Cluster Analysis

DANG THI THU HIEN
Bài giảng của DSLab
Viện nghiên cứu cao cấp về Toán (VIASM)

Vietnam Institute for
Advanced Study in Mathematics

Nội dung

1. Giới thiệu, Bài toán, Dữ liệu
2. Độ đo khác biệt/tương tự
3. Kỹ thuật phân cụm
    1. Phân hoạch, Phân cấp, Mật độ
    2. Thuật toán đặc thù: FCM, EM (Nhóm chuyên sâu)
4. Chất lượng và đặc trưng các cụm (Nhóm chuyên sâu)
5. Thực hành ứng dụng bài toán VNA

Công cụ thực hành

- Scikit-learn Data Clustering (Python) http: //scikit-learn.org/stable/modules/clustering.html
- Open Source Data Mining Software (WEKA Workbench) http://www.cs.waikato.ac.nz/ml/weka/
- Apache Mahout Machine Learning Library http://mahout.apache.org/users/clustering/
- R-archive network http://cran.r-project.org/
- Tanagra https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
- YALE (rapid-i.com)
- KNIME (www.knime.org)
- Orange (www.ailab.si/orange)

Công cụ thực hành…

Các công cụ thương mại

- Oracle 10g/11g DBMS và Oracle 10g/11g Data Mining
  www.oracle.com
- Microsoft data mining tools (MS SQL Server 2005/2008 DBMS và Business Intelligence Development Studio)
  www.microsoft.com
- Hỗ trợ từ Intelligent Miner (IBM)
- Hỗ trợ từ Enterprise Miner (SAS Institute)

Giới thiệu phân cụm

- Tình huống ngoại lai
  - Liệu đây có phải khách hàng tiềm năng?

Giới thiệu phân cụm

- Tình huống ngoại lai…
  - Có gì bất thường trong chuyến bay này k?

Giới thiệu phân cụm…

- Tình huống ngoại lai

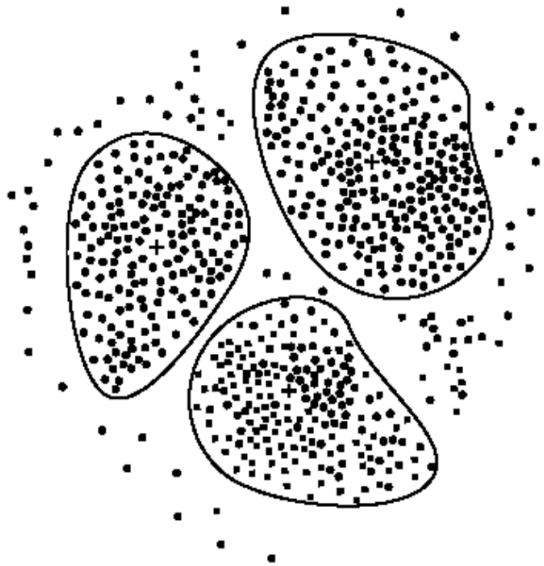
Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?

Giới thiệu phân cụm…

- Tình huống – Biên và nhiễu

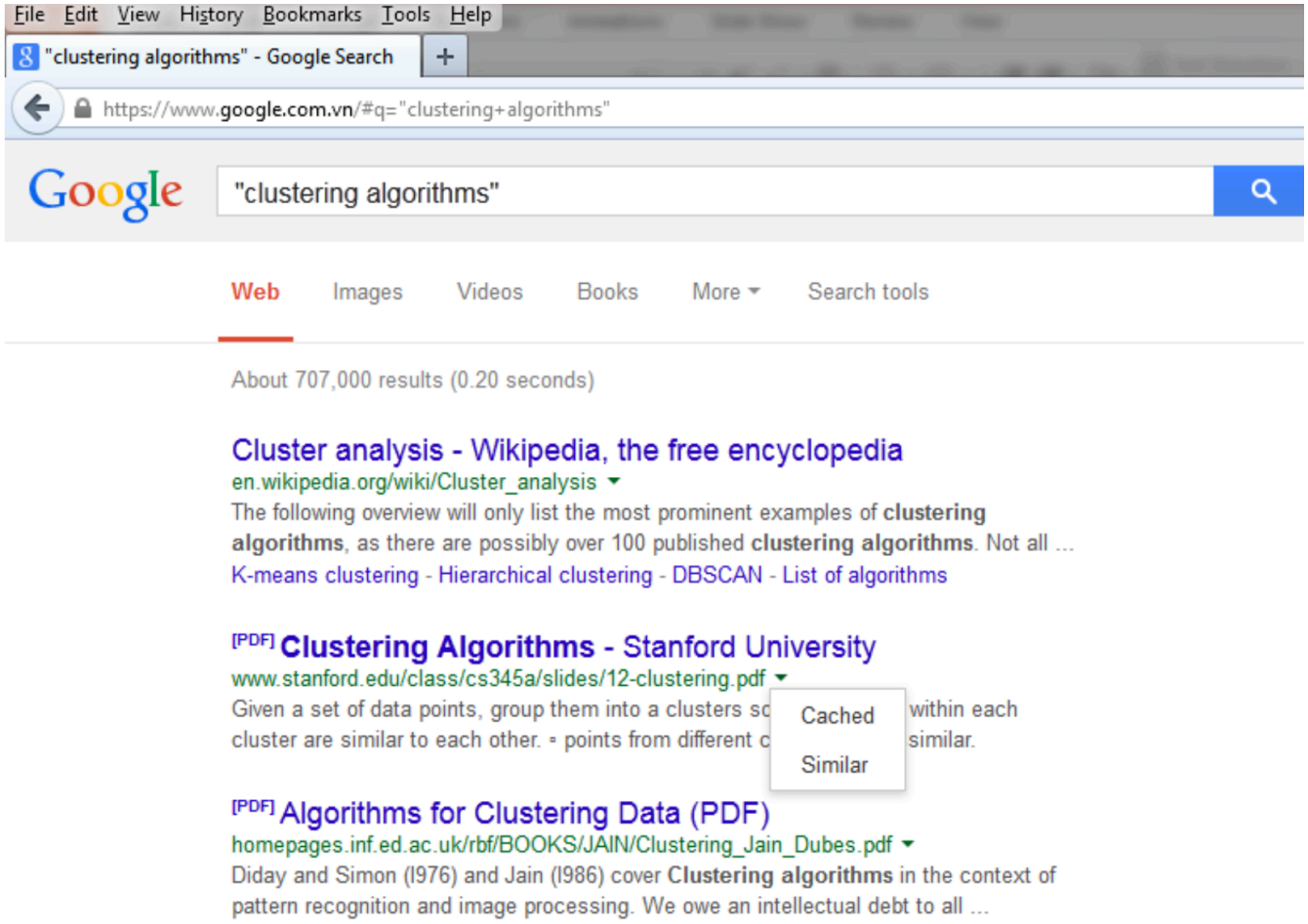
  Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)

  - Giải pháp giảm thiểu nhiễu
    - Phân tích cụm (cluster analysis)



Giới thiệu phân cụm…

- Tình huống – Tìm kiếm



Giới thiệu phân cụm…

- Tình huống – Phân cụm ảnh

http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeatures.data.html

Bài toán

- **Khách hàng Bông sen vàng**
  - Dự báo khách hàng tiềm năng
  - Dự báo khả năng nâng hạng của KH
  - Phân tích xu hướng của các KH (nâng, hạ, thời gian đi,..)
  - Xu hướng chọn hạng bay và các dịch vụ khác,…
  - ….
- **Điều hành chuyến bay**
  - Dự báo/Phân tích về việc Delay
  - Khuyến nghị/khuyến cáo về an toàn bay
  - Phân tích đặc điểm các chuyến bay theo mùa/thời tiết và các đặc điểm tình hình khác
  - …
- **Bài toán về nhiên liệu**
  - Tối ưu hóa lượng nhiên liệu
  - Trợ giúp phi công trong việc đặt nhiên liệu
  - …

Dữ liệu

- Dữ liệu thực hành:
  - DL do VNA cung cấp gồm BSV, FIMS
  - Hàng không USA: http://stat-computing.org/dataexpo
  - Dữ liệu khách hàng
  - Dữ liệu giao thông

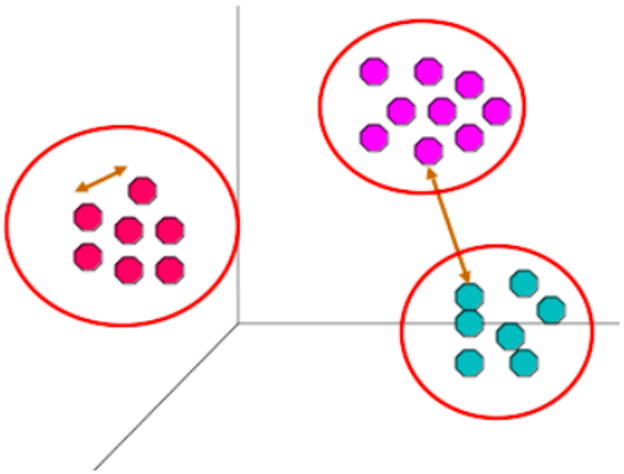
Phân/gom cụm là gì?

- *"PCDL là một kỹ thuật trong DATA MINING, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định"*

Mục tiêu của phân cụm

Chia các đối tượng thành các cụm *thuần nhất* và *phân biệt với nhau*, tức là các nhóm đối tượng thoả mãn:

- độ tương tự của các đối tượng trong mỗi nhóm cao nhất có thể (tiêu chuẩn **liên kết chặt**),
- các đối tượng trong các nhóm khác nhau phân biệt nhất có thể (tiêu chuẩn **tách rời**),
  cần một độ đo đánh giá độ tương tự hay độ khác biệt



Ứng dụng của phân cụm

- Hiểu dữ liệu (Understanding)
  - Gộp nhóm các tài liệu liên quan
  - Nhóm các gien và protein có chức năng tương tự
  - Phân cụm các cổ phiếu có biến động giá tương tự
  - ...
- Tóm tắt dữ liệu (summarization)
  - Giảm kích thước dữ liệu

Ứng dụng của phân cụm...

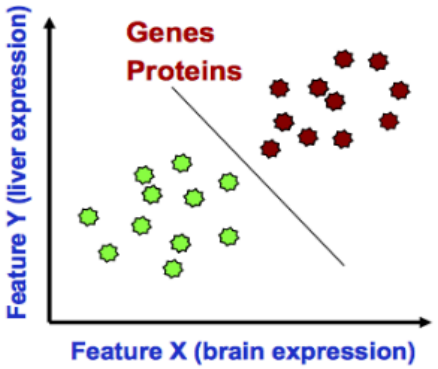
- Hỗ trợ giai đoạn tiền xử lý dữ liệu (data preprocessing)
- Mô tả sự phân bố dữ liệu/đối tượng (data distribution)
- Nhận dạng mẫu (pattern recognition)
- Phân tích dữ liệu không gian (spatial data analysis)
- Xử lý ảnh (image processing)
- Phân mảnh thị trường (market segmentation)
- Gom cụm tài liệu ((WWW) document clustering)
- ...

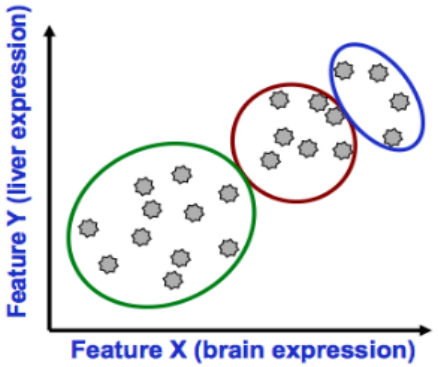
Phân cụm (Clustering) và Phân lớp (Classification)?

- Mục tiêu phân cụm: nhóm các đối tượng tương tự, nhờ đó **phát hiện cấu trúc ẩn** của dữ liệu.
- Mục tiêu phân lớp: Trích rút các đặc trưng từ dữ liệu cho phép **phân loại các phần tử mới** vào các **lớp đã xác định.**

Phân cụm và Phân lớp…

- Các đối tượng được mô tả bởi một hay nhiều đặc trưng (features)
- **Phân lớp (supervised learning)**
  - Có nhãn cho một số điểm dữ liệu
  - Cần một "quy tắc" cho phép gán nhãn chính xác cho các điểm dữ liệu mới
  - Bài toán con: chọn đặc trưng
  - Độ đo: độ chính xác phân lớp
- **Phân cụm (unsupervised learning)**
  - Không có nhãn sẵn
  - Nhóm các điểm vào cụm dựa vào độ "gần" của chúng
  - Xác định cấu trúc trong dữ liệu
  - Độ đo: các đặc trưng kiểm chứng độc lập



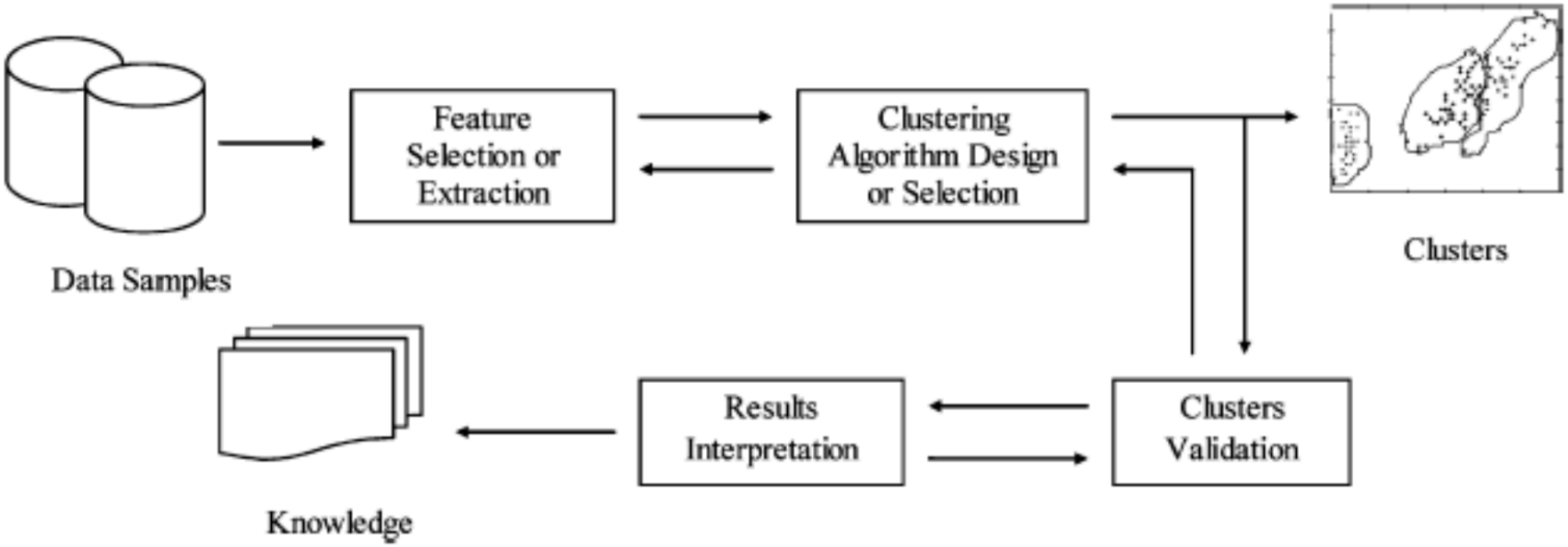


VIASM
Vietnam Institute for Advanced Study in Mathematics

Các kiểu phân cụm

- Biểu diễn cụm:
  - Phân hoạch
  - Cây phân cấp
- Đặc điểm phân cụm:
  - Mỗi đối tượng thuộc/không thuộc một cụm duy nhất
  - Phân cụm mờ/không mờ, có/không có trọng số xác suất Các cụm đều nhau/không đồng đều

Quá trình phân cụm dữ liệu



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

Các bước phân cụm tự động

1. Thu thập dữ liệu
2. Tính toán độ tương tự giữa *n* cá thể từ các bảng dữ liệu ban đầu
3. Chọn một thuật toán phân cụm và thực hiện
4. Diễn giải kết quả:
   - Đánh giá chất lượng phân cụm
   - Mô tả các cụm (lớp) đạt được.

Độ khác biệt/phi tương tự - tương tự

- Các phương pháp phân cụm cần có

  - Tiêu chuẩn đo độ khác biệt/phi tương tự (khoảng cách) giữa các đối tượng, hoặc đo độ tương tự giữa các đối tượng.
    - Với các dữ liệu định lượng thường sử dụng khoảng cách
    - Với dữ liệu văn bản thường sử dụng độ tương tự

Độ khác biệt/phi tương tự (khoảng cách)

Đo độ khác biệt giữa các đối tượng:

- Cho E là tập n đối tượng cần phân cụm
- Độ đo sự khác biệt d: $E \times E \rightarrow R+$
  - 1. $d(i,i)=0 \ \forall i \in E$
  - 2. $d(i,i')=d(i',i) \ \forall i,i' \in E \times E$

Độ đo khoảng cách thoả mãn các thuộc tính của một tiêu chuẩn đo độ khác biệt

Dữ liệu

	X_1	$\cdots$	X_p
1	x_{11}	$\cdots$	x_{p1}
$\vdots$			
i	x_{1i}	$\cdots$	x_{pi}
$\vdots$			
n	x_{1n}	$\cdots$	x_{pn}

- X_k $(1 \leq k \leq p)$: các biến tương ứng với các thuộc tính dữ liệu
- Các kiểu dữ liệu: định lượng/liên tục (*quantitative/continuous*), định tính/tên/phân loại/rời rạc (*qualitative/nominal/categorial/discrete*), nhị phân (*binary*), văn bản (*text*), chuỗi thời gian (*time series*), đồ thị (*graph*)

Biến nhị phân

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
P_1	1	0	0	1	1	0	1	1
P_2	1	1	0	1	0	1	0	1
P_3	1	1	1	1	0	0	1	1
P_4	0	1	0	0	0	1	0	1
P_5	0	0	1	1	0	1	0	1

n_{ij} = số các tương hợp dương (11)

$n_{\overline{ij}}$ = số các tương hợp âm (00)

q_{ij} = số các bất tương hợp (01) ou (10)

- Hàm đo độ tương tự: đo sự giống nhau giữa các đối tượng
  - tăng với các tương hợp
  - giảm với các bất tương hợp

$\forall e_i, e_j \in E \times E : \quad S(e_i, e_j) = f(n_{ij}, n_{\overline{ij}}, q_{ij})$

Biến nhị phân…

$$S_\theta(e_i, e_j) = \frac{n_{ij}}{\theta n_{ij} + q_{ij}}$$

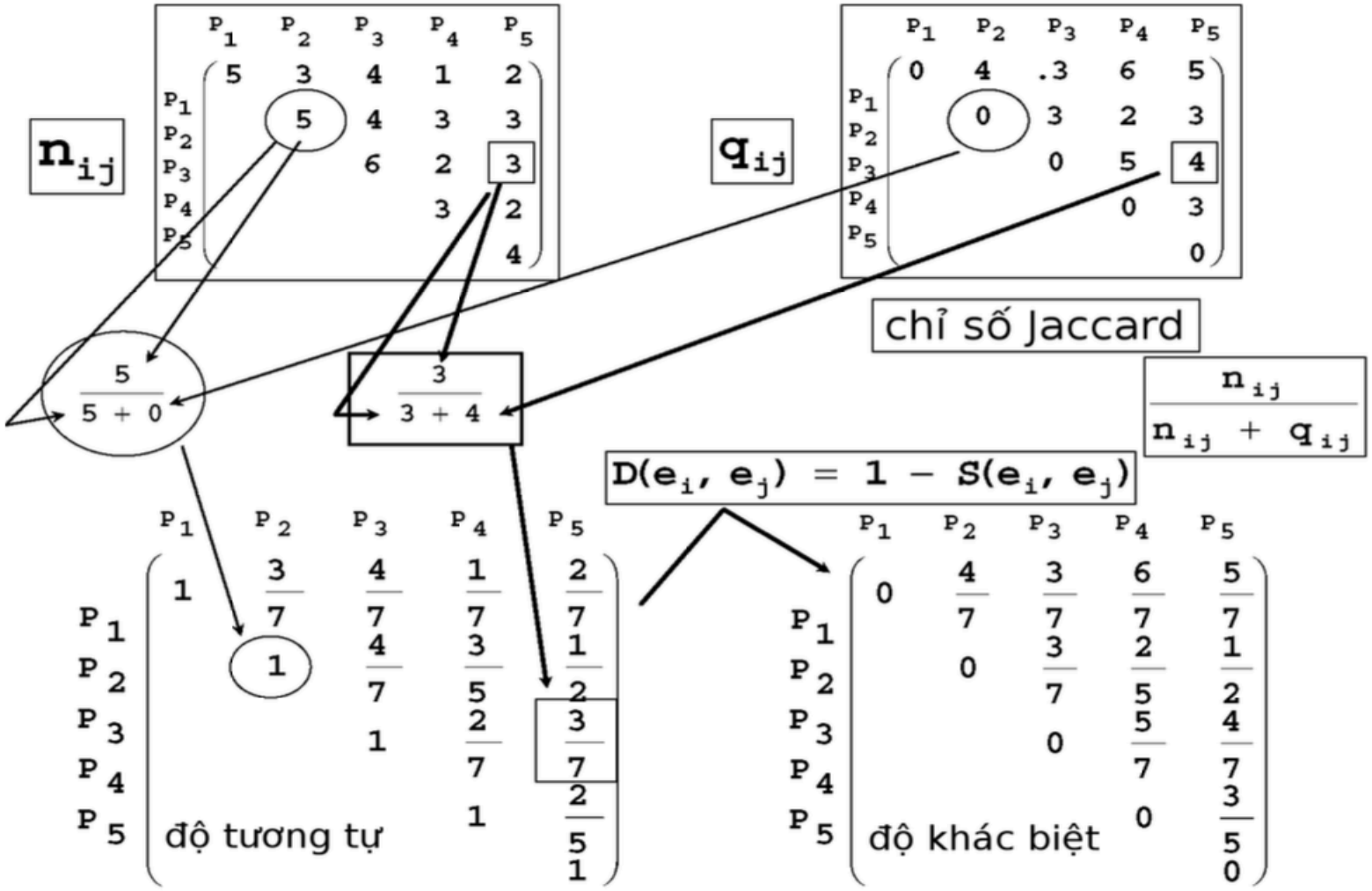
$\theta = 1$ – tiêu chuẩn Jaccard

$\theta = 2$ – tiêu chuẩn Dice

$$S_{\alpha,\beta}(e_i, e_j) = \frac{n_{ij} - \alpha q_{ij} + n_{\overline{ij}}}{n_{ij} + \beta q_{ij} + n_{\overline{ij}}}$$

