UCR time series benchmarks

Data Set	Number of Class K^*	Size of Data Set (Training+Testing)	Length
<i>Synthetic Control</i>	6	300+300	60
<i>Gun-Point</i>	2	50+150	150
<i>CBF</i>	3	30+900	128
<i>Face (all)</i>	14	560+1,690	131
<i>OSU Leaf</i>	6	200+242	427
<i>Swedish Leaf</i>	15	500+625	128
<i>50Words</i>	50	450+455	270
<i>Trace</i>	4	100+100	275
<i>Two Patterns</i>	4	1,000+4000	128
<i>Wafer</i>	2	1,000+6,174	152
<i>Face (four)</i>	4	24+88	350
<i>Lightning-2</i>	2	60+61	637
<i>Lightning-7</i>	7	70+73	319
<i>ECG</i>	2	100+100	96
<i>Adiac</i>	37	390+391	176
<i>Yoga</i>	2	300+3,000	426

External Index

- Application: RI values of clustering ensembles (Yang & Chen 2011)

Data Set	CE	HBGF	SDP-CE	WCE
<i>Syn Control</i>	68.8 + 2.1	74.8 ± 2.2	82.1+1.9	86.1 ± 2.5*
<i>Gun-Point</i>	51.8 + 1.4	53.8 ± 2.0	50.0±0.9	54.1 ± 1.8*
<i>CBF</i>	53.8 + 2.4	66.0±1.9	66.3±2.1	63.9 ± 2.1*
<i>Face (all)</i>	35.1 + 1.9	44.8±2.5	50.5±1.2	51.9 ± 1.4*
<i>OSU Leaf</i>	35.2 + 1.7	48.1±2.9	46.9±2.1	45.5 ± 3.5*
<i>Swedish Leaf</i>	41.2 + 0.8	52.8±2.3	62.6±1.8	59.8 ± 2.1*
<i>50Words</i>	39.6 + 1.6	39.1±2.1	38.9±1.9	37.2 ± 2.7
<i>Trace</i>	50.5 + 2.0	45.6±2.2	55.1±1.9	57.2 ± 2.3*
<i>Two Patterns</i>	33.1 + 1.8	33.0±1.9	36.9±2.3	37.7 ± 2.5*
<i>Wafer</i>	62.1 + 1.9	72.8±2.6	70.0±2.4	71.7 ± 2.4*
<i>Face (four)</i>	65.2 + 2.1	72.1±3.1	71.8±3.5	78.9 ± 3.0*
<i>Lightning-2</i>	60.1 + 1.3	59.3±2.1	66.2±1.6	77.9 ± 1.8*
<i>Lightning-7</i>	53.1 + 2.1	55.6±3.0	57.9±2.4	58.7 ± 3.4*
<i>ECG</i>	65.2 + 1.6	68.7±2.0	69.2±1.7	69.0 ± 1.7*
<i>Adiac</i>	36.2 + 2.3	41.4±2.5	45.9±1.9	36.8 ± 2.5
<i>Yoga</i>	50.6 + 2.3	60.0±2.2	68.2±2.2	62.6 ± 2.0 *

Summary

- **Cluster validation** is a process that evaluate clustering results with a pre-defined criterion.
- Two different types of cluster validation methods
 - Internal indexes
 - no “ground truth” available
 - defined based on “common sense” or “a priori knowledge”
 - Application: finding the “proper” number of clusters, ...
 - External indexes
 - “ground truth” known or reference given (“relative index”)
 - Application: performance evaluation of clustering, ...
- Still an active area in clustering analysis researches

K. Wang et al, “CVAP: Validation for cluster analysis,” Data Science Journal, vol. 8, May 2009.
[Code online available: <http://www.mathworks.com/matlabcentral/fileexchange/authors/24811>]