MANCHESTER 1824

The University of Manchester

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: Weighted clustering ensemble
- Summary

Motivation and Background

- Motivation

Supervised classification

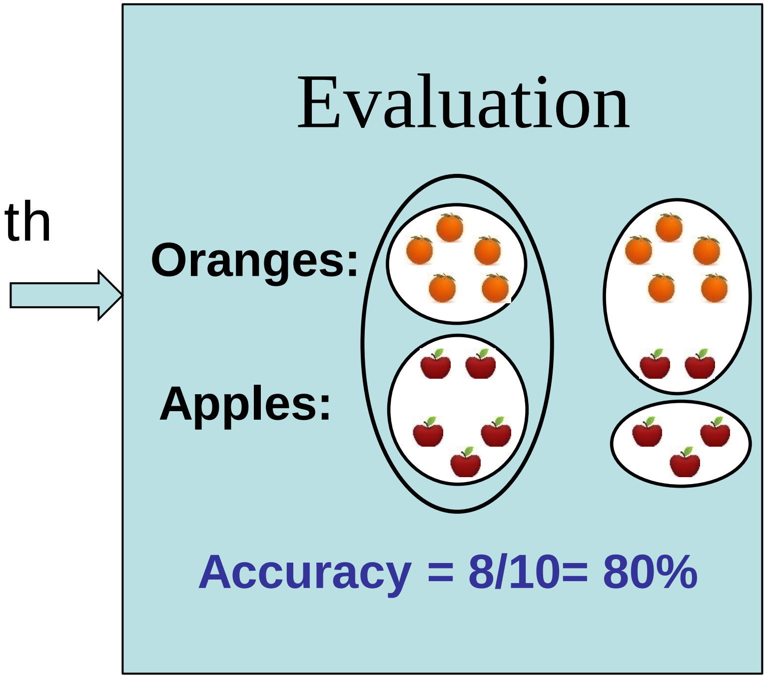
– Class labels known for ground truth
– Accuracy: based on labels given

