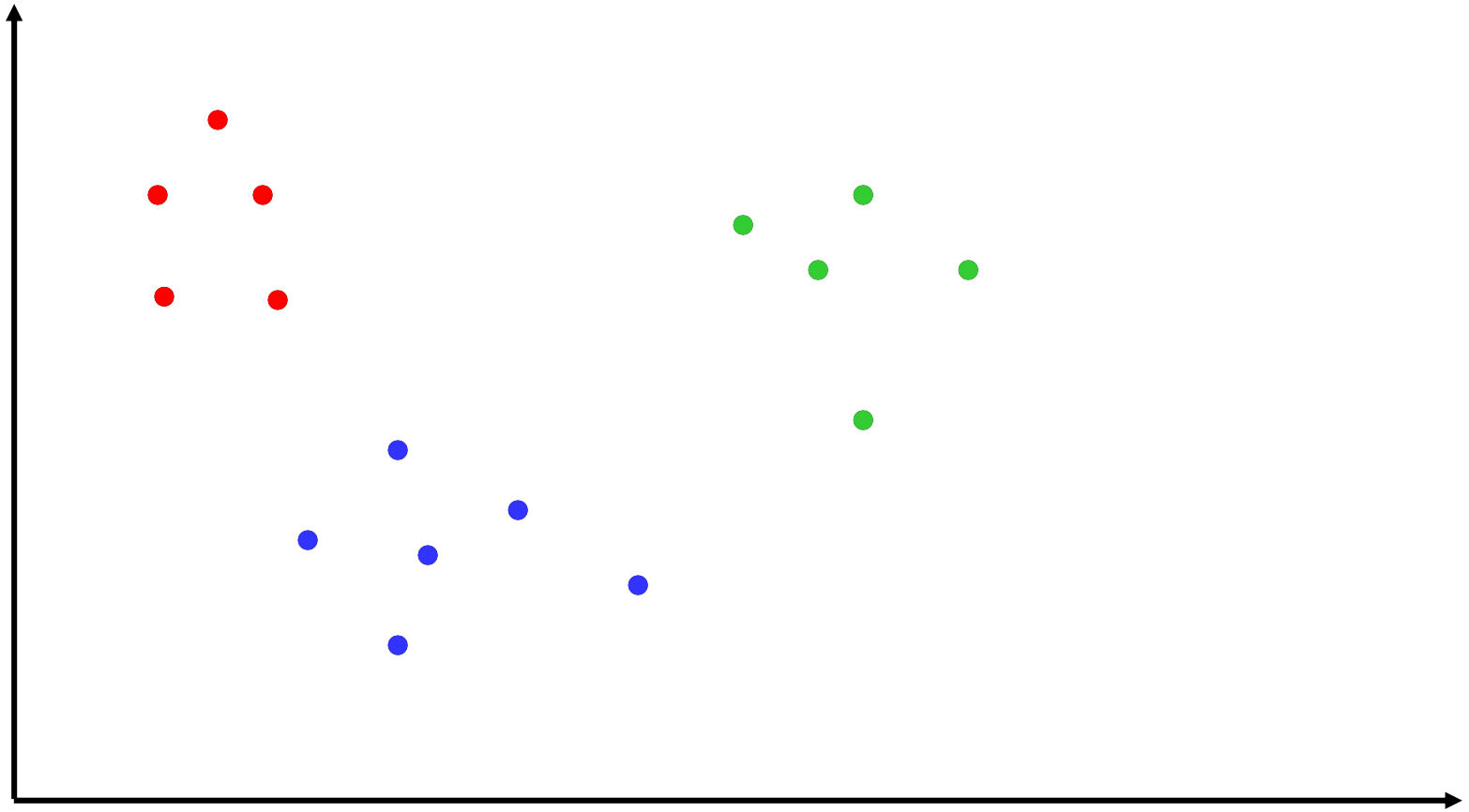
Phân cụm văn bản
(Text Clustering)

Nguyễn Mạnh Hiển
hiennm@tlu.edu.vn

Phân cụm (clustering)

