Phân tích cụm

Nguyễn Thị Minh Huyền

Bài giảng của DSLab
Viện Nghiên cứu Cao cấp về Toán (VIASM)

Nội dung

VIASM
Vietnam Institute for Advanced Study in Mathematics

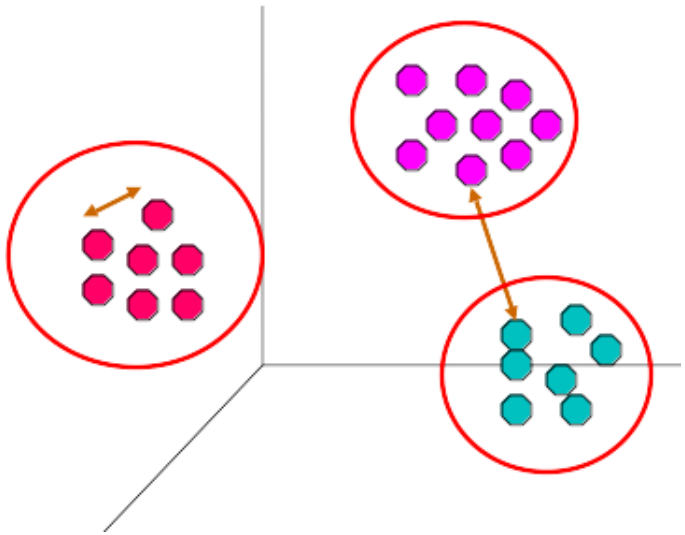
Nội dung

Vietnam Institute for Advanced Study in Mathematics

Mục tiêu 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



Vietnam Institute for Advanced Study in Mathematics

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

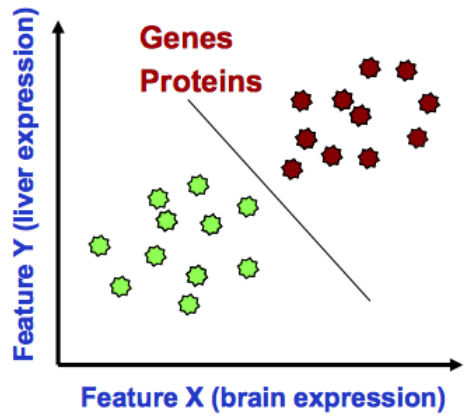
Phân cụm (Clustering) và Phân lớp (Classification)

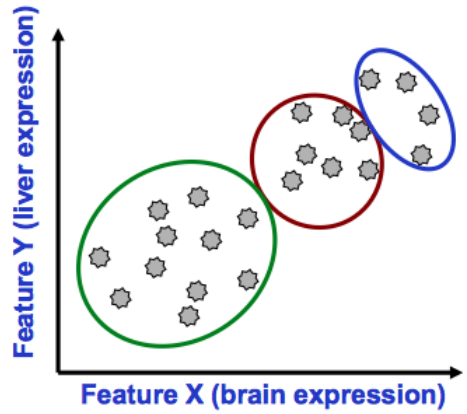
- Mục tiêu phân cụm: nhóm các đối tượng tương tự, nhờ đó **phát hiện cấu trúc ẩn** của dữ liệu
- Mục tiêu phân lớp: Trích rút các đặc trưng từ dữ liệu cho phép **phân loại các phần tử mới** vào các **lớp đã xác định**

Vietnam Institute for Advanced Study in Mathematics

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

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.

Vietnam Institute for Advanced Study in Mathematics

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.

Việc phân cụm được thực hiện với các biến và phương pháp được chọn một cách có chủ đích

Vietnam Institute for Advanced Study in Mathematics

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Độ khác biệt/tương tự

Các phương pháp phân cụm cần có

- Tiêu chuẩn đo độ khác biệt (khoảng cách) giữa các đối tượng, hoặc ngược lại là đo độ tương tự giữa các đối tượng. Chẳng hạn:
  - Với các dữ liệu định lượng thường sử dụng khoảng cách
  - Với dữ liệu văn bản thường sử dụng độ tương tựi

Vietnam Institute for Advanced Study in Mathematics

Độ khác biệt (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 $d : E \times E \rightarrow R^+$

$$
\begin{array}{lll}
1. & d(i,i) = 0 & \forall i \in E \\
2. & d(i,i') = d(i',i) & \forall i, i' \in E \times E
\end{array}
$$

Vietnam Institute for Advanced Study in Mathematics

Độ khác biệt (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 $d : E \times E \rightarrow R^+$

$$
\begin{array}{lll}
1. & d(i,i) = 0 & \forall i \in E \\
2. & d(i,i') = d(i',i) & \forall i, i' \in E \times E
\end{array}
$$

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

Vietnam Institute for Advanced Study in Mathematics

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*)

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Tập biến nhị phân (1/3)

Ví dụ:

	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

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

Vietnam Institute for Advanced Study in Mathematics

Tập biến nhị phân (2/3)