Clustering analysis

– No class labels
– Evaluation still demanded

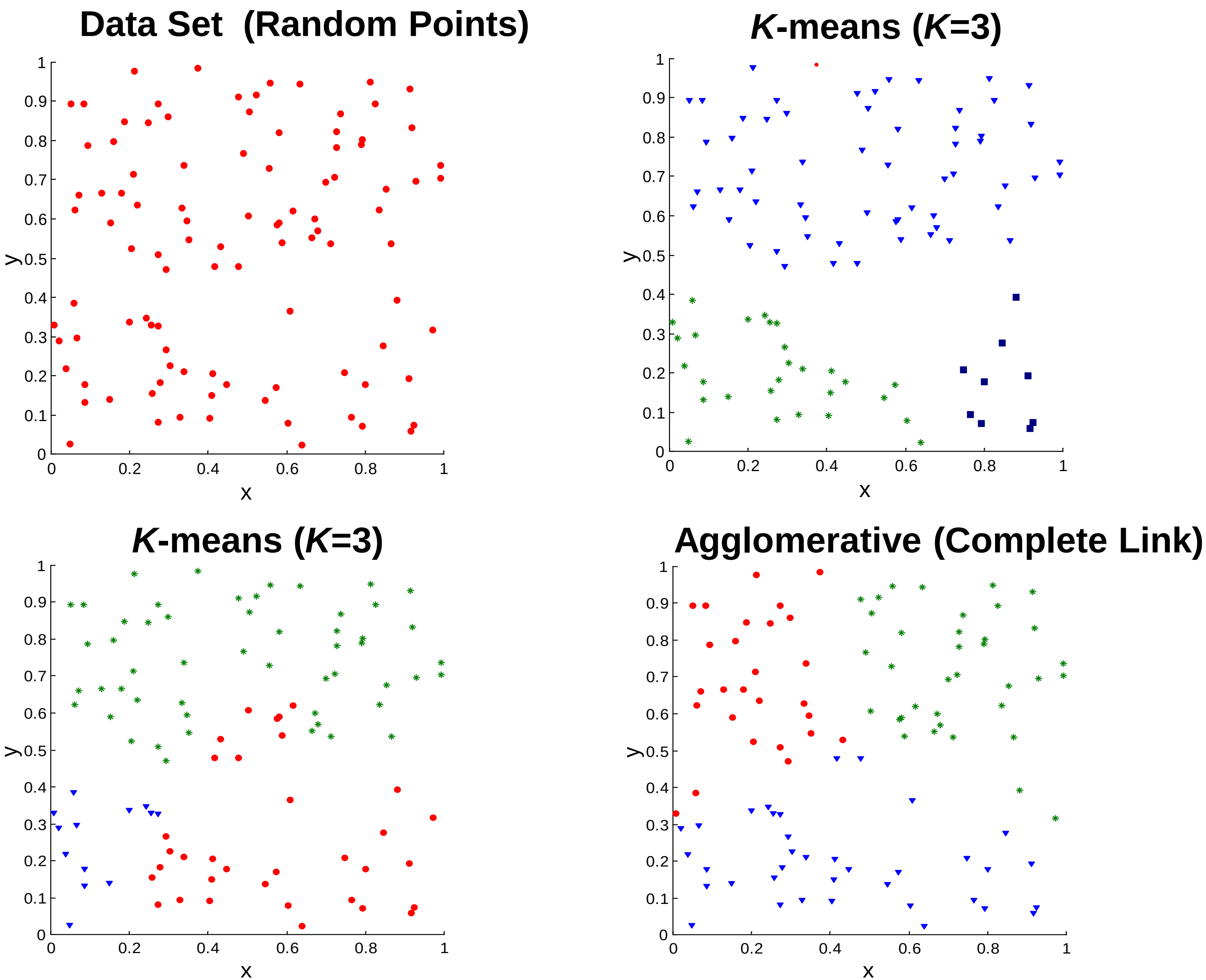
Validation needs to

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



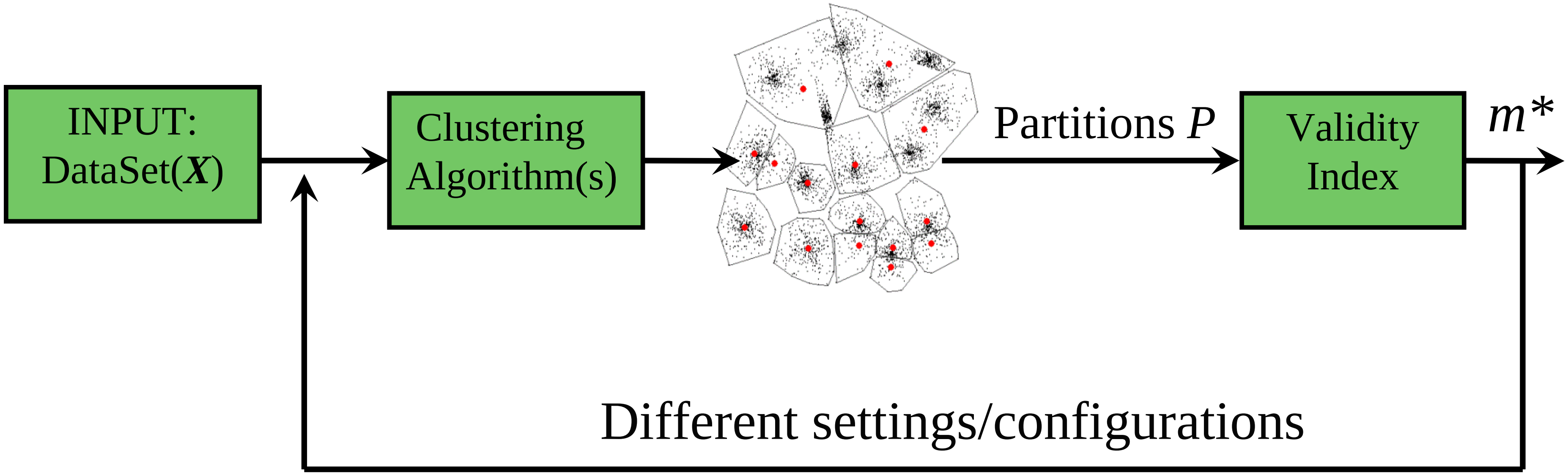
Motivation and Background

- Illustrative Example: which one is the "best"?



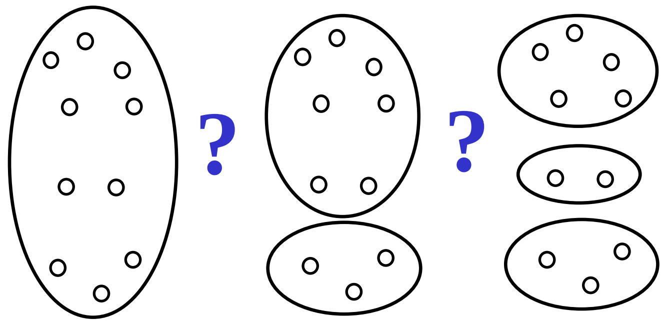
Motivation and Background

- *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!

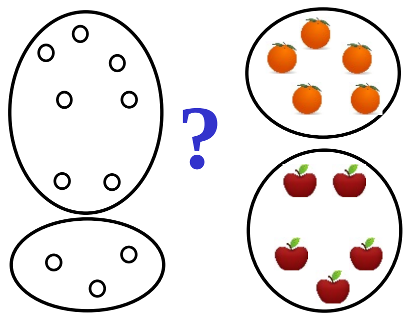


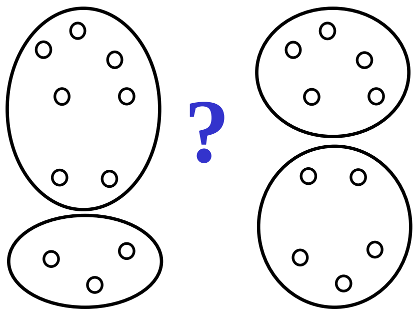
Motivation and Background

- 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)





