Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Chapter 5
Data Clustering

Assoc. Prof. TRAN MINH QUANG

quangtran@hcmut.edu.vn

http://researchmap.jp/quang

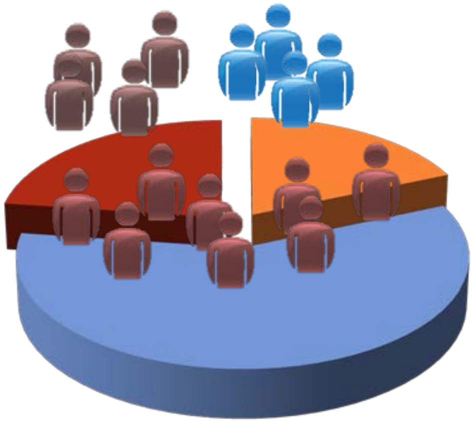
2020/6/30

CONTENT

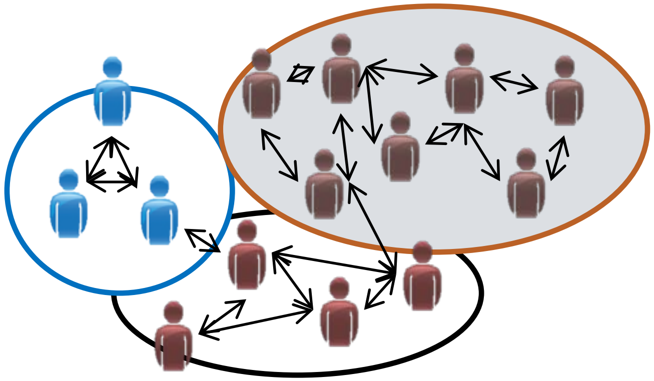
1. Overview on clustering
2. Partition-based clustering
3. Hierarchical clustering
4. Density-based clustering
5. Modeling-based clustering
6. Other clustering methods
7. Summary

1. OVERVIEW

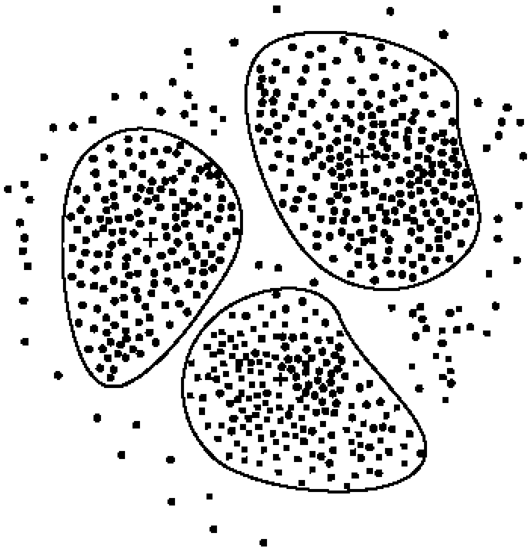
- Situations



Customer clustering



Clustering relationships from social network

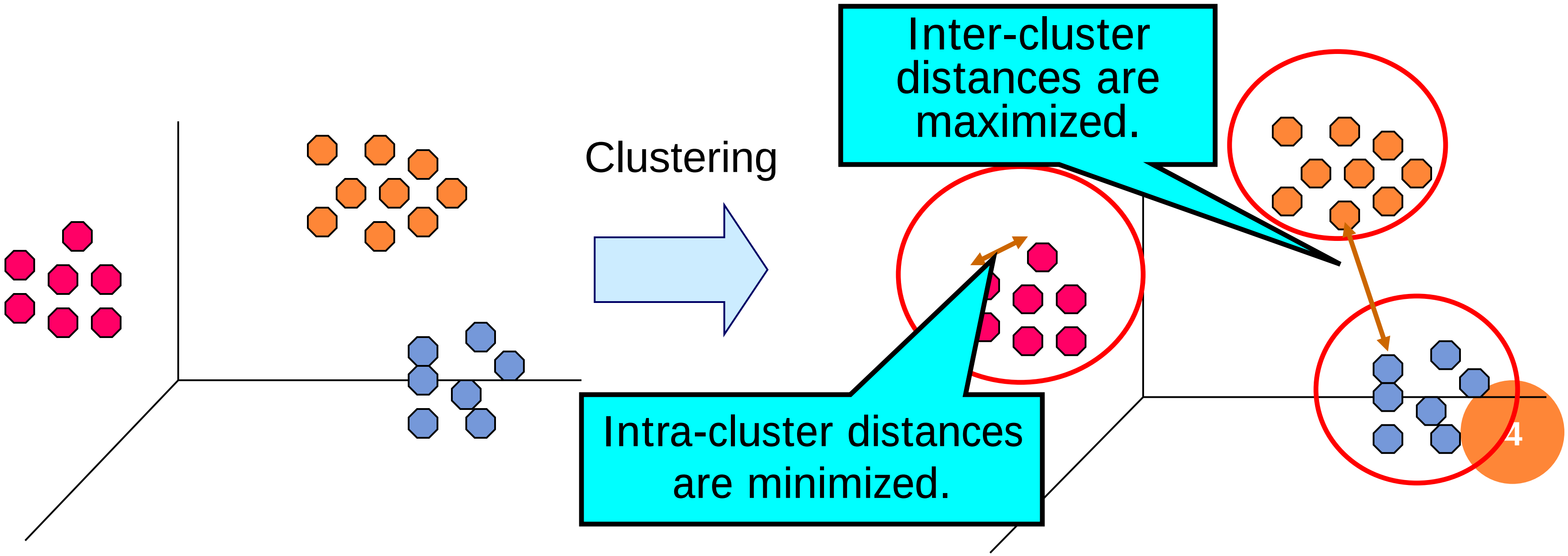


Clustering & outlier detections

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Overview

- Clustering: to cluster/group data objects
- Objects in a cluster are similar with each other compared to those from other clusters
  - *Obj1, Obj2 in C1; Obj3 in C2 → Obj1 is more similar to Obj2 than Obj3.*



1. Overview – Object Dissimilarity

- Problem on data types of data objects which are being clustered

Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- n objects

- p variables/attributes

Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- d(i,j) distance between object i and j, calculated based on the type of attributes/variables

<u>Note</u>: $d(i,i) = 0$; $d(i,j) = d(j,i) \geq 0$; $d(i,j) = d(i,k) + d(k,j)$

1. Overview – Object Dissimilarity

- Vector objects: i and j are presented as vectors x, y
- Similarity between i and j is calculated using cosine measure

$$s(x, y) = \frac{x^T . y}{|x||y|} \quad \text{where} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix}; y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{bmatrix}$$

Or

$s(x, y) = (x_1 * y_1 + \dots + x_p * y_p)/((x_1^2 + \dots + x_p^2)^{1/2} * (y_1^2 + \dots + y_p^2)^{1/2})$

1. OVERVIEW – OBJECT DISSIMILARITY

- Distance calculation: Based on the attribute type

✓ Interval-scaled variables/attributes

✓ Binary variables/attributes

✓ Categorical variables/attributes

✓ Ordinal variables/attributes

✓ Ratio-scaled variables/attributes

✓ Variables/attributes of mixed types

1. OVERVIEW – OBJECT DISSIMILARITY

- Interval-scaled variables/attributes

Mean absolute deviation
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

Mean
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

Z-score measurement
$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

<u>Note</u>: use z_{if} instead of x_{if}; i = 1..n, f = 1..p

1. Overview – Object Dissimilarity

- Euclidean

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

- Minkowski

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

- Manhattan

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Overview – Object Dissimilarity

- Binary variables/attributes

a: count of attrs whose values in i is 1 and in j is 1

		Object j		
		1	0	sum
	1	a	b	$a+b$
Object i	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p *(= a + b + c + d)*

- Use simple distance (if symmetric): $$d(i,j) = \frac{b+c}{a+b+c+d}$$

- Use Jaccard distance (if asymmetric): $$d(i,j) = \frac{b+c}{a+b+c}$$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Overview – Object Dissimilarity

- Binary variables/attributes
  - Ex

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - Gender: symmetric (probability an object gets “M”, “F” is similar)
  - Other binary attributes: asymmetric
  - Y, P → 1, N → 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

1. Overview – Object Dissimilarity

- Variables/attributes of mixed types

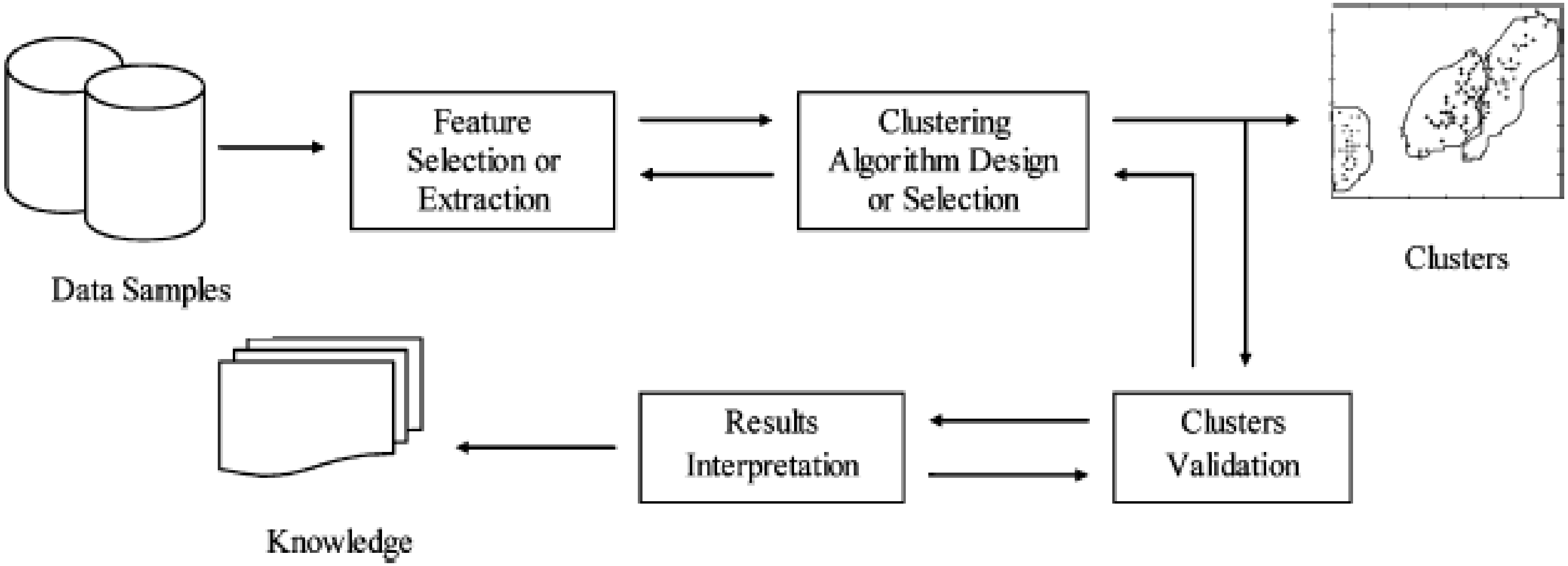
$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