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

$\theta = 1 -$ tiêu chuẩn Jaccard

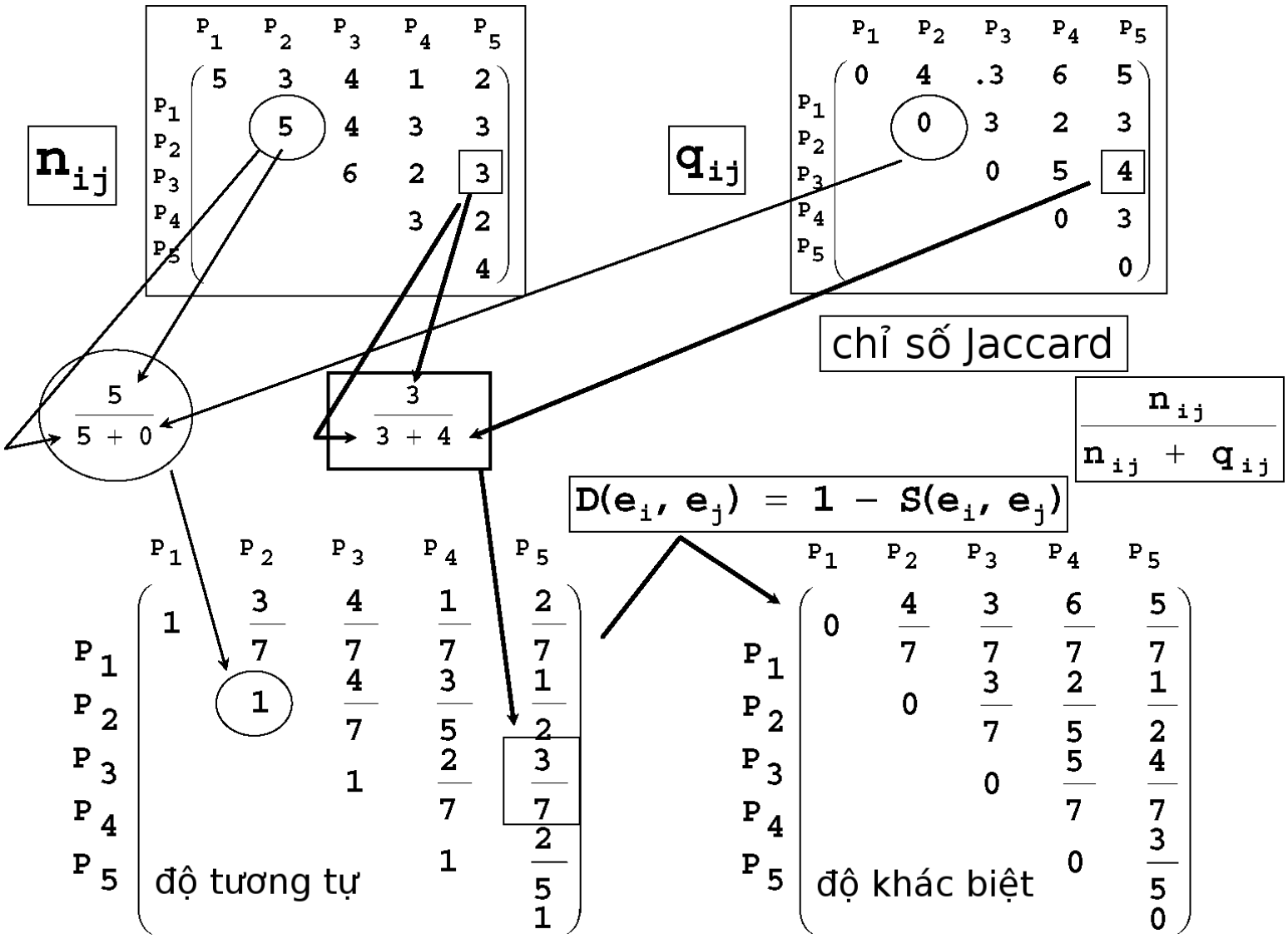
$\theta = 2 -$ tiêu chuẩn Dice

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

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

Vietnam Institute for Advanced Study in Mathematics

Tập biến nhị phân (3/3)



Nội dung

Vietnam Institute for Advanced Study in Mathematics

Tập biến định tính/phân loại (1/3)

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 (2/3)

Sử dụng độ đo tương tự

- 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}

Vietnam Institute for Advanced Study in Mathematics

Tập biến định tính/phân loại (3/3)

Sử dụng độ đo tương tự

- 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
  - Tần suất xuất hiện nghịch đảo
  - 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}

Vietnam Institute for Advanced Study in Mathematics

Nội dung

Vietnam Institute for Advanced Study in Mathematics

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 $= max_k(|x_{ik} - x_{jk}|)$

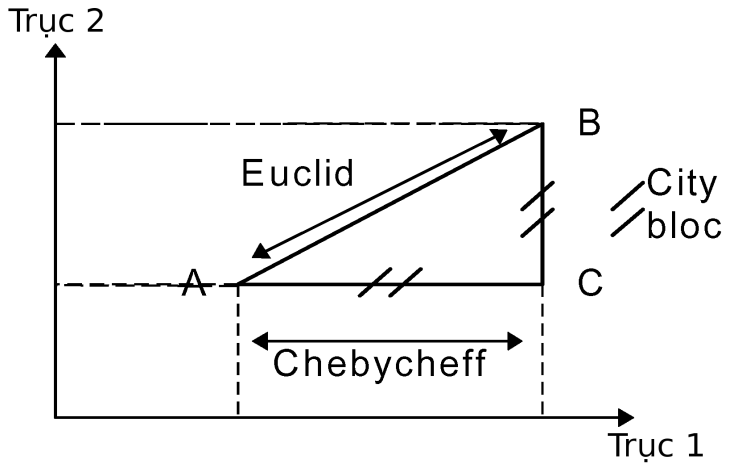
Vietnam Institute for Advanced Study in Mathematics

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)

Vietnam Institute for Advanced Study in Mathematics

Bẫy về số chiều

Curse of dimensionality

- Xét tỉ lệ thể tích giữa siêu cầu bán kính đơn vị với thể tích siêu khối hộp ngoại tiếp hình cầu ($\Rightarrow$ cạnh bằng 2 đơn vị)
  - Với số chiều $p = 1$, siêu cầu và siêu khối là đoạn thẳng đơn vị, tỉ lệ thể tích 1/1 = 1.
  - $p = 2$, có tỉ lệ $3.14/2^2 = 0.798$
  - $p = 3$, tỉ lệ $(4/3 \times 3.14)/2^3 = 0.52$
  - $p \to \infty$, tỉ lệ thể tích $\to 0$

Với số chiều lớn, khoảng cách giữa các điểm trở nên bằng nhau

Vietnam Institute for Advanced Study in Mathematics

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):

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

Nội dung

Vietnam Institute for Advanced Study in Mathematics

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

Vietnam Institute for Advanced Study in Mathematics

Độ đo khoảng cách/độ tương tự cho dữ liệu 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

Vietnam Institute for Advanced Study in Mathematics

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:

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

Vietnam Institute for Advanced Study in Mathematics

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Các phương pháp phân cụm

Các thuật toán chính

- Phân hoạch trực tiếp
- Cây phân cấp: số cụm được xác định sau bằng cách cắt cây phân cấp tại 1 mức nào đó.
- Dựa vào đồ thị

Phân hoạch trực tiếp

- 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 dựa vào mô hình xác suất (trộn phân bố)
- Các phương pháp dựa vào mật độ (DBSCAN, OPTICS, DENCLUE)
- Phương pháp nơ-ron (sơ đồ Kohonen)

Cách thực hiện nói chung

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?

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Tạo phân hoạch trực tiếp: *k*-means và các phương pháp mở rộng

- Thực hiện theo 1 tiêu chuẩn đã cho nhằm tìm một cách phân nhóm tốt nhất các cá thể thành **1 số lớp định trước** (rời nhau).
- Nguyên lí của các phương pháp này: xây dựng *k* nhóm (*k* là số được chọn trước) từ *n* cá thể dựa vào 1 thuật toán lặp « Định vị tâm/phân nhóm » nhằm tối ưu hoá 1 tiêu chuẩn đo chất lượng phân cụm.