The University of Manchester

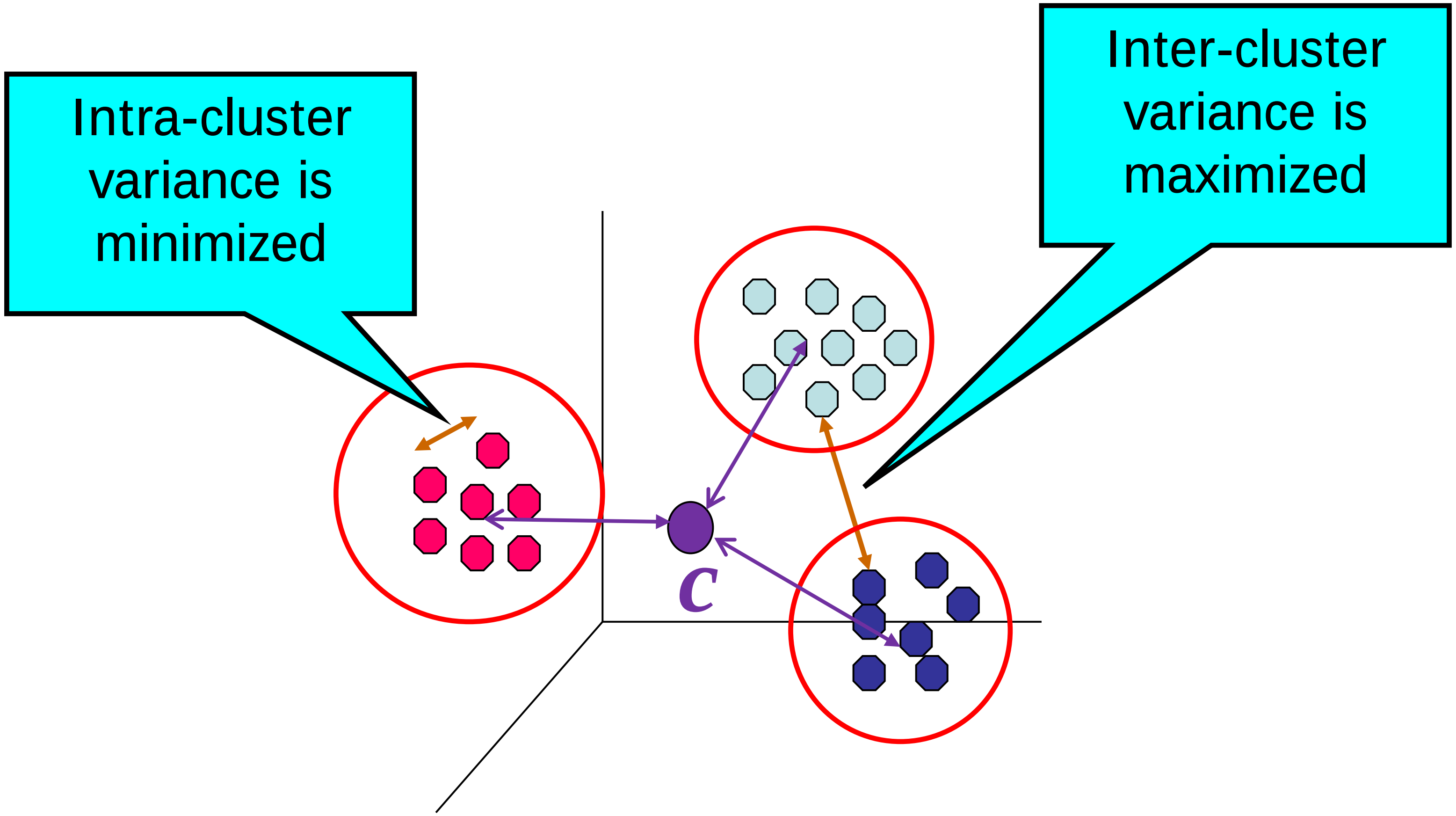
MANCHESTER 1824

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 (MHΓ) 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 K clusters where cluster i has n_i data points and $\boldsymbol{c}_i$ is its centroid. $d(.,.)$ is a distance used in this algorithm.

  – Within cluster variance (SSW)

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

  – Between cluster variance (SSB)

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

  where $\boldsymbol{c}$ 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 K clusters

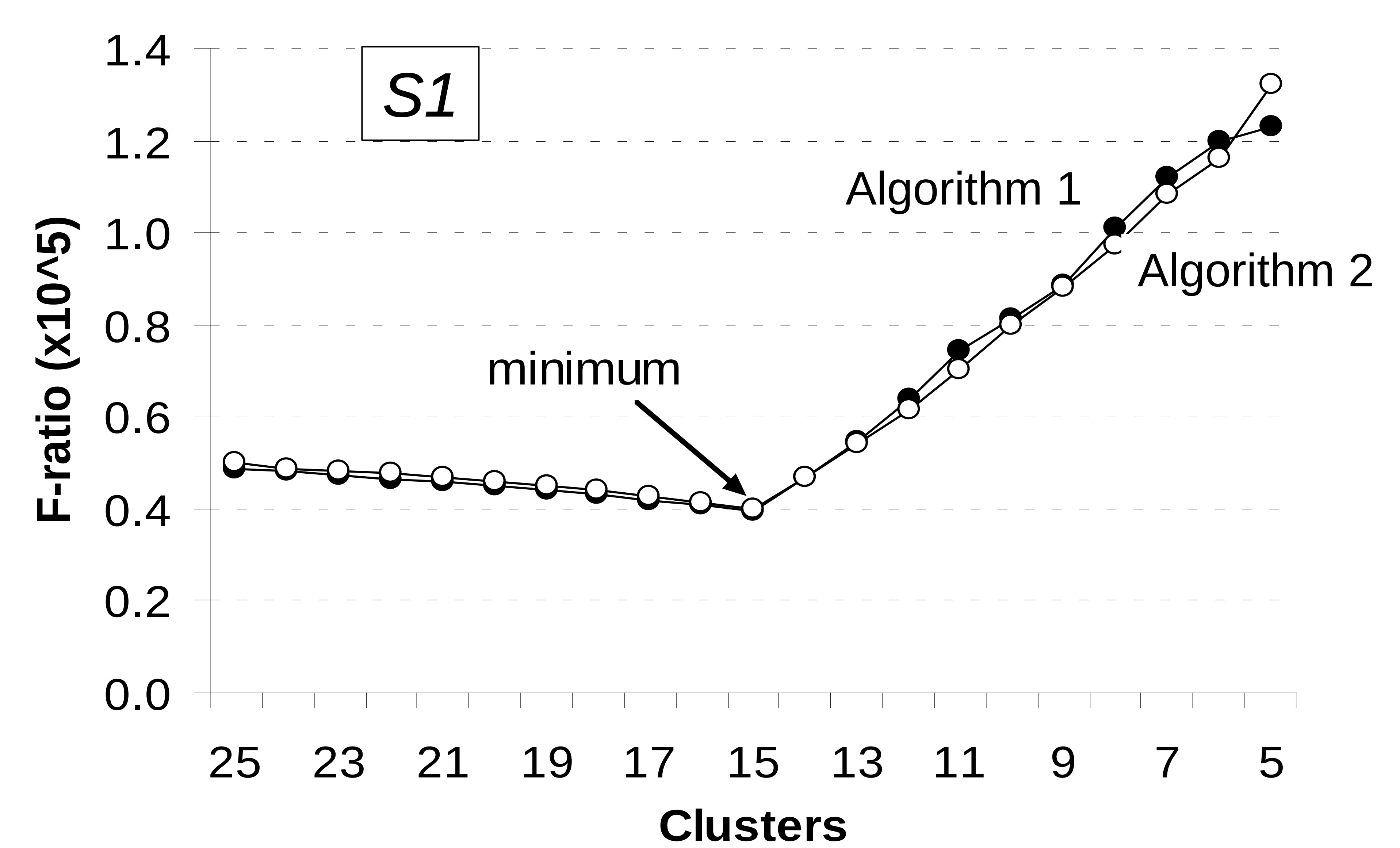
$$F(K) = \frac{K * SSW(K)}{SSB(K)} = \frac{K \sum_{i=1}^{K} \sum_{j=1}^{n_i} d^2(\mathbf{x}_{ij}, \mathbf{c}_i)}{\sum_{i=1}^{K} n_i d^2(\mathbf{c}_i, \mathbf{c})}$$