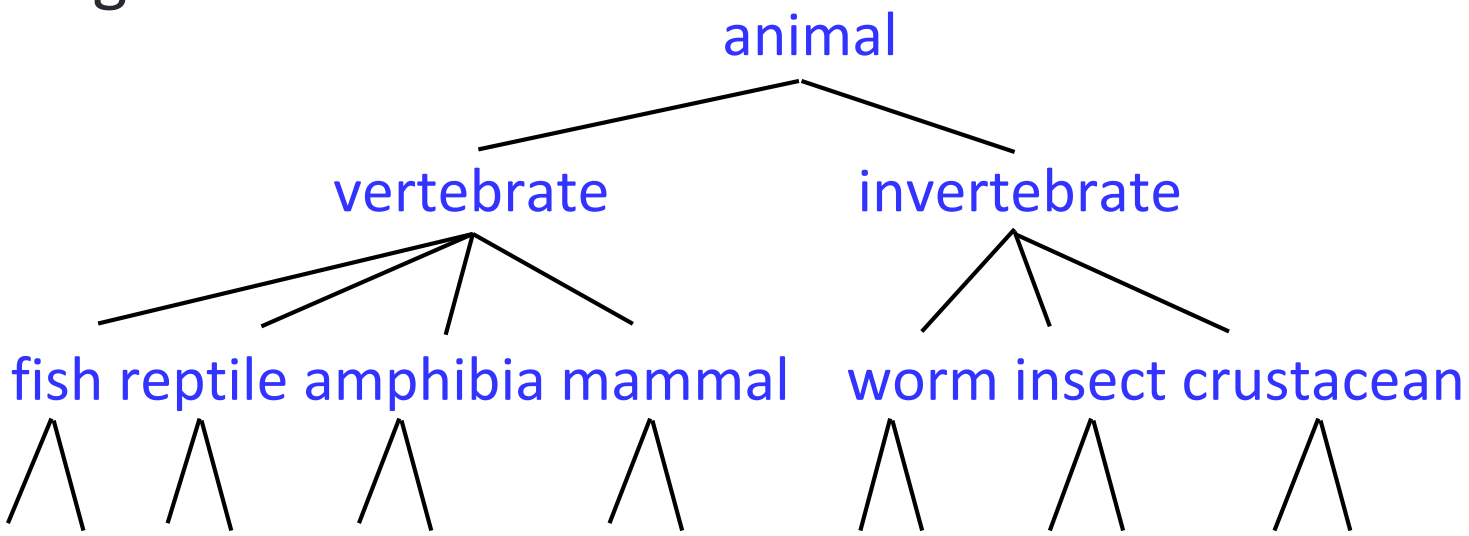
- Phân chia các mẫu không nhãn vào các tập con rời nhau, gọi là các ***cụm*** (cluster), sao cho:
  - Các mẫu trong cùng một cụm rất giống nhau.
  - Các mẫu khác cụm rất khác nhau.
- Phát hiện các lớp mới theo kiểu ***không giám sát*** (unsupervised), trong đó không có sẵn nhãn lớp cho các mẫu.

Ví dụ phân cụm



Phân cụm phân cấp (hierarchical clustering)

- Xây dựng cây phân cụm (dendogram) từ một tập mẫu không nhãn.



- Áp dụng đệ quy một thuật toán phân cụm chuẩn để tạo ra sự phân cụm theo kiểu phân cấp.

Phân cụm tích tụ (agglomerative) và phân cụm phân chia (divisive)

- Phương pháp tích tụ (từ dưới lên) bắt đầu bằng cách coi mỗi mẫu đơn lẻ là một cụm riêng, và lặp đi lặp lại việc kết hợp hai cụm thành những cụm ngày càng lớn hơn.
- Phương pháp phân chia (từ trên xuống) bắt đầu bằng cách coi tất cả các mẫu lập thành một cụm duy nhất, và lặp đi lặp lại việc tách một cụm thành hai cụm rời nhau.

Phân cụm tích tụ phân cấp (Hierarchical Agglomerative Clustering – HAC)

- Giả sử đã có hàm tính độ tương tự giữa hai mẫu.
- Bắt đầu với mỗi mẫu lập thành một cụm riêng biệt, sau đó lặp đi lặp lại việc hợp hai cụm giống nhau nhất cho đến khi chỉ còn lại một cụm duy nhất.
- Lịch sử hợp các cụm sẽ lập thành một cây nhị phân biểu diễn sự phân cấp của các cụm.