Thuật toán *k*-means

- Mục tiêu: Phân hoạch dữ liệu thành *k* lớp bất kì khác rỗng
- Thuật toán cơ sở:
  - Chọn ngẫu nhiên *k* điểm làm tâm của *k* cụm
  - **repeat**
    - Tạo *k* cụm bằng cách xếp mỗi cá thể vào cụm có tâm gần nó nhất
    - Xác định tâm mới của mỗi cụm
  - **until** các tâm cụm không thay đổi

Vietnam Institute for Advanced Study in Mathematics

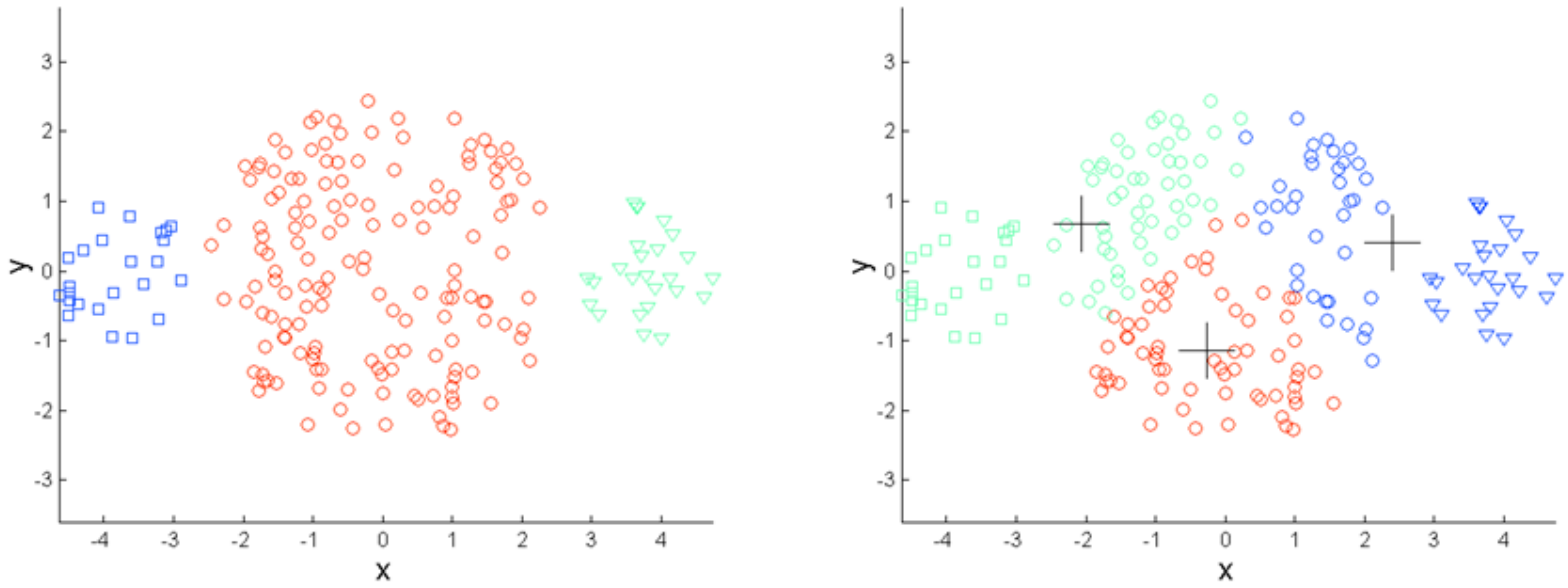
Thuật toán *k*-means

- Các tâm cụm ban đầu được chọn ngẫu nhiên: các cụm sinh ra thay đổi giữa các lần chạy
- Tâm cụm thường là trung bình các điểm trong cụm
- Độ đo khoảng cách có thể là khoảng cách Euclid, độ tương tự cosin, độ tương quan ...
- Thuật toán *k*-means hội tụ thường sau một số ít vòng lặp đầu tiên: điều kiện dừng thường được đổi thành "cho đến khi có tương đối ít điểm thay đổi cụm"
- Độ phức tạp $O(n * k * I * d)$, n là số điểm (cá thể), k là số cụm, I là số bước lặp, d là số chiều dữ liệu (số thuộc tính).

Hạn chế của phương pháp k-means

- k-means có vấn đề khi các cụm khác nhau về kích thước, mật độ và hình dạng không phải hình cầu
- k-means cũng gặp vấn đề khác là khi dữ liệu có chứa ngoại lệ.
- Cách giải quyết? Tăng k, hậu xử lí gộp các cụm

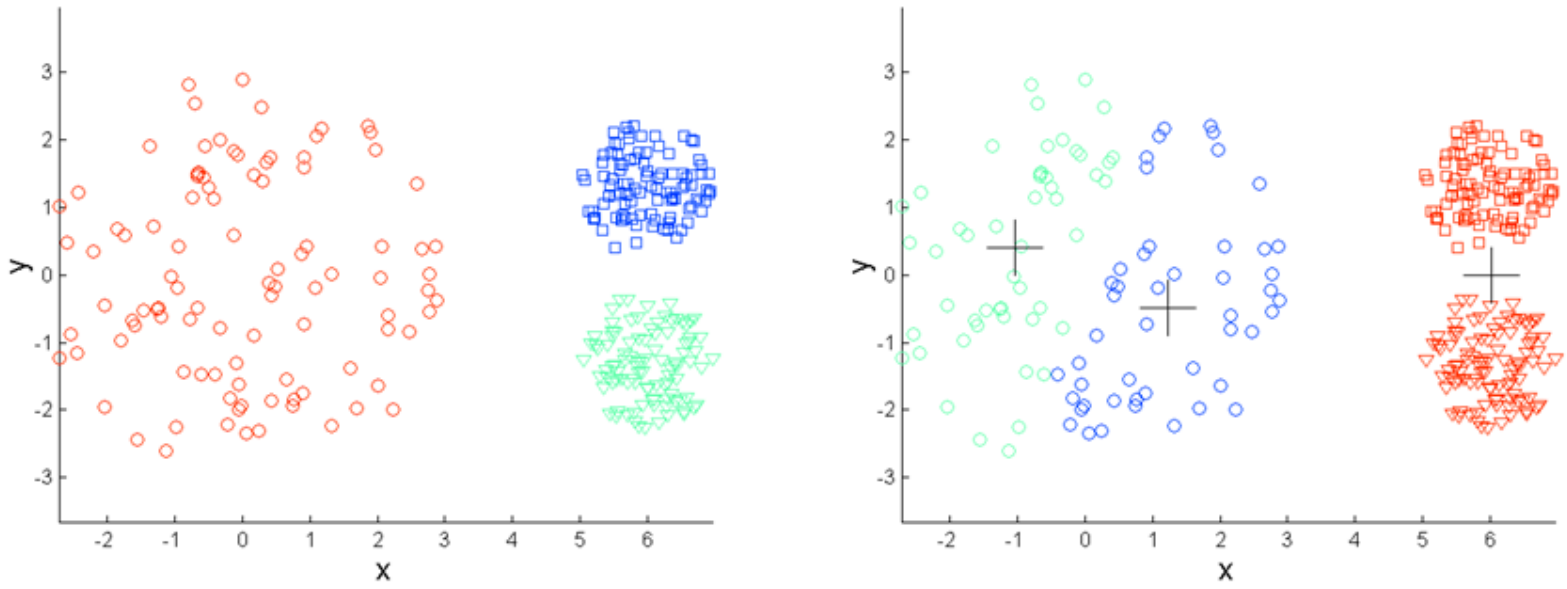
Hạn chế của phương pháp k-means - kích thước cụm