- If x_{if} or x_{jf} is missed then $\delta_{ij}^{(f)} = 0$, else $\delta_{ij}^{(f)} = 1$
- For $d_{ij}^{(f)}$
  - *f (variable/attribute)* a binary (nominal): $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; $d_{ij}^{(f)} = 1$ in other cases
  - *f* : interval-scaled (Minkowski, Manhattan, Euclidean)
  - *f* : ordinal (ordered data, e.g. gold, silver, bronze medals) or ratio-scaled: let $r_{if} = \{1,..,M_f\}$, then use z_{if} instead of x_{if} and use Minkowski, Manhattan, or Euclidean for distance calculation

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

1. OVERVIEW – CLUSTERING

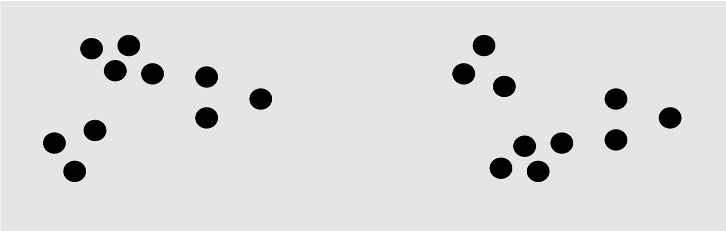
- Clustering process



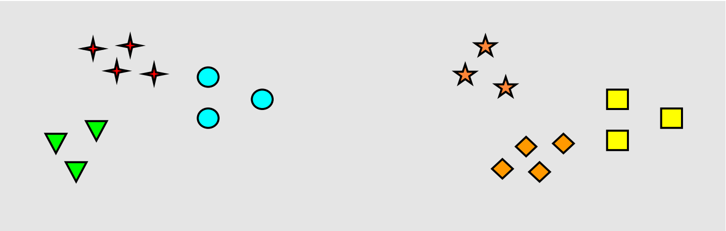
R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), May 2005, pp. 645-678.

1. OVERVIEW – CLUSTERING

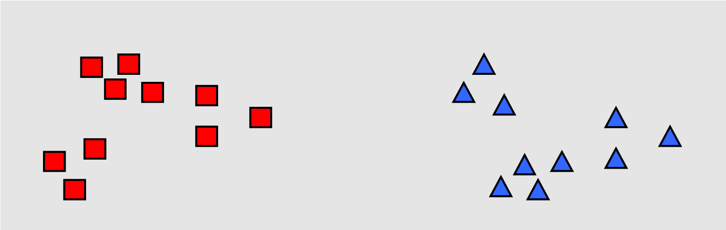
- How many clusters should be created?
- How many objects in each clusters?
- A particular object should belong to how many clusters?



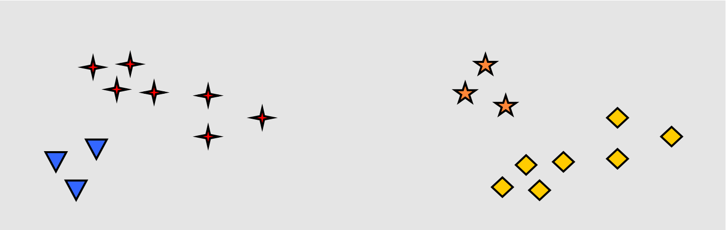
How many clusters



6 clusters?



2 clusters?



4 clusters?

1. OVERVIEW - CLUSTERING

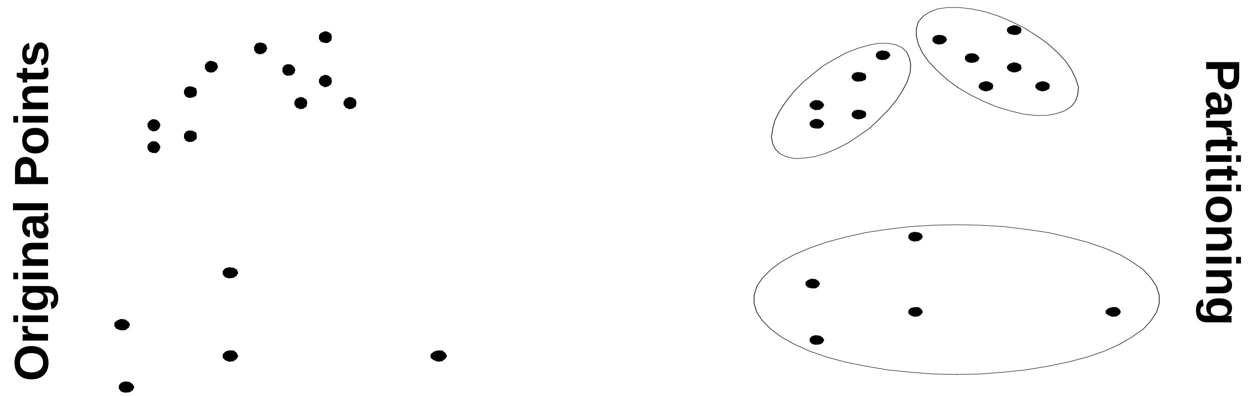
- Essential requirements in a clustering method
  - Scalability: work with the change of data size, type,…
  - Ability to process with multiple data types (high dimensionality)
  - Ability to provide clusters with arbitrary shapes
  - Require minimum input parameters/instructions
  - Ability to work with noisy data
  - Ability of incremental clustering and insensitivity to the order of input records
  - Interpretability and usability

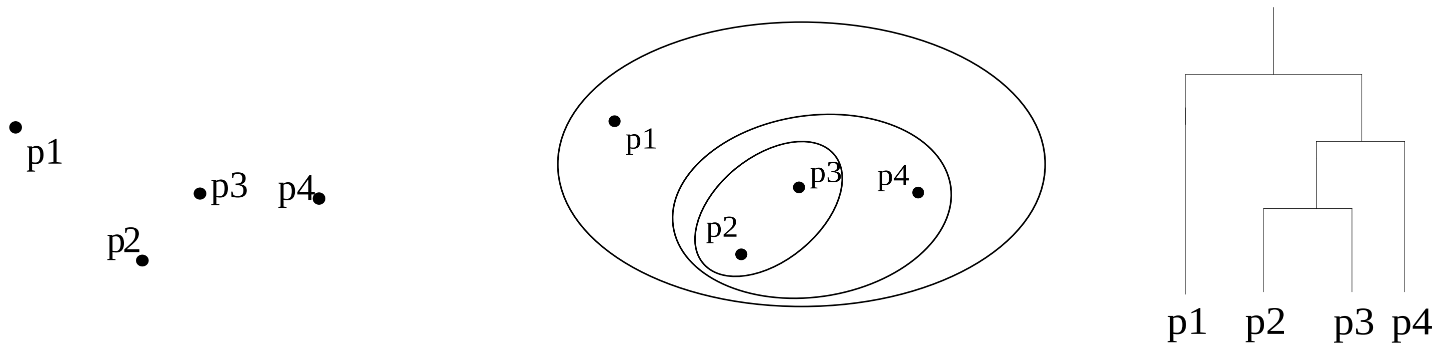
1. Overview - Clustering

- Common clustering methods
  - Partitioning: partitions are created and evaluated based on some criteria
  - Hierarchical: the data set is hierarchically divided (ordering) based on some criteria
  - Density-based clustering: based on connectivity and density of data distribution
  - Model-based: a hypothesis model (distribution model) of a cluster (a subset of data) is proposed, then parameters are modified to make the model best fit with the data set.

