Học Máy
(IT 4862)

Nguyễn Nhật Quang

quangnn-fit@mail.hut.edu.vn

Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ thông tin và truyền thông

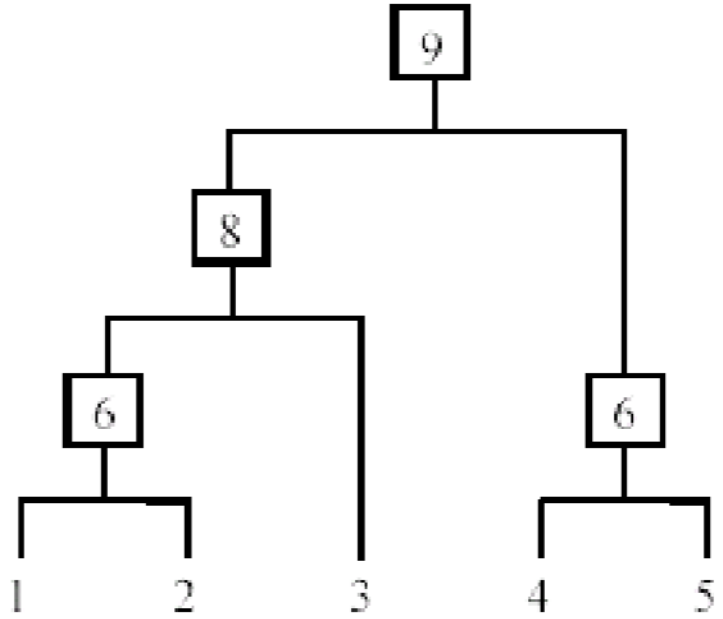
Năm học 2011-2012

Nội dung môn học:

- Giới thiệu chung
- Đánh giá hiệu năng hệ thống học máy
- Các phương pháp học dựa trên xác suất
- Các phương pháp học có giám sát
- **Các phương pháp học không giám sát**
  - **Phân cụm dựa trên tích tụ phân cấp: HAC (Hierarchical agglomerative clustering)**
- Lọc cộng tác
- Học tăng cường

HAC (1)

- Sinh ra một chuỗi lồng nhau của các cụm, được gọi là **dendrogram**
  - Cũng được gọi là một phân loại (*taxonomy*)/phân cấp (*hierarchy*)/cây (*tree*) của các ví dụ