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

Vietnam Institute for Advanced Study in Mathematics

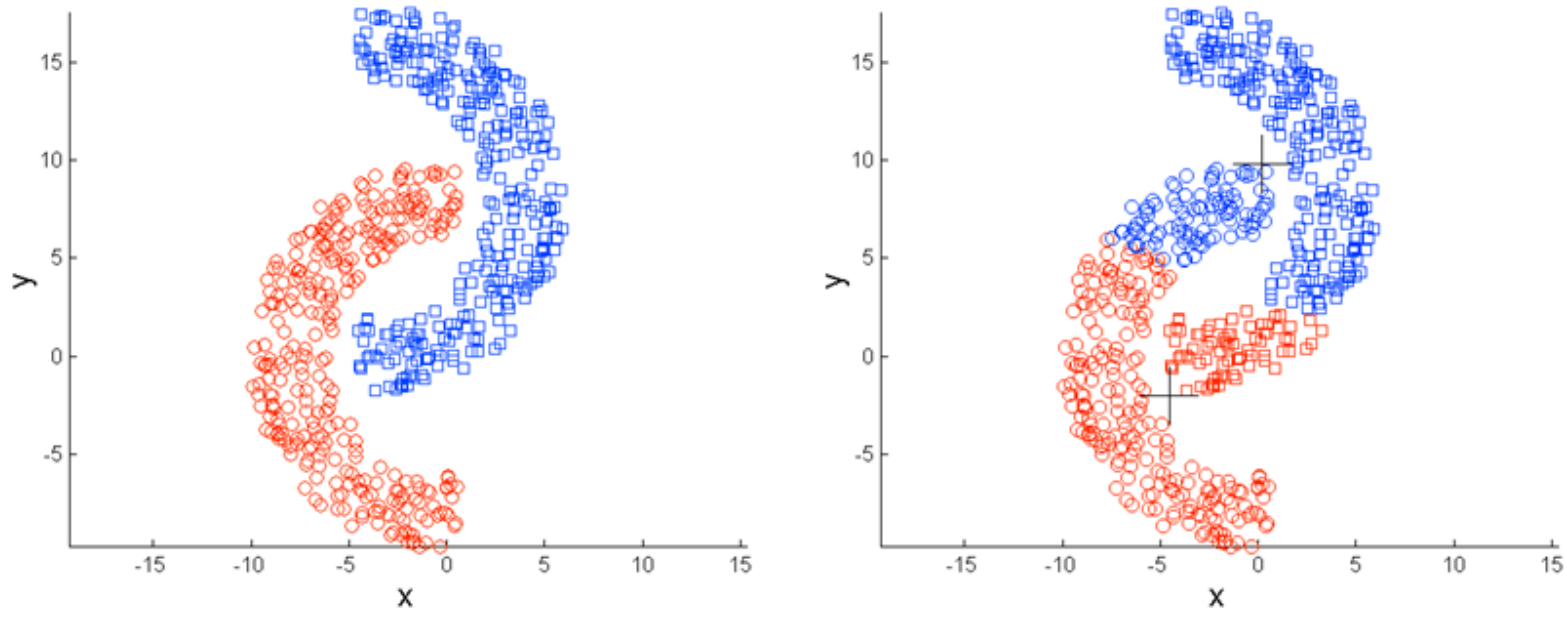
Hạn chế của phương pháp k-means - mật độ cụm



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

Vietnam Institute for Advanced Study in Mathematics

Hạn chế của phương pháp k-means - hình dạng cụm



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

Vietnam Institute for Advanced Study in Mathematics

Vấn đề khởi tạo các tâm cụm

- Số cách chọn tâm cụm ngẫu nhiên là rất lớn, cách chọn tâm cụm ban đầu đóng vai trò quan trọng, ảnh hưởng tới kết quả
- 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

k-means và vấn đề sinh cụm rỗng

- Thuật toán *k*-means cơ bản có thể dẫn tới cụm rỗng
- Thuật toán *k*-means mở rộng thực hiện việc cập nhật tâm cụm sau mỗi bước gán cá thể vào cụm
  Tốn thời gian hơn tuy nhiên đảm bảo các cụm luôn khác rỗng

k-means: tiền xử lí và hậu xử lí

- Tiền xử lí
  - Chuẩn hoá dữ liệu
  - Loại bỏ các ngoại lệ (*outliers*)
- Hậu xử lí
  - Loại bỏ các cụm nhỏ có thể tương ứng với dữ liệu ngoại lệ
  - Chia nhỏ các cụm "lỏng", tức là các cụm có tổng phương sai lớn
  - Trộn các cụm gần nhau và có phương sai nhỏ.

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Các phương pháp phân cấp

2 lớp rời nhau hoặc lớp này chứa lớp kia

- Phân cụm từ trên xuống
- Phân cụm từ dưới lên
- Các phương pháp hỗn hợp: ví dụ phân cụm mờ (*fuzzy clustering*)

Thuật toán phân cụm từ dưới lên

- Bước đầu: Tính độ tương tự giữa các đối tượng từng đôi một
- Đầu vào: $n(n-1)/2$ độ tương tự
- Lặp cho đến khi ghép được tất cả các đối tượng thành 1 nhóm: $n-1$ bước
  1. Ghép 2 đối tượng/cụm gần nhau nhất (sử dụng tiêu chuẩn kết nhập)
  2. Tính lại các tiêu chuẩn đo độ tương tự giữa các thành phần (các nhóm hay các đối tượng riêng lẻ còn lại)
- Ta thu được 1 cây có chỉ số độ tương tự
- Trong loại này có các phương pháp: Cure, Rock, Birch, Cameleon

Các tiêu chuẩn kết nhập (1/2)

- Khi gộp nhóm 2 hay nhiều các thể, cần tính lại khoảng cách giữa nhóm này và các đối tượng khác.
- Để làm điều đó, cần có tiêu chuẩn kết nhập.