1. OVERVIEW - CLUSTERING

- Intuitive examples





Original Points **Hierarchical**

1. Overview - Clustering

- Evaluation methods
  - External validation: evaluate the clustering results based on pre-defined/given structures of clusters
  - Internal validation: evaluation based on proximity matrix calculated from given dataset
  - Relative validation: Based on the comparison with other approaches/methods

  → Criteria to evaluated and select clustering methods

    - ***Compactness****: objects in a cluster should be close with each other*
    - ***Separation****: Clusters should be far apart*

1. Overview - Clustering

- External validation
  - Measure: Rand statistic, Jaccard coefficient, Folkes and Mallows index, …
- Iinternal validation
  - Measure: Silhouette index, Dunn’s index, …
- Relative validation: compare different approaches on effectiveness and efficiency

1. Overview - Clustering

- external validation measures – contingency matrix

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$	$[0, \log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i[2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$[0, 2\log\max(K, K')]$
5	Mutual Information	MI	$\sum_i\sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$(0, \log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}{\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}][\binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i \binom{n_{i\cdot}}{2} - 2\sum_j \binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j \binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n}\max_\sigma \sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n.$

J. Wu and et al. Adapting the Right Measures for K-means Clustering. KDD'09, Paris, France, July 2009.

2. Partition-based Clustering

- Evaluating the clustering results

Partition C

Partition P		C_1	C_2	$\cdots$	$C_{K'}$	$\sum$
	P_1	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1\cdot}$
	P_2	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2\cdot}$
	$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$
	P_K	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K'}$	n

Contingency matrix

- Partition P: clustering result from n objects

- Partition C: actual clusters of n objects

- $n_{ij} = |P_i \cap C_j|$: number of object in P_i extracted from C_j

2. Partition-based Clustering

- Evaluating the clustering results

I	C_1	C_2	C_3	$\sum$
P_1	3	4	12	19
P_2	8	3	12	23
P_3	12	12	0	24
$\sum$	23	19	24	66

II	C_1	C_2	C_3	$\sum$
P_1	0	7	12	19
P_2	11	0	12	23
P_3	12	12	0	24
$\sum$	23	19	24	66

Examine the results from 2 methods, namely I and II

- Partition P: clustering results from n (=66) objects

- Partition C: actual clusters of n (=66) objects

- $n_{ij} = |P_i \cap C_j|$: number of object in P_i extracted from C_j

2. Partition-based Clustering

- Evaluating the clustering results
  - Entropy (smaller is better)

$$Entropy(I) = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$$

$$= -\sum_i \frac{n_i}{n} (\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i})$$

$$= -\frac{19}{66}(\frac{3}{19}\log\frac{3}{19} + \frac{4}{19}\log\frac{4}{19} + \frac{12}{19}\log\frac{12}{19})$$

$$-\frac{23}{66}(\frac{8}{23}\log\frac{8}{23} + \frac{3}{23}\log\frac{3}{23} + \frac{12}{23}\log\frac{12}{23})$$

$$-\frac{24}{66}(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24})$$

$$= ???$$

$$Entropy(II) = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$$

$$= -\sum_i \frac{n_i}{n} (\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i})$$

$$= -\frac{19}{66}(\frac{0}{19}\log\frac{0}{19} + \frac{7}{19}\log\frac{7}{19} + \frac{12}{19}\log\frac{12}{19})$$

$$-\frac{23}{66}(\frac{11}{23}\log\frac{11}{23} + \frac{0}{23}\log\frac{0}{23} + \frac{12}{23}\log\frac{12}{23})$$

$$-\frac{24}{66}(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24})$$

$$= ???$$

→ What method (I or II) is better?

2. Partition-based: K-Means

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

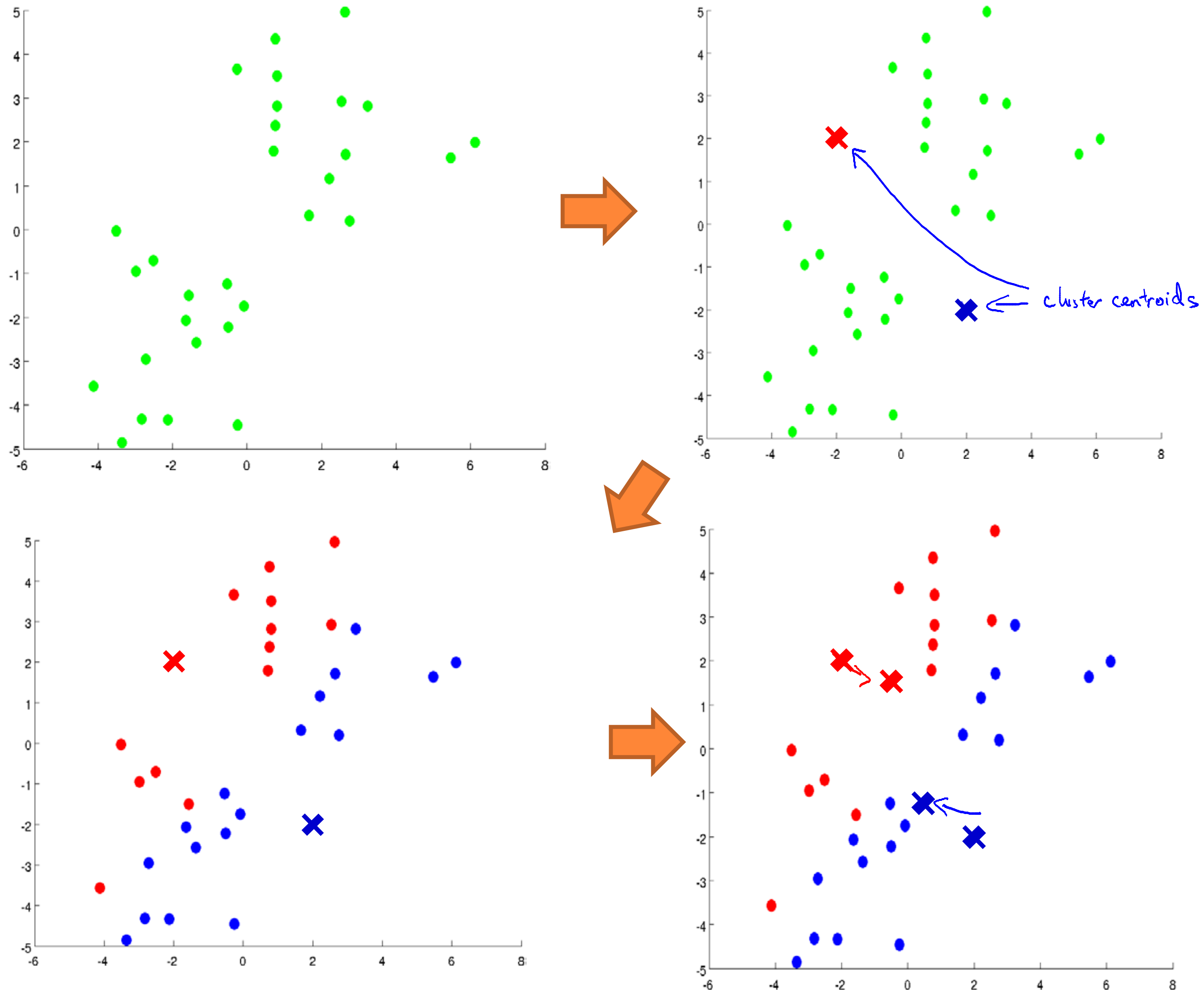
Input:

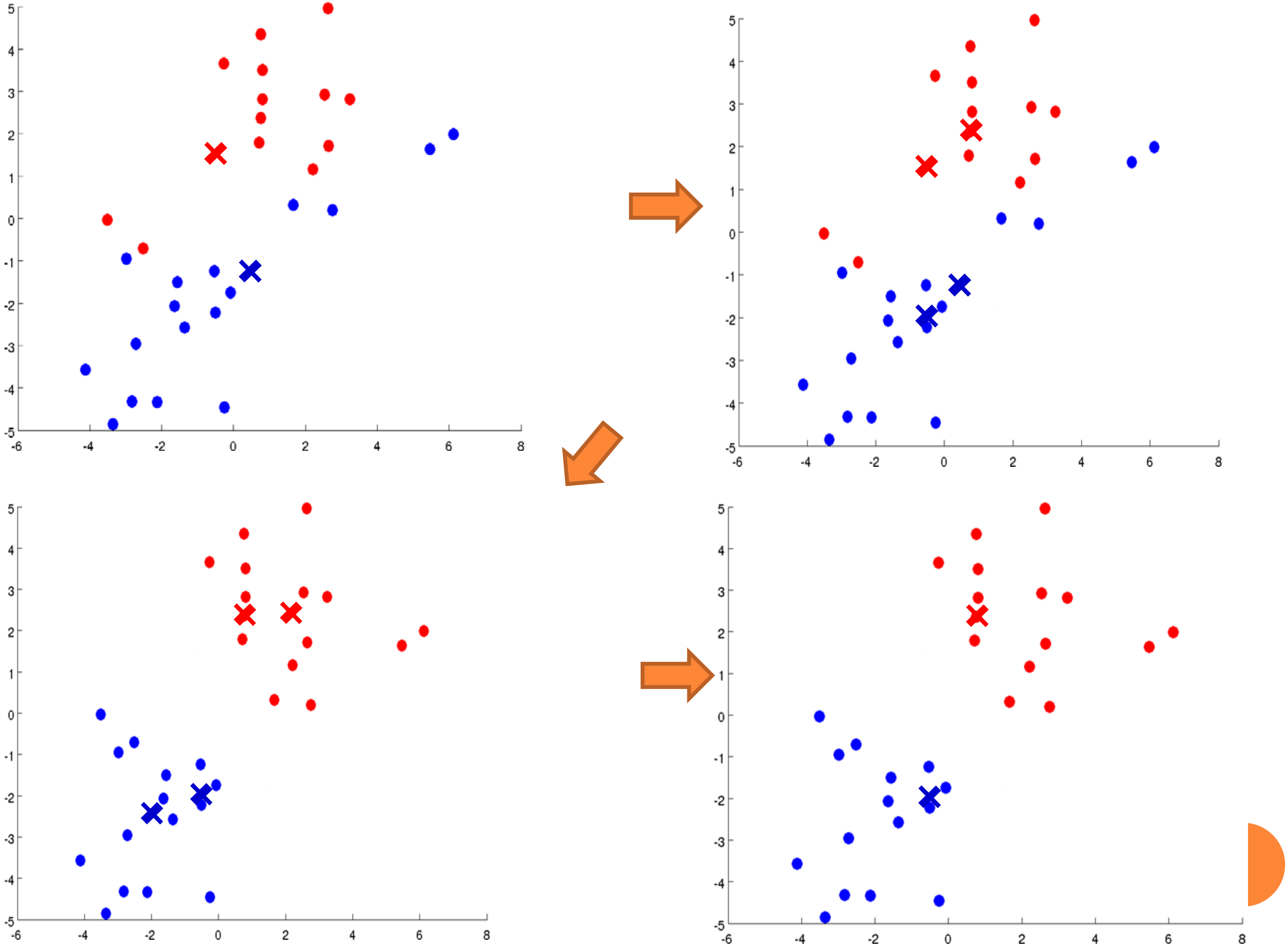
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