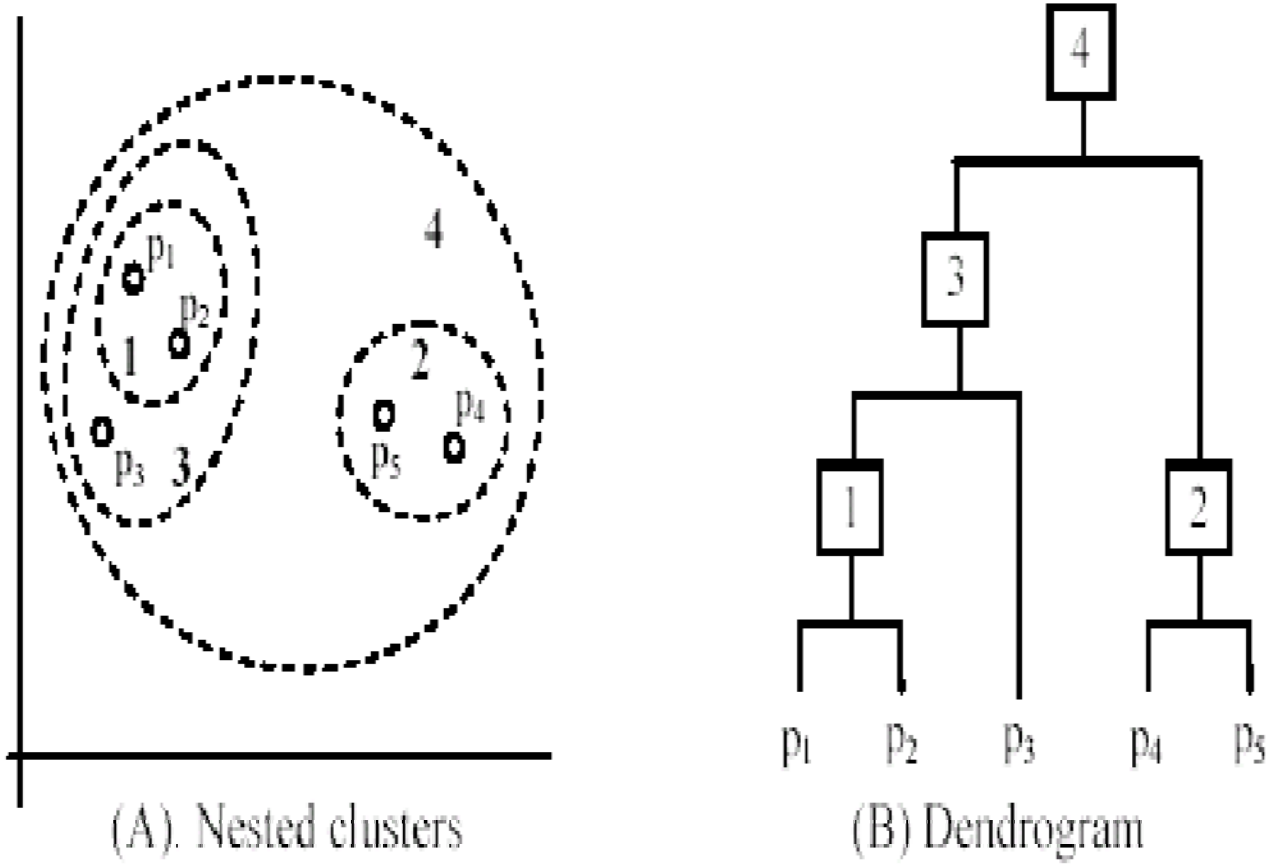


[Liu, 2006]

HAC (2)

- Phân cụm dựa trên tích tụ phân cấp (Hierarchical Agglomerative Clustering – HAC) sẽ xây dựng dendrogram từ mức đáy (cuối) dần lên (bottom-up)
- Giải thuật HAC
  - Bắt đầu, mỗi ví dụ chính là một cụm (là một nút trong dendrogram)
  - Hợp nhất 2 cụm có mức độ tương tự (gần) nhau nhất
    - Cặp gồm 2 cụm có khoảng cách nhỏ nhất trong số các cặp cụm
  - Tiếp tục quá trình hợp nhất
  - Giải thuật kết thúc khi tất cả các ví dụ được hợp nhất thành một cụm duy nhất (là nút gốc trong dendrogram)

HAC – Ví dụ



(Venn diagram)

[Liu, 2006]

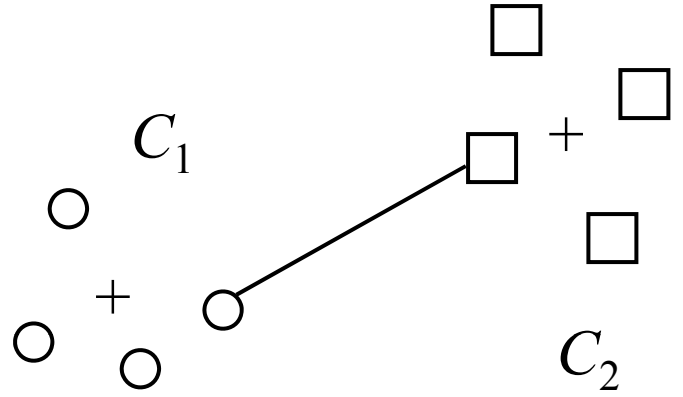
Khoảng cách giữa 2 cụm

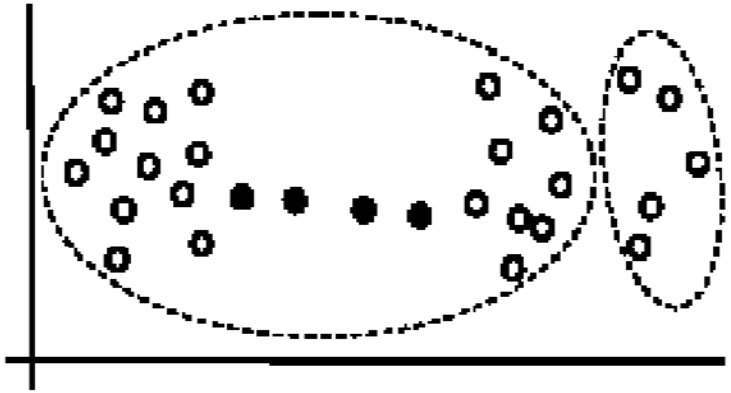
- Giải thuật HAC cần định nghĩa việc tính toán khoảng cách giữa 2 cụm
  - Trước khi hợp nhất, cần tính khoảng cách giữa mỗi cặp 2 cụm có thể
- Có nhiều phương pháp để đánh giá khoảng cách giữa 2 cụm – đưa đến các biến thể khác nhau của giải thuật HAC
  - Liên kết đơn (Single link)
  - Liên kết hoàn toàn (Complete link)
  - Liên kết trung bình (Average link)
  - Liên kết trung tâm (Centroid link)
  - …