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

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

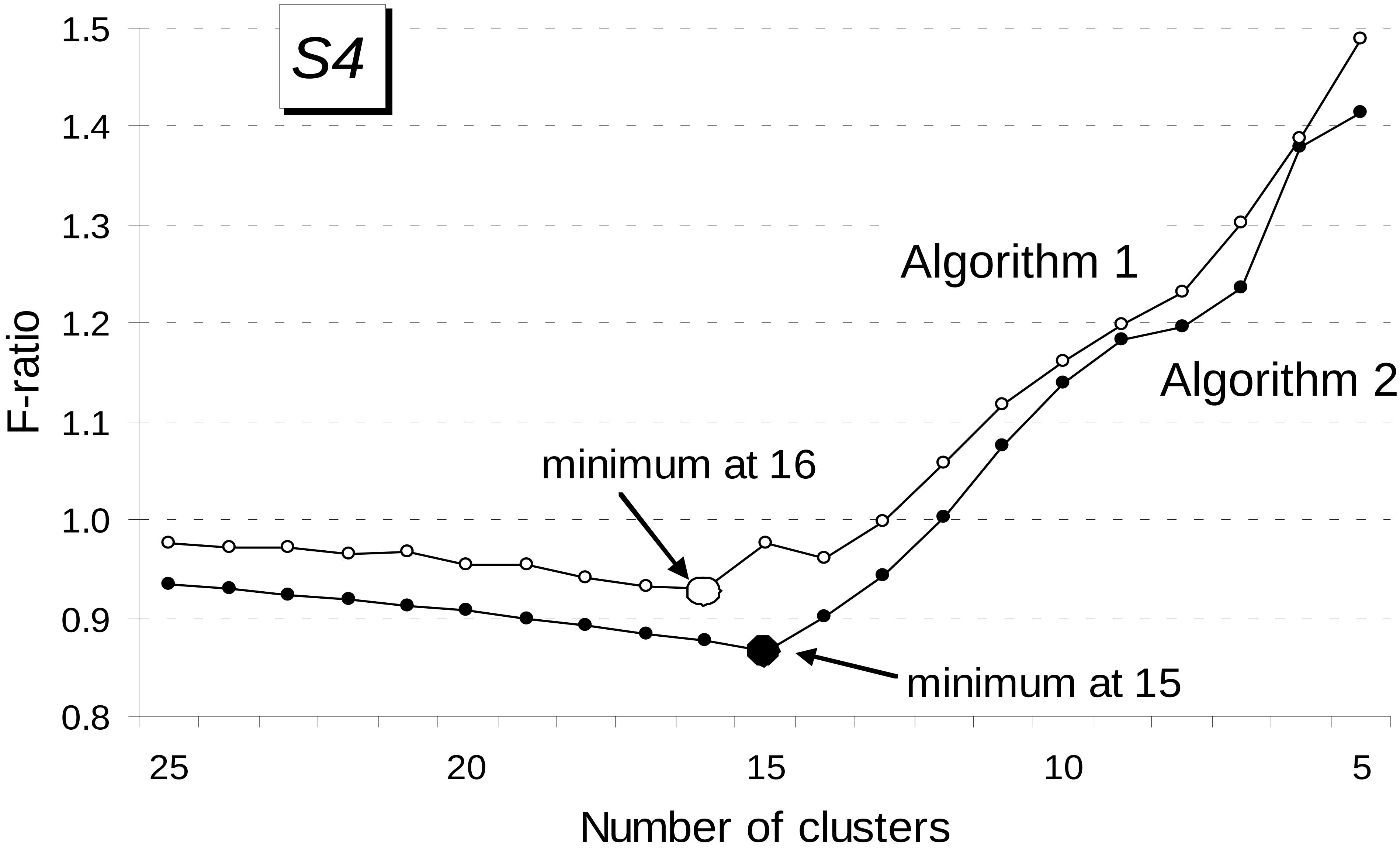
Internal Index

- 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 external indexes:
  - ➢ Rand Index
  - ➢ Adjusted Rand Index
  - ➢ Pair counting index
  - ➢ Information theoretic index
  - ➢ Set matching index
  - ➢ DVI index
  - ➢ Normalised mutual information (NMI) index