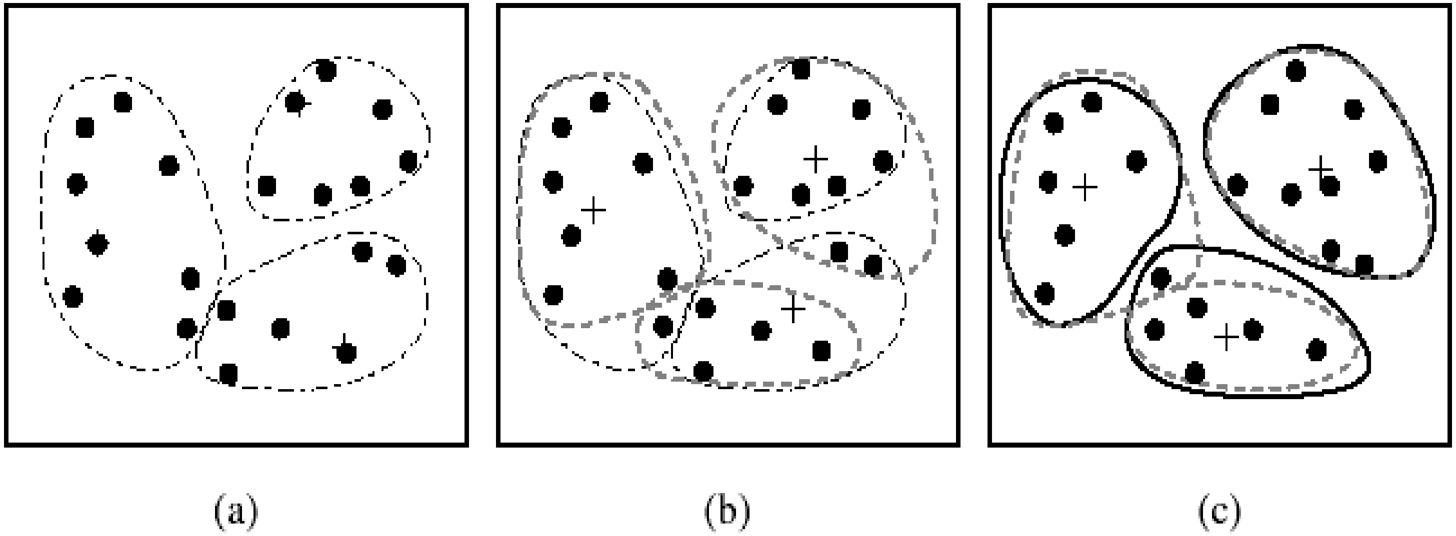
Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;


cluster centroids


5 4 3 2 1 0 -1 -2 -3 -4 -5
-6 -4 -2 0 2 4 6 8
5 4 3 2 1 0 -1 -2 -3 -4 -5
-6 -4 -2 0 2 4 6 8
5 4 3 2 1 0 -1 -2 -3 -4 -5
-6 -4 -2 0 2 4 6 8
5 4 3 2 1 0 -1 -2 -3 -4 -5
-6 -4 -2 0 2 4 6 8

2. Partition-based Clustering



Clustering of a set of objects based on the *k*-means method. (The mean of each cluster is marked by a "+".)

2. Partition-based Clustering

- Quality of cluster C_i is measured by

$$s_i = \sum_{x \in C_i} dist(x, r_i)^2$$

- Quality of the clustering method (k clusters)

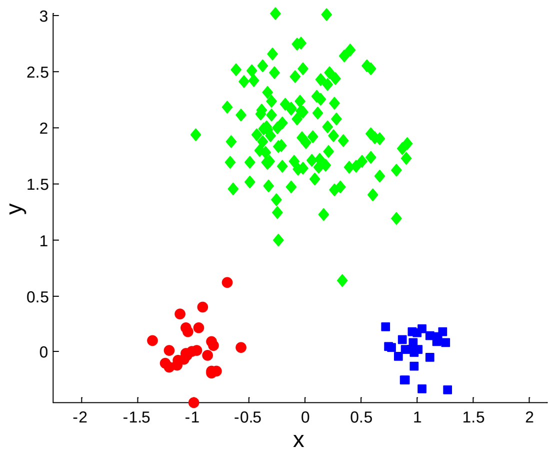
$$s = \sum_{i=1}^{k} s_i$$

- Note: $dist(x, r_i)$: the distance from object x in C_i to the centroid of r_i the cluster
  => smaller or larger is better?

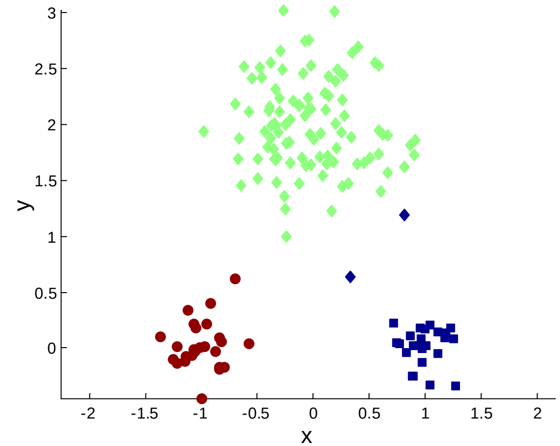
2. PARTITION-BASED CLUSTERING

- K-means' characteristics
  - Local optimization problem
  - A cluster is characterized by its center (i.e. the "mean" object)
    - ✓ How is the good value for k?
    - ✓ Complexity: *O(nkt)*, where n is the dataset size, k is the number of clusters, t is the number of iterations (k << n, t << n)
  - Affected by noise, outliers
  - Not appropriate for clustering nonconvex clusters of those with different sizes
    - ✓ Results: hyperspherial shape clusters
    - ✓ Relatively uniform sizes

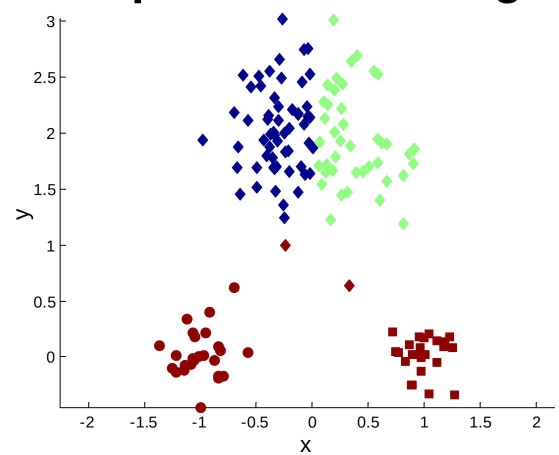
2. PARTITION-BASED CLUSTERING



Original Points



Optimal Clustering



Sub-optimal Clustering

Note: refer to other partition-based methods such as PAM(k-medoids) method

3. HIERARCHICAL CLUSTERING

- Hierarchical clustering: group objects into clusters based on hierarchy
  - Agglomerative: bottom-up
  - Divisive: top-down

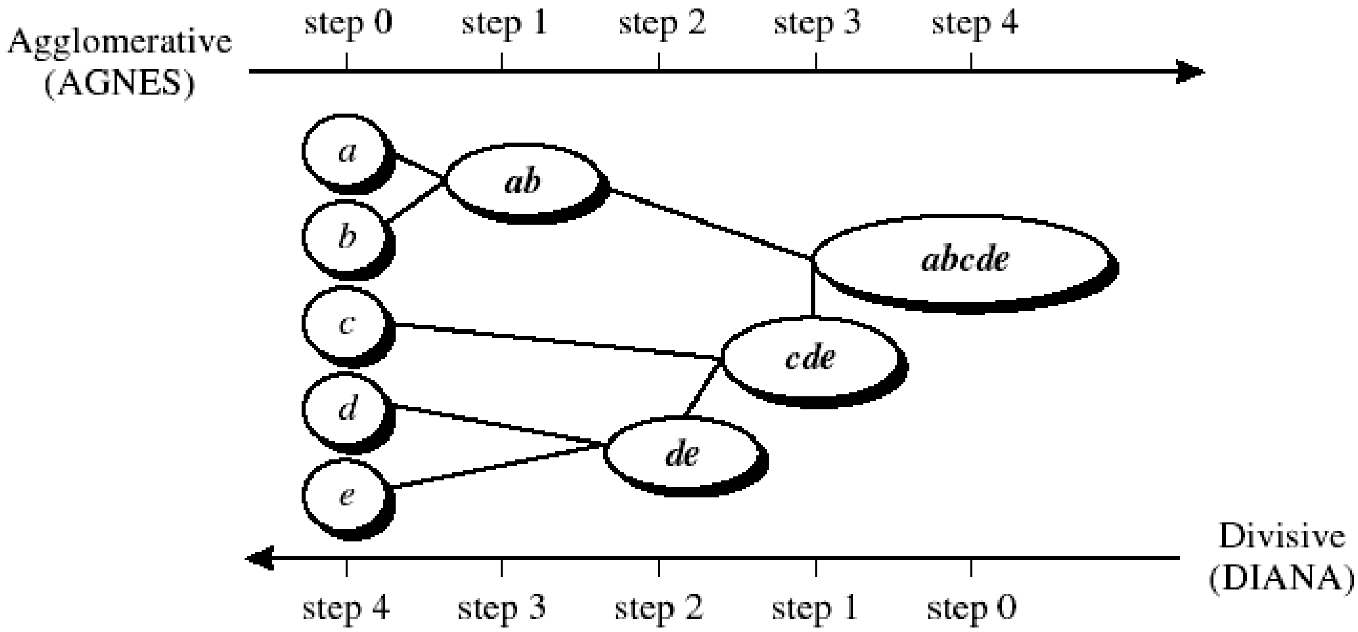
→ Don't need number of clusters (i.e., k in K-Means)

→ Need a stop condition when building the tree

→ Can't turn back at each agglomeration/division step

3. HIERARCHICAL CLUSTERING

- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting) → bottom-up
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis) → top-down