Thuật toán HAC

1. Bắt đầu với mỗi mẫu lập thành một cụm riêng.
2. Lặp cho đến khi chỉ còn lại một cụm:
    2.1. Trong số các cụm hiện tại, tìm hai cụm c_i và c_j giống nhau nhất.
    2.2. Thay c_i và c_j bằng cụm $c_i \cup c_j$

Độ tương tự giữa các cụm

- Gọi sim(x, y) là độ tương tự giữa hai mẫu x và y.
  - Trong trường hợp x và y là các véctơ văn bản, ta có thể dùng độ đo côsin.
- Cách tính độ tương tự giữa hai cụm:
  - Liên kết đơn (single link): Tính bằng độ tương tự giữa hai mẫu giống nhau nhất (mỗi mẫu thuộc một cụm khác nhau).
  - Liên kết đầy đủ (complete link): Tính bằng độ tương tự giữa hai mẫu khác nhau nhất.
  - Liên kết trung bình (average link): Tính bằng độ tương tự trung bình giữa hai mẫu.

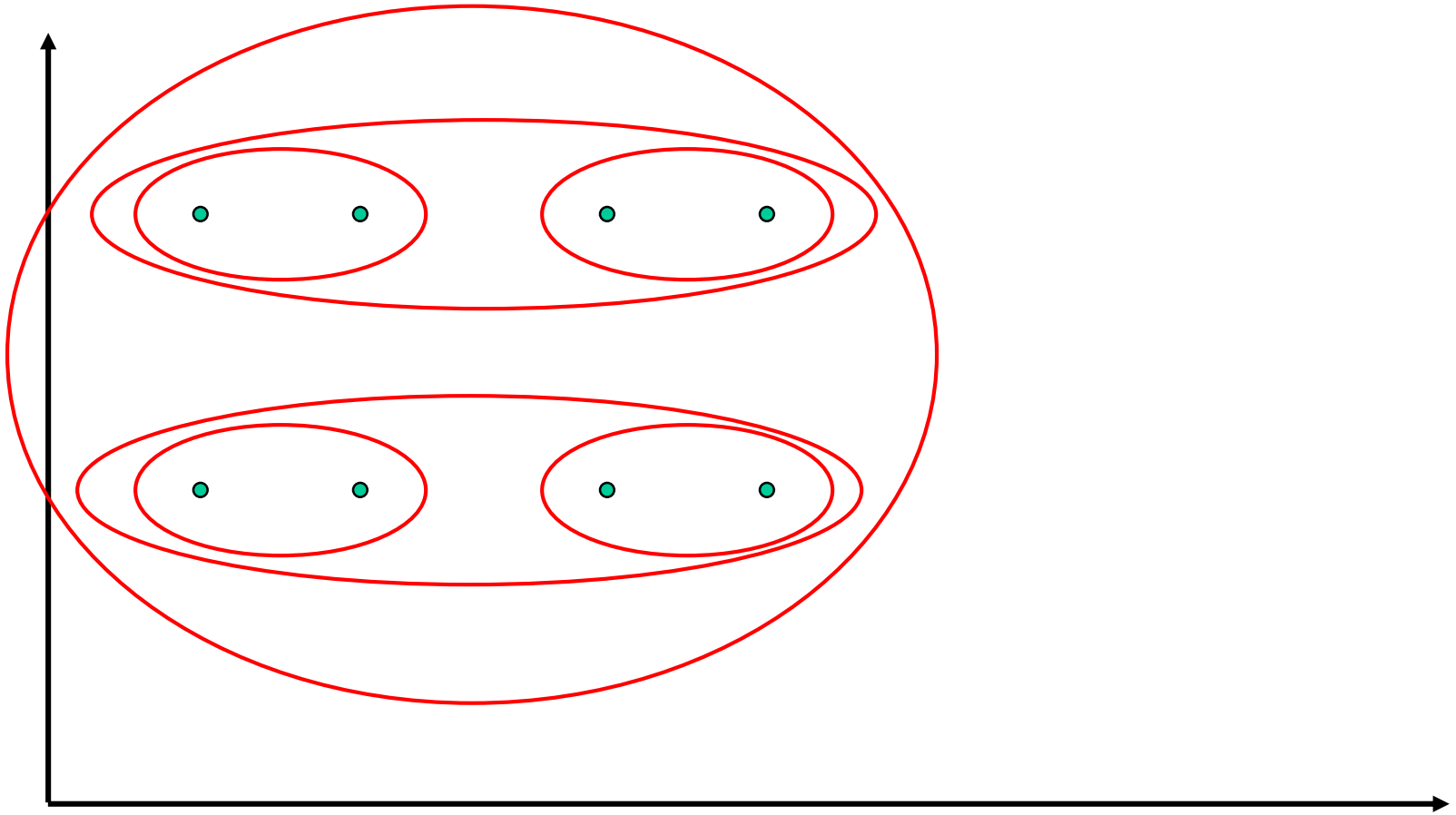
Phân cụm tích tụ dùng liên kết đơn

- Dùng độ tương tự lớn nhất của các cặp mẫu:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Có thể dẫn đến những cụm mảnh và trải dài.
  - Phù hợp trong một số ứng dụng, như phân cụm các hòn đảo.