HAC – Liên kết đơn

HAC liên kết đơn (Single link):

- Khoảng cách giữa 2 cụm là **khoảng cách nhỏ nhất** giữa các ví dụ (các thành viên) của 2 cụm đó
- Có xu hướng sinh ra các cụm có dạng "chuỗi dài" (long chain)



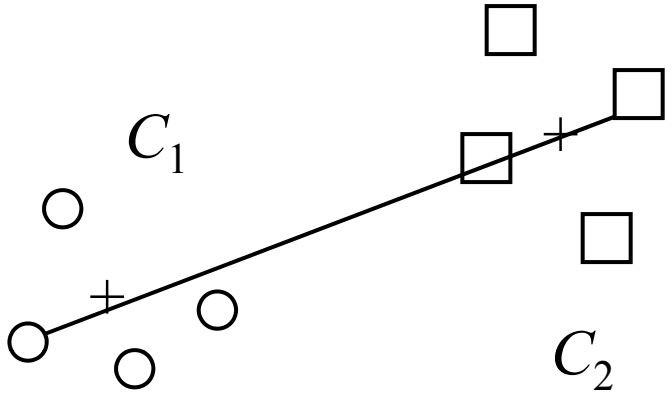


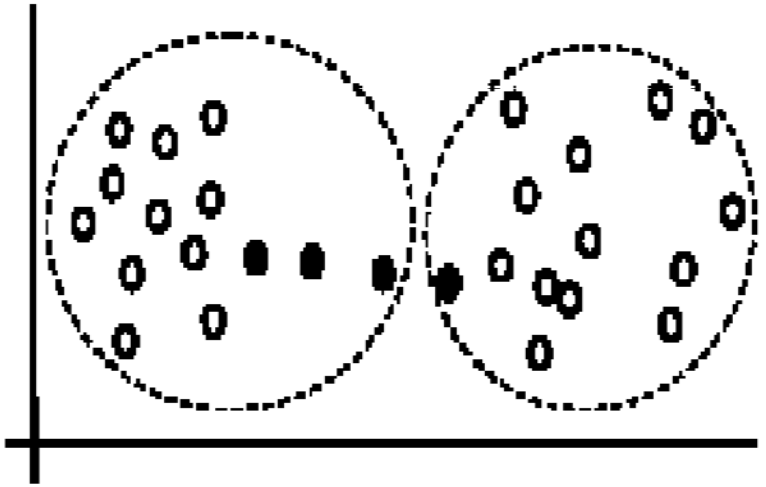
[Liu, 2006]

HAC – Liên kết hoàn toàn

HAC liên kết hoàn toàn (Complete link):

- Khoảng cách giữa 2 cụm là **khoảng cách lớn nhất** giữa các ví dụ (các thành viên) của 2 cụm đó
- Nhạy cảm (gặp lỗi phân cụm) đối với các ngoại lai (outliers)
- Có xu hướng sinh ra các cụm có dạng "bụi cây" (clumps)





[Liu, 2006]

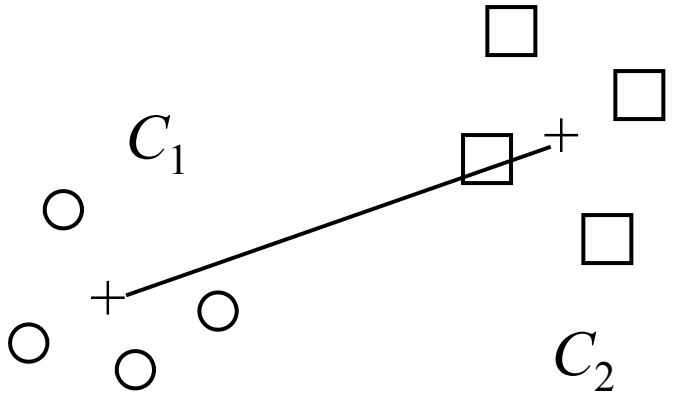
HAC – Liên kết trung bình

- Khoảng cách trong liên kết trung bình (Average-link) là sự thỏa hiệp giữa các khoảng cách trong liên kết hoàn toàn (Complete-link) và liên kết đơn (Single-link)
  - Để giảm mức độ nhạy cảm (khả năng lỗi) của phương pháp phân cụm dựa trên liên kết hoàn toàn đối với các ngoại lai (outliers)
  - Để giảm xu hướng sinh ra các cụm có dạng "chuỗi dài" của phương pháp phân cụm dựa trên liên kết đơn (dạng "chuỗi dài" không phù hợp với khái niệm tự nhiên của một cụm)
- Khoảng cách giữa 2 cụm là khoảng cách trung bình của tất cả các cặp ví dụ (mỗi ví dụ thuộc về một cụm)