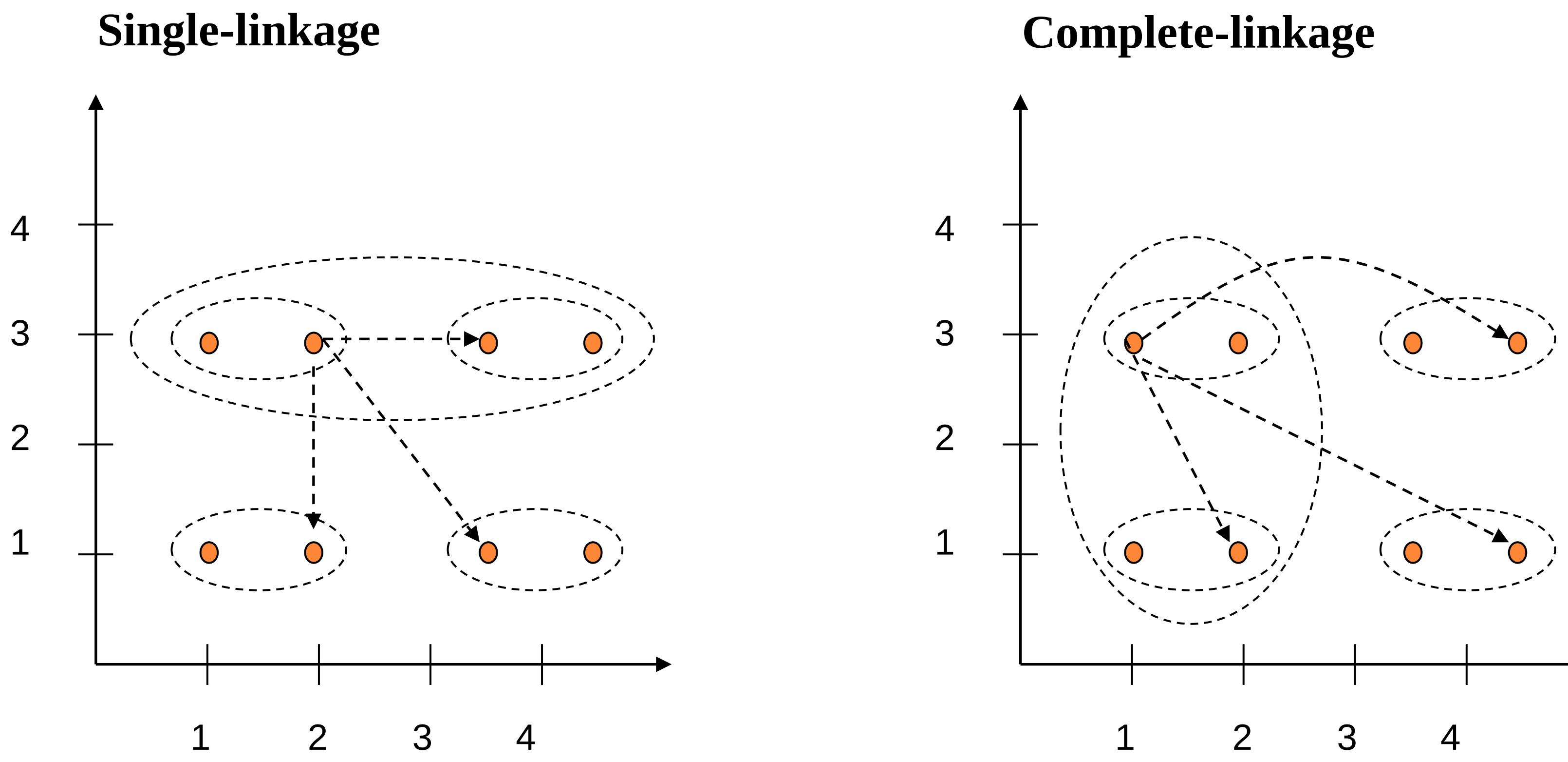


Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

3. Hierarchical clustering

- AGNES (Agglomerative NESting)
  - Initiation: each object = cluster (n clusters)
  - Merge objects based on some criteria
    - Single-linkage approach: Based on the shortest distance between two objects in the two clusters C1, C2
    - Complete-linkage: Based on the longest distance between two objects in the two clusters C1, C2
  - Merging is iterated until all objects are in the same cluster

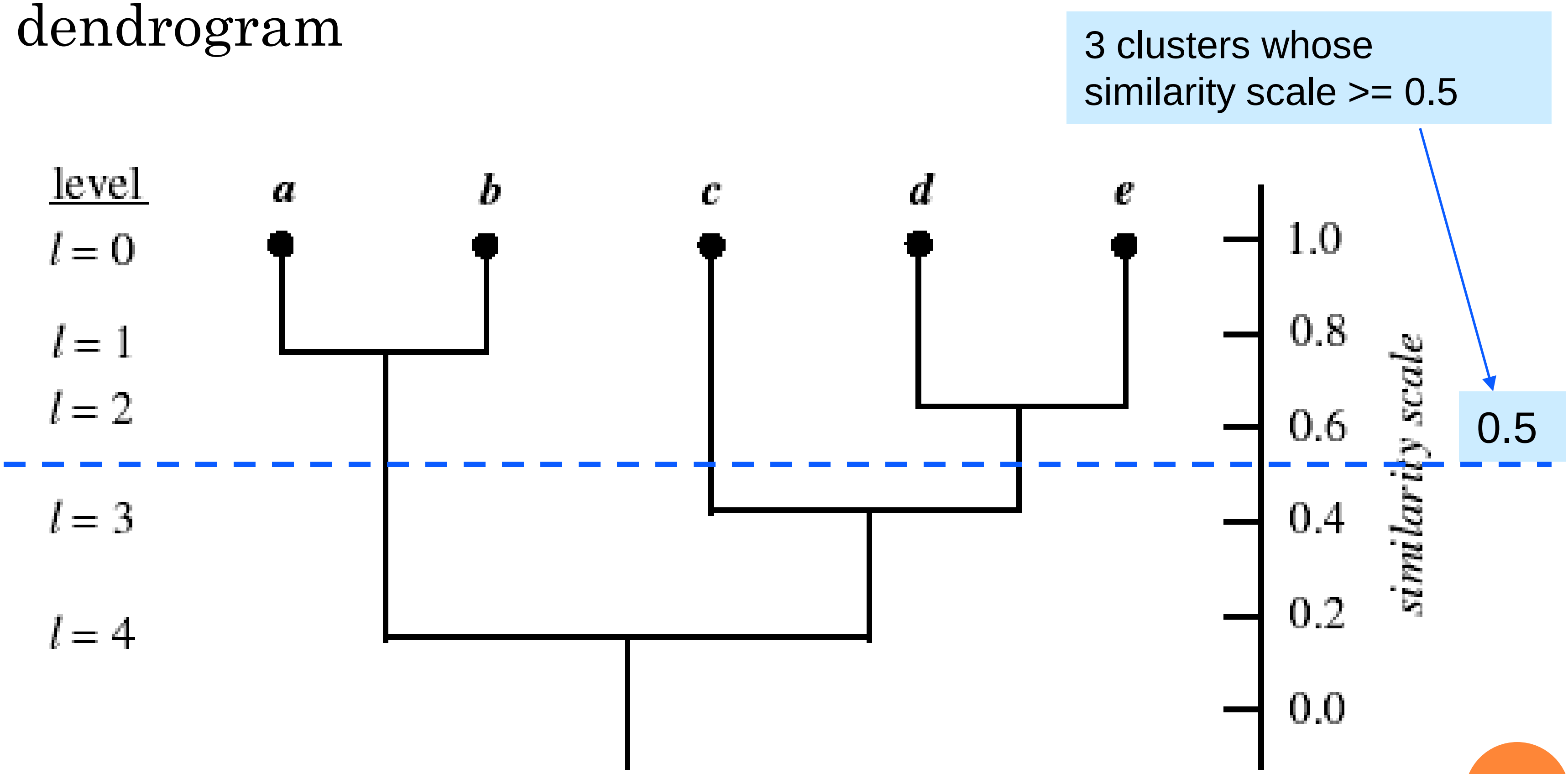
3. HIERARCHICAL CLUSTERING



Agglomeration criteria: single-linkage and complete-linkage

3. Hierarchical clustering

- The process of hierarchical clustering is presented by dendrogram



3. Hierarchical clustering

- Measuring the distance between 2 clusters C_i and C_j

$$\text{Minimum distance}: \quad d_{min}(C_i, C_j) = \min_{\boldsymbol{p} \in C_i,\, \boldsymbol{p}' \in C_j} |\boldsymbol{p} - \boldsymbol{p}'|$$

$$\text{Maximum distance}: \quad d_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i,\, \boldsymbol{p}' \in C_j} |\boldsymbol{p} - \boldsymbol{p}'|$$

$$\text{Mean distance}: \quad d_{mean}(C_i, C_j) = |\boldsymbol{m_i} - \boldsymbol{m_j}|$$

$$\text{Average distance}: \quad d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i} \sum_{\boldsymbol{p}' \in C_j} |\boldsymbol{p} - \boldsymbol{p}'|$$

p, p’: objects in the two clusters

|p-p’|: distance between p and p’

m_i, m_j: the “mean” objects of C_i, C_j

n_i, n_j: number of objects in C_i, C_j

3. HIERARCHICAL CLUSTERING

- Some hierarchical clustering algorithms
  - BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
  - ROCK (Robust Clustering using linKs): applied to categorical/discrete attributes
  - Chameleon: using a dynamic model to identify the similarity between pair of clusters

3. Hierarchical Clustering

- Issues in hierarchical clustering methods
  - Need to appropriately identify the agglomeration/division point
  - Scalability: each decision on agglomeration/division must evaluate many objects

  → We can integrate hierarchical clustering with other methods to improve the performance, scalability,…

    → Ex., multiple-phase clustering (a kind of divide and conquer method)

