



**IT4853**

# **Tìm kiếm và trình diễn thông tin**

---

Bài 14. Phân cụm văn bản (2)

IIR. C16. Flat clustering

*Bộ môn Hệ thống thông tin  
Viện CNTT & TT*



# Nội dung chính

---

- **Tính hội tụ của K-means**
- **Đánh giá kết quả chia cụm**



# K-means luôn hội tụ

---

- **RSS: Residual Sum of Squares;**
- **RSS tổng bình phương khoảng cách giữa các văn bản và trọng tâm gần nhất;**
- **RSS giảm dần sau mỗi bước chia cụm**
  - **Vì mỗi văn bản được gán với trọng tâm gần nhất;**
- **RSS giảm sau mỗi bước xác định lại tâm cụm**
  - **Xem slides tiếp theo**
- **Số cách chia cụm là hữu hạn;**

# RSS giảm khi xác định lại tâm cụm

- $RSS = \sum_{k=1..K} RSS_k$
- $RSS_k(\vec{\mu}) = \sum_{\vec{x} \in \omega_k} \|\vec{\mu} - \vec{x}\|^2$
- $RSS_k(\vec{\mu}) = \sum_{\vec{x} \in \omega_k} \sum_{i=1..M} (\mu_i - x_i)^2$
- $\frac{\partial RSS_k(\vec{\mu})}{\partial \mu_i} = \sum_{\vec{x} \in \omega_k} 2(\mu_i - x_m)$
- $\mu_i = \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} x_i$

RSS đạt cực tiểu tại  $\vec{\mu}$  là tâm cụm



# Tính tối ưu của K-means

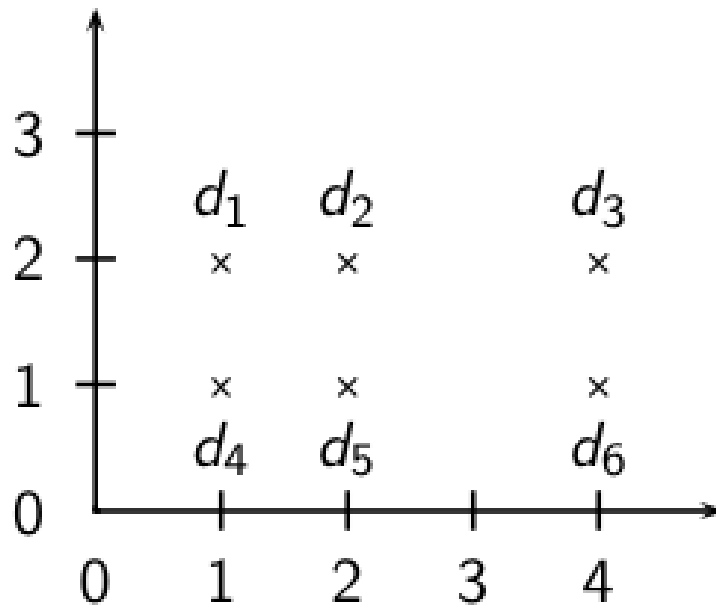
---

- **Hội tụ không đồng nhất với cách chia cụm tối ưu;**
- **Nếu lựa chọn tâm cụm ban đầu không tốt, chất lượng chia cụm có thể rất thấp.**



# Hội tụ, cận tối ưu

---



- **Kết quả chia cụm tối ưu cho  $K = 2$ ?**
- **Luôn hội tụ với các tập mầm  $\{d_i, d_j\}$  bất kỳ?**



# Khởi tạo K-means

---

- **Nhược điểm của khởi tạo ngẫu nhiên là không ổn định: kết quả chia cụm có thể không tối ưu**
- **Hiệu chỉnh:**
  - **Lựa chọn tập mẫu tốt;**
  - **V.D., thực hiện nhiều lượt sinh ngẫu nhiên rồi chọn kết quả tốt nhất.**



# Độ phức tạp giải thuật K-means

---

- Tính khoảng cách giữa hai vec-tơ  $O(M)$
- Gắn văn bản với trọng tâm:  $O(KNM)$
- Xác định lại trọng tâm:  $O(NM)$
- Giả sử giải thuật hội tụ sau  $I$  bước
- Độ phức tạp tổng quát:  $O(IKNM)$

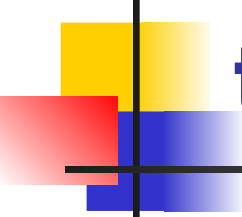




# Nội dung chính

---

- Tính hội tụ của K-means
- **Đánh giá kết quả chia cụm**



# Đánh giá kết quả chia cụm dựa trên dữ liệu phân lớp

---

- **Ý tưởng:** Coi kết quả phân lớp là phương án chia cụm tối ưu, đáp ứng tốt nhất các tiêu chí chia cụm.
  - Đánh giá kết quả chia cụm bằng cách so sánh với kết quả phân lớp mẫu.
- **Các độ đo:**
  - Purity
  - Rand Index