Ví dụ liên kết đơn



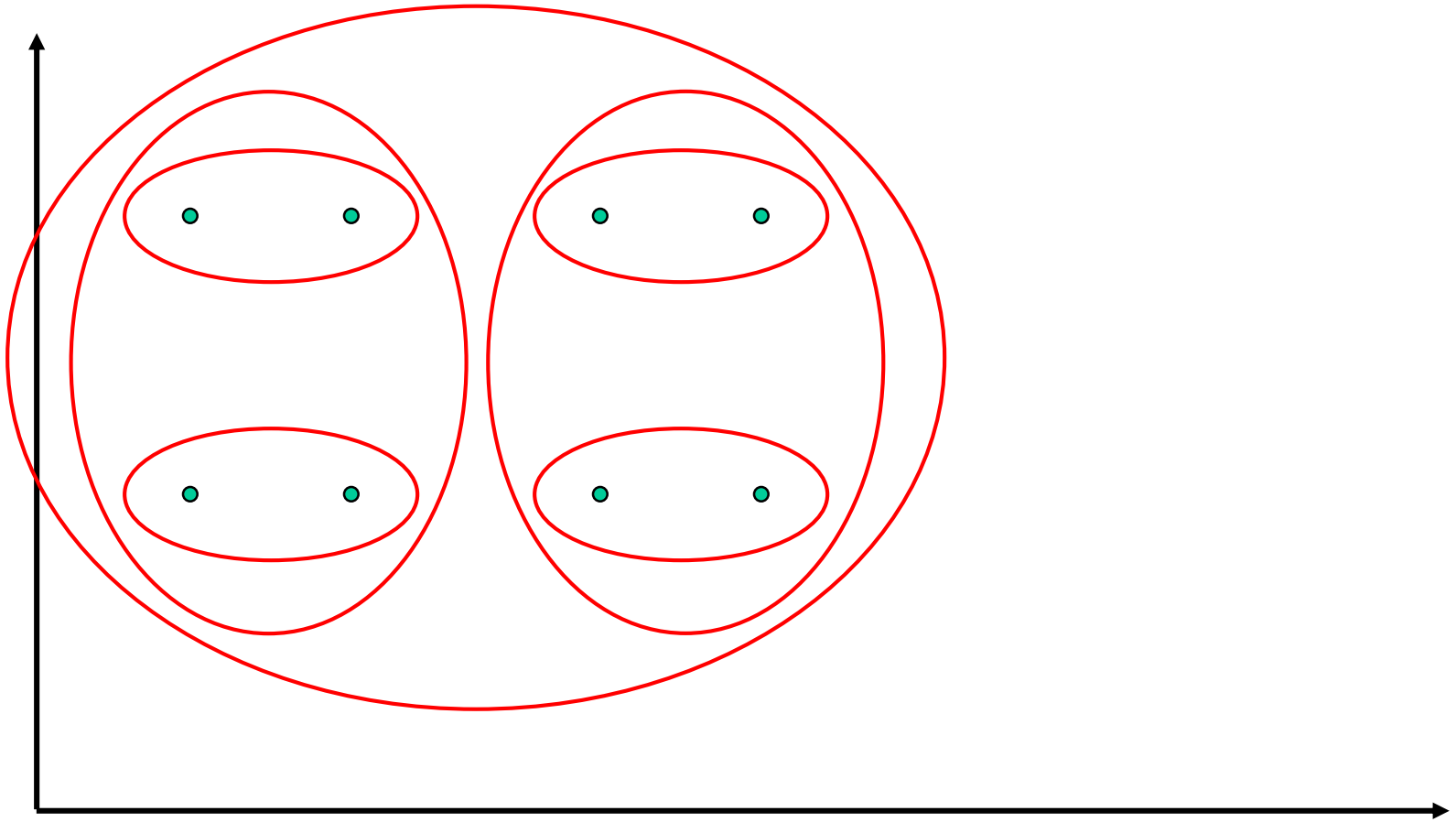
Phân cụm tích tụ dùng liên kết đầy đủ

- Dùng độ tương tự nhỏ nhất của các cặp mẫu:

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Tạo ra những cụm chặt hơn và có hình cầu.
  - Thường ta thích những cụm như vậy hơn.

Ví dụ liên kết đầy đủ



Tính toán độ tương tự giữa hai cụm

- Sau khi hợp hai cụm c_i và c_j, độ tương tự của cụm hợp với một cụm c_k nào đó có thể tính nhanh hơn như sau:
  - Liên kết đơn:

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

  - Liên kết đầy đủ:

$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$

Phân cụm phẳng (không phân cấp)

- Thường phải cung cấp trước số cụm k mong muốn.
- Chọn ngẫu nhiên k mẫu làm các hạt giống.
- Tạo ra k cụm khởi đầu dùng các hạt giống đó.
- Lặp lại việc gán các mẫu cho các cụm gần nhất để cải thiện dần dần chất lượng phân cụm.
- Ngừng khi phân cụm hội tụ (các cụm không thay đổi nữa) hoặc sau một số bước lặp định trước.

Thuật toán phân cụm k-means

- Giả sử các mẫu là những véctơ các giá trị thực.
- ***Trọng tâm*** (centroid) của mỗi cụm *c*:

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Việc gán lại các mẫu vào các cụm dựa trên khoảng cách tới trọng tâm của các cụm.