Các tiêu chuẩn kết nhập (2/2)

- Kết nối gần nhất

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

Vietnam Institute for Advanced Study in Mathematics

Các tiêu chuẩn kết nhập (2/2)

- Kết nối gần nhất

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Đường kính hợp 2 tập

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

Vietnam Institute for Advanced Study in Mathematics

Các tiêu chuẩn kết nhập (2/2)

- Kết nối gần nhất

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Đường kính hợp 2 tập

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Tiêu chuẩn trung bình (liên kết trung bình)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

Vietnam Institute for Advanced Study in Mathematics

Các tiêu chuẩn kết nhập (2/2)

- Kết nối gần nhất

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Đường kính hợp 2 tập

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Tiêu chuẩn trung bình (liên kết trung bình)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Tiêu chuẩn Ward (độ tăng của phương sai)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$

VIASM
Vietnam Institute for Advanced Study in Mathematics

Các tiêu chuẩn kết nhập (2/2)

- Kết nối gần nhất

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Đường kính hợp 2 tập

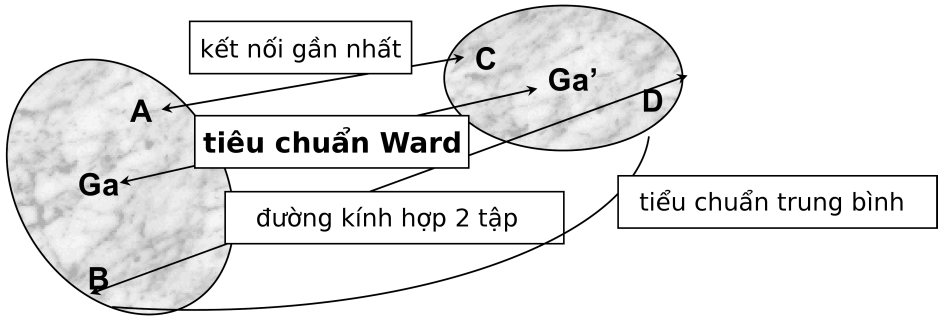
$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Tiêu chuẩn trung bình (liên kết trung bình)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

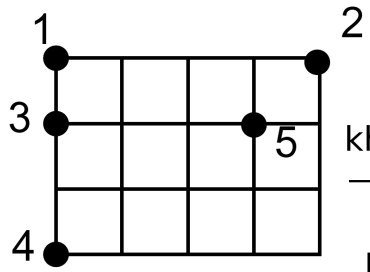
- Tiêu chuẩn Ward (độ tăng của phương sai)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$



Vietnam Institute for Advanced Study in Mathematics

Ví dụ áp dụng phân cụm (1/2)



khoảng cách Euclid bình phương

	1	2	3	4	5
1	0	16	**1**	9	10
2		0	17	25	2
3			0	4	9
4				0	13
5					0

Kết nhập 1 và 3 thành nút 6
Tiêu chuẩn kết nhập: **kết nối gần nhất**

$$nv(6,\{2\}) = inf\{d(\{1\},\{2\}), d(\{3\},\{2\})\} = inf\{16,17\} = 16$$
$$nv(6,\{4\}) = inf\{d(\{1\},\{4\}), d(\{3\},\{4\})\} = inf\{9,4\} = 4$$
$$nv(6,\{5\}) = inf\{d(\{1\},\{5\}), d(\{3\},\{5\})\} = inf\{10,9\} = 9$$

	2	4	5	6
2	0	25	**2**	16
4		0	13	4
5			0	9
6				0

bước tiếp theo kết nhập 2 và 5 thành nút 7

$$nv(7,\{4\}) = inf\{d(\{2\},\{4\}), d(\{5\},\{4\})\} = inf\{25,13\} = 13$$
$$nv(7,6) = inf\{d(\{2\},\{1\}), d(\{5\},\{1\}), d(\{2\},\{3\}), d(\{5\},\{3\})\} = inf\{16,10,17,9\} = 9$$

	4	6	7
4	0	**4**	13
6		0	9
7			0

bước tiếp theo kết nhập 4 và 6 thành nút 8

$$nv(7,8) = inf\begin{Bmatrix} d(\{2\},\{1\}), d(\{5\},\{1\}), \\ d(\{2\},\{3\}), d(\{5\},\{3\}), \\ d(\{2\},\{4\}), d(\{5\},\{4\}) \end{Bmatrix} = inf\begin{Bmatrix} 16,10, \\ 17,9, \\ 25,13 \end{Bmatrix} = 9$$