HAC – Liên kết trung tâm

HAC liên kết trung tâm (Centroid link):

- Khoảng cách giữa 2 cụm là khoảng cách giữa 2 điểm trung tâm (centroids) của 2 cụm đó



Giải thuật HAC – Độ phức tạp

- Tất cả các biến thể của giải thuật HAC đều có độ phức tạp tối thiểu mức $O(r^2)$
  - r: Tổng số các ví dụ (kích thước của tập dữ liệu)
- Phương pháp phân cụm HAC liên kết đơn (Single-link) có độ phức tạp mức $O(r^2)$
- Các phương pháp phân cụm HAC liên kết hoàn toàn (Complete-link) và liên kết trung bình (Average-link) có độ phức tạp mức $O(r^2 \log r)$
- Do độ phức tạp cao, giải thuật HAC khó có thể áp dụng được đối với các tập dữ liệu có kích thước (rất) lớn

Các hàm khoảng cách

- Một thành phần quan trọng của các phương pháp phân cụm
  - Cần xác định các hàm tính độ khác biệt (dissimilarity/distance functions), hoặc các hàm tính độ tương tự (similarity functions)
- Các hàm tính khoảng cách khác nhau đối với
  - Các kiểu dữ liệu khác nhau
    - Dữ liệu kiểu số (Numeric data)
    - Dữ liệu kiểu định danh (Nominal data)
  - Các bài toán ứng dụng cụ thể

Hàm khoảng cách cho thuộc tính số

- Họ các hàm khoảng cách hình học (khoảng cách Minkowski)
- Các hàm được dùng phổ biến nhất
  - Khoảng cách Euclid
  - Khoảng cách Manhattan (khoảng cách City-block)
- Ký hiệu $d(\mathbf{x}_i, \mathbf{x}_j)$ là khoảng cách giữa 2 ví dụ (2 vectơ) $\mathbf{x}_i$ và $\mathbf{x}_j$
- Khoảng cách Minkowski (với p là một số nguyên dương)

$$d(\mathbf{x_i}, \mathbf{x_j}) = [(x_{i1} - x_{j1})^p + (x_{i2} - x_{j2})^p + ... + (x_{in} - x_{jn})^p]^{1/p}$$

Hàm k/c cho thuộc tính nhị phân

- Sử dụng một ma trận để biểu diễn hàm tính khoảng cách
  - *a*: Tổng số thuộc tính có giá trị là 1 trong cả $\mathbf{x}_i$ và $\mathbf{x}_j$
  - *b*: Tổng số các thuộc tính có giá trị là 1 trong $\mathbf{x}_i$ và có giá trị là 0 trong $\mathbf{x}_j$
  - *c*: Tổng số các thuộc tính có giá trị là 0 trong $\mathbf{x}_i$ và có giá trị là 1 trong $\mathbf{x}_j$
  - *d*: Tổng số các thuộc tính có giá trị là 0 trong cả $\mathbf{x}_i$ và $\mathbf{x}_j$

ví dụ $\mathbf{x}_i$ \ ví dụ $\mathbf{x}_j$	1	0
1	a	b
0	c	d

- **Hệ số phù hợp đơn giản (Simple matching coefficient)**. Tỷ lệ sai lệch giá trị của các thuộc tính giữa 2 ví dụ:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{b+c}{a+b+c+d}$$

Hàm k/c cho thuộc tính định danh

- Hàm khoảng cách cũng dựa trên phương pháp đánh giá tỷ lệ khác biệt giá trị thuộc tính giữa 2 ví dụ
- Với 2 ví dụ $\mathbf{x}_i$ và $\mathbf{x}_j$, ký hiệu p là tổng số các thuộc tính (trong tập dữ liệu), và q là số các thuộc tính mà giá trị là như nhau trong $\mathbf{x}_i$ và $\mathbf{x}_j$

$$d(\mathbf{x_i}, \mathbf{x_j}) = \frac{p - q}{p}$$

Tài liệu tham khảo

- B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.