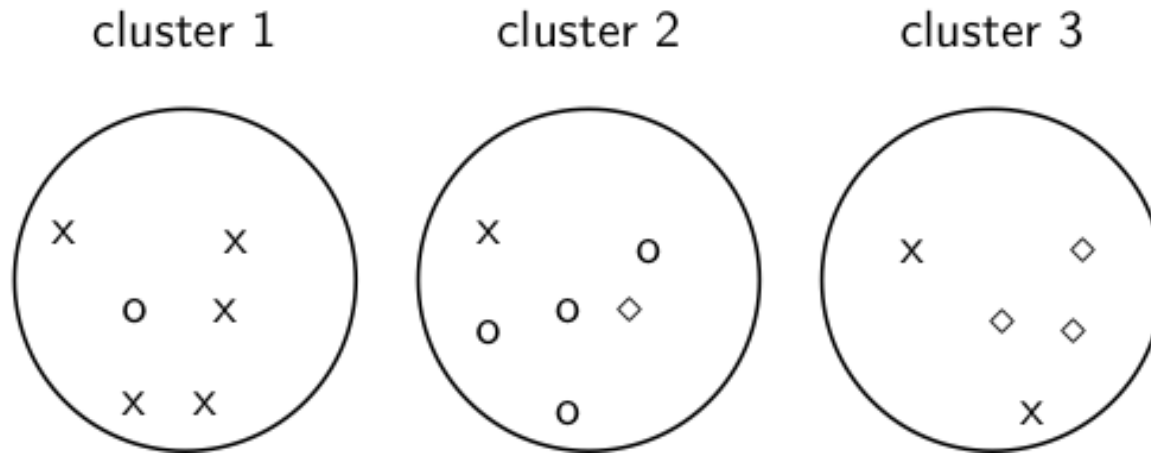
# Độ đo Purity

---

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

- $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$  là tập cụm,
- $C = \{c_1, c_2, \dots, c_J\}$  là tập lớp.

# Ví dụ Purity



- **Tính purity:**

- $\max_j |\omega_1 \cap c_j| = 5; \max_j |\omega_2 \cap c_j| = 4;$

- $\max_j |\omega_3 \cap c_j| = 3$

- **Purity =  $(1/17) \times (5 + 4 + 3) \approx 0.71$ .**



# Rand Index

---

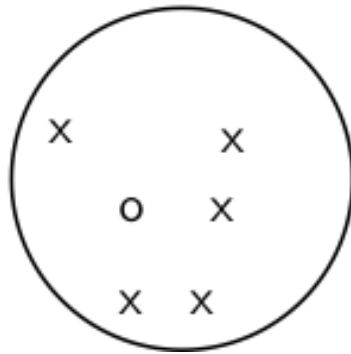
$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

	<b>Cùng lớp</b>	<b>Khác lớp</b>
<b>Cùng cụm</b>	<b>TP</b>	<b>FP</b>
<b>Khác cụm</b>	<b>FN</b>	<b>TN</b>

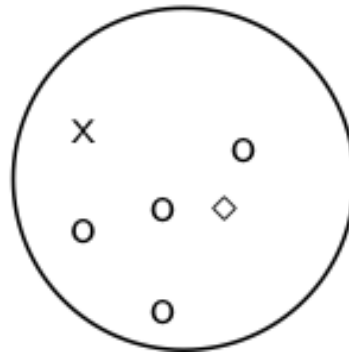
- **TP+ FN + FP + TN = N là tổng số cặp văn bản.**