$\alpha = 0, \quad \beta = 1$ – tiêu chuẩn so sánh đơn giản

Biến nhị phân…



Tập biến định tính/phân loại

Sử dụng độ đo tương tự

- Cách 1: nhị phân hoá biến, sử dụng độ tương tự dùng cho tập biến nhị phân

- Cách 2: phân tích tương ứng, chiếu biến định tính vào không gian liên tục

Tập biến định tính/phân loại…

- Cách 3: tính tổng độ tương tự trên từng biến

$$S(e_i, e_j) = \sum_{k=1}^{p} s(x_{ki}, x_{kj})$$

  - So sánh đơn giản:

$$s(x_{ki}, x_{kj}) = \begin{cases} 1 & x_{ki} = x_{kj} \\ 0 & x_{ki} \neq x_{kj} \end{cases}$$

  - Tần suất xuất hiện nghịch đảo

$$s(x_{ki}, x_{kj}) = \begin{cases} 1/p_k(x_{ki})^2 & x_{ki} = x_{kj} \\ 0 & x_{ki} \neq x_{kj} \end{cases}$$

$p_k(x_{ki})$ = xác suất biến thứ k nhận giá trị x_{ki}

Tập biến định tính/phân loại…

Cách 3:…

- Độ đo Goodall

$$s(x_{ki}, x_{kj}) = \begin{cases} 1 - p_k(x_{ki})^2 & x_{ki} = x_{kj} \\ 0 & x_{ki} \neq x_{kj} \end{cases}$$

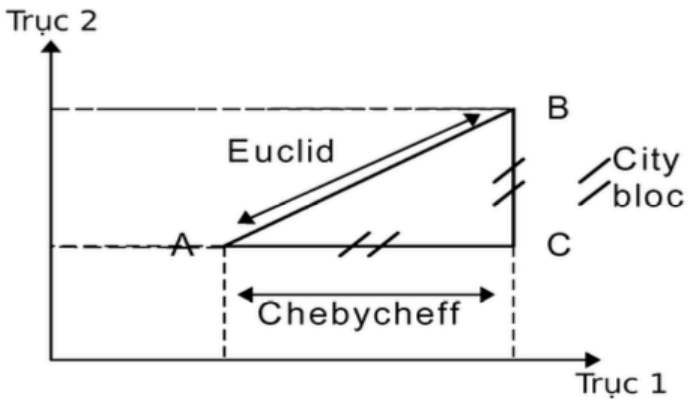
$p_k(x_{ki})$ = xác suất biến thứ k nhận giá trị x_{ki}

Tập biến định lượng

Khoảng cách Minkowski

$$D(e_i, e_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right]^{1/q}$$

- $q = 2$: Euclid, $q = 1$: city-block (Manhattan)
- $q \to +\infty$: khoảng cách Chebycheff $= max_k(|x_{ik} - x_{jk}|)$



(Khoảng cách city-block và Chebycheff thay đổi khi xoay các trục)

Tập biến hỗn hợp định tính/ định lượng

- Tổng có trọng số độ tương tự trên tập biến định tính và tập biến định lượng:

$$S(e_i, e_j) = \lambda S_C(e_{iC}, e_{jC}) + (1 - \lambda) S_N(e_{iN}, e_{jN})$$

trong đó e_{iC}, e_{jC} là véc-tơ giá trị các biến phân loại, e_{iN}, e_{jN} là véc-tơ giá trị các biến định lượng, $\lambda \in [0, 1]$ là trọng số căn chỉnh độ quan trọng của các biến định lượng so với biến định tính

- Công thức chuẩn hoá (chia cho độ lệch chuẩn):

$$S(e_i, e_j) = \lambda S_C(e_{iC}, e_{jC})/\sigma_C + (1 - \lambda) S_N(e_{iN}, e_{jN})/\sigma_N$$

Biểu diễn dữ liệu văn bản

- Sử dụng mô hình túi từ (*bag of words*), mỗi tài liệu là véc-tơ d chiều, d là số lượng từ trong từ điển, mỗi thành phần véc-tơ là tần suất của từ tương ứng trong tài liệu (hoặc một giá trị nhị phân)

$$\overline{X} = (x_1, \cdots, x_d)$$

- Dữ liệu thưa

Độ đo khoảng cách/tương tự cho DL văn bản

- Dùng trực tiếp khoảng cách Minkowski không phù hợp, tài liệu dài hơn thì khoảng cách lớn hơn.
- Cách 1: Giảm số chiều bằng LSA (*Latent Semantic Analysis*), trước khi dùng khoảng cách Minkowski
- Cách 2: Dùng độ tương tự cosine

Độ tương tự cosine

- Cho 2 tài liệu $\overline{X} = (x_1, \cdots, x_d)$, $\overline{Y} = (y_1, \cdots, y_d)$
- Độ tương tự cosine

$$cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \sqrt{\sum_{i=1}^{d} y_i^2}}$$

  - Cách tính này bỏ qua tần suất tương đối của các từ: chẳng hạn 2 tài liệu chứa từ "khoa học" thì ít tương tự hơn 2 tài liệu cùng chứa thuật ngữ "khai phá dữ liệu"

Độ tương tự với tf-idf

- idf (*Inverse Document Frequency*) $id_i = \log(N/n_i)$
  trong đó n_i là số tài liệu chứa từ thứ i, N là tổng số tài liệu
- Hàm "giảm xóc" (*damping function*) áp dụng trên tần suất từ, hạn chế ảnh hưởng của những từ xuất hiện nhiều lần
  $f(x_i) = \sqrt{x_i}$, hoặc
  $f(x_i) = \log(x_i)$
- Thay x_i bằng tần suất chuẩn hoá: $h(x_i) = f(x_i).id_i$
- Dùng độ tương tự cosine hoặc hệ số Jaccard:

$$J(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i)h(y_i)}{\sum_{i=1}^{d} h(x_i)^2 + \sum_{i=1}^{d} h(y_i)^2 - \sum_{i=1}^{d} h(x_i)h(y_i)}$$

Các phương pháp đánh giá việc phân cụm

- Đánh giá ngoại (external validation)
  - Đánh giá kết quả gom cụm dựa vào cấu trúc được chỉ định trước cho tập dữ liệu
- Đánh giá nội (internal validation)
  - Đánh giá kết quả gom cụm theo số lượng các vector của chính tập dữ liệu (ma trận gần – proximity matrix)
- Đánh giá tương đối (relative validation)
  - Đánh giá kết quả gom cụm bằng việc so sánh các kết quả gom cụm khác ứng với các bộ trị thông số khác nhau

→ Tiêu chí cho việc đánh giá và chọn kết quả gom cụm tối ưu

- Độ nén (compactness): các đối tượng trong cụm nên gần nhau.
- Độ phân tách (separation): các cụm nên xa nhau.

Các phương pháp đánh giá việc phân cụm…

- Đánh giá ngoại (external validation)
  - Độ đo: Rand statistic, Jaccard coefficient, Folkes and Mallows index, …
- Đánh giá nội (internal validation)
  - Độ đo: Hubert's Γ statistic, Silhouette index, Dunn's index, …
- Đánh giá tương đối (relative validation)

Các phương pháp đánh giá việc phân cụm…

- Các độ đo đánh giá ngoại (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.

Các phương pháp đánh giá việc phân cụm…

- Các độ đo đánh giá nội (internal validation measures)

Measure	Notation	Definition	Optimal value
Root-mean-square std dev	$RMSSTD$	$\{\sum_i \sum_{x \in C_i} \parallel x - c_i \parallel^2 / [P \sum_i (n_i - 1)]\}^{\frac{1}{2}}$	Elbow
R-squared	RS	$(\sum_{x \in D} \parallel x - c \parallel^2 - \sum_i \sum_{x \in C_i} \parallel x - c_i \parallel^2) / \sum_{x \in D} \parallel x - c \parallel^2$	Elbow
Modified Hubert Γ statistic	Γ	$\frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x,y) d_{x \in C_i, y \in C_j}(c_i, c_j)$	Elbow
Calinski-Harabasz index	CH	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x \in C_i} d^2(x,c_i)/(n-NC)}$	Max
I index	I	$(\frac{1}{NC} \cdot \frac{\sum_{x \in D} d(x,c)}{\sum_i \sum_{x \in C_i} d(x,c_i)} \cdot \max_{i,j} d(c_i,c_j))^p$	Max
Dunn's indices	D	$\min_i \{\min_j (\frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_k \{\max_{x,y \in C_k} d(x,y)\}})\}$	Max
Silhouette index	S	$\frac{1}{NC} \sum_i \{\frac{1}{n_i} \sum_{x \in C_i} \frac{b(x)-a(x)}{\max[b(x),a(x)]}\}$ $a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i, y \neq x} d(x,y), b(x) = \min_{j, j \neq i} [\frac{1}{n_j} \sum_{y \in C_j} d(x,y)]$	Max
Davies-Bouldin index	DB	$\frac{1}{NC} \sum_i \max_{j, j \neq i} \{[\frac{1}{n_i} \sum_{x \in C_i} d(x,c_i) + \frac{1}{n_j} \sum_{x \in C_j} d(x,c_j)] / d(c_i,c_j)\}$	Min
Xie-Beni index	XB	$[\sum_i \sum_{x \in C_i} d^2(x,c_i)] / [n \cdot min_{i,j \neq i} d^2(c_i,c_j)]$	Min
SD validity index	SD	$Dis(NC_{max}) Scat(NC) + Dis(NC)$ $Scat(NC) = \frac{1}{NC} \sum_i \parallel \sigma(C_i) \parallel / \parallel \sigma(D) \parallel, Dis(NC) = \frac{max_{i,j} d(c_i,c_j)}{min_{i,j} d(c_i,c_j)} \sum_i (\sum_j d(c_i,c_j))^{-1}$	Min
S_Dbw validity index	S_Dbw	$Scat(NC) + Dens\_bw(NC)$ $Dens\_bw(NC) = \frac{1}{NC(NC-1)} \sum_i [\sum_{j, j \neq i} \frac{\sum_{x \in C_i \cup C_j} f(x, u_{ij})}{max\{\sum_{x \in C_i} f(x,c_i), \sum_{x \in C_j} f(x,c_j)\}}]$	Min

): data set; n: number of objects in D; c: center of D; P: attributes number of D; NC: number of clusters; C_i: the i-th cluster; n_i: number of objects in C_i; center of C_i; $\sigma(C_i)$: variance vector of C_i; $d(x,y)$: distance between x and y; $\parallel X_i \parallel = (X_i^T \cdot X_i)^{\frac{1}{2}}$

Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu. Understanding of internal clustering validation measures. In: *Proc.*

Các kỹ thuật phân cụm

- 1. Phân cụm phân hoạch
- 2. Phân cụm phân cấp
- 3. Phân cụm dựa trên mật độ
- 4. Một số thuật toán đặc thù: FCM, EM,…

Phân cụm phân hoạch

- Thuật toán K-means
- Thuật toán PAM, CLARA, CLARANS

Phân cụm phân hoạch

- Phân hoạch quanh các tâm: Chia dữ liệu thành *k* nhóm với giá trị *k* định trước.
- *k*-means: phương pháp phân cụm động (*dynamic cluster*), tâm di động (*mobile center*)
- Các phương pháp mở rộng: *k*-modes, *k*-medoids (PAM, CLARA, CLARANS, ...)

Phân cụm phân hoạch…

- Nguyên tắc
  - Giảm số biến (ví dụ: phân tích thành phần)
  - Thực hiện phân cụm
- Lựa chọn tham số và đánh giá
  - Số cụm?
  - Chất lượng phân cụm?

Thuật toán k-means

- MacQeen đề xuất năm 1967.
- Sinh ra *k* cụm DL {C1, C2, …,Ck} từ một tập DL chứa *n* đối tượng trong không gian *d* chiều Xi = (xi1, xi2, …, xid) (i=1…n).
- Sao cho hàm tiêu chuẩn:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x - m_i)$$

đạt giá trị tối thiểu.

  - m_i là trọng tâm của cụm *Ci*.
  - *D* là khoảng cách giữa hai đối tượng.

Thuật toán k-means…

- Trọng tâm của một cụm là một véc tơ, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng vectơ DL trong cụm đang xét.
- Tham số đầu vào của thuật toán là số cụm *k*.
- Đầu ra là các trọng tâm của các cụm.
- Độ đo khoảng cách *D* giữa các đối tượng thường dùng khoảng cách Euclide, vì đây là mô hình khoảng cách dễ để lấy đạo hàm và xác định các cực trị tối thiểu.
- Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tuỳ vào ứng dụng.

Thuật toán k-means…

InPut : Số cụm k và các trọng tâm cụm $\{m_j\}_{j=1}^{k}$;

OutPut : Các cụm C_i ($i = \overline{1,k}$) và hàm tiêu chuẩn E đạt giá trị tối thiểu;

Begin

Bước 1: Khởi tạo :

Chọn k trọng tâm $\{m_j\}_{j=1}^{k}$ ban đầu trong không gian R^d (d là số chiều của dữ liệu) . Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2 : Tính toán khoảng cách :

Đối với mỗi điểm X_i (1<=i<=n), tính toán khoảng cách của nó tới mỗi trọng tâm m_j j=1,k. Và sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

Bước 3 : Cập nhật lại trọng tâm :

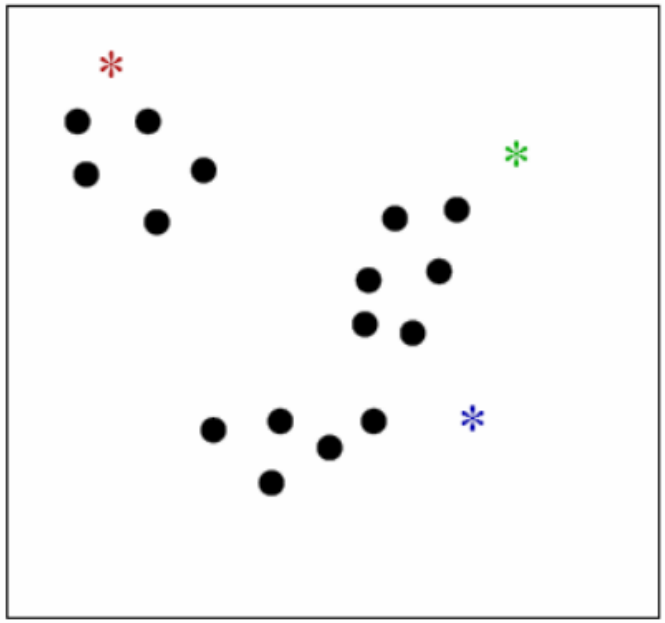
Đối với mỗi j=1,k , cập nhật trọng tâm cụm m_j bằng các xác định trung bình cộng của các vectơ đối tượng dữ liệu.

Bước 4 : Điều kiện dừng

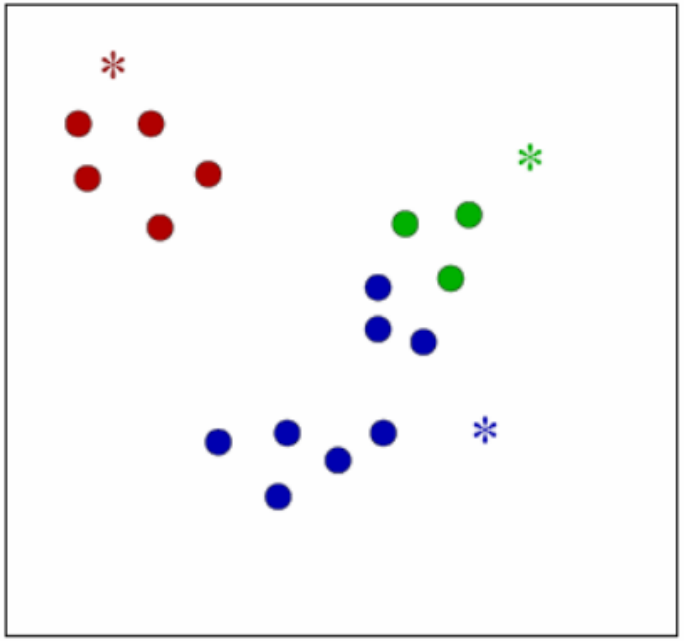
Lặp các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

End.

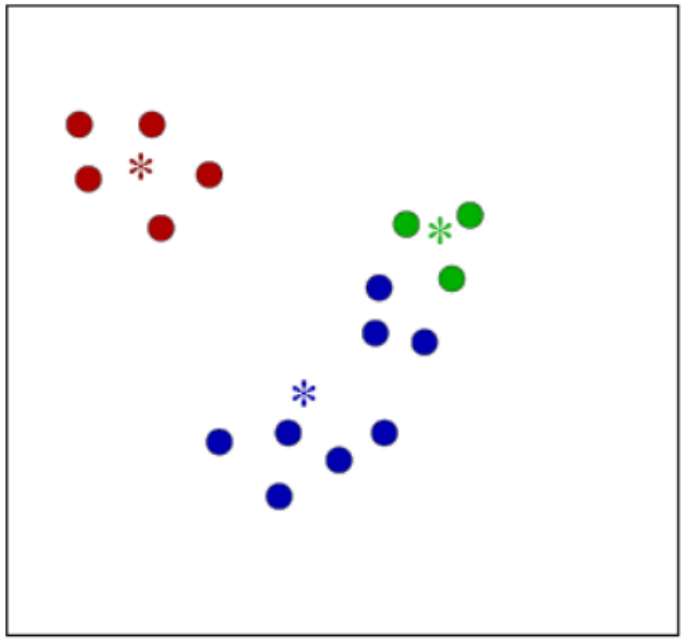
Hình 7 : Các bước thực hiện của thuật toán k-means



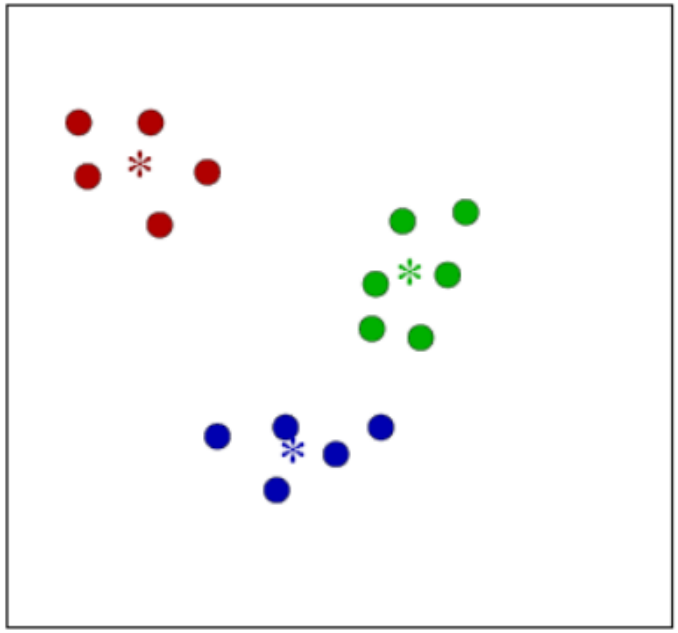
Initialize representatives (“means”)