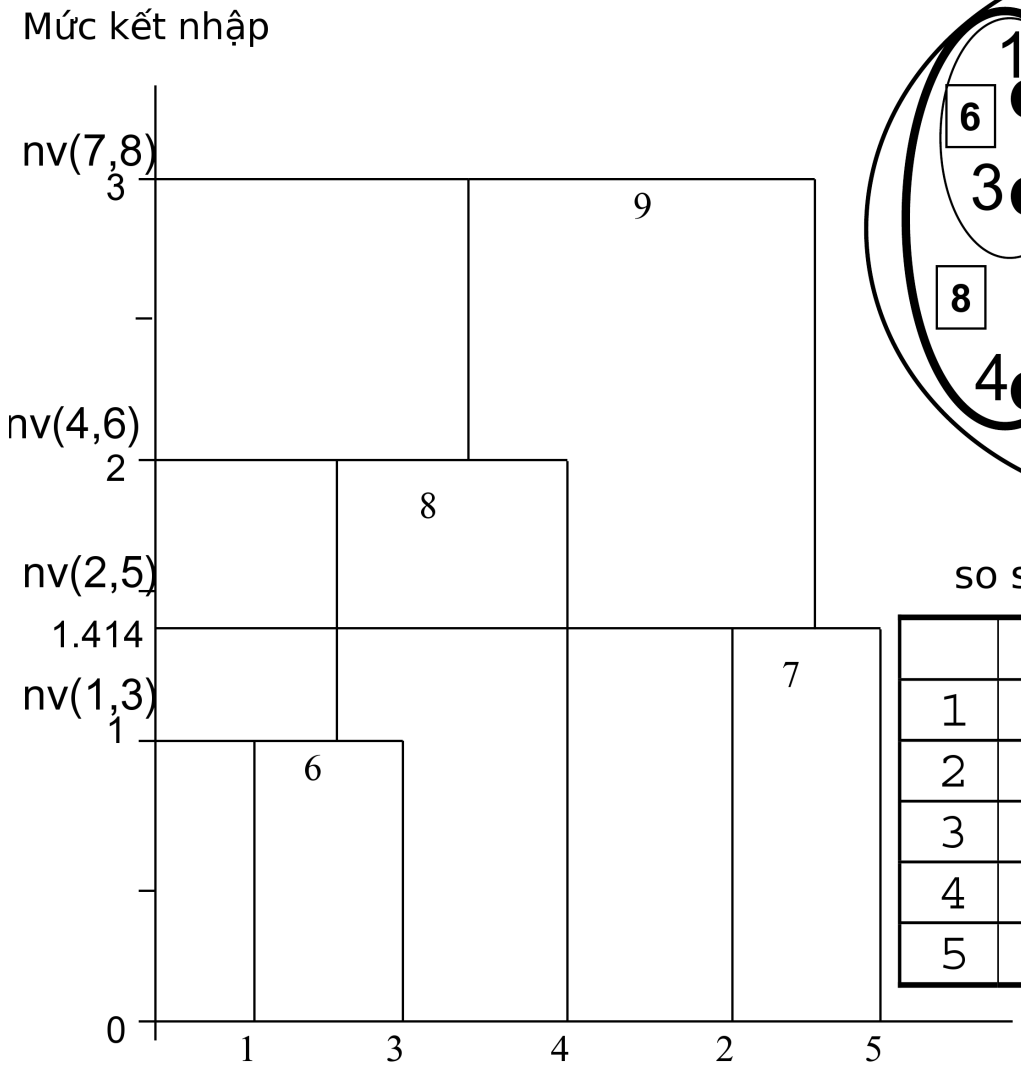
	7	8
7	0	9
8		0

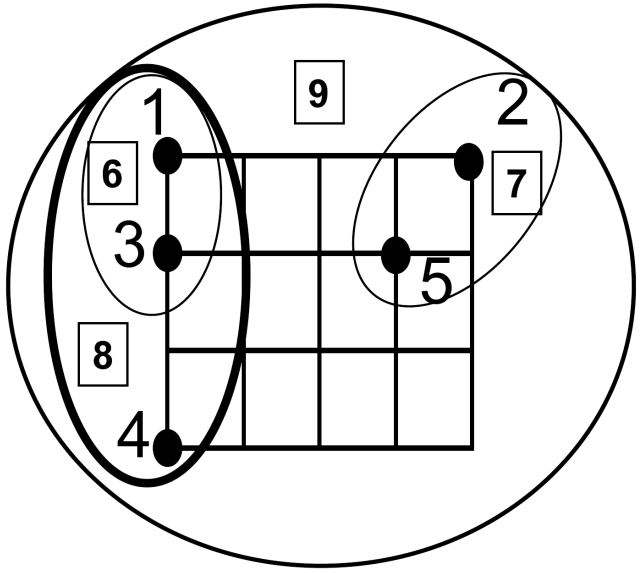
bước sau kết nhập 7 và 8 thành nút 9

Kết thúc phân cụm

Vietnam Institute for Advanced Study in Mathematics

Ví dụ áp dụng phân cụm (2/2)





so sánh các khoảng cách

	1	2	3	4	5
1	0	16/9	1/1	9/4	10/9
2		0	17/9	25/9	2/2
3			0	4/4	9/9
4				0	13/9
5					0

Vietnam Institute for Advanced Study in Mathematics

Tệp dữ liệu lớn trong các không gian có số chiều lớn

- Thích nghi thuật toán k-means
- BIRCH - Balanced Iterative Reducing and Clustering using Hierachies
  `http://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf`
- PDDP - Principal Direction Divisive Partitioning
- DBSCAN - Density-Based Spatial Clustering of Applications with Noise

Nội dung

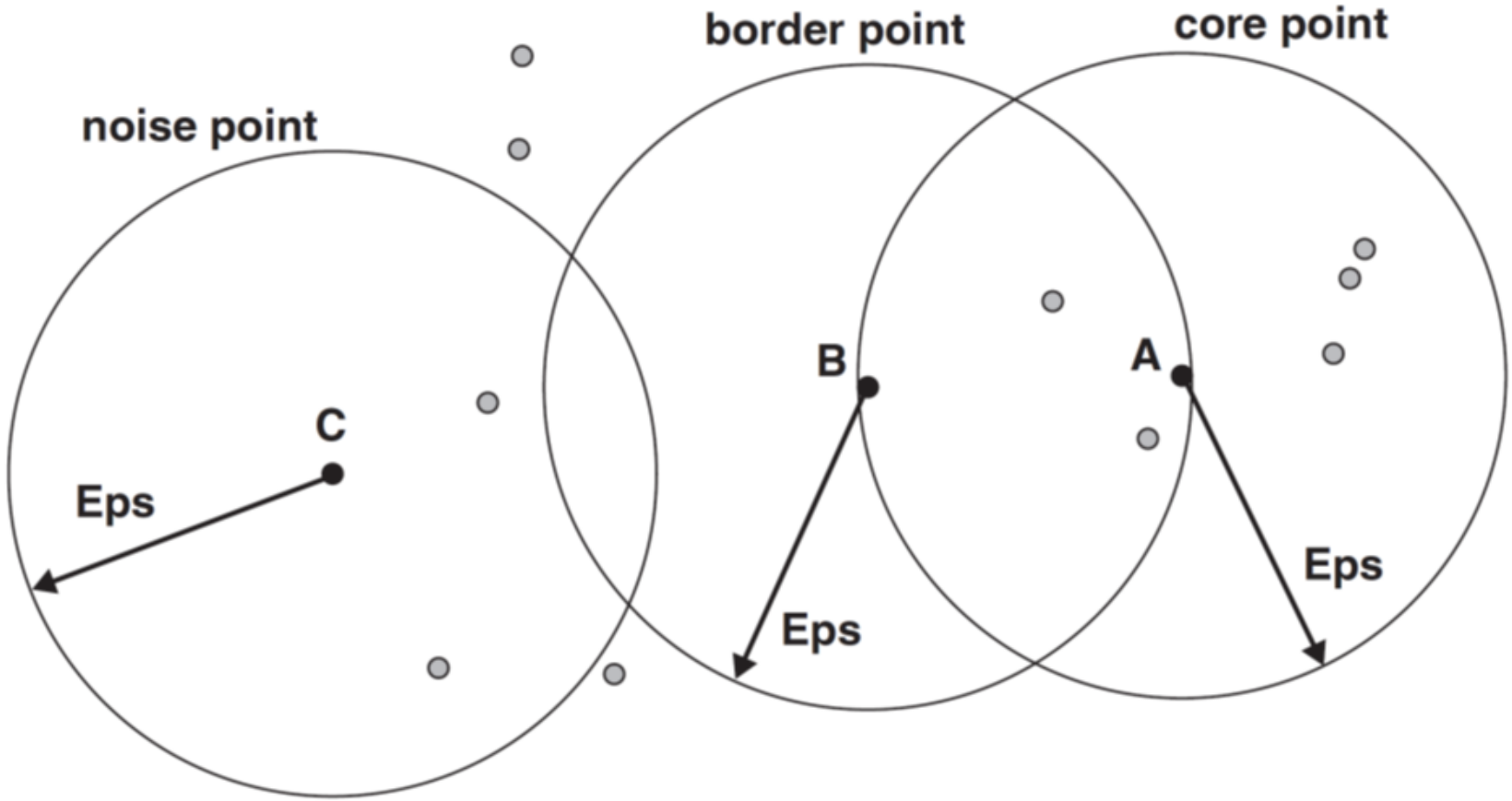
Vietnam Institute for Advanced Study in Mathematics