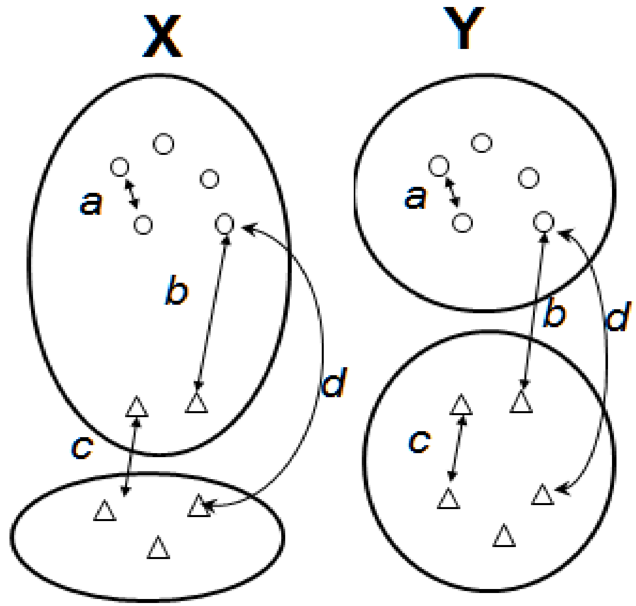
External Index

- 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	*a*	*b*
X: Pairs in different clusters	*c*	*d*



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) –> in different clusters

Y: [1, 2] labelled by (q, p) –> 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) –> in different clusters

Y: [1, 3] labelled by (q, q) –> 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) –> in the same cluster

Y: [1, 4] labelled by (q, p) –> 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) –> in the same cluster

Y: [1, 5] labelled by (q, q) –> 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

$$
\begin{array}{c|c}
\boxed{N} & \cdot \\
\hline
\cdot & n_{\cdot\cdot}
\end{array}
=
\begin{array}{|cccc|c}
\hline
n_{11} & n_{12} & \cdots & n_{1l} & n_{1\cdot} \\
n_{21} & n_{22} & \cdots & n_{2l} & n_{2\cdot} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
n_{k1} & n_{k2} & \cdots & n_{kl} & n_{k\cdot} \\
\hline
n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot l} & n_{\cdot\cdot}
\end{array}
$$

Assume there are k clusters in **X** and l classes in **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*

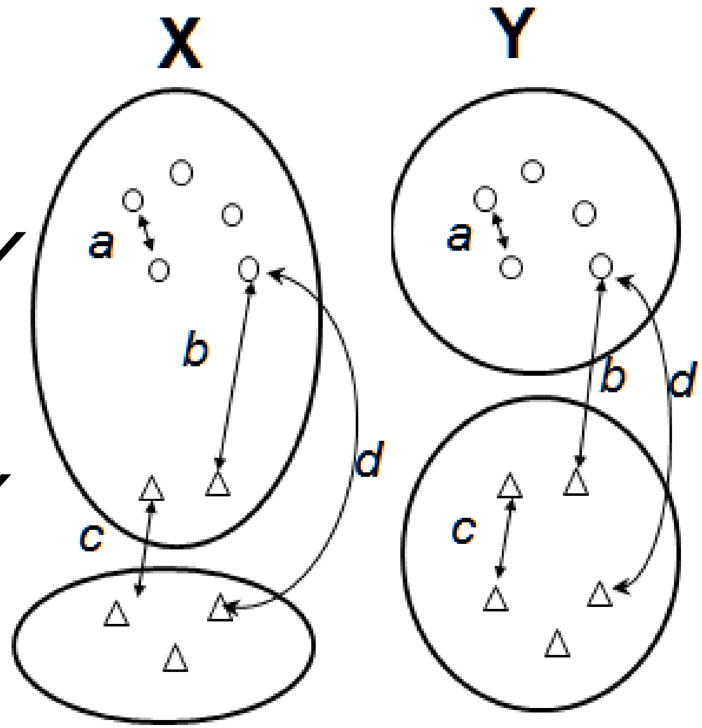
$$b = \frac{1}{2}\left(\sum_{j=1}^{l} n_{.j}^2 - \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}^2\right)$$