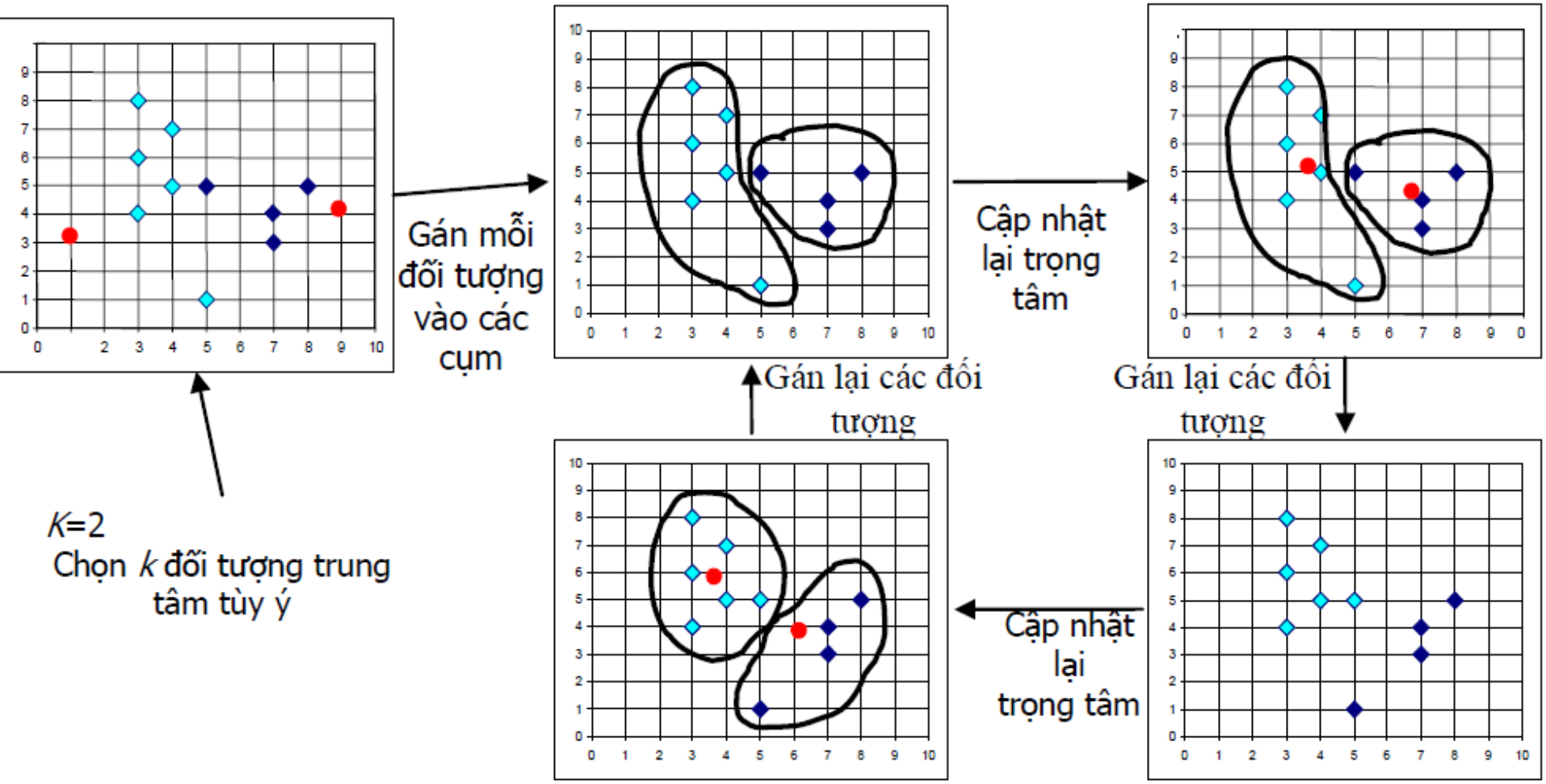


Assign to nearest representative



Re-estimate means





Hình 2.2. Hình dạng cụm dữ liệu được khám phá bởi k-means

Thuật toán chi tiết

BEGIN

1. Write ("Nhập số đối tượng dữ liệu");readln(n);
2. Nhập n đối tượng dữ liệu;
3. Write ("Nhập số cụm dữ liệu");readln(k);
4. MSE = + ∞;
5. For i = 1 to k do $m_i = x_{i+(i-1)*[n/k]}$; //Khởi tạo k trọng tâm
6. Do{
7. OldMSE = MSE;
8. MSE' = 0;
9. For j = 1 to k do
10. {m'_j =0; n'_j = 0;}
11. Endfor;
12. For i = 1 to n do
12. For j = 1 to k do
Tính toán khoảng cách Euclide
14. bình phương : $D^2(x_i, m_j)$;
15. Endfor
16. Tìm trọng tâm gần nhất m_h tới X_i.
17. $m'_h = m'_h + X_i$; $n'_h = n'_h + 1$;
18. MSE' = MSE' + $D^2(x_i, m_h)$;
19. Endfor
20. n_j = max (n'_j, 1); $m_j = m'_j / n_j$;
21. Endfor
22. MSE = MSE';
23} while (MSE < OldMSE)

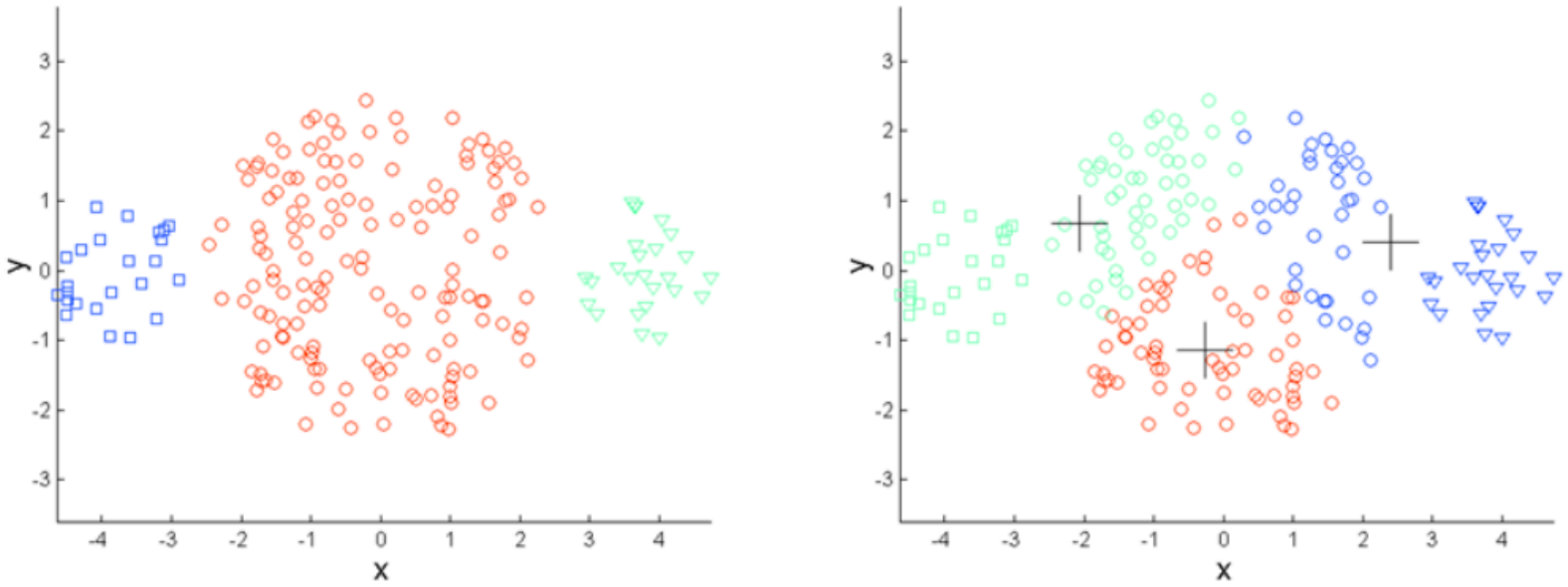
END;

Thuật toán k-means…

- Thuật toán hội tụ và độ phức tạp: $O((3nkd)\tau T^{flop})$
  - n là số đối tượng, k là số cụm, d là số chiều, τ là số vòng lặp, T^{flop} là thời gian thực hiện một phép tính cơ sở.
- Ưu:
  - K-means phân tích phân cụm đơn giản có thể áp dụng đối với tập DL lớn.
- Nhược:
  - Có vấn đề khi các cụm khác nhau về kích thước, mật độ và hình dạng không phải hình cầu.
  - Nhạy cảm với nhiễu và các phần tử ngoại lai.
  - Cách giải quyết, tăng *k*, hậu xử lý gộp các cụm?

Thuật toán k-means…

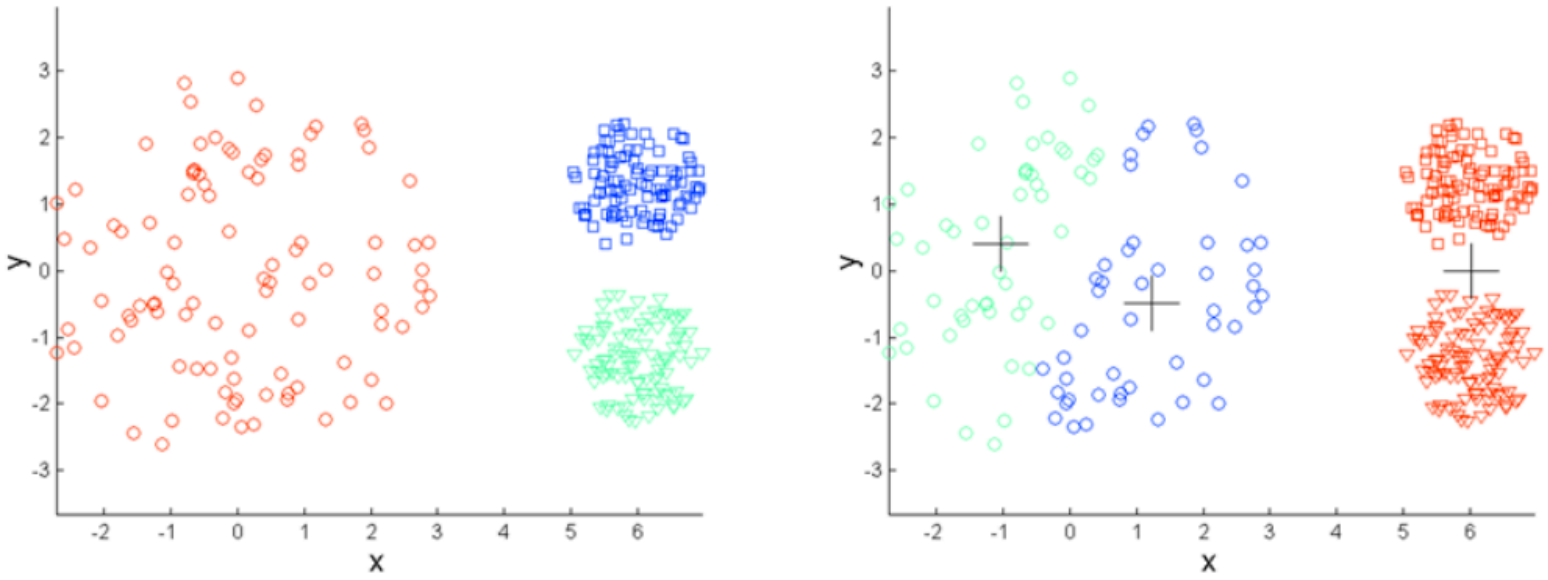
- Hạn chế - kích thước cụm



Các điểm ban đầu và kết quả phân cụm ($k = 3$)

Thuật toán k-means…

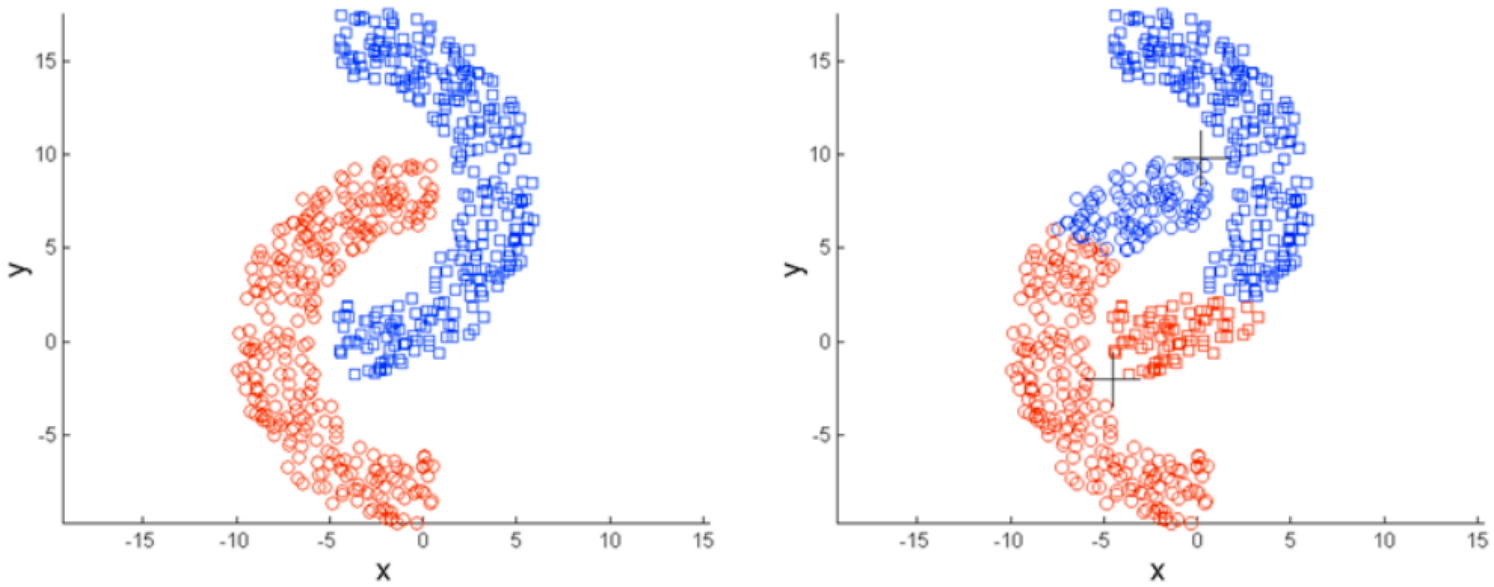
- Hạn chế - mật độ cụm



Các điểm ban đầu và kết quả phân cụm ($k = 3$)

Thuật toán k-means…

- Hạn chế - hình dạng cụm



Các điểm ban đầu và kết quả phân cụm ($k = 2$)

Thuật toán k-means…

Vấn đề khởi tạo các tâm cụm

- Cách chọn tâm ban đầu ảnh hưởng lớn tới kq phân cụm, thường được chọn ngẫu nhiên.
- Một số giải pháp:
  - Chạy nhiều lần (tuy nhiên không thể thử hết mọi cách chạy)
  - Lấy mẫu và dùng phương pháp phân cấp để xác định các tâm khởi tạo
  - Chọn nhiều hơn *k* tâm rồi lựa chọn thu gọn *k* tâm tách biệt
  - Phân cụm với số cụm lớn rồi thực hiện phân cụm phân cấp

- Thực tế chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào

Thuật toán k-means…

- Bài toán Khách hàng Bông sen vàng của VNA
  - Có mối liên hệ hạng thẻ với việc đặt hạng vé không?
  - Các hạng thẻ có xu hướng đặt hạng vé nào?

Thuật toán k-means…

- Bài toán Khách hàng Bông sen vàng của VNA?
  - Có mối liên hệ hạng thẻ với việc đặt hạng vé không?
  - Các hạng thẻ có xu hướng đặt hạng vé nào?

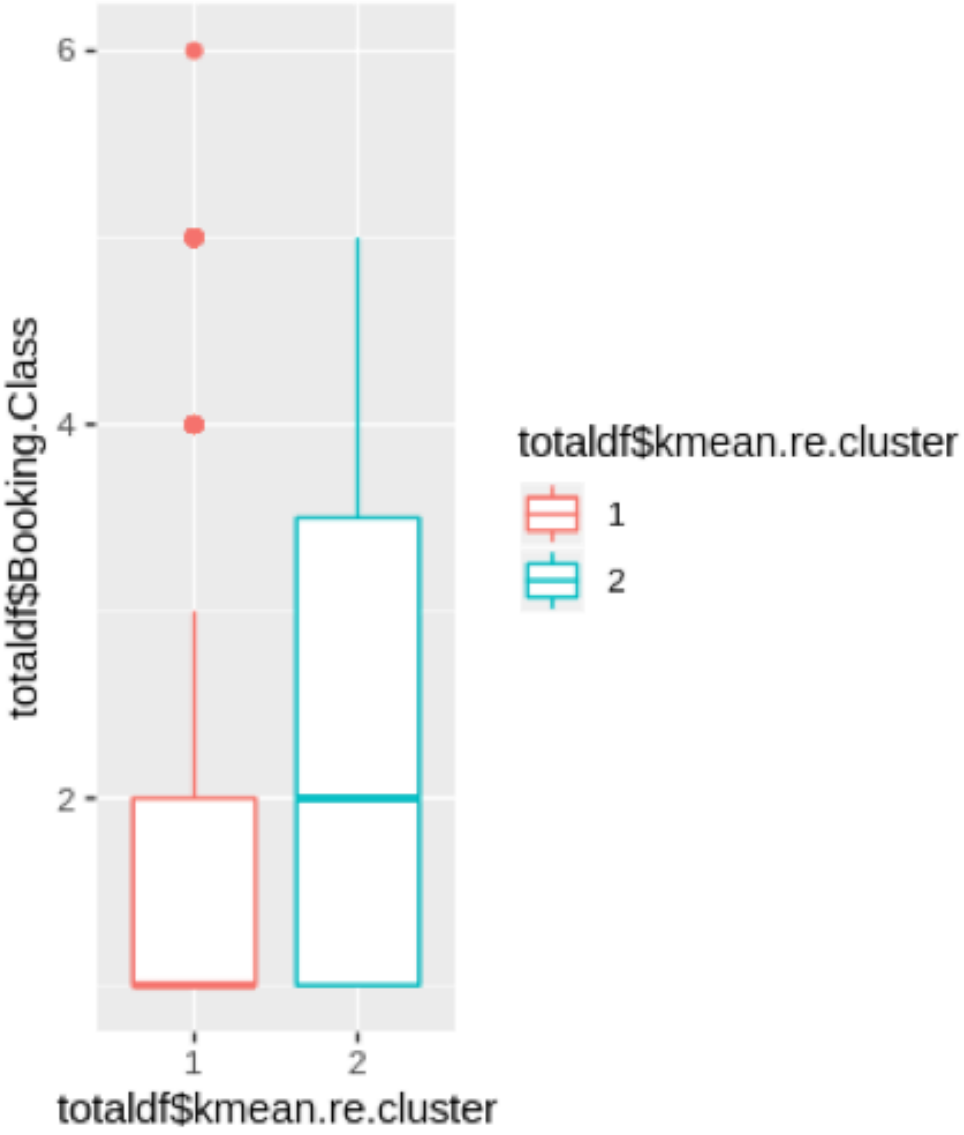
- 1. Hạng thẻ Platium và Gold thường sử dụng Hạng vé Phổ thông linh hoạt
- 2. Nhóm Silver và Register thường sử dụng Phổ thông tiết kiệm
- 3. Nhóm 3: Số ít Platium, Gold … sử dụng hạng vé thương gia

```
> summary(d1)
 Booking.Class     Tier.Level    kmean.re.cluster
 Min.   :1.000   Min.   :1.000   Min.   :1
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1
 Median :1.000   Median :2.000   Median :1
 Mean   :1.716   Mean   :2.183   Mean   :1
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1
 Max.   :6.000   Max.   :3.000   Max.   :1
> summary(d2)
 Booking.Class     Tier.Level    kmean.re.cluster
 Min.   :1.000   Min.   :3.000   Min.   :2
 1st Qu.:1.000   1st Qu.:4.000   1st Qu.:2
 Median :2.000   Median :4.000   Median :2
 Mean   :2.251   Mean   :4.427   Mean   :2
 3rd Qu.:3.500   3rd Qu.:5.000   3rd Qu.:2
 Max.   :5.000   Max.   :5.000   Max.   :2
>
```

Thuật toán k-means…

- Bài toán Khách hàng Bông sen vàng của VNA?
  - Có mối liên hệ hạng thẻ với việc đặt hạng vé không?
  - Các hạng thẻ có xu hướng đặt hạng vé nào?

- 1. Hạng thẻ Platium và Gold thường sử dụng Hạng vé Phổ thông linh hoạt
- 2. Nhóm Silver và Register thường sử dụng Phổ thông tiết kiệm
- 3. Nhóm 3: Số ít Platium, Gold … sử dụng hạng vé thương gia



Thuật toán k-means…

- Đến nay, có nhiều thuật toán kế thừa tư tưởng của k-means như thuật toán k-modes, PAM, CLARA, CLARANS, k- prototypes, …

Thuật toán PAM

- PAM (Partitioning Around Medoids) là mở rộng của k-means, đề xuất bởi Kaufman và Rousseeuw. Nhằm xử lý hiệu quả DL nhiễu/phần tử ngoại lai.
- PAM sử dụng các đối tượng *medoid* để biểu diễn cho các cụm DL. *medoid* là đối tượng đặt tại vị trí trung tâm nhất bên trong của mỗi cụm. Nên *medoid* ít bị ảnh hưởng của các đối tượng ở rất xa trung tâm, còn các trọng tâm của k-means lại bị tác động.
  - Ban đầu, PAM khởi tạo k đối tượng *medoid*.
  - phân phối các đối tượng còn lại vào các cụm tương ứng với các *medoid* đại diện, sao cho chúng tương tự với *medoid* trong cụm nhất.