DBSCAN

- Mật độ = số điểm nằm trong một bán kính (*Eps*) xác định
- Một điểm là điểm lõi (*core point*) nếu nó có ít nhất *MinPts* điểm trong bán kính *Eps*
  - Đây là các điểm nằm trong một cụm
  - Điểm đang xét cũng tính trong cụm
- Một điểm biên (*border point*) là điểm không phải lõi, nhưng nằm trong lân cận điểm lõi
- Các điểm khác (không phải là điểm lõi hay biên) còn lại gọi là điểm nhiễu (*noise point*)

DBSCAN - các loại điểm lõi, biên và nhiễu

$MinPts = 7$



Vietnam Institute for Advanced Study in Mathematics

DBSCAN - thuật toán

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

1: *current_cluster_label* $\leftarrow$ 1
2: **for** all core points **do**
3: **if** the core point has no cluster label **then**
4: *current_cluster_label* $\leftarrow$ *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**

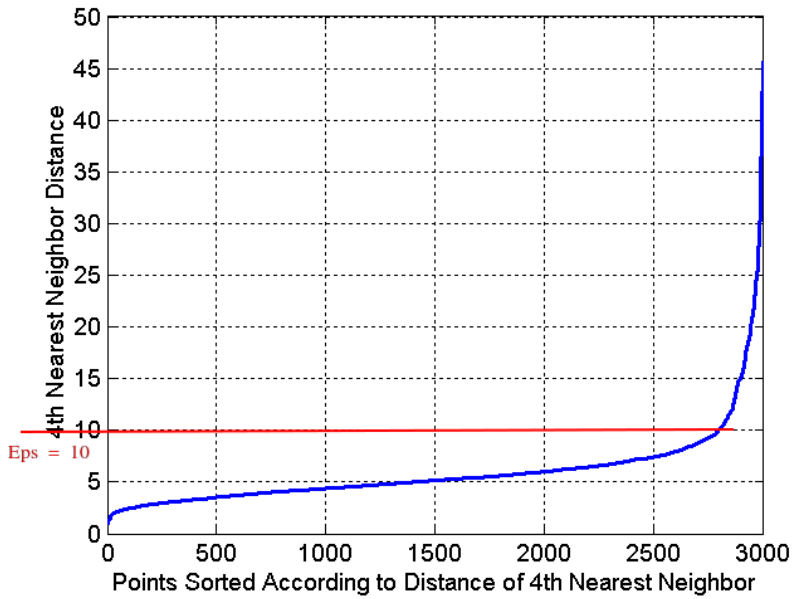
Vietnam Institute for Advanced Study in Mathematics

DBSCAN

- Ưu điểm
  - Chống được nhiễu
  - Có thể xử lí các cụm có hình dạng và kích thước khác nhau
- Nhược điểm
  - Không tốt khi mật độ có nhiều biến động
  - Không tốt khi số chiều lớn

DBSCAN: Xác định *Eps* và *MinPts*

- Ý tưởng là với các điểm ở trong cùng một cụm, khoảng cách tới điểm gần nhất thứ *k* là gần giống nhau
- Các điểm nhiễu có hàng xóm gần nhất thứ *k* ở khoảng cách xa hơn
- Xác định *Eps*: Vẽ đồ thị trên đó khoảng cách từ từng điểm tới điểm gần nhất thứ *k* tới các điểm được sắp xếp tăng dần. *Eps* được chọn ở vị trí gãy của đường cong



Vietnam Institute for Advanced Study in Mathematics

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Hai kiểu phân cụm

- Phân cụm cứng (*hard clustering*)
  - Mỗi điểm dữ liệu được gán vào một cụm cụ thể
- Phân cụm mềm (*soft clustering*)
  - Mỗi điểm dữ liệu có thể được gắn vào nhiều cụm với trọng số/xác suất tương ứng, tổng trọng số/xác suất bằng 1.

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Phân cụm mờ

- Phân k cụm, tổng quát hoá hàm mục tiêu

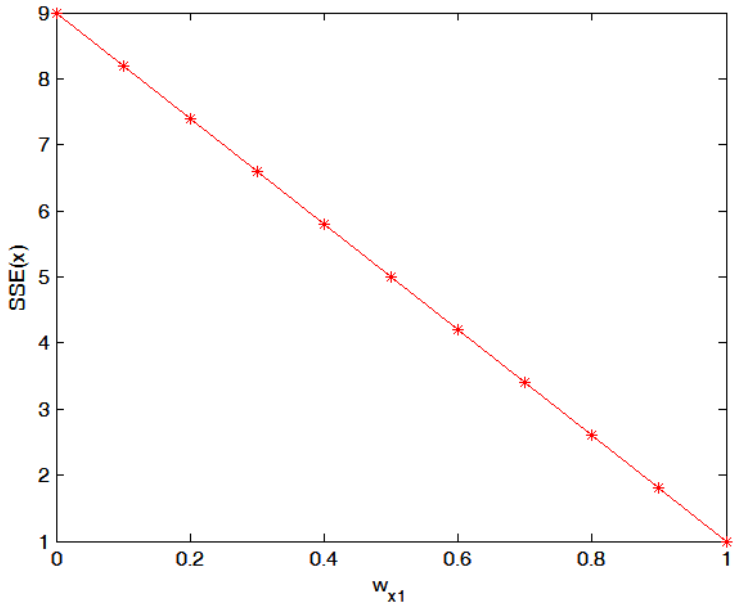
$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{m} w_{ij} dist(x_i, c_j)^2 \qquad \sum_{j=1}^{k} w_{ij} = 1$$