4. DENSITY-BASED CLUSTERING

- Each cluster is a dense region of objects
- Objects in the sparse regions are outliers
- The size and shape of clusters are diverse
- Some well-known algorithms
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - OPTICS (Ordering Points To Identify the Clustering Structure)
  - DENCLUE (DENsity-based CLUstEring): *based on distribution functions*

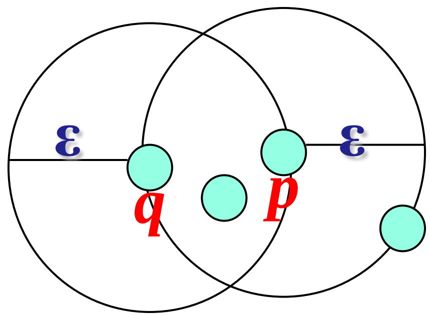
Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

4. DENSITY-BASED CLUSTERING

- Concepts
  - ε: neighborhood radius of an object
  - ε-neighborhood: Number of objects in the neighborhood region (defined by ε)
  - Core object: is an object that satisfies (*MinPts is given)*

$$\varepsilon\text{-neighborhood} \geq \boldsymbol{MinPts}$$

  - ***Directly density-reachable***: *q directly density-reachable* from *p* if *q* in *p*'s ε-neighborhood and *p* is a core object.



p: core object (***MinPts*** = 3)

q: is not a core object

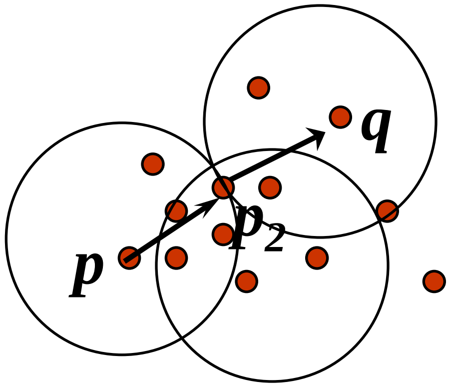
p: directly density-reachable from *q*? X

q: directly density-reachable from *p*? √

4. Density-based clustering

- ***Density-reachable***:
  - Given a data set $\boldsymbol{D}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{MinPts}$
  - q ***density-reachable*** from p if $\exists$ a chain of objects p_1, ..., $p_n \in \boldsymbol{D}$, where $p_1 = p$ and $p_n = q$ so that p_{i+1} ***directly density-reachable*** from p_i (in accordance with $\boldsymbol{\varepsilon}$, $\boldsymbol{MinPts}$, $1 \leq i \leq n$).
  - Note:
    - Transitive closure of the "directly density-reachable"
    - Asymmetric

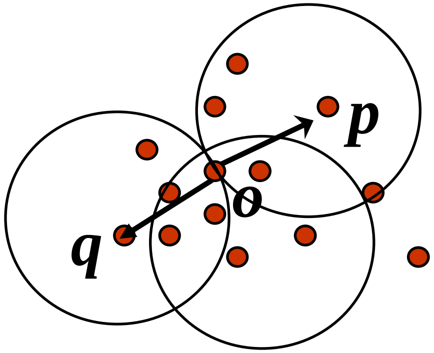
MinPts = 5

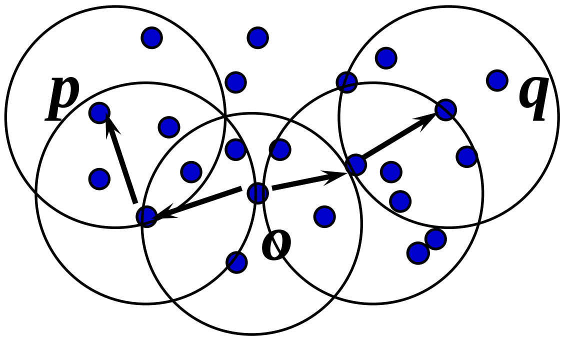


Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

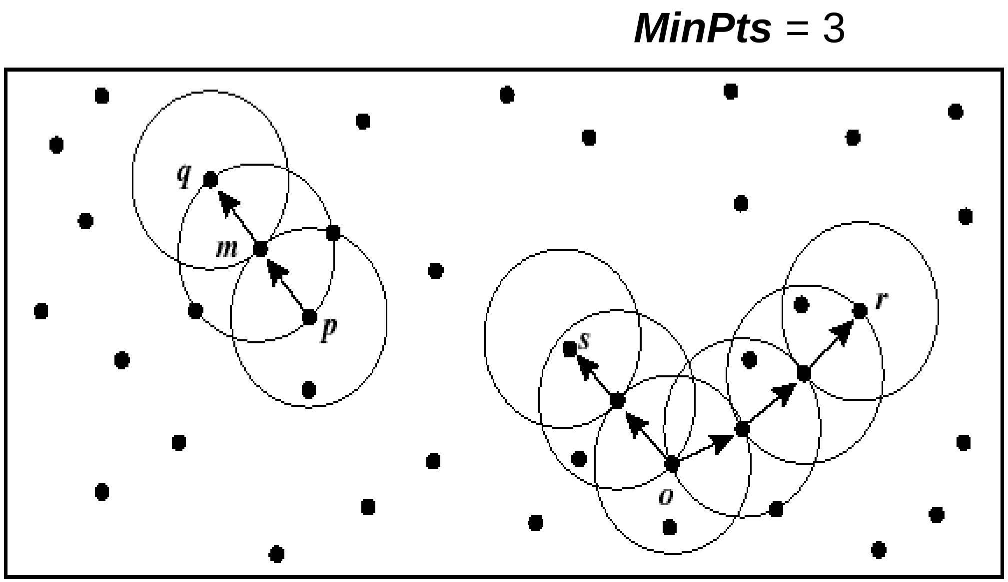
4. DENSITY-BASED CLUSTERING

- ***Density-connected*** :
  - Given $\boldsymbol{D}$, $\boldsymbol{\varepsilon}$, and ***MinPts***
  - $p, q \in \boldsymbol{D}$
  - q ***density-connected*** with p if $\exists\ o \in \boldsymbol{D}$ so that q and p are ***density-reachable*** from o.
  - **Symmetric**

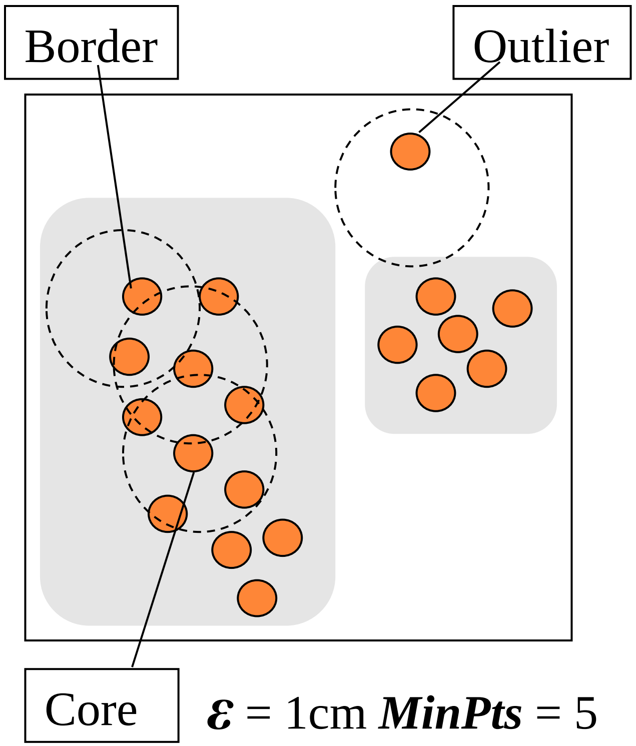




4. Density-based clustering



Density reachability and density connectivity in density-based clustering



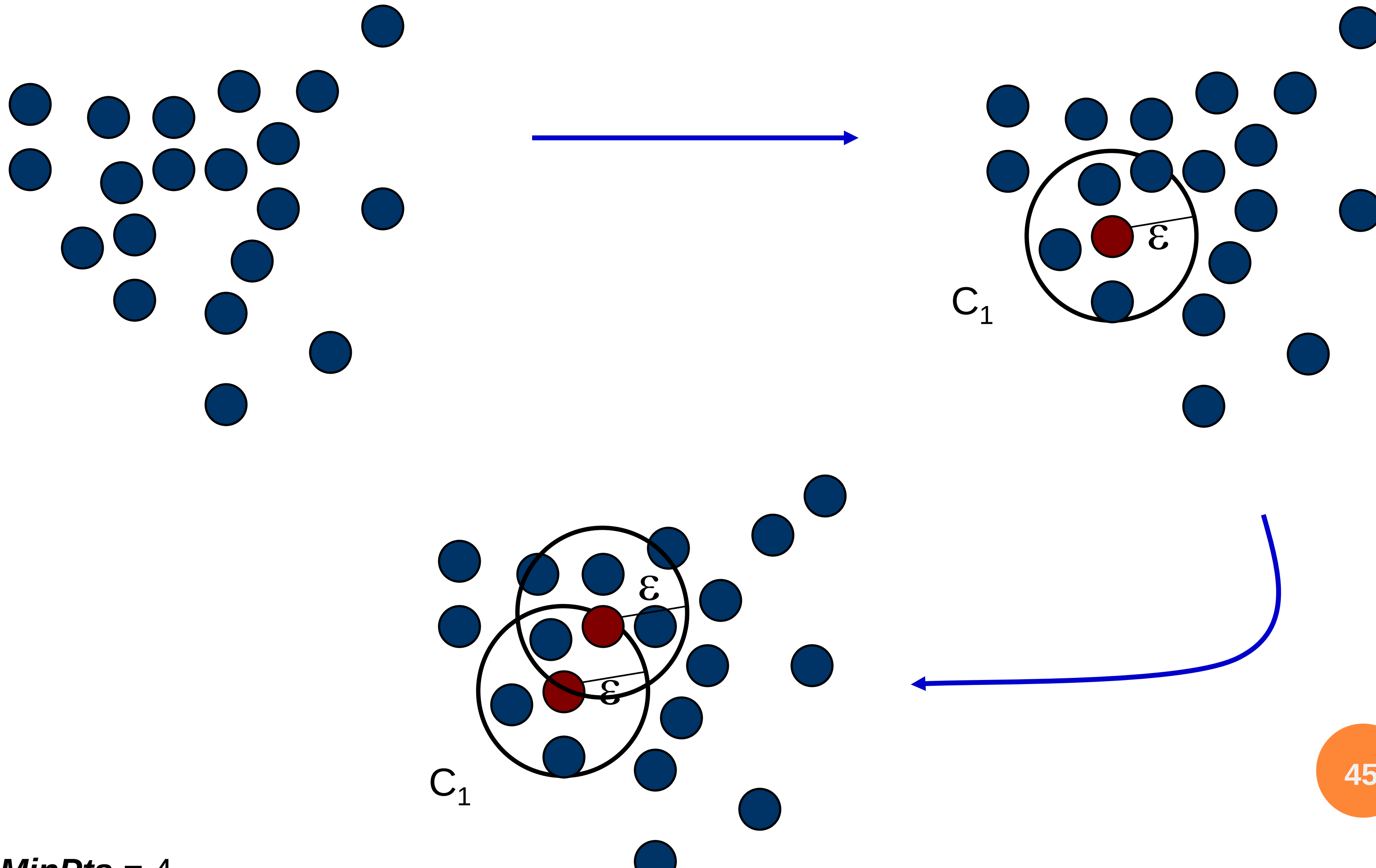
- Density based cluster: a set of objects connected with each other based on density. It consists of core objects border objects
- Noises/outliers are those do not belong to a any cluster