Thuật toán PAM…

- Nếu Oj là đối tượng không phải là **medoid** và Om là đối tượng **medoid,** ta nói Oj thuộc về cụm có **medoid** là Om làm đại diện nếu: $d(Oj, Om) = \min_{Oe} d(Oj,Oe)$.
  - d(Oj,Oe) là độ phi tương tự giữa Oj và Oe.
  - $\min_{Oe}$ là giá trị nhỏ nhất của độ phi tương tự giữa Oj và tất cả các đối tượng **medoid** của các cụm DL.
- Chất lượng của mỗi cụm được đánh giá qua độ phi tương tự trung bình giữa một đối tượng và **medoid** tương ứng với cụm của nó, nghĩa là được đánh giá qua chất lượng của tất cả các **medoid.**

Thuật toán PAM…

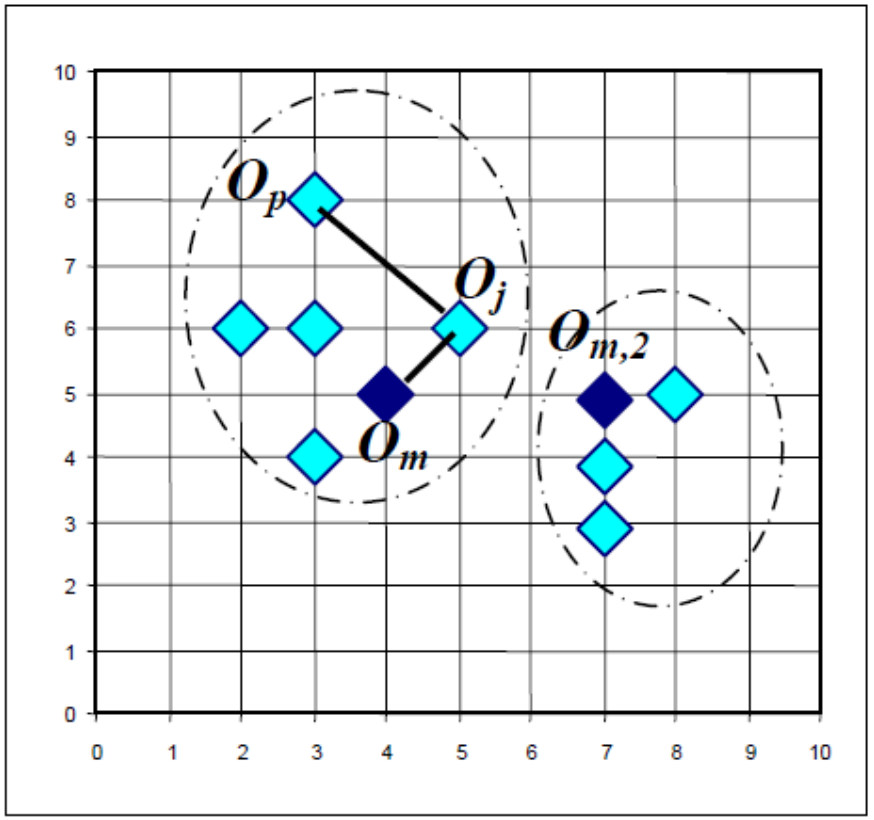
- Độ phi tương tự thường được xác định bằng độ đo khoảng cách. PAM được áp dụng cho DL không gian.
- Xác định các *medoid:*
  - PAM bắt đầu lựa chọn k đối tượng *medoid* bất kỳ.
  - Hoán chuyển *medoid* Om và một đối tượng Op không phải là *medoid*, nếu chất lượng phân cụm tốt hơn.
  - Quá trình này kết thúc khi chất lượng phân cụm không thay đổi.
- Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn. Tốt khi hàm tiêu chuẩn đạt giá trị tối thiểu.

Thuật toán PAM…

- Để quyết định hoán chuyển hai đối tượng *Om* và *Op* hay không, thuật toán PAM sử dụng giá trị tổng chi phí hoán chuyển *Cjmp* làm căn cứ:
- - *Om*: Là đối tượng medoid hiện thời cần được thay thế
- - *Op*: Là đối tượng medoid mới thay thế cho *Om*
- - *Oj*: Là đối tượng dữ liệu (không phải là medoid) có thể được di chuyển sang cụm khác.
- - *Om,2*: Là đối tượng medoid hiện thời khác với *Om* mà gần đối tượng *Oj* nhất

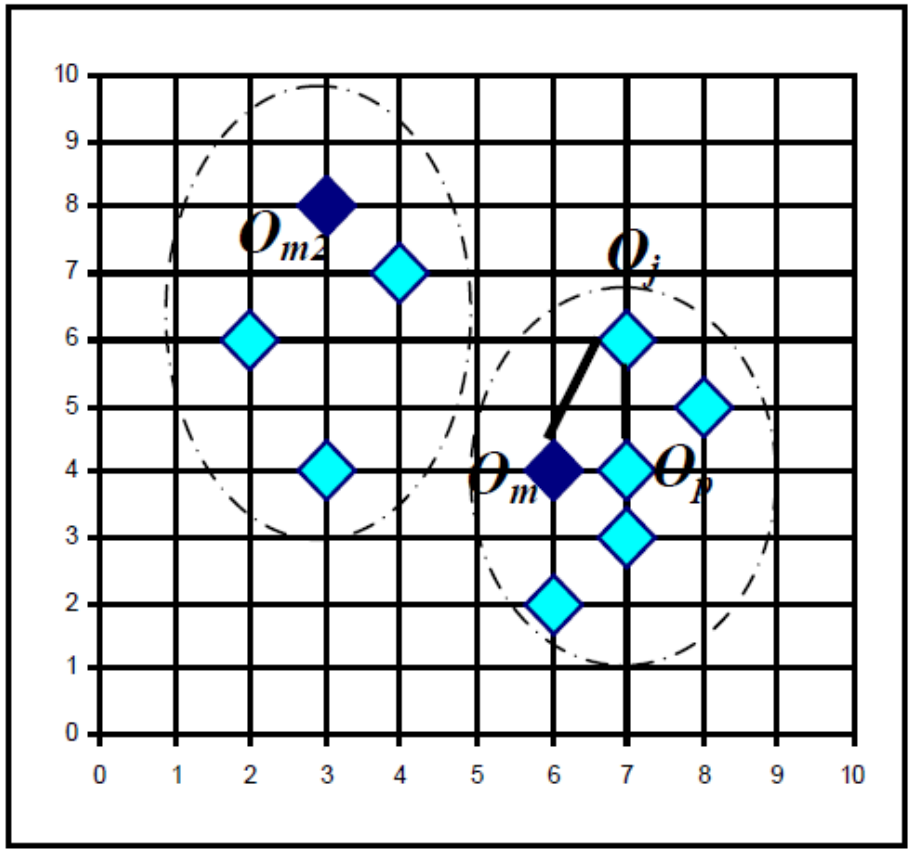
Thuật toán PAM…

- PAM tính giá trị hoán đổi *Cjmp* cho tất cả các đối tượng *Oj*. *Cjmp* được tính với 4 cách khác nhau như sau:
- **Trường hợp 1**: Giả sử *Oj* hiện thời thuộc về cụm có đại diện là *Om* và *Oj* tương tự với *Om,2* hơn *Op* (d(*Oj*, *Op*) ≥ d(*Oj*, *Om,2*))
- Nếu ta thay thế *Om* bởi đối tượng medoid mới *Op* thì *Oj* sẽ thuộc về cụm có đối tượng đại diện là *Om,2*.
- Vì vậy, giá trị hoán chuyển *Cjmp* = *d*(*Oj*, *Om,2*) – *d*(*Oj*, *Om*). *Cjmp* là không âm.



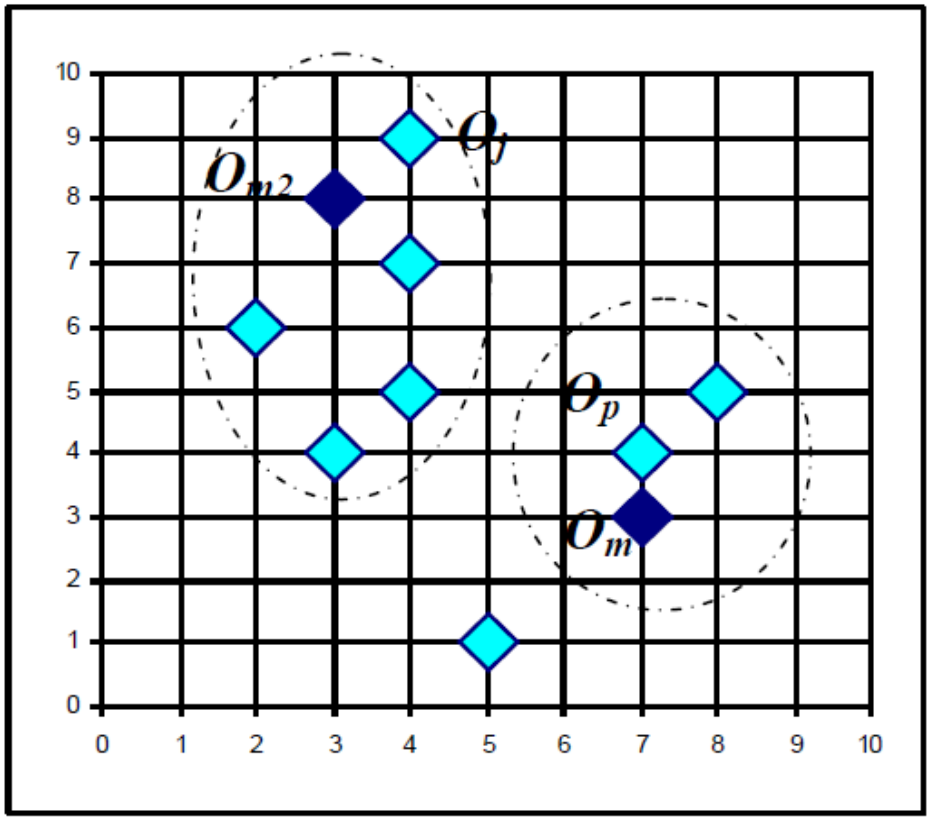
Thuật toán PAM…

- **Trường hợp 2:** Oj hiện thời thuộc về cụm có đại diện là *Om*, nhưng *Oj* ít tương tự với *Om,2* so với *Op* (*d*(*Oj*,*Op*)< *d*(*Oj*,*Om,2*)).
- Nếu thay thế *Om* bởi *Op* thì *Oj* sẽ thuộc về cụm có đại diện là *Op*.
- Vì vậy, giá trị *Cjmp* =(*Oj*,*Op*)- d(*Oj*, *Om*). *Cjmp* có thể âm hoặc dương.



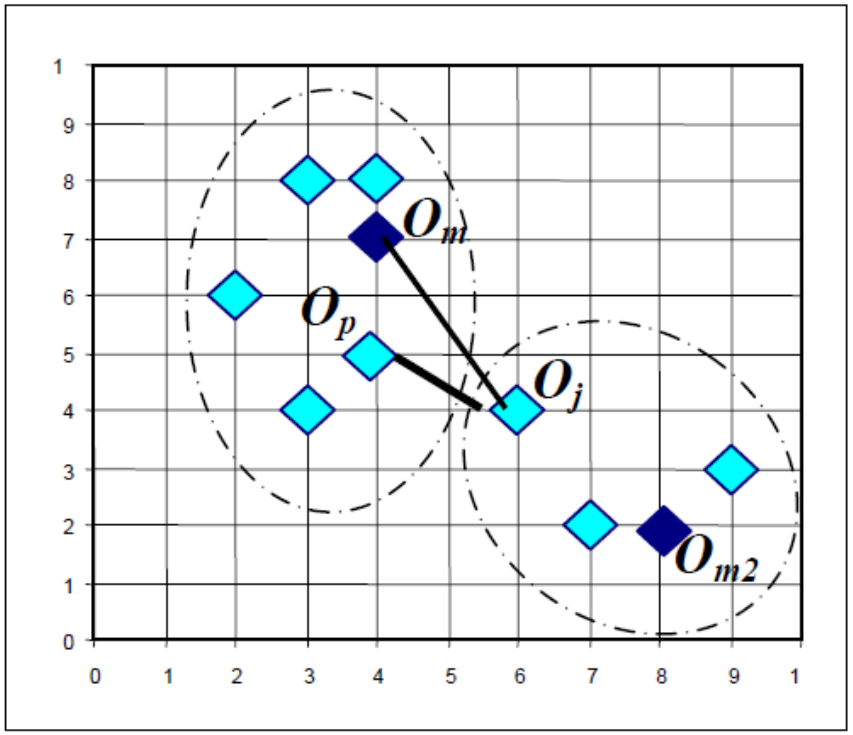
Thuật toán PAM…

- **Trường hợp 3:** Giả sử *Oj* hiện thời không thuộc về cụm có đối tượng đại diện là *Om* mà thuộc về cụm có đại diện là *Om,2*.
- Giả sử *Oj* tương tự với *Om,2* hơn so với *Op*.
- Nếu *Om* được thay thế bởi *Op* thì *Oj* vẫn sẽ ở lại trong cụm có đại diện là *Om,2*. *Cjmp*= 0.



Thuật toán PAM…

- **Trường hợp 4:** *Oj* hiện thời thuộc về cụm có đại diện là *Om,2* nhưng *Oj* ít tương tự tới *Om,2* hơn so với *Op*.
- Nếu thay thế *Om* bởi *Op* thì *Oj* sẽ chuyển từ cụm *Om,2* sang cụm *Op*.
- Giá trị hoán chuyển *Cjmp* = (*Oj*,*Op*)-*d*(*Oj*, *Om,2*). *Cjmp* luôn âm.



- **Tổng giá trị hoán chuyển *Om* bằng *Op* được xác định:**

$$TC_{mp} = \Sigma_j C_{jmp}$$

Thuật toán PAM…

- **Thuật toán K-Medoid có thể được mô tả cụ thể như sau:**
- **Input**:Tập DL có n phần tử, k phân cụm.
- **Output**: k cụm DL sao cho chất lượng phân cụm tốt nhất
- *Bước 1*: Chọn k đối tượng medoid bất kỳ
- *Bước 2*: Gán mỗi đối tượng còn lại vào một cụm mà nó tương tự (gần) với đối tượng medoid của cụm nhất.
- *Bước 3*: Chọn ngẫu nhiên một đối tượng Op không là đối tượng medoid.
- *Bước 4*:Tính tổng chi phí TC_{mp} để đổi từ medoid cũ Om sang medoid mới Op.
- *Bước 5*: Nếu $TC_{mp}<0$ thì thay thế Om với Op để tạo ra một tập với đối tượng medoid mới.
- *Bước 6*: Quay lại bước 2 cho đến khi không có sự thay đổi medoid nào nữa thì dừng.

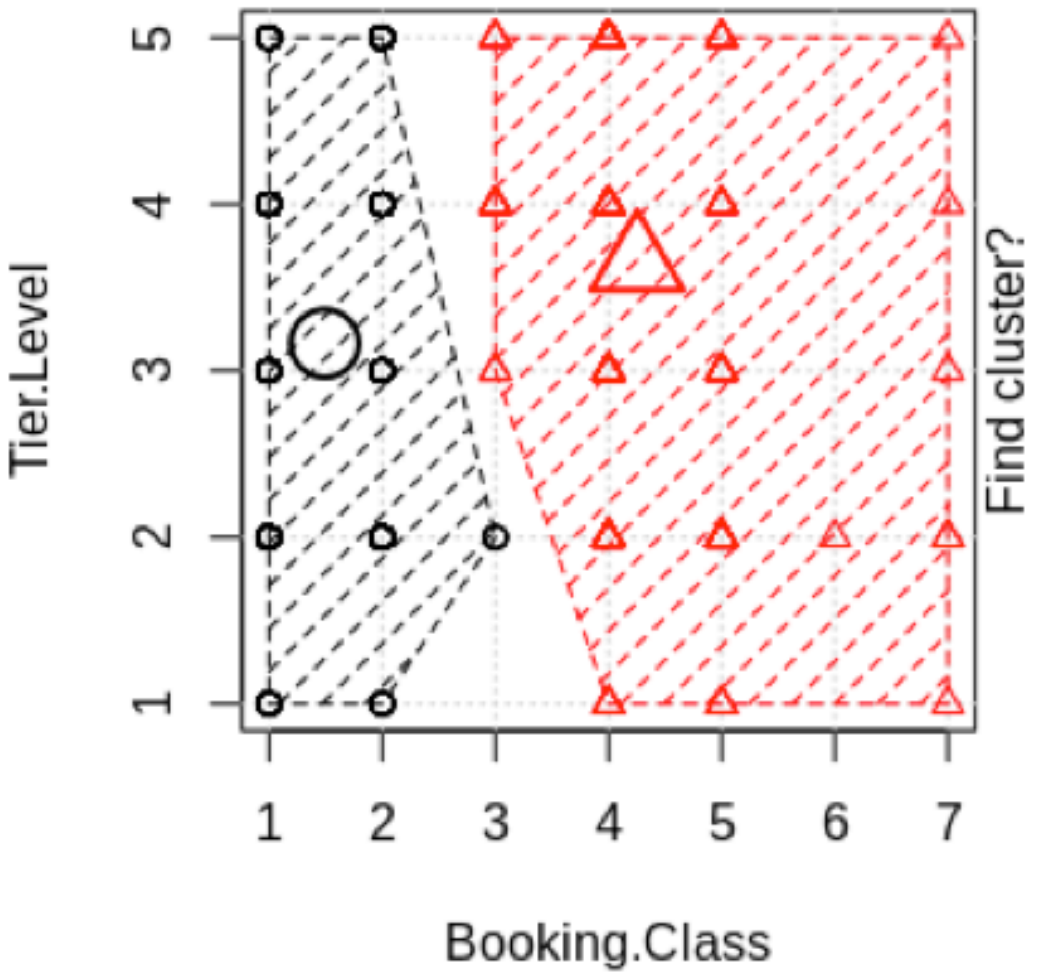
Thuật toán PAM…

- Trong bước 2 và 3, PAM phải duyệt tất cả k(n-k) cặp Om, Op.
- Với mỗi cặp, tính TC_{mp} yêu cầu kiểm tra n-k đối tượng.
- Nên, độ phức tạp $O(Ik\ (n-k)^2)$, I là số vòng lặp.
- ☞PAM kém hiệu quả về thời gian tính toán khi k và n lớn.

Thuật toán PAM…

- Bài toán Khách hàng Bông sen vàng của VNA?
  - Có mối liên hệ hạng thẻ với việc đặt hạng vé không?
  - Các hạng thẻ có xu hướng đặt hạng vé nào?

- 1. Hạng thẻ Platium và Gold thường sử dụng Hạng vé Phổ thông linh hoạt
- 2. Nhóm Silver và Register thường sử dụng Phổ thông tiết kiệm
- 3. Nhóm 3: Số ít Platium, Gold … sử dụng hạng vé thương gia



Thuật toán CLARA

- CLARA (Clustering LARge Application), Kaufman đề xuất 1990 [10][19]. Khắc phục nhược điểm của PAM khi k, n lớn.
  - CLARA trích mẫu cho tập dl, áp dụng PAM tìm ra các tâm *medoid* cho mẫu này.
  - Nếu mẫu dl được trích ngẫu nhiên, thì các *medoid* của nó xấp xỉ với các *medoid* của toàn bộ tập dl ban đầu.
  - CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗi trường hợp và chọn kết quả tốt nhất khi phân cụm trên các mẫu này.
- Chất lượng của cụm được đánh giá qua độ phi tương tự trung bình của toàn bộ các đối tượng trong tập dl ban đầu.

Thuật toán CLARA…

CLARA (5);

BEGIN

1. For i = 1 to 5 do

2. Lấy một mẫu có 40 + 2k đối tượng dữ liệu ngẫu nhiên từ tập dữ liệu và áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các đối tượng medoid đại diện cho các cụm.

3. Đối với mỗi đối tượng O_j trong tập dữ liệu ban đầu, xác định đối tượng medoid tương tự nhất trong số k đối tượng medoid.

4. Tính độ phi tương tự trung bình cho phân hoạch các đối tượng dành ở bước trước, nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.

5. Quay về bước 1.

END;

Thuật toán CLARA…

- Độ phức tạp $O(k(40+k)^2 + k(n-k))$.
- CLARA có thể thực hiện với tập dl lớn.

- Với kỹ thuật tạo mẫu trong PCDL: kết quả phân cụm có thể không phụ thuộc vào dl khởi tạo nhưng nó chỉ đạt tối ưu cục bộ.

- VD: Nếu các đối tượng *medoid* của dl khởi tạo không nằm trong mẫu, khi đó kết quả không đảm bảo là tốt nhất.

Thuật toán CLARANS

- CLARANS được Ng & Han đề xuất năm 1994. Để cải tiến chất lượng, áp dụng cho tập dữ liệu lớn. CLARANS cũng sử dụng các đối tượng trung tâm medoids làm đại diện cho các cụm dl.
- PAM là thuật toán phân hoạch có kiểu k-medoids. Nó bắt đầu khởi tạo k tâm đại diện **medoid** và liên tục thay thế mỗi tâm bởi một đối tượng khác trong cụm cho đến khi tổng khoảng cách của các đối tượng đến tâm cụm không giảm.
- CLARANS là thuật toán PCDL kết hợp thuật toán PAM với chiến lược tìm kiếm kinh nghiệm mới.

Thuật toán CLARANS…

- **Ý tưởng** CLARANS không xem xét tất cả các khả năng có thể thay thế các đối tượng tâm medoids, ngay lập tức thay thế các tâm này nếu việc thay thế là tốt hơn cho phân cụm chứ không cấn tối ưu nhất.
- Một phân hoạch cụm phát hiện được sau khi thay thế đối tượng trung tâm gọi là láng giềng (Neighbor) của phân hoạch cụm trước đó.

  - Số các láng giềng được hạn chế bởi tham số do người dùng đưa vào là **Maxneighbor** được lựa chọn ngẫu nhiên.
  - Tham số **Numlocal** xác định số vòng lặp tối ưu cục bộ được tìm kiếm.
  - Không phải tất các láng giềng được duyệt mà chỉ có **Maxneighbor** số láng giềng được duyệt.

Thuật toán CLARANS…

- Khi chọn các trung tâm **medoid**, CLARANS lựa chọn một giải pháp tốt hơn bằng cách:
  - Lấy ngẫu nhiên một đối tượng của k đối tượng trung tâm **medoid** của cụm.
  - Cố gắng thay thế nó với một đối tượng được chọn ngẫu nhiên trong (n-k) đối tượng còn lại.
  - Nếu không có giải pháp nào tốt hơn sau một số cố gắng lựa chọn ngẫu nhiên, thuật toán dừng và cho kết quả phân cụm tối ưu cục bộ.