  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
$w_{x1} = 1, w_{x2} = 0$



Vietnam Institute for Advanced Study in Mathematics

Fuzzy C-means

- Hàm mục tiêu

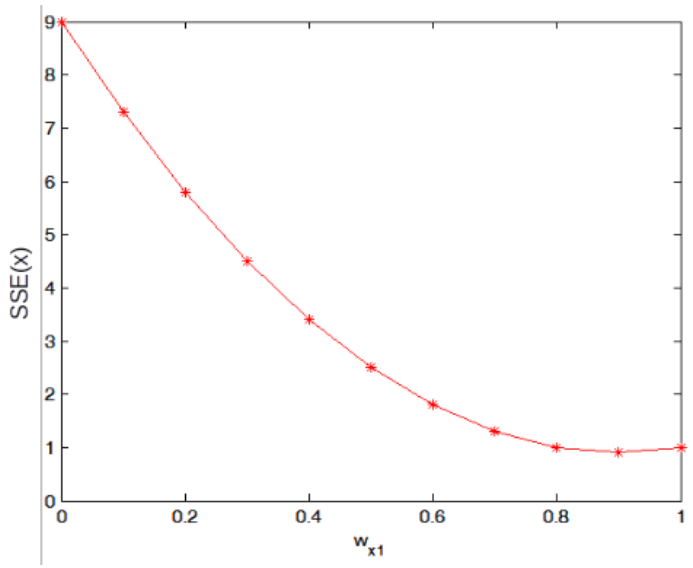
$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{m} w_{ij}^{p} dist(x_i, c_j)^2 \qquad \sum_{j=1}^{k} w_{ij} = 1$$

  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 $w_{ij} \in [0, 1]$

Vietnam Institute for Advanced Study in Mathematics

Fuzzy C-means: 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
$w_{x1} = 0.9, w_{x2} = 0.1$

VIASM
Vietnam Institute for Advanced Study in Mathematics

Fuzzy C-means

- 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

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

Nội dung

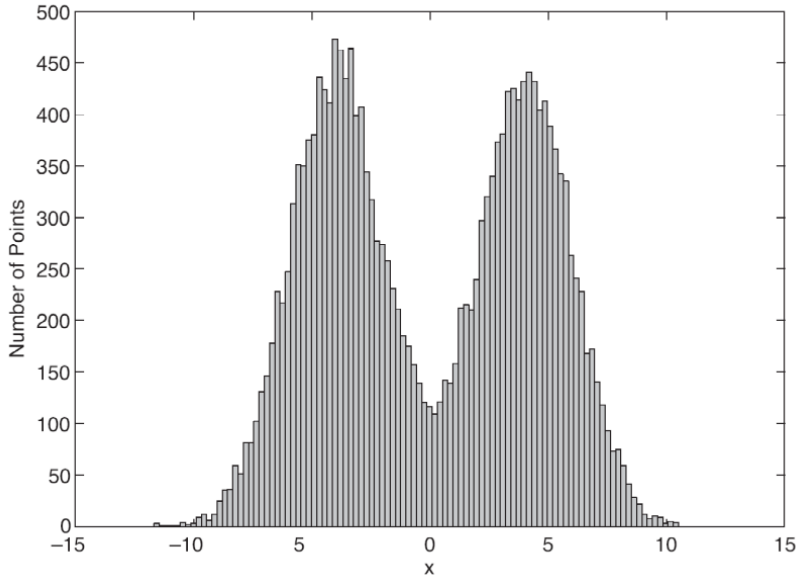
Vietnam Institute for Advanced Study in Mathematics

Phân cụm 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

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

Vietnam Institute for Advanced Study in Mathematics

Phân cụm xác suất: Thuật toán 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

Vietnam Institute for Advanced Study in Mathematics

Phân cụm xác suất: 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

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

EM: Thuật toán chi tiết hơn

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)

Các vấn đề với EM

- 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 $O(d^2)$, 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 $O(d)$

Nội dung

Vietnam Institute for Advanced Study in Mathematics

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

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

Tiêu chuẩn/độ đo đánh giá chất lượng phân cụm

- 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

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'

Vietnam Institute for Advanced Study in Mathematics

Độ đo đánh giá trong: Tính gắn kết và tính 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

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

Vietnam Institute for Advanced Study in Mathematics

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

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

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Tìm đặc trưng các lớp

- **Giúp cho việc diễn giải 1 phân hoạch:**
  1 phân hoạch sẽ được nâng giá trị 1 cách đáng kể nếu nó đi kèm với một mô tả về các lớp theo các thuộc tính và các cá thể.

Diễn giải phân hoạch theo các 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.

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.

Vietnam Institute for Advanced Study in Mathematics

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 " $\Rightarrow n_{kj}/n_j$

Vietnam Institute for Advanced Study in Mathematics

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.

Giá trị kiểm tra cho các 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

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

Vietnam Institute for Advanced Study in Mathematics

Giá trị kiểm tra cho các biến tên

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

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

Vietnam Institute for Advanced Study in Mathematics

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.

Nội dung

Vietnam Institute for Advanced Study in Mathematics

Phân cụm trong thực hành (1/2)

- 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.

VIASM
Vietnam Institute for Advanced Study in Mathematics

Phân cụm trong thực hành (2/2)

- 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.

VIASM
Vietnam Institute for Advanced Study in Mathematics

Phân cụm trên các biến

- Phương pháp phân cụm dựa trên các cá thể nhằm mục đích ghép nhóm các cá thể này thành một số hạn chế các lớp tiêu biểu là phương pháp hay được dùng nhất. Tuy nhiên người ta cũng có thể thực hiện phân cụm trên các biến (sau khi chuyển vị tệp dữ liệu) nhằm giảm số biến và cũng có thể nghiên cứu về sự dư thừa biến.

Bài tập thực hành

- Phân tích cụm trên một số bộ dữ liệu: vertebrate, cars, golf
- Một số công cụ có thể sử dụng
  - 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`

Vietnam Institute for Advanced Study in Mathematics

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
- 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