Các độ đo khoảng cách

- Khoảng cách Euclide (chuẩn L_2):

$$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

- Chuẩn L_1:

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m}|x_i - y_i|$$

- Độ tương tự côsin (biến đổi sang khoảng cách bằng cách lấy 1 trừ đi nó):

$$1 - \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

Các bước của thuật toán *k*-means

1. Gọi d là độ đo khoảng cách giữa các mẫu.
2. Chọn ngẫu nhiên k mẫu s_1, s_2, …, s_k làm các hạt giống.
3. Cho đến khi phân cụm hội tụ hoặc một tiêu chuẩn dừng nào đó được thỏa mãn:
    3.1. Với mỗi mẫu x_i:
        Gán x_i vào cụm c_j với $d(x_i, s_j)$ nhỏ nhất.
    3.2. Với mỗi cụm c_j:
        $s_j = \mu(c_j)$

Phân cụm văn bản

- Có thể áp dụng trực tiếp các thuật toán phân cụm HAC và k-means vào dữ liệu văn bản.
- Thông thường, ta sẽ dùng cách biểu diễn văn bản tf-idf và độ đo tương tự côsin.
- Các ứng dụng:
  - Phân cụm các văn bản trả về để kết quả tìm kiếm có tính tổ chức tốt hơn.
  - Tìm theo các trọng tâm cụm trước tiên, sau đó mới tìm các văn bản cụ thể trong những cụm có trọng tâm gần nhất.
  - Tạo cây phân cụm văn bản để tiện cho việc duyệt văn bản.