Thuật toán CLARANS…

- Trong trường hợp tệ nhất, CLARANS so sánh một đối tượng với tất các đối tượng **Medoid**. Vì vậy, độ phức tạp của CLARANS là $O(kn^2)$, nên CLARANS không thích hợp với tập dl lớn (trong trường hợp xấu nhất).
- CLARANS có ưu điểm là không gian tìm kiếm không bị giới hạn như CLARA, và trong cùng một lượng thời gian thì chất lượng của các cụm tốt hơn CLARA.

Nhận xét họ thuật toán phân hoạch

- ☞ k-means thích hợp với các cụm DL có dạng hình cầu. Khi các cụm khá gần nhau thì một số đối tượng của một cụm có thể nằm trong các cụm khác.

- ☞ PAM là cải tiến của k-means để khắc phục trường hợp DL chứa nhiễu/phần tử ngoại lai.

Nhận xét họ thuật toán phân hoạch…

- ☞CLARA và CLARANS là các thuật toán dựa trên hàm tiêu chuẩn của thuật toán PAM, có khả năng áp dụng với tập DL lớn, nhưng hiệu quả phụ thuộc vào kích thước của các mẫu được phân. CLARANS hiệu quả hơn CLARA.

- ☞Hạn chế: các thuật toán phân cụm phân hoạch là chỉ thích hợp đối với DL số, ít chiều, và chỉ khám phá ra các cụm dạng hình cầu, nhưng lại áp dụng tốt với DL có các cụm phân bố độc lập và trong mỗi cụm có mật độ phân bố cao.

Phân cụm phân cấp

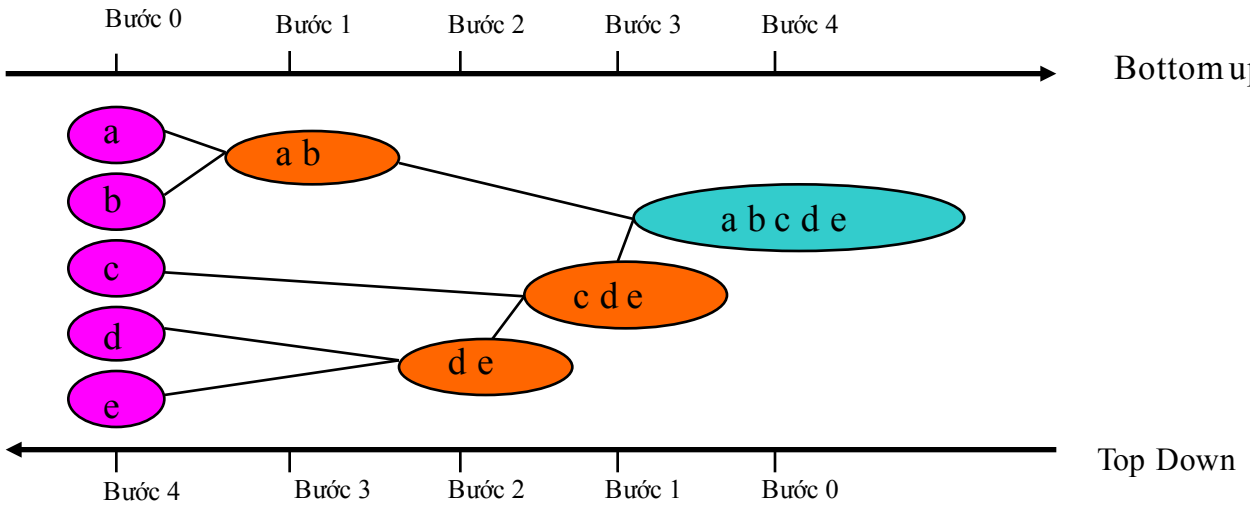
- Thuật toán BIRCH
- Thuật toán CURE

Phân cụm phân cấp

- Sắp xếp một tập DL đã cho thành cấu trúc có dạng hình cây:
  - Cây này xây dựng theo kỹ thuật đệ quy.
  - Cây phân cụm xây dựng theo hai phương pháp:
    - Dưới lên (Bottom up)
    - Trên xuống (Top down)
- Phương pháp "dưới lên" (*Bottom up*):
  - Bắt đầu mỗi đối tượng khởi tạo tương ứng với các cụm riêng biệt.
  - Nhóm các đối tượng theo một độ đo tương tự (như khoảng cách giữa hai trung tâm của hai nhóm).
  - Thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm (mức cao nhất của cây phân cấp), hoặc đến khi thỏa mãn điều kiện kết thúc.
- ☞ Sử dụng chiến lược ăn tham.

Phân cụm phân cấp…

- Phương pháp "trên xuống" (*Top Down*):
  - Bắt đầu tất cả các đối tượng được xếp trong cùng một cụm.
  - Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự.
  - Đến khi mỗi đối tượng là một cụm, hoặc điều kiện dừng thỏa mãn.
- ☞ Sử dụng chiến lược chia để trị



Phân cụm phân cấp…

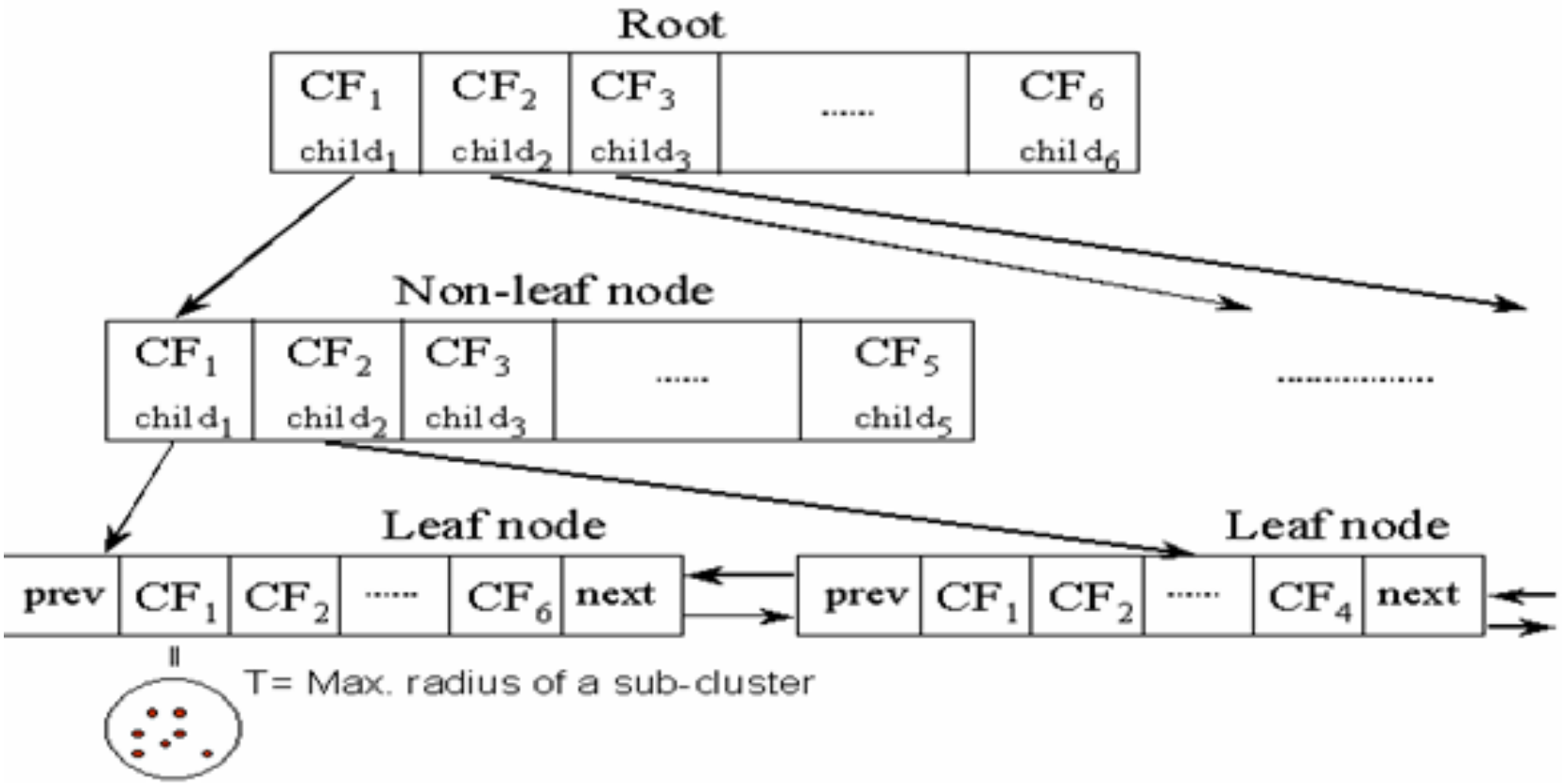
- Một số thuật toán phân cụm phân câp điển hình như CURE, BIRCH…
- ☞ Thực tế khi áp dụng, kết hợp cả hai phương pháp phân hoạch và phân cấp
- ☞Kết quả thu được của phương pháp phân cấp có thể cải tiến thông qua bước phân cụm phân hoạch.
- ☞Là hai phương pháp PCDL cổ điển, hiện nay có nhiều thuật toán cải tiến.

Thuật toán BIRCH

- BIRCH *(Balanced Iterative Reducing and Clustering Using Hierarchies)* sử dụng chiến lược phân cụm trên xuống (top down).
- http://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf

- **Ý tưởng**: không lưu toàn bộ các đối tượng DL của các cụm trong bộ nhớ mà chỉ lưu các đại lượng thống kê. Đó là bộ ba (n, LS, SS)
  - n là số đối tượng trong cụm
  - LS là tổng các giá trị thuộc tính của các đối tượng trong cụm
  - SS là tổng bình phương của các giá trị thuộc tính của các đối tượng trong cụm.
- Các bộ ba được gọi là đặc trưng của cụm(*Cluster Features - CF*), được lưu giữ trong một cây được gọi là cây CF (CF-tree).

Thuật toán BIRCH…



Hình 14 : Cây CF được sử dụng bởi thuật toán BIRCH

Thuật toán BIRCH…

- Cây CF là cây cân bằng. Cây CF chứa các nút trong và nút lá. Một cây CF được đặc trưng bởi hai tham số :
  - *Yếu tố nhánh (Branching Factor -B)*: xác định số tối đa các nút con của mỗi nút trong.
  - *Ngưỡng (Threshold - T)* : Khoảng cách tối đa giữa bất kỳ một cặp đối tượng trong nút lá của cây, khoảng cách này còn gọi là đường kính của các cụm con được lưu tại các nút lá.

Thuật toán BIRCH…

- Thuật toán BIRCH thực hiện qua giai đoạn sau :
- *Giai đoạn 1* : BIRCH duyệt tất cả các đối tượng trong CSDL và xây dựng một cây CF khởi tạo.
  - Các đối tượng lần lượt được chèn vào nút lá gần nhất của cây CF (nút lá của cây đóng vai trò là cụm con).
  - Nếu đường kích của cụm con sau khi chèn lớn hơn ngưỡng T, thì nút lá được tách.
  - Quá trình này lặp đến khi tất cả các đối tượng đều được chèn vào trong cây. Thấy rằng, mỗi đối tượng trong cây chỉ được đọc một lần
  - Để lưu toàn bộ cây CF trong bộ nhớ điều chỉnh kích thước của cây CF thông qua điều chỉnh ngưỡng T.
- *Giai đoạn 2* : BIRCH lựa chọn một thuật toán PCDL (như thuật toán phân cụm phân hoạch chẳng hạn) để thực hiện PCDL cho các nút lá của cây.

Thuật toán BIRCH…

- **INPUT**: CSDL gồm n đối tượng, ngưỡng T, k
- **OUTPUT**: k cụm dữ liệu
- ***Bước 1*:** Duyệt tất cả các đối tượng trong CSDL và xây dựng một cây CF khởi tạo. Một đối tượng được chèn vào nút lá gần nhất tạo thành cụm con. Nếu đường kính của cụm con này lớn hơn T thì nút lá được tách. Khi một đối tượng thích hợp được chèn vào nút lá, tất cả các nút trỏ tới gốc của cây được cập nhật với các thông tin cần thiết.

Thuật toán BIRCH…

- ***Bước 2:*** Nếu cây CF hiện thời không có đủ bộ nhớ trong thì tiến hành xây dựng một cây CF nhỏ hơn bằng cách điều khiển bởi tham số *T* (vì tăng T sẽ làm hoà nhập một số các cụm con thành một cụm, điều này làm cho cây CF nhỏ hơn). Bước này không cần yêu cầu bắt đầu đọc dữ liệu lại từ đầu nhưng vẫn đảm bảo hiệu chỉnh cây dữ liệu nhỏ hơn.

Thuật toán BIRCH…

- ***Bước 3:*** Thực hiện phân cụm: Các nút lá của cây CF lưu giữ các đại lượng thống kê của các cụm con. Trong bước này, BIRCH sử dụng các đại lượng thống kê này để áp dụng một số kỹ thuật phân cụm thí dụ như k-means và tạo ra một khởi tạo cho phân cụm

Thuật toán BIRCH…

- ***Bước 4:*** Phân phối lại các đối tượng dữ liệu bằng cách dùng các đối tượng trọng tâm cho các cụm đã được khám phá từ bước 3: Đây là một bước tuỳ chọn để duyệt lại tập dữ liệu và gán nhãn lại cho các đối tượng dữ liệu tới các trọng tâm gần nhất. Bước này nhằm để gán nhãn cho các dữ liệu khởi tạo và loại bỏ các đối tượng ngoại lai

Thuật toán BIRCH…

- Khi hòa nhập 2 cụm ta có :

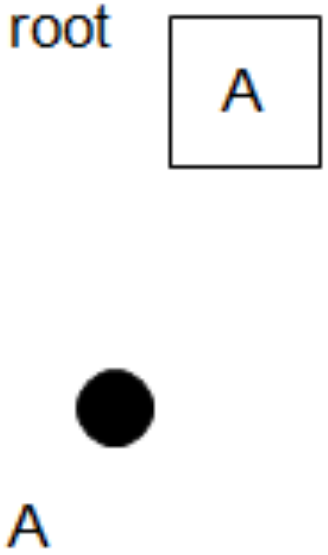
  CF=CF1+CF2=(n1+n2 ;LS1+LS2, SS1+SS2)

- Khoảng cách giữa các cụm có thể đo bằng khoảng cách Euclid, Manhatta,....

Thuật toán BIRCH…

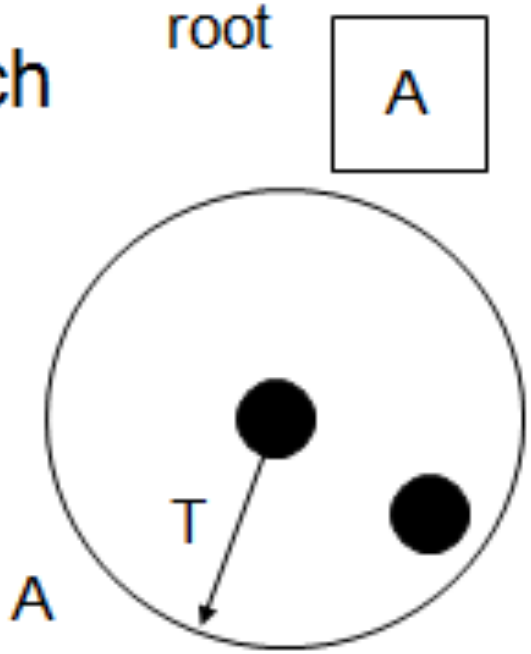
- Ví dụ về cây CF

  Ban đầu, các điểm dữ liệu trong một cụm



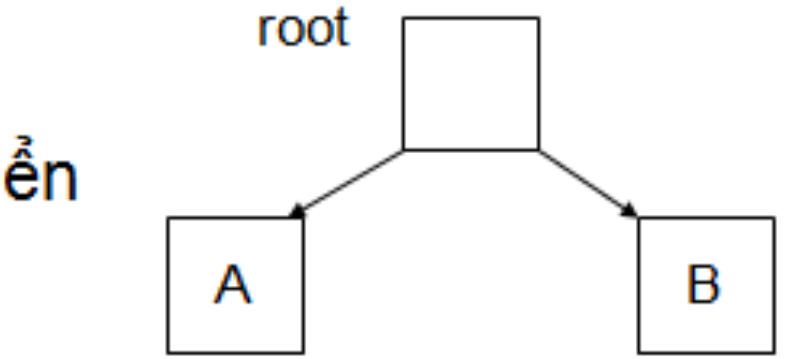
Thuật toán BIRCH…

Dữ liệu đến sẽ thuộc cụm A nếu kích thước của cụm không vượt quá ngưỡng T.



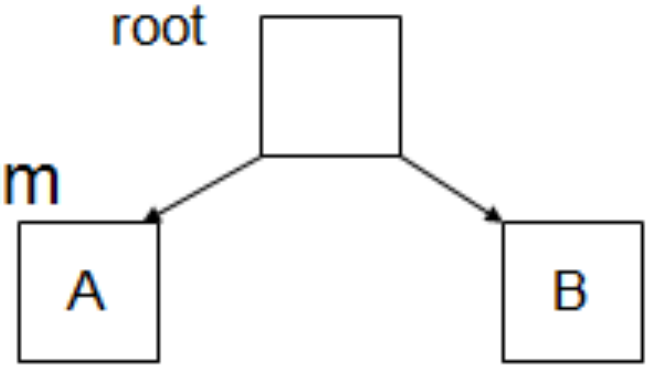
Thuật toán BIRCH…

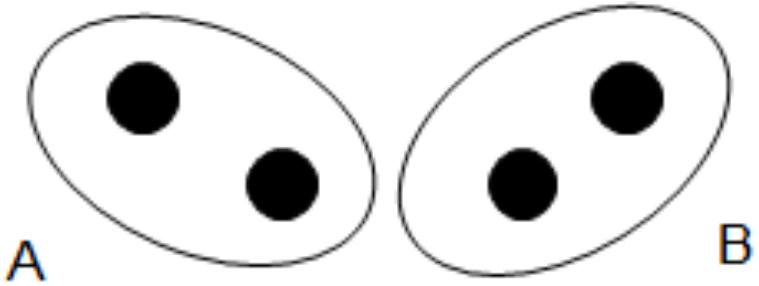
Nếu kích thước cụm phát triển quá lớn, cụm được chia thành hai cụm, và các điểm được phân phối.



Thuật toán BIRCH…

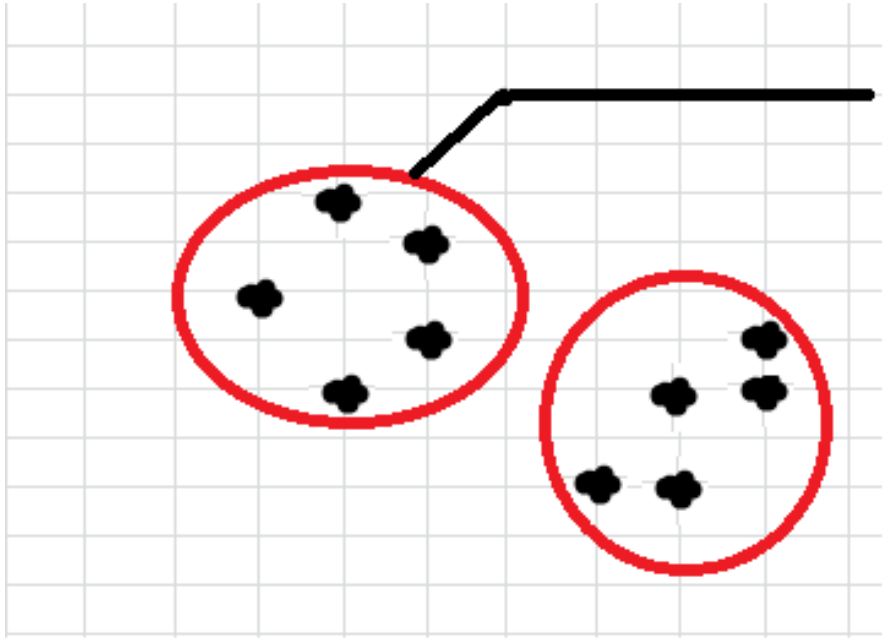
Tại mỗi nút của cây, cây CF giữ thông tin các giá trị đặc trưng của cụm