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*

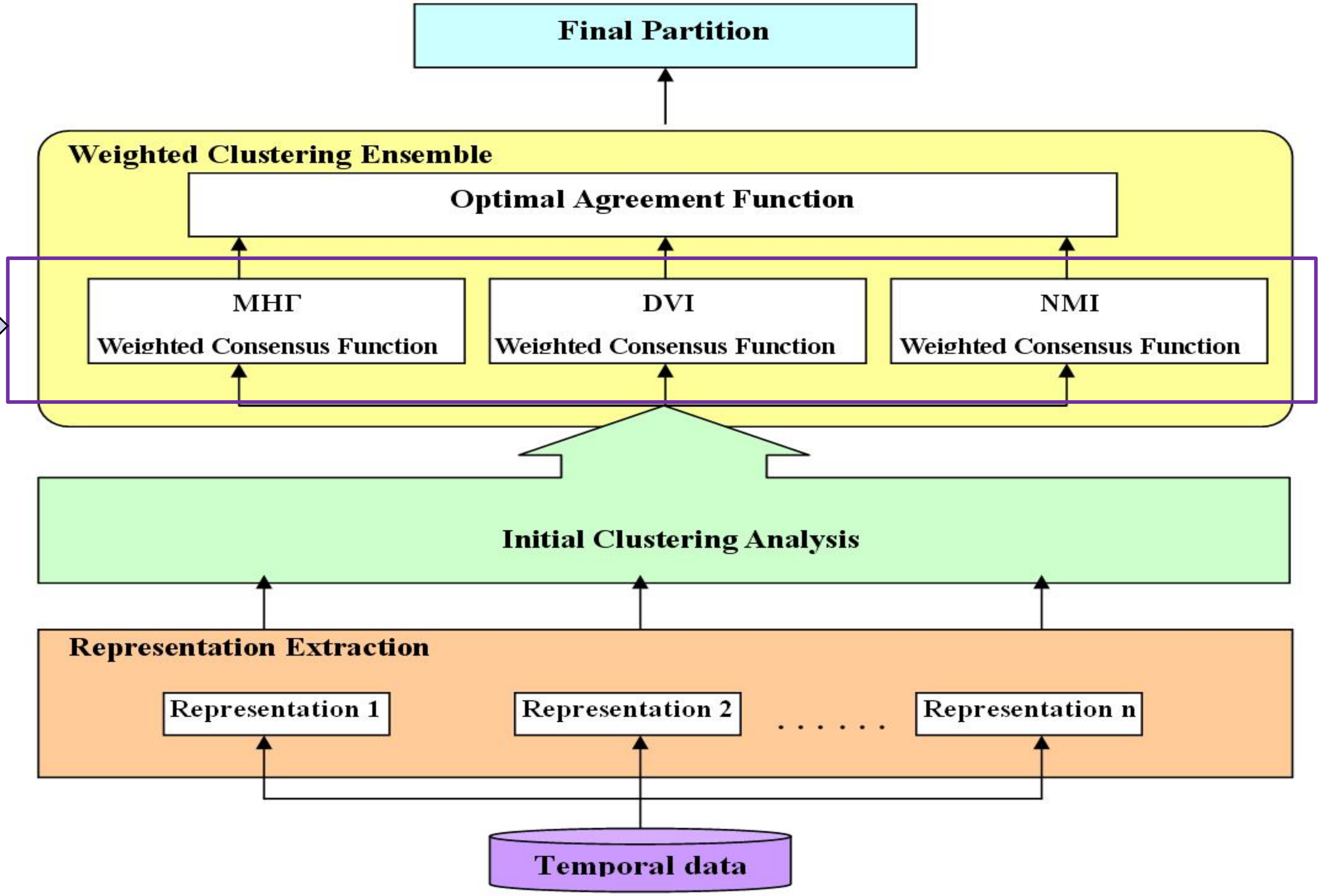
$$d = \frac{1}{2}\left(N^2 + \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}^2 - \left(\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{l} n_{.j}^2\right)\right)$$