4. DENSITY-BASED CLUSTERING

- DBSCAN
  - Input: $\boldsymbol{D}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{MinPts}$
  - Output: density-based clusters (and noises/outliers)

  - Algorithm
    - 1. Identify $\boldsymbol{\varepsilon}$–neighborhood for each $p \in \boldsymbol{D}$
    - 2. If p is a core object -> **create a cluster** for **p**
    - 3. From any core object p, find all ***density-reachable*** objects (or clusters) and put them to p's cluster
      - 3.1. Density-reachable clusters can be merged
      - 3.2. Stop when there is no object can be put into clusters

4. DENSITY-BASED CLUSTERING



MinPts = 4

4. Density-based clustering

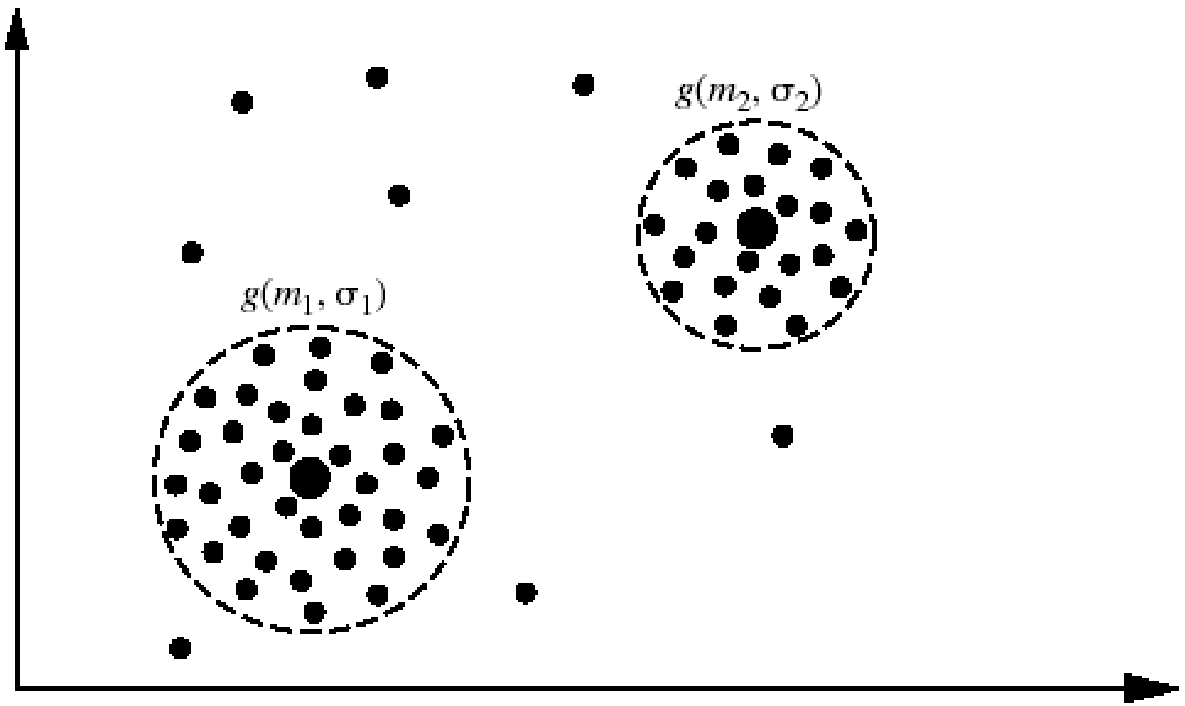
- DBSCAN characteristics
  - Clusters' sizes and shapes are diverse
    - No consumption about the object distribution
    - Don't need initial *k* (number of clusters)
    - Initialization doesn't affect the result
    - Need to defined the "density", i.e., ε and ***MinPts***
  - It can help to identify noise and outliers effectively
  - Complexity: O(nlogn) → O(n^2)

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

5. MODEL-BASED CLUSTERING

- Optimize the fit between data and some math models
  - Assumption: Data is generated based on some probability distribution models
- Methods
  - Statistical approaches: Extension of the k-means: Expectation-Maximization (EM)
  - Machine learning approaches: conceptual clustering
  - ANN based approaches: Self-Organizing Feature Map (SOM)

5. Model-based clustering



Each cluster can be represented by a probability distribution, centered at a mean, and with a standard deviation. Here, we have two clusters, corresponding to the Gaussian distributions $g(m_1, \sigma_1)$ and $g(m_2, \sigma_2)$, respectively, where the dashed circles represent the first standard deviation of the distributions.

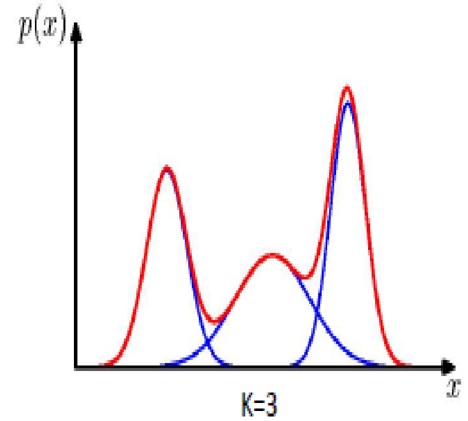
5. Model-based clustering

- Assume that data is generated based on some Gaussian models
- Each Gaussian model is parameterized by ϴ (μ_i, Σ_i)
  - Center: μ_i
  - Variance: Σ_i (ignore)
- Find the cluster (in k ones) where x_i belongs

Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{Mixing coefficient}} \underbrace{\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}_{\text{Component}}$$

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



z_{ij} : if x_i belongs to j-th cluster

5. Model-based clustering

- Probability that x is x_i $\quad p(x = x_i)$

$$p(x = x_i) = \sum_{\mu_j} p(x = x_i, \mu = \mu_j) = \sum_{\mu_j} p(\mu = \mu_j) p(x = x_i \mid \mu = \mu_j)$$

$$= \sum_{\mu_j} p(\mu = \mu_j) \frac{1}{\left(2\pi\sigma^2\right)^{d/2}} \exp\left(- \frac{\left\| x_i - \mu_j \right\|_2}{2\sigma^2} \right)$$

- Log-likelihood of data

$$\sum_i \log p(x = x_i) = \sum_i \log \left[\sum_{\mu_j} p(\mu = \mu_j) \frac{1}{\left(2\pi\sigma^2\right)^{d/2}} \exp\left(- \frac{\left\| x_i - \mu_j \right\|_2}{2\sigma^2} \right) \right]$$

- Find algorithm to Maximize Log-likelihood

5. MODEL-BASED CLUSTERING

- Expectation-Maximization (EM) algorithm
  - Is an iterative algorithm to find the ***Maximum Likelihood (ML)*** – workable even there is missing data
  - **EM** consists of 2 steps:
    - **Expectation step:** the (missing) data are estimated given the observed data and current estimates of model parameters
    - **Maximization step:** The likelihood function is maximized under the assumption that the (missing) data are known

5. MODEL-BASED CLUSTERING

E-Step

M-Step

$$E[z_{ij}] = p(\mu = \mu_j \mid x = x_i)$$

$$= \frac{p(x = x_i \mid \mu = \mu_j)\, p(\mu = \mu_j)}{\sum_{n=1}^{k} p(x = x_i \mid \mu = \mu_n)\, p(\mu = \mu_j)}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}\, p(\mu = \mu_j)}{\sum_{n=1}^{k} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}\, p(\mu = \mu_n)}$$

$$\mu_j \leftarrow \frac{1}{\sum_{i=1}^{m} E[z_{ij}]} \sum_{i=1}^{m} E[z_{ij}]\, x_i$$

$$p(\mu = \mu_j) \leftarrow \frac{1}{m} \sum_{i=1}^{m} E[z_{ij}]$$

5. MODEL-BASED CLUSTERING

- Summarize the EM:
  - Input: **D** (**n** object), **K** clusters
  - Output: Optimal parameter Θ (μ_i, Σ_i) describing the model
  - Algorithm:

    1. Initiation

       1.1. Randomly select **K** objects as K clusters' centers

       1.2. Estimate the initial values for Θ (μ_i, Σ_i) (if needed)

    2. Iterate the process of modifying Θ (μ_i, Σ_i) (i.e., clusters):

       2.1. **E-step**: assign x_i to C_k with the probability $P(x_i \in C_k)$, where k=1..**K**