Thuật toán BIRCH…

- Ví dụ CF = (n , LS , SS) , n là số đối tượng của DL



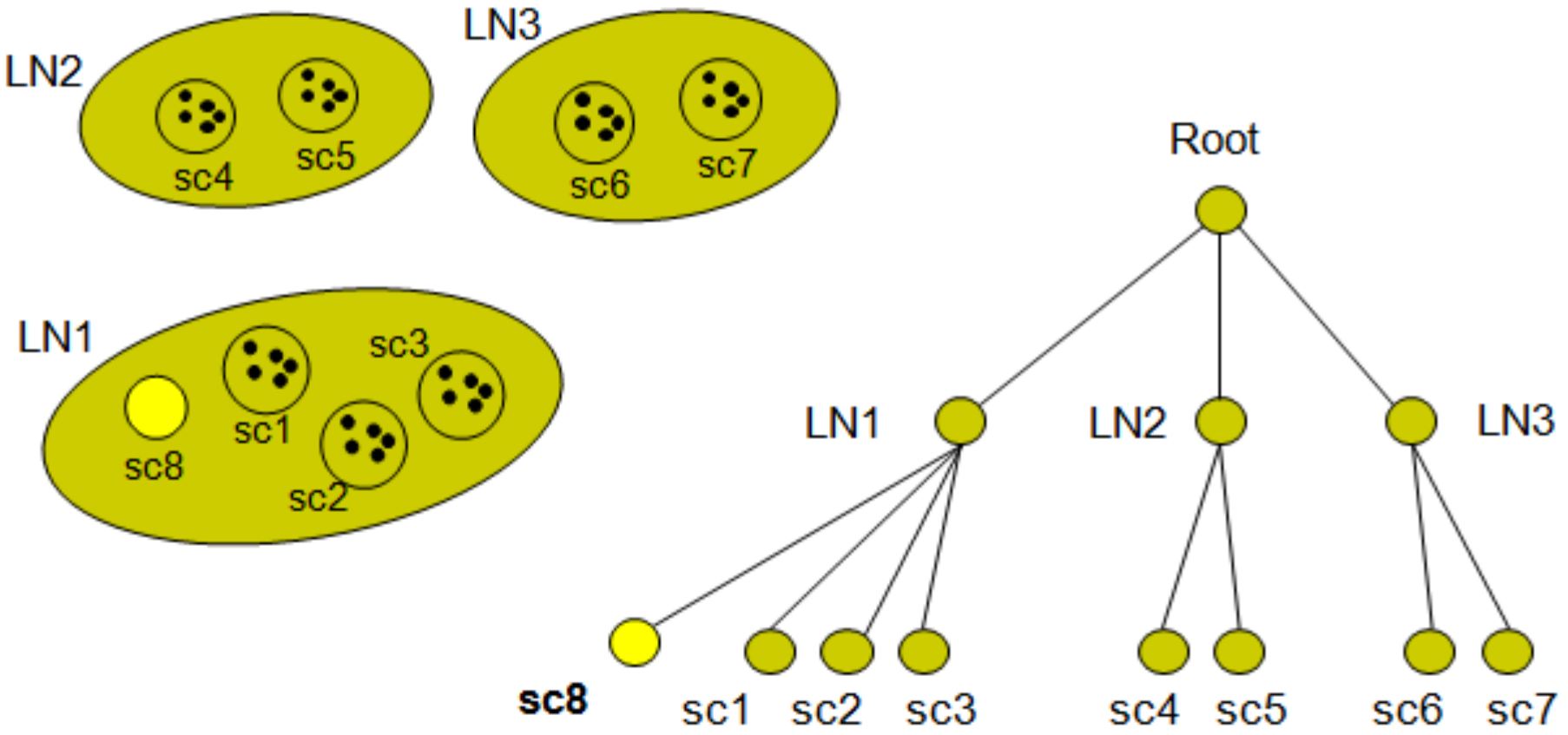
CF = (5 , (16,30) , (54,190))

(3 ,4)
(2,6)
(4,5)
(4,7)
(3,8)

Ví dụ về kết quả phân cụm bằng thuật toán BIRCH

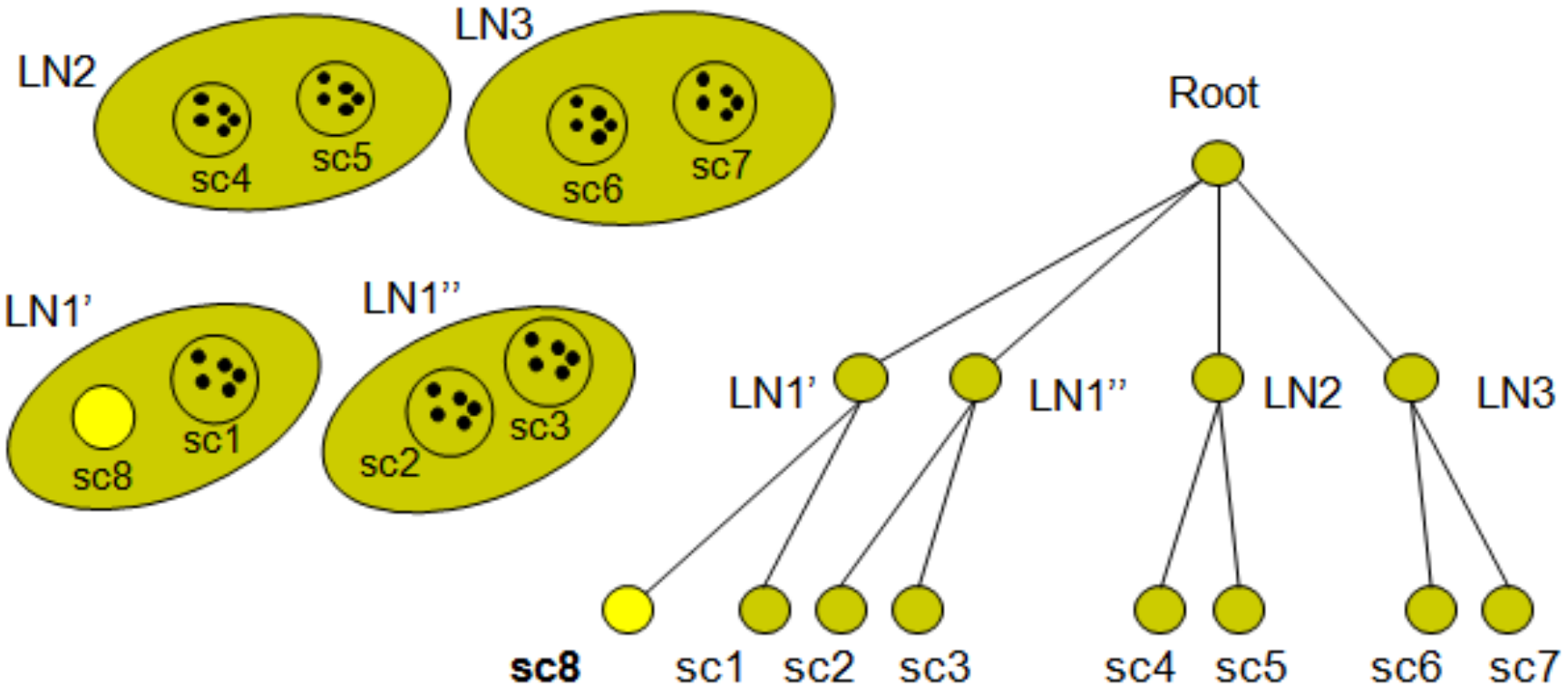
Thuật toán BIRCH…

- Ví dụ về hoạt động chèn nút lá



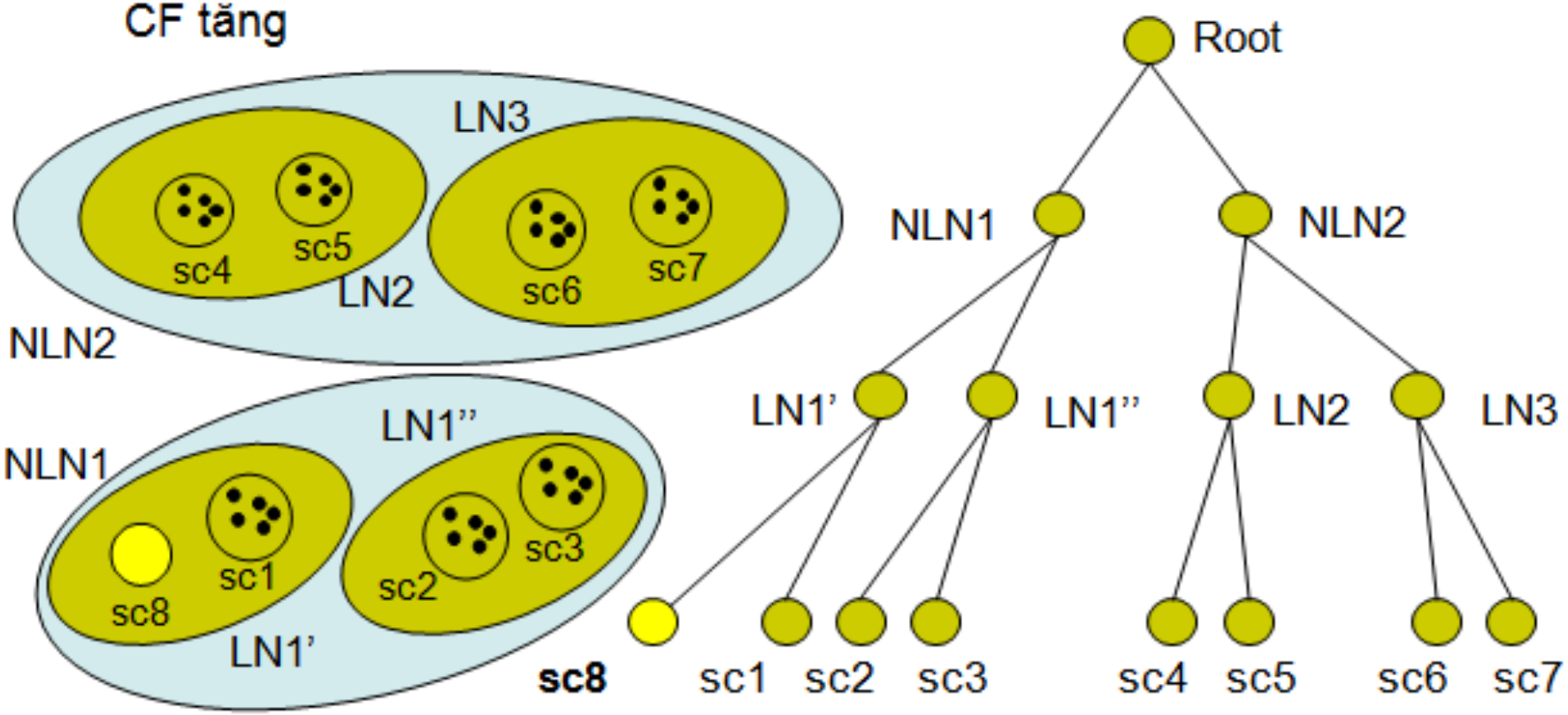
Thuật toán BIRCH…

Vì các yếu tố phân nhánh của một nút lá không thể vượt quá 3 nên sau đó LN1 được phân chia.



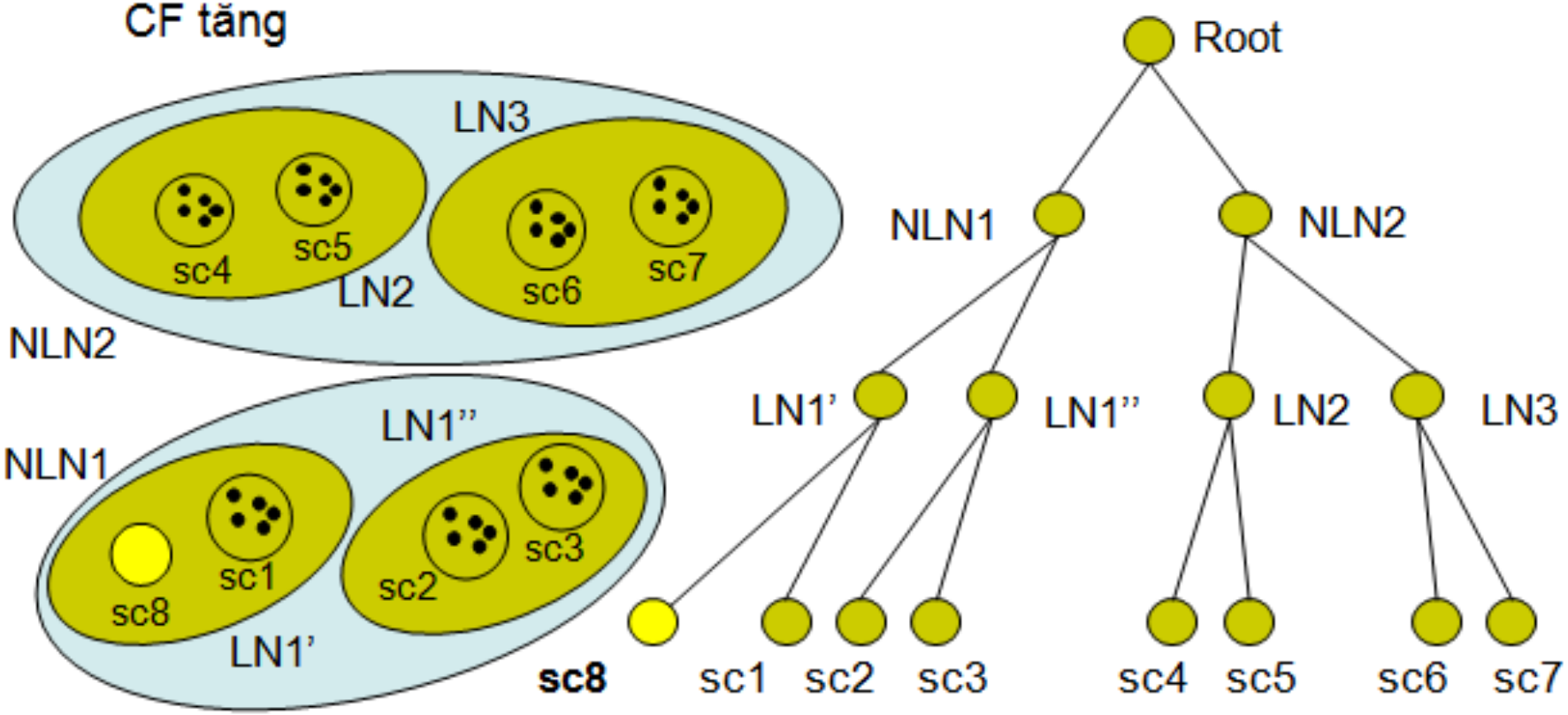
Thuật toán BIRCH…

Các yếu tố phân nhánh của một nút lá không thể vượt quá 3,nên sau đó gốc được phân chia và chiều cao của cây CF tăng



Thuật toán BIRCH…

Các yếu tố phân nhánh của một nút lá không thể vượt quá 3,nên sau đó gốc được phân chia và chiều cao của cây CF tăng



Thuật toán BIRCH…

- Ưu:
  - Với cấu trúc cây CF, BIRCH có tốc độ thực hiện PCDL nhanh.
  - Tốt với tập DL lớn
  - Hiệu quả khi tập DL tăng trưởng theo thời gian.
  - Chỉ duyệt toàn bộ DL một lần với một lần quét thêm tuỳ chọn.
  - Độ phức tạp là O(n), với n là số đối tượng DL.
- Nhược:
  - Chất lượng của các cụm không được tốt.
  - Nếu dùng khoảng cách Euclide, nó chỉ tốt với các DL số.
  - Tham số T có ảnh hưởng rất lớn tới kích thước và tính tự nhiên của cụm.
  - Không thích hợp với DL đa chiều.

Phân cụm dựa trên mật độ

- Thuật toán DBSCAN
- Thuật toán OPTICS
- Thuật toán DENCLUE

Phân cụm dựa trên mật độ

- Nhóm các đối tượng theo hàm mật độ xác định.
  - Mật độ được định nghĩa như là số các đối tượng lân cận của một đối tượng DL theo một ngưỡng nào đó.
  - Khi một cụm DL đã xác định thì nó tiếp tục được phát triển thêm các đối tượng DL mới miễn là số các lân cận của chúng lớn hơn một ngưỡng đã xác định.
- PP này có thể phát hiện ra các cụm DL với hình thù bất kỳ.
- Việc xác định các tham số mật độ của thuật toán rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả PCDL.

Phân cụm dựa trên mật độ…

- Một số thuật toán điển hình như DBSCAN, OPTICS, DENCLUE, …

Hình 5: Một số hình dạng cụm dữ liệu khám phá được bởi kỹ thuật PCDL dựa trên mật độ với 3 CSDL.

Thuật toán DBSCAN

- DBSCAN (*Density-Based Spatial Clustering of Applications with noise*).
- Thuật toán tìm các đối tượng mà có số đối tượng láng giềng lớn hơn một ngưỡng tối thiểu.

- Tìm tất cả các đối tượng mà các láng giềng của nó thuộc về lớp các đối tượng đã xác định ở trên.

- Một cụm được xác định bằng một tập tất cả các đối tượng liên thông mật độ với các láng giềng của nó.
- DBSCAN có thể tìm ra các cụm với hình thù bất kỳ.

Thuật toán DBSCAN…

- Khi có một đối tượng được chèn vào chỉ tác động đến một láng giềng xác định.
- DBSCAN yêu cầu xác định bán kính Eps của các láng giềng và số các láng giềng tối thiểu Minpts, thường xác định bằng ngẫu nhiên/kinh nghiệm.
- Áp dụng chỉ số không gian để xác định các láng giềng của một đối tượng DL, nên độ phức tạp của DBSCAN đã được cải tiến là O(nlogn). Nếu không áp dụng cấu trúc chỉ số thì là $O(n^2)$.

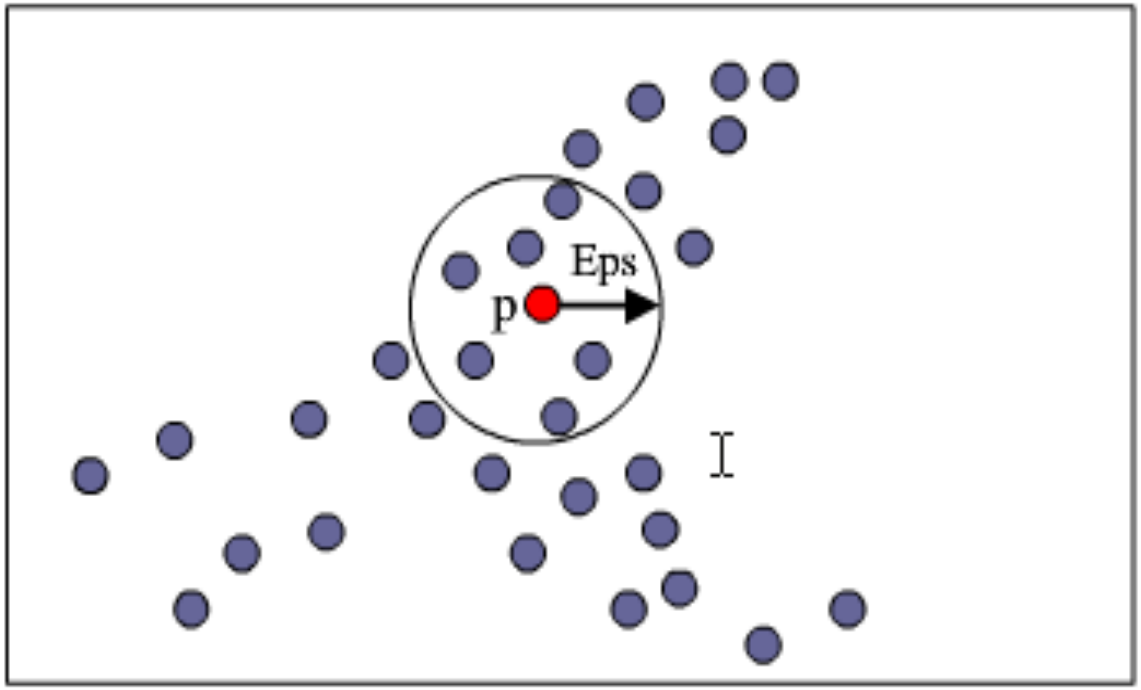
Thuật toán DBSCAN…

- Khoảng cách Euclide được dùng đo sự tương tự giữa các đối tượng nhưng không hiệu quả đối với DL đa chiều.

- ☞DBSCAN dựa trên các khái niệm mật độ có thể áp dụng cho các tập DL không gian lớn đa chiều.

Thuật toán DBSCAN…

- **Định nghĩa 1**: Lân cận của một điểm p với ngưỡng Eps (Eps - Neighborhood of a point):
  - Ký hiệu NEps(p), NEps(p) = {q ∈ D| khoảng cách Dist(p,q) ≤ Eps }, D là tập DL cho trước.



Thuật toán DBSCAN…

- Điểm p muốn nằm trong cụm C thì NEps(p) phải có tối thiểu MinPts điểm.

- Số điểm tối thiểu được chọn là bao nhiêu cũng là bài toán khó, vì: Nếu số điểm tối thiểu lớn thì chỉ những điểm nằm thực sự trong cụm C mới đạt đủ tiêu chuẩn, trong khi đó những điểm nằm ngoài biên của cụm không thể đạt được điều đó. Ngược lại, nếu số điểm tối thiểu là nhỏ thì mọi điểm sẽ rơi vào một cụm.

Thuật toán DBSCAN…

- Theo ĐN trên, chỉ những điểm thực sự nằm trong cụm mới thoả ĐK là điểm thuộc vào cụm. Những điểm nằm ở biên của cụm không thoả.

- Để tránh được điều này, có thể đưa ra một tiêu chuẩn khác để ĐN: Nếu một điểm p muốn thuộc một cụm C phải tồn tại một điểm q mà $p \in NEps(q)$ và số điểm trong NEps(q) phải lớn hơn số điểm tối thiểu như ĐN sau:

Thuật toán DBSCAN…

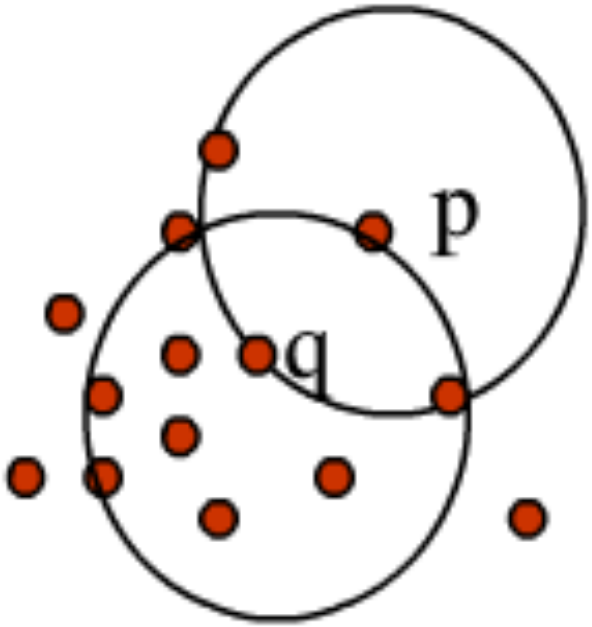
- **Định nghĩa 2**: *Đến được trực tiếp theo mật độ (Directly Density - reachable)*

- Một điểm p được gọi là đến được trực tiếp từ điểm q với ngưỡng Eps nếu :
  - 1. p ∈NEps(q)
  - 2. || NEps(q)|| ≥ MinPts (Điều kiện nhân)

- Điểm q gọi là điểm nhân (Core point).
- Đến được trực tiếp là một hàm phản xạ và đối xứng đối với hai điểm nhân và bất đối xứng nếu một trong hai điểm đó không phải là điểm nhân.