Ex. 4

Weighted Clustering Ensemble

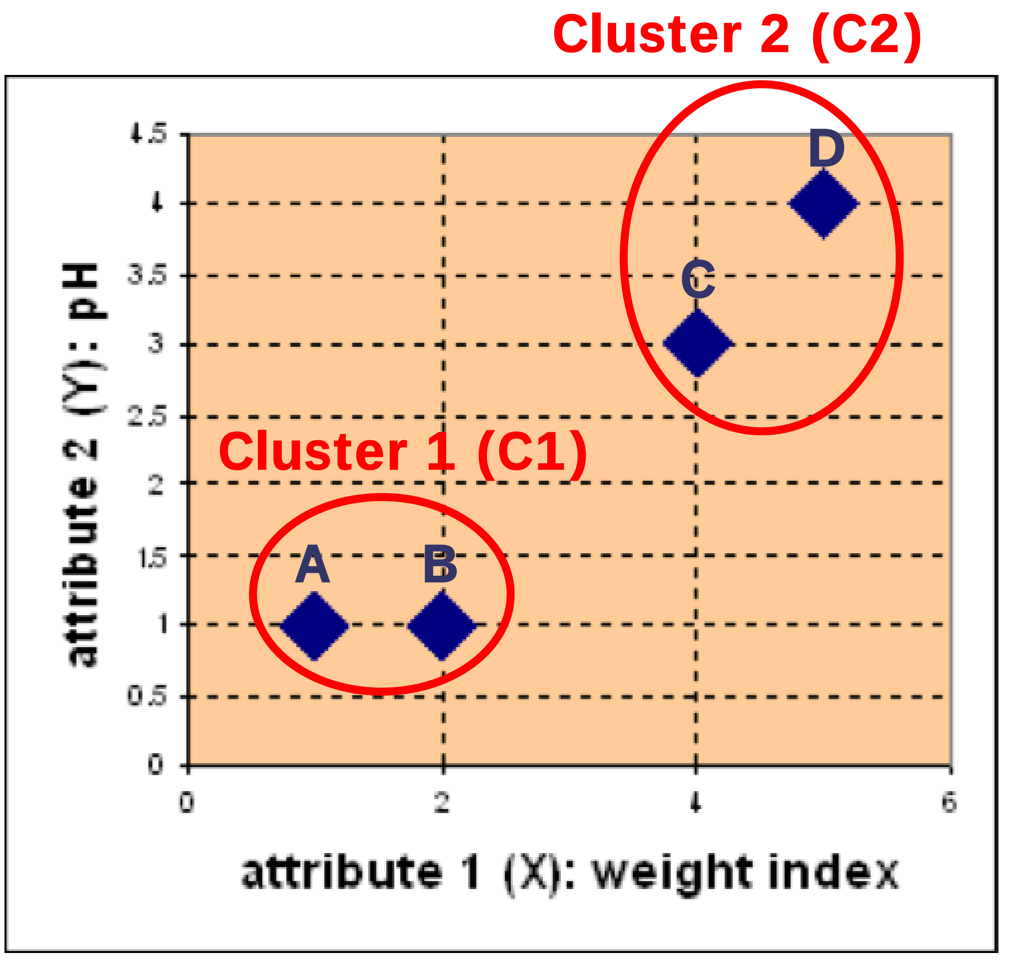
Use validity index values as weights



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.

Evidence Collection

❑ "Distance" Matrix from the clustering result

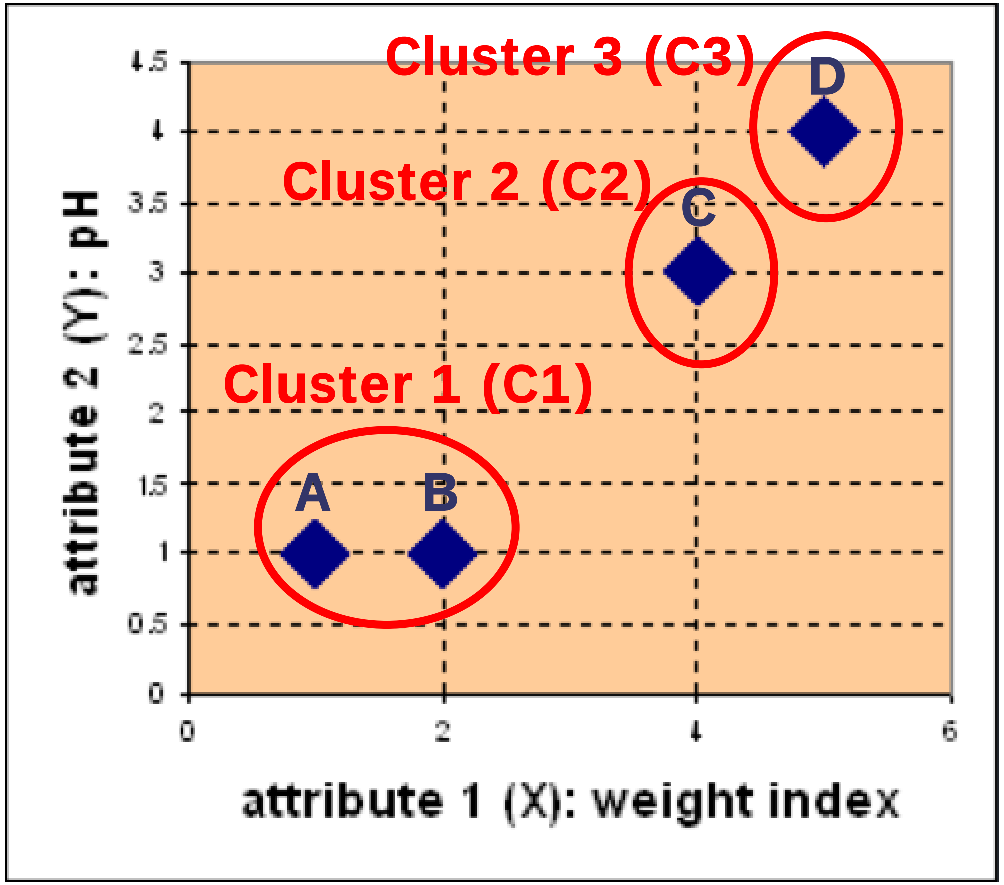


"distance" Matrix

$$
\begin{array}{cc}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
D_1 = & \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{array}{c} A \\ B \\ C \\ D \end{array}
\end{array}
$$

Evidence Collection

- ❑ "Distance" Matrix from the clustering result



"distance Matrix"

$$D_2 = \begin{array}{c@{}c} \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \end{array}$$

Optimal Agreement

- ❑ **Ensembled "distance" Matrix (evidence-accumulation)**

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

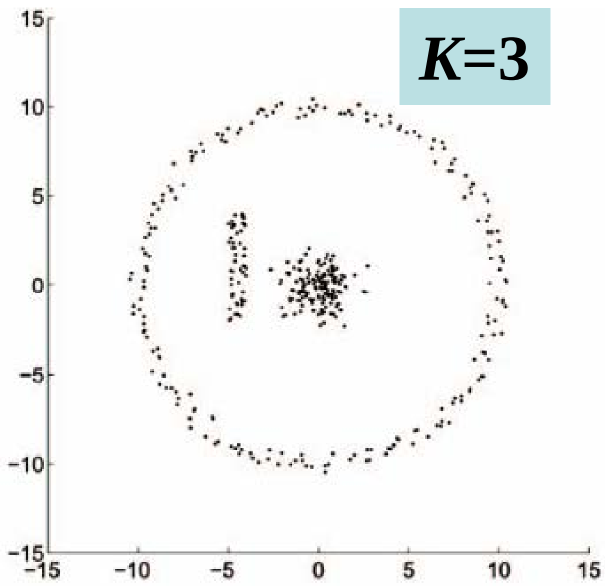
$$D_E = \frac{1}{3}[(w_{M1} + w_{D1} + w_{N1})D_1 + (w_{M2} + w_{D2} + w_{N2})D_2] = \begin{bmatrix} 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 \\ 2 & 2 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix}$$