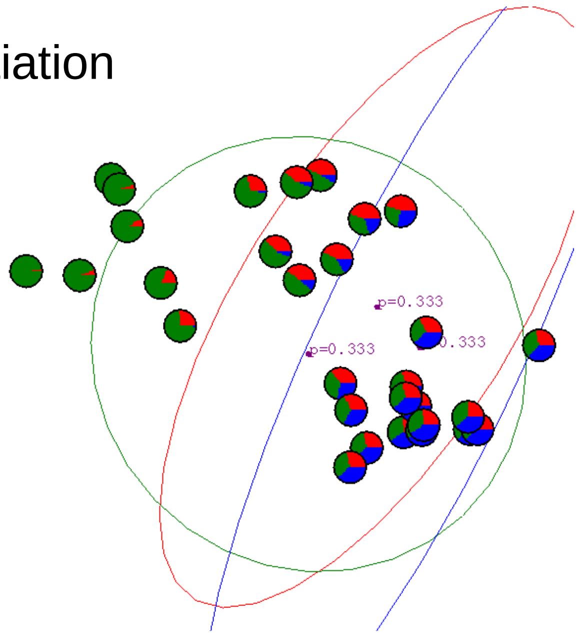
       2.2. **M-step**: Estimate Θ (μ_i, Σ_i)

       2.3. Stop when given condition/criteria is reached (e.g. ML)

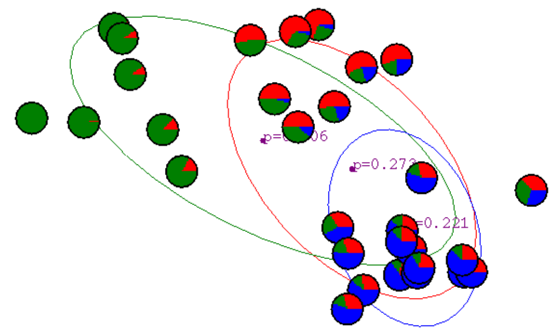
Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

5. MODEL-BASED CLUSTERING

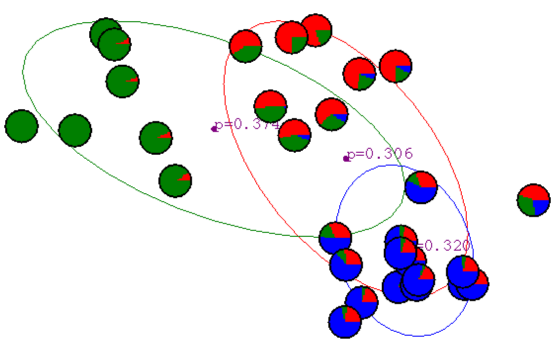
Initiation



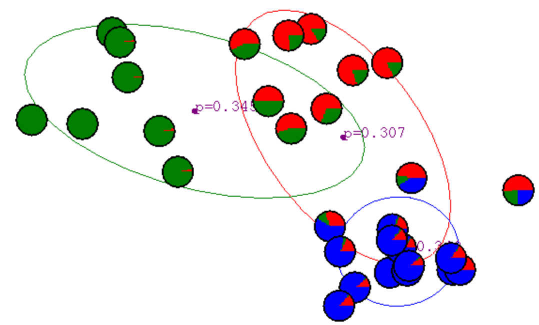
After 1st step



After 2nd step

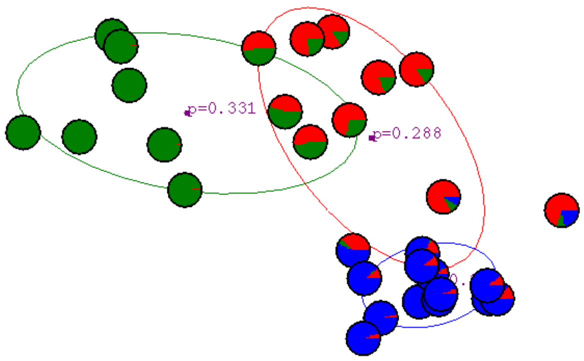


After 3rd step

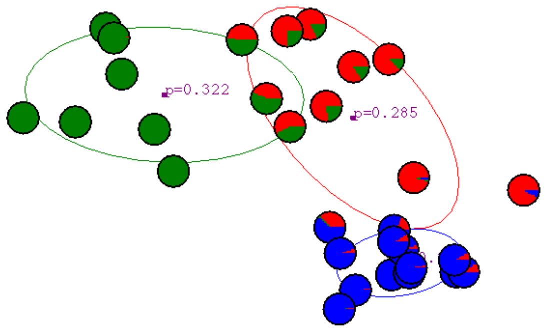


5. Model-based clustering

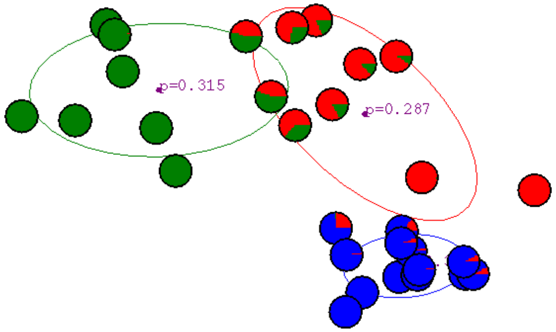
After 4th step



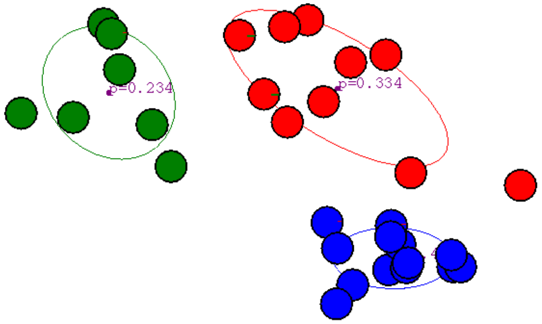
After 5th step



After 6th step

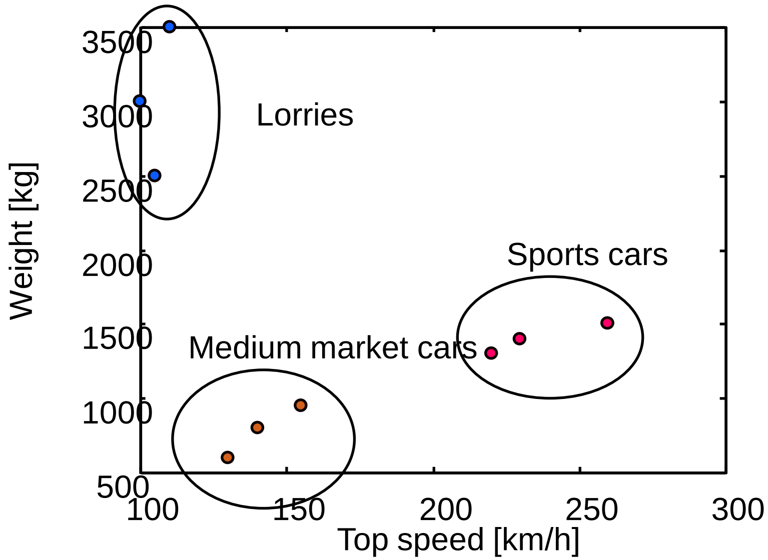


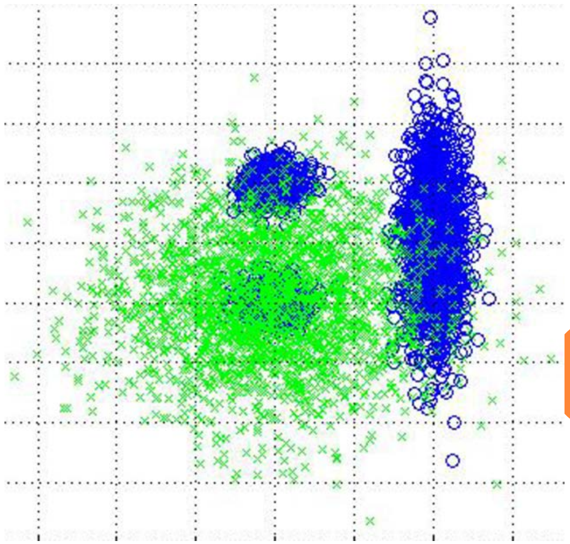
After 20th step



6. OTHER CLUSTERING METHODS

- Hard clustering
  - Each object belongs to only 1 cluster
  - Degree of membership (DoM): each object to each cluster is 0 or 1
  - Boundary: between clusters is clear
- Fuzzy clustering
  - An object can be belong to more than one cluster with the DoM from 0 to 1
  - Boundary: is vague/fuzzy





7. SUMMARY

- Clustering: to group objects into cluster based on their similarity
- Measuring similarity is based on data types
- Common approaches: partition-based, hierarchy, density-based, model-based, …

7. SUMMARY

Cluster algorithm	Complexity	Capability of tackling high dimensional data
K-means	$O(NKd)$ (time) $O(N+K)$ (space)	No
Fuzzy c-means	Near $O(N)$	No
Hierarchical clustering*	$O(N^2)$ (time) $O(N^2)$ (space)	No
CLARA	$O(K(40+K)^2+K(N-K))^+$ (time)	No
CLARANS	Quadratic in total performance	No
BIRCH	$O(N)$ (time)	No
DBSCAN	$O(N \log N)$ (time)	No
CURE	$O(N_{sample}^2 \log N_{sample})$ (time) $O(N_{sample})$ (space)	Yes
WaveCluster	$O(N)$ (time)	No
DENCLUE	$O(N \log N)$ (time)	Yes
FC	$O(N)$ (time)	Yes
CLIQUE	Linear with the number of objects, Quadratic with the number of dimensions	Yes
OptiGrid	Between $O(Nd)$ and $O(Nd \log N)$	Yes
ORCLUS	$O(K_0^3+K_0Nd+K_0^2d^3)$ (time) $O(K_0d^2)$ (space)	Yes

R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), May 20
pp. 645-678.

REFERENCE

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

Q&A

quangtran@hcmut.edu.vn