Thuật toán DBSCAN…

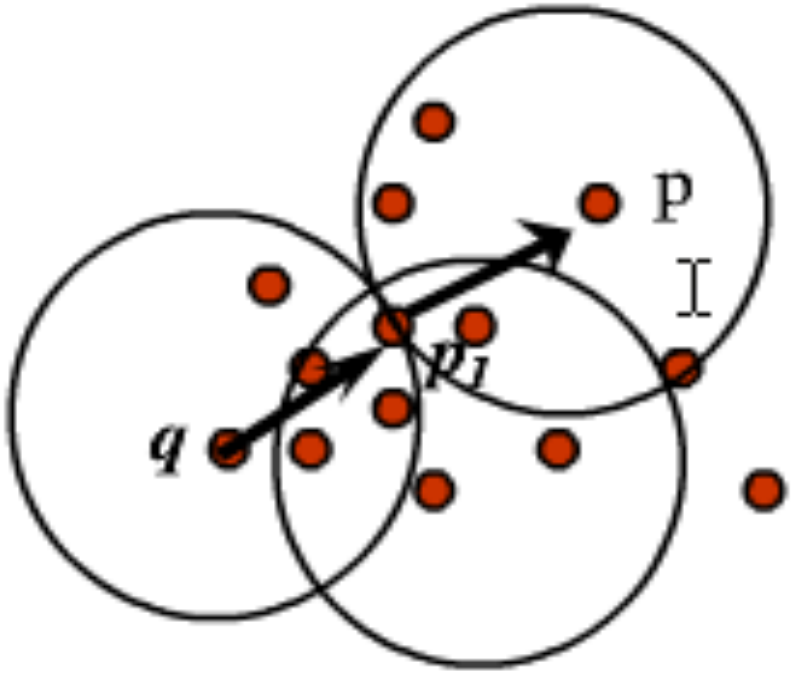
- Hình: Đến được trực tiếp theo mật độ



MinPts = 5
Eps = 1 cm

Thuật toán DBSCAN…

- **Định nghĩa 3**: *Đến được mật độ (Density - Reachable)*
- Một điểm p được gọi là đến được từ một điểm q với hai tham số Eps và MinPts nếu tồn tại một dãy p = p1, p2, …, pn =q thoả mãn pi+1 là có thể đến được trực tiếp từ pi với i=1…n-1.

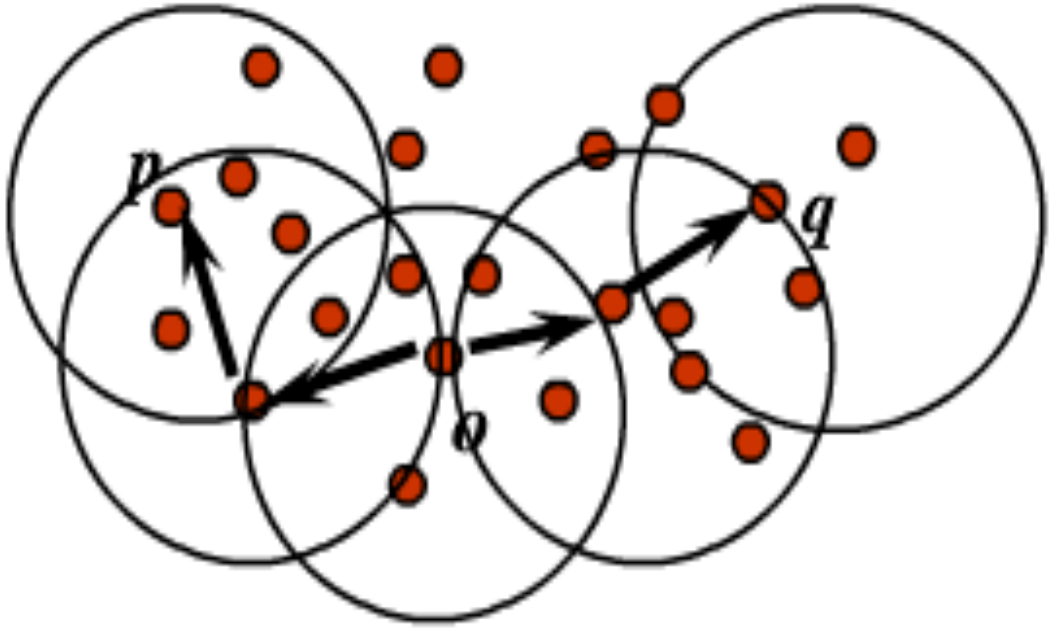


Thuật toán DBSCAN…

- Hai điểm biên của một cụm C có thể không đến được nhau vì cả hai có thể đều không thoả mãn điều kiện nhân.
- Phải tồn tại một điểm nhân trong C mà cả hai điểm đều có thể đến được từ điểm đó.
- Để cho thuận tiện chúng ta có định nghĩa liên thông mật độ (Density - Connectivity).

Thuật toán DBSCAN…

- **Định nghĩa 4**: *Liên thông mật độ (Density - Reachable)*
  - Điểm p được gọi là điểm liên thông với q theo hai tham số Eps, MinPts nếu tồn tại một điểm O mà cả hai điểm p, q đều có thể đến được theo tham số Eps và MinPts.
  - Liên thông mật độ có tính chất đối xứng và phản xạ.

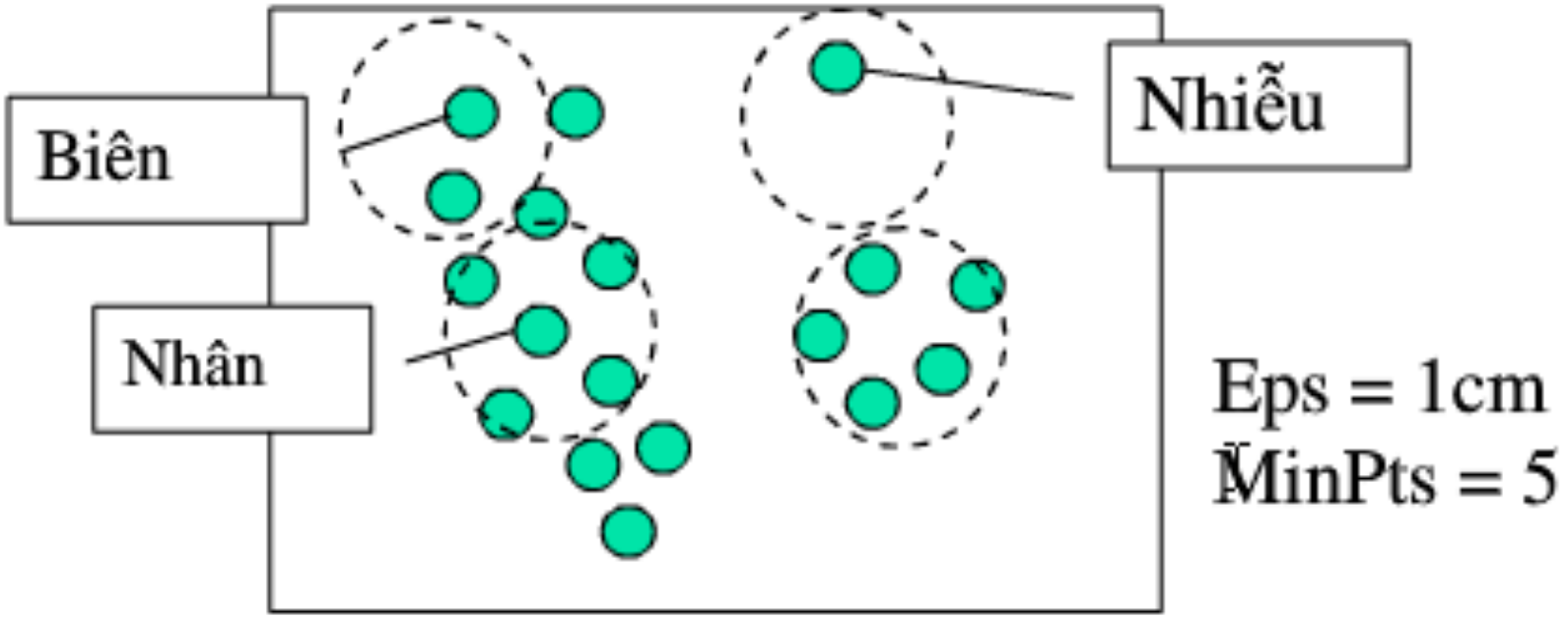


Thuật toán DBSCAN…

- **Định nghĩa 5** : *Cụm (Clustering)*
- Một tập con C khác rỗng của D được gọi là một cụm (cluster) theo Eps và MinPts nếu thoả hai điều kiện :
  - ∀ p,q ∈ D, nếu p ∈ C, q có thể đến được từ p theo Eps và MinPts thì q∈C.
  - ∀ p,q ∈ C, p liên thông mật độ với q theo Eps và MinPts.

- **Định nghĩa 6**: *Dữ liệu nhiễu (Noise)*
  - Điểm nhiễu là điểm không thuộc vào cụm nào trong các cụm C1, C2, …, Ck, tức là Noise={p|∀i=1..k,p∉Ci}

Thuật toán DBSCAN…



Thuật toán DBSCAN…

- Với hai tham số Eps và MinPts cho trước, có thể khám phá các cụm theo hai bước:
  - *Bước 1*: Chọn một điểm bất kỳ từ tập DL ban đầu thoả mãn điều kiện nhân.
  - *Bước 2*: Lấy tất cả các điểm đến được mật độ với điểm nhân đã chọn ở trên để tạo thành cụm.

Thuật toán DBSCAN...

- Loại bỏ các điểm nhiễu
- Thực hiện phân cụm trên các điểm còn lại

```
1:  current_cluster_label ← 1
2:  for all core points do
3:      if the core point has no cluster label then
4:          current_cluster_label ← current_cluster_label + 1
5:          Label the current core point with cluster label
            current_cluster_label
6:      end if
7:      for all points in the Eps-neighborhood, except the point itself
        do
8:          if the point does not have a cluster label then
9:              Label the point with cluster label current_cluster_label
10:         end if
11:     end for
12: end for
```

Thuật toán DBSCAN…

- **Thuật toán DBSCAN**
- B1: Khởi tạo điểm p tuỳ ý và lấy tất cả các điểm đến được mật độ từ p với Eps, MinPts.
- B2: Nếu p không phải là nhân (có thể là biên, nhiễu, hoặc số lân cận của p ít hơn Minpts) thì DBSCAN sẽ đi thăm điểm tiếp theo.
- B3: Ngược lại (p là nhân) thì một cụm được hình thành và chứa tất cả các đối tượng trong lân cận của p. Sau đó, lân cận của những điểm này sẽ được khảo sát để xem liệu nó có được thêm tiếp vào cụm này hay không.
- B4: Nếu cụm không thể mở rộng thêm được nữa, thuật toán chọn ngẫu nhiên một đối tượng p khác chưa xét và lặp lại quá trình trên.
- B5: Lặp cho đến khi mọi đối tượng đều được gom cụm hay được đánh dấu là ngoại lai/nhiễu. Với một CSDL có n mẫu tin, cần phải truy vấn vùng n lần.

Thuật toán DBSCAN…

- Nếu sử dụng giá trị toàn cục Eps và MinPts, DBSCAN có thể hoà nhập hai cụm theo định nghĩa 5 thành một cụm nếu mật độ của hai cụm gần bằng nhau.

- Giả sử khoảng cách giữa hai tập DL S1, S2 là Dist (S1,S2) = min{dist(p,q)| p∈ S1 và q∈ S2}.

Thuật toán DBSCAN…

- **Ưu điểm:**
- • Có thể phát hiện các cluster có hình dạng bất kỳ
- • Chỉ yêu cầu một hàm đo khoảng cách và hai tham số đầu vào: Eps và MinPts.
- • Cho ra kết quả tốt và thực thi hiệu quả trên nhiều tập dữ liệu.
- **Nhược điểm:**
- • Không thích hợp cho việc tìm các cluster trong CSDL cực lớn.
- • Nếu tập dữ liệu có mật độ thay đổi lớn, thuật toán quản lý kém hiệu quả.
- **Độ phức tạp**
- Độ phức tạp trung bình của mỗi truy vấn là O(logn).
- Độ phức tạp của thuật toán là O(nlogn), n là kích thước tập dữ liệu.

Thuật toán OPTICS

- Là mở rộng của DBSCAN, bằng cách giảm bớt các tham số đầu vào.
- OPTICS *(Ordering Points To Identify the Clustering Structure)* sắp xếp các cụm theo thứ tự tăng dần nhằm tự động phân cụm dữ liệu.
- Thứ tự này diễn tả cấu trúc DL phân cụm dựa trên mật độ chứa thông tin tương đương với phân cụm dựa trên mật độ với một dãy các tham số đầu vào.
- OPTICS xem xét bán kính tối thiểu nhằm xác định các láng giềng phù hợp.
- DBSCAN và OPTICS tương tự với nhau về cấu trúc và có cùng độ phức tạp: O(nLogn) (n là kích thước của tập DL).

Thuật toán OPTICS…

- **Ưu điểm**:
- OPTICS là mở rộng của DBSCAN nên có cùng các ưu điểm như DBSCAN.
- Kết quả cụm không khác nhiều so với DBSCAN nhưng khắc phục được những nhược điểm mà DBSCAN mắc phải.
- Độ phức tạp của OPTICS tương đương với DBSCAN.
- **Nhược điểm**:
- Có vài đối tượng biên có thể bị thiếu khi trích rút cụm nếu chúng được xử lý bởi OPTICS trước khi một đối tượng nhân của cụm tương ứng được tìm thấy.

Thuật toán DENCLUE

- DENCLUE*(DENsity - Based CLUstEring)* là thuật toán PCDL dựa trên một tập các hàm phân phối mật độ.
- Ý tưởng:
  - Sự tác động của một đối tượng tới láng giềng của nó xác định bởi hàm ảnh hưởng (*Influence Function*).
  - Mật độ toàn cục của không gian các đối tượng được mô hình là tổng tất cả các hàm ảnh hưởng.
  - Các cụm được xác định bởi các đối tượng mật độ cao (*density attactors)*, là các điểm cực đại của hàm mật độ toàn cục.

Thuật toán DENCLUE…

- Hàm ảnh hưởng: Cho x,y là hai đối tượng trong không gian d chiều F^d, hàm ảnh hưởng của đối tượng y lên x được xác định f_B^y:Fd $\rightarrow R_o^+$, y $= f_B^y(x)$.

- Hàm ảnh hưởng là hàm tuỳ chọn, miễn là nó được xác định bởi khoảng cách d(x,y) của các đối tượng, như khoảng cách Euclide.

Thuật toán DENCLUE…

- DENCLUE phụ thuộc nhiều vào ngưỡng nhiễu ξ (*Noise Threshold*) và tham số mật độ δ.
- Ưu điểm
  - Có cơ sở toán học vững chắc
  - Có khả năng xử lý các phần tử ngoại lai.
  - Cho phép khám phá ra các cụm với hình thù bất kỳ ngay cả đối với các DL đa chiều.
- Độ phức tạp là O(nlogn).

Nhận xét về các thuật toán dựa trên mật độ

- ☞Các thuật toán dựa trên mật độ không thực hiện kỹ thuật phân mẫu trên tập DL như trong các thuật toán phân cụm phân hoạch

- ☞điều này có thể làm tăng thêm độ phức tạp do có sự khác nhau giữa mật độ của các đối tượng trong mẫu với mật độ của toàn bộ DL.

Phân cụm đặc thù

- Phân cụm mờ FCM, εFCM
- Phân cụm xắc suất EM
- Thuật toán STING
- Thuật toán CLIQUE

Phân cụm mờ

- Trong thực tế, các cụm dl lại có thể chồng lên nhau.
- Người ta đã áp dụng lý thuyết về tập mờ trong PCDL để giải quyết cho trường hợp này.
- Cách kết hợp này được gọi là phân cụm mờ.
- Trong phương pháp phân cụm mờ, độ phụ thuộc của đối tượng dl x_k tới cụm thứ i (u_{ik}) có giá trị thuộc khoảng [0,1].

Phân cụm mờ…

- Ý tưởng trên đã được giới thiệu bởi Ruspini (1969), được Dunn áp dụng năm 1973 nhằm xây dựng phương pháp phân cụm mờ dựa trên tối thiểu hoá hàm tiêu chuẩn.
- Bezdek (1982) đã tổng quát hoá phương pháp này và xây dựng thành thuật toán phân cụm mờ c-means có sử dụng trọng số mũ.
- K-means là PCDL rõ và c-means là phân cụm mờ tương ứng, cùng sử dụng một chiến lược phân cụm dl.

Phân cụm mờ…

- c-means mờ gọi tắt là FCM (Fuzzy C means-FCM) đã thành công trong nhận dạng mẫu, xử lý ảnh, y học, …
- Nhược: của FCM là nhạy cảm với nhiễu/phần tử ngoại lai, nghĩa là các trung tâm cụm có thể nằm xa so với trung tâm thực của cụm.
- Có nhiều cải tiến: Phân cụm dựa trên xác suất (keller, 1993), phân cụm nhiễu mờ (Dave, 1991), Phân cụm dựa trên toán tử L_P Norm (Kersten, 1999). Thuật toán ε-Insensitive Fuzzy C-means (εFCM).

- ☞*Đi sâu vào hai thuật toán FCM và εFCM*

Phân cụm mờ...

- Phân k cụm, tổng quát hoá hàm mục tiêu

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{m} w_{ij} dist(x_i, c_j)^2 \qquad \sum_{j=1}^{k} w_{ij} = 1$$

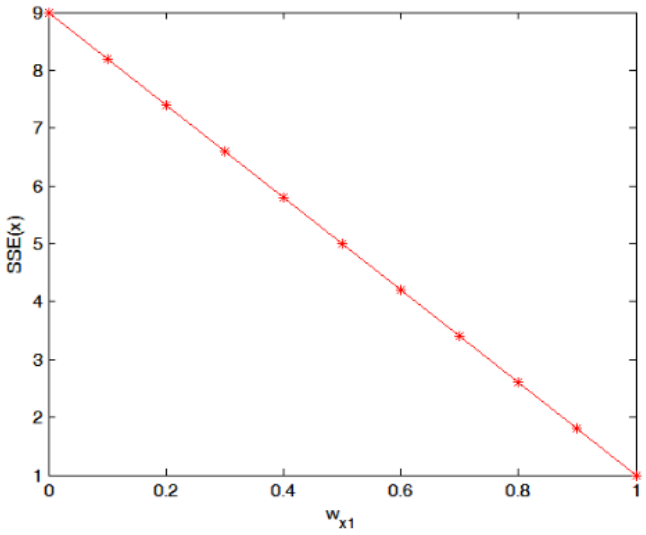
  c_j là tâm cụm, w_{ij} là trọng số mà đối tượng x_i thuộc vào cụm j
- Phân cụm cứng $w_{ij} \in \{0, 1\}$
- Để cực tiểu hoá SSE, lặp các bước sau:
  - Cố định c_j và xác định w_{ij} (gán cụm)
  - Cố định w_{ij} và tính lại c_j

Phân cụm mờ…

- Ước lượng trọng số

$$SSE(x) = w_{x1}(2-1)^2 + w_{x2}(5-2)^2$$
$$= w_{x1} + 9w_{x2} \qquad (1)$$

$SSE(x)$ cực tiểu khi
$w_{x1} = 1, w_{x2} = 0$



FCM

- Hàm mục tiêu

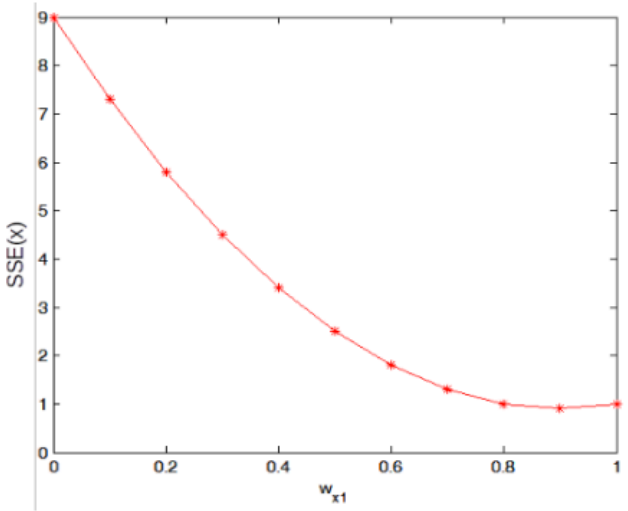
$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{m} w_{ij}^{p} dist(x_i, c_j)^2 \qquad \sum_{j=1}^{k} w_{ij} = 1$$

  trong đó $p > 1$ là số mũ điều khiển độ mờ
- Để cực tiểu hàm mục tiêu, lặp lại các bước sau:
  - Cố định c_j và xác định w_{ij}
  - Cố định w_{ij} và tính lại c
- Phân cụm Fuzzy C-means $w_{ij} \in [0, 1]$

FCM...