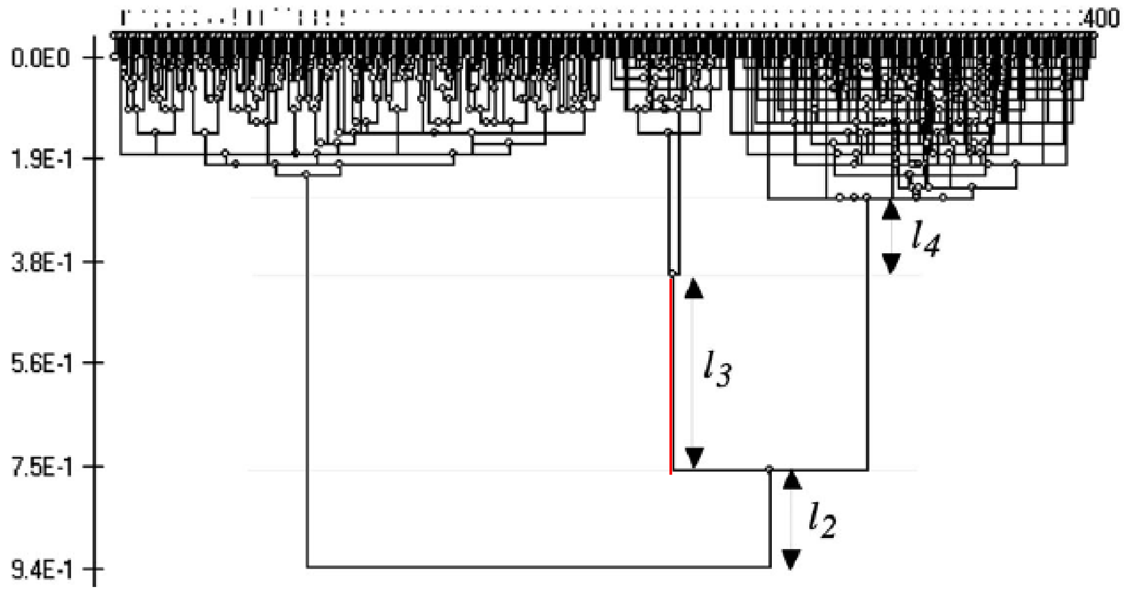
$$\text{when } w_{M1} = w_{M2} = 1,\ w_{D1} = w_{D2} = 1,\ w_{N1} = w_{N2} = 1.$$

Application

❑ Application to "non-convex" dataset

– Data set of 400 data points (shown below, also used in last lecture)
– Initial clustering analysis: *K*-mean (k=2,...,15), 3 initial settings →42 partitions
– Converting clustering results to "distance" matrices to achieve the ensembled "distance matrix" without weighting (setting all weights to be one).
– Applying the Agglomerative algorithm to the ensemble "distance matrix"
– Cut the dendrogram tree with the maximum *K*-cluster life time to decide *K*

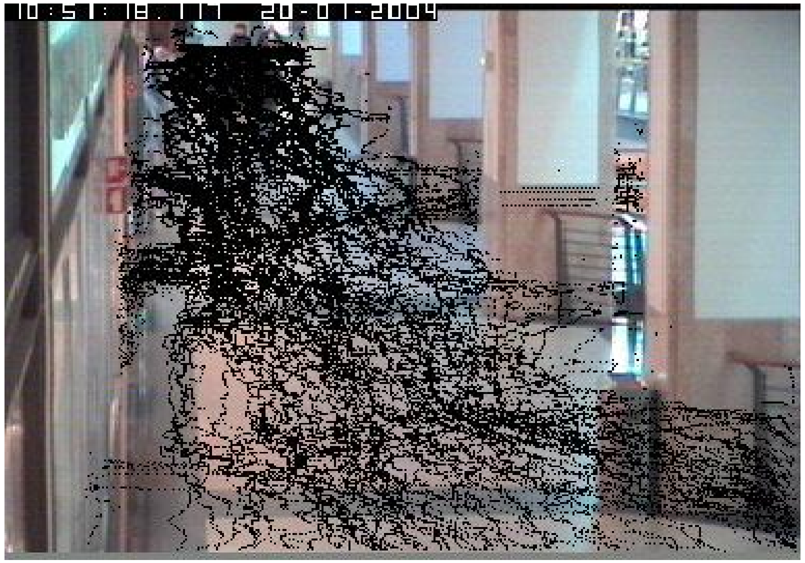




Application

❑ Clustering analysis on CAVIAR database

– annotated video sequences of pedestrians, a set of 222 high-quality moving trajectories
– clustering analysis of trajectories is useful for many applications



❑ Experimental setting

– **Representations:** PCF, DCF, PLS and PDWT
– **Initial clustering analysis:** K-mean algorithm (4<K<20), 5 initial settings
– **Ensemble:** 300 partitions totally (75 partitions/representation)

Application

- Clustering result on CAVIAR database

Application

- Application: UCR time series benchmarks

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

Application

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

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