- Ví dụ:

$$SSE(x) = w_{x1}^2(2-1)^2 + w_{x2}^2(5-2)^2$$
$$= w_{x1}^2 + 9w_{x2}^2 \qquad (2)$$



$SSE(x)$ cực tiểu khi
$w_{x1} = 0.9, w_{x2} = 0.1$

FCM...

- Hàm mục tiêu

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{m} w_{ij}^{p} dist(x_i, c_j)^2 \qquad \sum_{j=1}^{k} w_{ij} = 1$$

- Khởi tạo: chọn ngẫu nhiên các trọng số w_{ij}
- Lặp lại:
  - Cập nhật tâm cụm

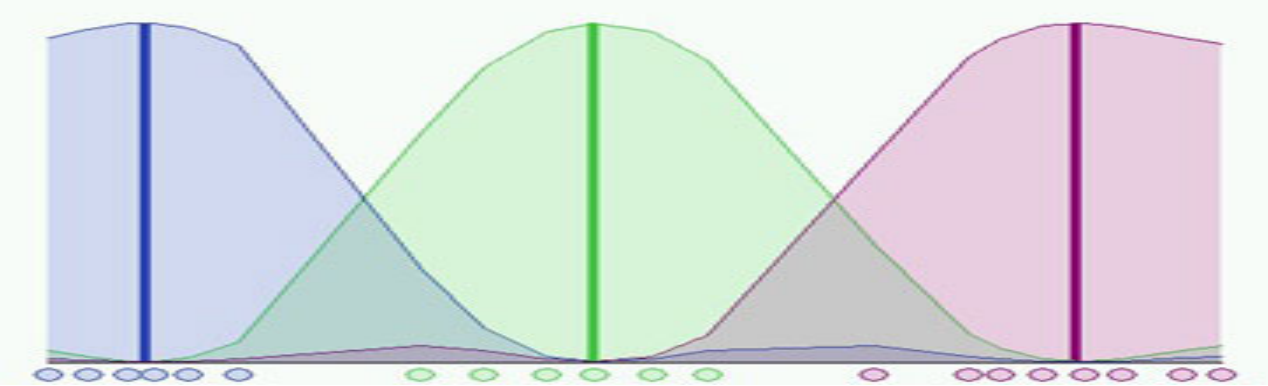
$$c_j = \sum_{i=1}^{m} w_{ij} x_i / \sum_{i=1}^{m} w_{ij}$$

  - Cập nhật trọng số

$$w_{ij} = (1/dist(x_i, c_j)^2)^{\frac{1}{p-1}} / \sum_{j=1}^{k} (1/dist(x_i, c_j)^2)^{\frac{1}{p-1}}$$

FCM…

- Mô phỏng các cụm khám phá được theo FCM:



Hình 25: Các cụm khám phá được bởi thuật toán phân cụm mờ

- Độ phức tạp của FCM cũng như k-means là O(Ikn).
- FCM là mở rộng của k-means nhằm khám phá ra các cụm chồng lên nhau.
- FCM vẫn có các nhược điểm của k-means với các phần tử ngoại lai/nhiễu.

Phân cụm xác suất

- Ý tưởng là mô hình hoá tập điểm dữ liệu như là mô hình trộn phân bố
  - Phân bố được dùng điển hình là phân bố chuẩn (Gausse), nhưng cũng có thể chọn các phân bố khác
- Các cụm được tìm thấy bằng cách ước lượng các tham số của các phân bố trong mô hình trộn
  - Có thể sử dụng một thuật toán EM (*Expectation-Maximization*) để ước lượng tham số
    - *k*-means thực chất là một dạng đặc biệt của cách tiếp cận này
  - Cung cấp một biểu diễn gọn gàng cho các cụm
  - Xác suất một điểm thuộc vào một cụm có chức năng tương tự như phân cụm mờ.

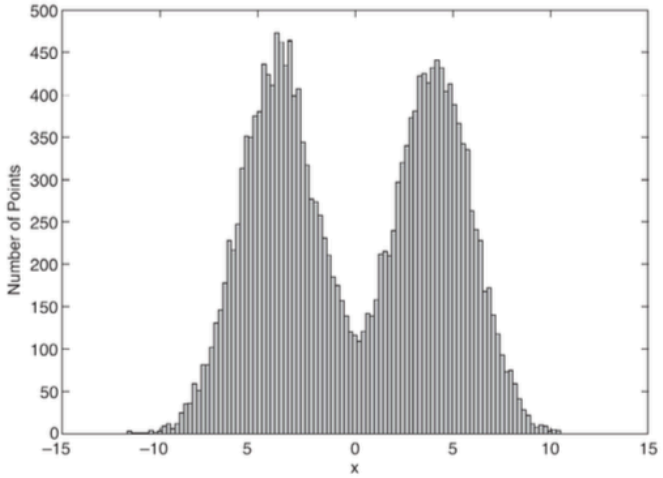
Phân cụm xác suất

- Dựa vào trộn phân bố
  - Ý tưởng là mô hình hoá tập điểm dữ liệu như là mô hình trộn phân bố
    - Phân bố được dùng điển hình là phân bố chuẩn (Gausse), nhưng cũng có thể chọn các phân bố khác
  - Các cụm được tìm thấy bằng cách ước lượng các tham số của các phân bố trong mô hình trộn
    - Có thể sử dụng một thuật toán EM (*Expectation-Maximization*) để ước lượng tham số
      - *k*-means thực chất là một dạng đặc biệt của cách tiếp cận này
    - Cung cấp một biểu diễn gọn gàng cho các cụm
    - Xác suất một điểm thuộc vào một cụm có chức năng tương tự như phân cụm mờ.

Phân cụm xác suất…

- Ví dụ

  - Ví dụ: xét mô hình hoá các điểm có đồ hình bên phải
  - Trong hình, các điểm có vẻ tuân theo mô hình trộn của 2 phân bố chuẩn



  - Giả sử có thể ước lượng được θ là các tham số trung bình và độ lệch chuẩn của mỗi phân bố trên
    - Các tham số này xác định rõ 2 cụm
    - Có thể tính được xác suất mỗi điểm thuộc vào một cụm
    - Có thể gán mỗi điểm vào cụm mà xác suất thu được lớn nhất

$$P(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Phân cụm xác suất - TT EM

Khởi tạo các tham số
Repeat
  Với mỗi điểm, tính xác suất của nó cho từng phân bố
  Dùng các xuất suất này cập nhật tham số của mỗi phân bố
Until không còn thay đổi

- Tương tự với *k*-means, gồm bước gán cụm và cập nhật
- Có thể dùng khởi tạo ngẫu nhiên
- Có vấn đề về cực tiểu địa phương
- Với phân bố chuẩn, điển hình dùng *k*-means để khởi tạo
- Nếu sử dụng phân bố chuẩn, có thể tìm được dạng cầu lẫn dạng e-líp

Phân cụm xác suất - TT EM...

- Cập nhật tâm cụm
  - Cập nhật trọng số

$$c_j = \sum_{i=1}^{m} x_i P(C_j|x_i) / \sum_{i=1}^{m} P(C_j|x_i)$$

x_i là một điểm, C_j là một cụm, c_j là tâm cụm đó

  - Tương tự như công thức phân cụm k-means mờ:
    - Trọng số là các xác suất, nhưng không luỹ thừa
    - Các xác suất được tính dựa vào quy tắc Bayes

$$P(C_j|x_i) = \frac{P(x_i|C_j)P(C_j)}{\sum_{l=1}^{k} P(x_i|C_l)P(C_l)}$$

  - Cần gán trọng số cho mỗi cụm
    - Tương tự như xác suất tiên nghiệm

$$P(C_j) = \frac{1}{m} \sum_{i=1}^{m} P(C_j|x_i)$$

Phân cụm xác suất - TT EM...

- Thuật toán chi tiết

1: Khởi tạo các tham số mô hình
2: **repeat**
3: **Bước kì vọng EStep** (*Expectation Step*) Với mỗi đối tượng, tính xác suất đối tượng đó thuộc vào mỗi phân bố $P(\textit{distribution } j | x_i, \theta)$
4: **Bước cực đại MStep** (*Maximization Step*) Cho các xác suất tính ở bước EStep, tìm ước lượng mới của các tham số sao cho giá trị kì vọng đạt cực đại
5: **until** Các tham số không thay đổi (hoặc thay đổi nhỏ hơn ngưỡng nào đó)

Phân cụm xác suất - TT EM...

- Nhận xét
  - Có thể hội tụ chậm
  - Chỉ đảm bảo tìm được cực đại địa phương
  - Dùng các giả thiết thống kê quan trọng (về phân bố)
  - Số tham số cho phân bố chuẩn $O(d2)$, với d là số chiều:
    - Các tham biến gắn với ma trận hiệp phương sai.
    - k-means chỉ ước lượng các trung bình cụm, số lượng $O(d)$

Phân cụm xác suất - TT EM...

- Nhận xét
  - Có thể hội tụ chậm
  - Chỉ đảm bảo tìm được cực đại địa phương
  - Dùng các giả thiết thống kê quan trọng (về phân bố)
  - Số tham số cho phân bố chuẩn $O(d2)$, với d là số chiều:
    - Các tham biến gắn với ma trận hiệp phương sai.
    - k-means chỉ ước lượng các trung bình cụm, số lượng $O(d)$

Chất lượng và đặc trưng phân cụm

Đánh giá chất lượng phân cụm

- Với các phương pháp phân lớp (học có hướng dẫn), có sẵn nhiều tiêu chí đánh giá như độ chính xác, độ phủ ...
- Với bài toán phân cụm, cũng cần đánh giá chất lượng các cụm thu được nhằm
  - Tránh tìm mẫu trong dữ liệu nhiễu
  - So sánh 2 thuật toán phân cụm
  - So sánh 2 tập cụm
  - So sánh 2 cụm

Đánh giá chất lượng phân cụm…

- Các vấn đề đánh giá chất lượng phân cụm

1. Xác định "xu hướng phân cụm" của tập dữ liệu: có cấu trúc trong dữ liệu hay không?
2. Đánh giá ngoài: So sánh kết quả phân cụm với các cấu trúc nhóm đã biết, ví dụ dựa vào một thuộc tính phân lớp đã có
3. Đánh giá trong: phân tích sự phù hợp của kết quả phân cụm, tìm đặc trưng cụm
4. So sánh, xác định kết quả tốt hơn trong 2 kết quả phân cụm
5. Xác định số cụm

Đánh giá chất lượng phân cụm…

- Tiêu chuẩn/độ đo đánh giá chất lượng PC
  - Chỉ số ngoài: dùng để so sánh các cụm thu được với các lớp đã có sẵn
    - Entropy
  - Chỉ số trong: Độ đo chất lượng cấu trúc cụm thu được
    - Tổng bình phương lỗi SSE (*Sum of Squared Error*)
    - Hệ số dáng điệu
  - Chỉ số tương đối: So sánh 2 kết quả phân cụm
    - Dùng Entropy hoặc SSE

Đánh giá chất lượng phân cụm…

- Xác định số cụm
  - Đối với cây phân cấp sử dụng chỉ số đo độ tương tự
  - Với phân hoạch trực tiếp thì sử dụng các tiêu chí gắn với tổng bình phương lỗi
  - Đồ thị 'dáng điệu'

Đánh giá chất lượng phân cụm…

- Độ đo đánh giá trong:Tính gắn kết và tách biệt
  - Tính gắn kết trong cụm: đo sự liên kết của các đối tượng trong cùng một cụm. Ví dụ sử dụng tổng phương sai trong cụm

$$SSE = WSS(Within\ Sum\ of\ Squares) = \sum_i sum_{x \in C_i}(x - m_i)^2$$

  - Tính tách biệt của các cụm: đo sự phân biệt của một cụm so với các cụm khác. Ví dụ dùng tổng phương sai liên cụm

$$BSS(Between\ Sum\ of\ Squares) = \sum_i |C_i|(m - m_i)^2$$

Đánh giá chất lượng phân cụm…

- Độ đo đánh giá trong: hệ số dáng điệu (*silhouette coefficient*)
  - Hệ số dáng điệu kết hợp cả 2 yếu tố gắn kết và tách biệt, nhưng dùng cho cả từng điểm cũng như từng cụm và tổng thể
  - Cho một điểm *i*
    - Tính *a* = trung bình khoảng cách từ điểm *i* tới các điểm khác trong cụm
    - Tính *b* = min(trung bình khoảng cách từ điểm *i* tới các điểm ở trong cụm khác)
    - Hệ số dáng điệu của một điểm $s = (b - a)/\max(a, b)$
    - Hệ số này thuộc khoảng [0, 1], càng gần 1 chất lượng phân cụm càng tốt
  - Có thể tính hệ số dáng điệu trung bình cho 1 cụm hoặc cả một phân cụm

Đánh giá chất lượng phân cụm…

- Độ đo đánh giá ngoài

  - Xây dựng bảng thống kê chéo số phần tử của từng cụm thuộc từng lớp đã có sẵn (K cụm, L lớp)
  - Entropy
    - Với mỗi cụm j tính p_{ij} là xác suất một phần tử của j thuộc lớp i: $p_{ij} = m_{ij}/m_j$, m_j là số giá trị trong cụm j, m_{ij} là số các phần tử của lớp i nằm trong cụm j
    - Entropy của cụm j

$$e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$$

    - Tổng entropy của tất cả các lớp

$$e = \sum_{i=1}^{K} \frac{m_j}{m} e_j$$

  - Độ thuần nhất $purity_j = \max p_{ij}$

$purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$

Đặc trưng của cụm/lớp

- **Giúp cho việc diễn giải một phân hoạch:**
  - Một phân hoạch sẽ được nâng giá trị một cách đáng kể nếu nó đi kèm với một mô tả về các lớp theo các thuộc tính và các cá thể.

Đặc trưng của cụm/lớp…

- **Diễn giải phân hoạch theo cá thể**
- Với mỗi lớp người ta xét các yếu tố sau:
  - số phần tử,
  - đường kính (khoảng cách giữa 2 điểm xa nhất),
  - sự tách biệt (khoảng cách cực tiểu giữa lớp đang xét với lớp gần nó nhất),
  - tên của các cá thể nằm gần trọng tâm của lớp nhất,
  - tên của các cá thể nằm xa trọng tâm của lớp nhất.

Đặc trưng của cụm/lớp…

- **Diễn giải phân hoạch theo các biến liên tục**

So sánh giá trị trung bình $\bar{x}_k$ và độ lệch chuẩn s_k của 1 biến X trong lớp k với giá trị trung bình và độ lệch chuẩn tổng thể.

Đặc trưng của cụm/lớp…

- **Diễn giải phân hoạch theo các biến liên tục**

  So sánh giá trị trung bình $\bar{x}_k$ và độ lệch chuẩn s_k của 1 biến X trong lớp k với giá trị trung bình và độ lệch chuẩn tổng thể.

- **Diễn giải phân hoạch theo các biến định tính**

	Lớp k	Các lớp khác	Quần thể
Giá trị j	n_{kj}	*	n_j
Các giá trị khác	*	*	*
Quần thể	n_k	*	n

Tỉ lệ phần trăm tổng thể $\Rightarrow n_j/n$

Tỉ lệ phần trăm " giá trị / lớp " $\Rightarrow n_{kj}/n_k$

Tỉ lệ phần trăm " lớp / giá trị " $\Rightarrow n_{kj}/n_j$

Đặc trưng của cụm/lớp…

- **Giá trị kiểm tra (*test-value*)**
  - **Các thống kê trên các biến ở trên có thể được chuyển thành 1 tiêu chuẩn gọi là " giá trị kiểm tra ".**
  - Giá trị kiểm tra cho phép chọn lọc các biến liên tục hoặc các giá trị của các biến rời rạc đặc trưng nhất của mỗi lớp

Đặc trưng của cụm/lớp…

- Giá trị kiểm tra cho biến liên tục

  Giá trị kiểm tra bằng khoảng cách giữa giá trị trung bình trong 1 lớp và giá trị trung bình tổng thể biểu diễn theo số các độ lệch chuẩn:

$$\text{v-test} = \frac{\bar{x}_k - \bar{x}}{s_k(X)}$$

  với

$$s_k^2(X) = \frac{n - n_k}{n - 1} \cdot \frac{s^2(X)}{n_k}$$

- Giá trị kiểm tra cho các biến tên

  Giá trị kiểm tra của giá trị k của biến j :

$$\text{v-test} = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n - 1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

Đặc trưng của cụm/lớp…

- **Diễn giải giá trị kiểm tra**
- **Nếu |v-test| > 2, giá trị trung bình trong toàn bộ quần thể phân biệt đáng kể với giá trị trung bình của lớp.**
  - Diễn giải này chỉ có nghĩa cho các biến bổ sung không tham gia vào quá trình xây dựng các lớp: có sự phụ thuộc giữa các lớp của một phân hoạch và các biến được dùng cho định nghĩa phân hoạch.
  - Với các biến dùng trong phân cụm, các giá trị kiểm tra là các độ đo độ tương tự đơn giản giữa các biến và các lớp

Đặc trưng của cụm/lớp…

- **Diễn giải giá trị kiểm tra**
- **Nếu |v-test| > 2, giá trị trung bình trong toàn bộ quần thể phân biệt đáng kể với giá trị trung bình của lớp.**
  - Diễn giải này chỉ có nghĩa cho các biến bổ sung không tham gia vào quá trình xây dựng các lớp: có sự phụ thuộc giữa các lớp của một phân hoạch và các biến được dùng cho định nghĩa phân hoạch.
  - Với các biến dùng trong phân cụm, các giá trị kiểm tra là các độ đo độ tương tự đơn giản giữa các biến và các lớp

Phân cụm trong thực tế

- Đối với 1 phương pháp phân cụm từ dưới lên, người ta thường cắt cây phân cấp sao cho thu được các lớp thuần nhất nhất có thể mà vẫn phân biệt tốt lẫn nhau bằng cách dựa vào chỉ số ở các tầng (cf. ví dụ).
- Chiến lược "Phân tích thành tố + Phân cụm" cho phép loại bỏ các dao động ngẫu nhiên và thu được các lớp ổn định hơn, vì các trục thành tố thường rất ổn định đối với việc chọn mẫu.

Phân cụm trong thực tế…

- Chiến lược "Phân cụm hỗn hợp", nghĩa là thực hiện việc phân cụm từ dưới lên bắt đầu từ vài chục nhóm thuần nhất thu được từ 1 thuật toán kết nhập quanh các tâm động kiểu k-means, là chiến lược rất thích hợp cho việc phân hoạch 1 tập hợp chứa 1 số lượng lớn (hàng nghìn hoặc hàng chục nghìn) các cá thể.
- Tính thuần nhất của các lớp thu được có thể tối ưu hoá bằng một thủ tục củng cố các lớp, tức là thực hiện lại 1 quá trình kết nhập quanh các tâm động của các lớp.

Phân cụm trong thực tế…

- **Một số bài toán cụ thể trong thực tế**
  - Phân cụm trong hàng không
  - Phân cụm học sinh
  - Phân cụm khách hàng
  - Phân cụm/loại bệnh nhân
  - Phân cụm trong Giao thông
  - Phân cụm/nhóm văn bản
  - Hệ thống tài chính ngân hàng

Thực hành phân cụm

Thực hành phân cụm

- **Demo một số bài toán cụ thể trong thực tế**
  - Phân cụm trong hàng không
  - Phân cụm học sinh/SV
  - Phân cụm khách hàng
  - Phân cụm trong y tế
  - Phân cụm trong Giao thông
  - Hệ thống tài chính ngân hàng

Thực hành phân cụm…

- **Một số công cụ**
- Scikit-learn Data Clustering (Python) http: //scikit-learn.org/stable/modules/clustering.html
- Open Source Data Mining Software (WEKA Workbench) http://www.cs.waikato.ac.nz/ml/weka/
- Apache Mahout Machine Learning Library http://mahout.apache.org/users/clustering/
- R-archive network http://cran.r-project.org/
- Tanagra https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
- YALE (rapid-i.com)
- KNIME (www.knime.org)
- Orange (www.ailab.si/orange)

Thực hành phân cụm…

- **Một số công cụ…**

Các công cụ thương mại

- Oracle 10g/11g DBMS và Oracle 10g/11g Data Mining
  www.oracle.com
- Microsoft data mining tools (MS SQL Server 2005/2008 DBMS và Business Intelligence Development Studio)
  www.microsoft.com
- Hỗ trợ từ Intelligent Miner (IBM)
- Hỗ trợ từ Enterprise Miner (SAS Institute)

Tài liệu tham khảo

- Charu Aggarwal, Data Mining, Springer 2015.
- J. Han and M. Kamber, Data Mining: Concepts and Techniques, 3rd ed. http://www.cs.illinois.edu/~hanj/bk3/
- Tan, Steinbach, Karpatne and Kumar, Introduction to Data Mining, 2nd ed. https://www-users.cs.umn.edu/~kumar001/dmbook/index.php
- David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.
- Hillol Kargupta, Jiawei Han, Philip S.Yu, Rajeev Motwani, and Vipin Kumar, "Next Generation of Data Mining", Taylor & Francis Group, LLC, 2009.
- Ian H.Witten, Frank Eibe, Mark A. Hall, "Data mining : practical machine learning tools and techniques", Third Edition, Elsevier Inc, 2011.
- Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", Second Edition, Springer Science + Business Media, LLC 2005,2010.

Tutorials

- https://eric.univ-lyon2.fr/~ricco/cours/didacticiels/Python/en/cah_kmeans_avec_python.pdf
- http://data-mining-tutorials.blogspot.com/search/label/Clustering