# Ví dụ Rand Index

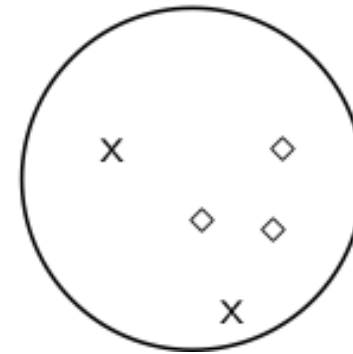
cluster 1



cluster 2



cluster 3



$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

**FP = 40 - 20 = 20, FN và TN được xác định tương tự.**



# Ví dụ Rand Index

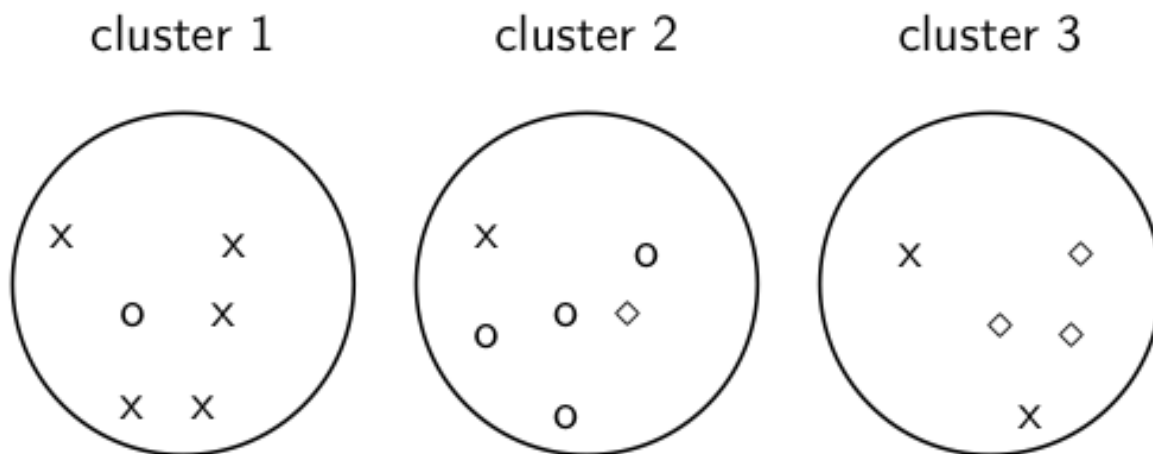
---

	Cùng lớp	Khác lớp
Cùng cụm	<b>TP = 20</b>	<b>FP = 20</b>
Khác cụm	<b>FN = 24</b>	<b>TN = 72</b>

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

$$RI = (20 + 72)/136$$

# Tổng hợp



	purity	NMI	RI	$F_5$
lower bound	0.0	0.0	0.0	0.0
maximum	1.0	1.0	1.0	1.0
value for example	0.71	0.36	0.68	0.46





## Bài tập 19.1

---

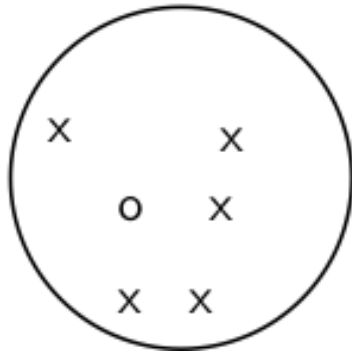
**Hai điều kiện dừng của giải thuật k-means:  
(i) kết quả phân cụm không thay đổi; (ii) tâm cụm không thay đổi.**

**Từ điều kiện (i) có suy ra được điều kiện (ii) hay không?**

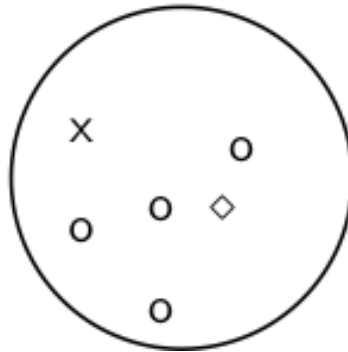
**Từ điều kiện (ii) có suy ra được điều kiện (i) hay không?**

## Bài tập 19.2

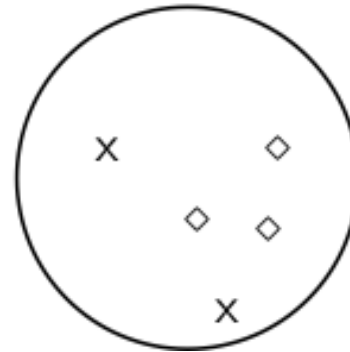
cluster 1



cluster 2



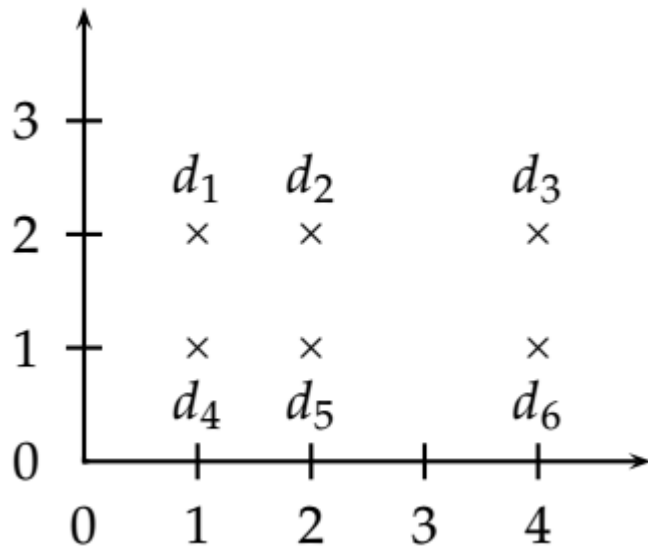
cluster 3



**Thay thế mỗi văn bản trên hình vẽ bằng hai văn bản. Sau đó hãy tính Purity và RI.**

**Thêm các văn bản trùng lặp có làm quá trình chia cụm khó hơn không? Đại lượng nào thay đổi/không thay đổi?**

# Bài tập 19.3



$$\{\{d_1, d_2, d_3\}, \{d_4, d_5, d_6\}\}$$

$$\{\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}\}$$

**Hãy tính RSS cho kết quả chia cụm trong cả hai trường hợp.**



## Bài tập 19.5

---

**Hãy lấy ví dụ một tập điểm và 3 trọng tâm ban đầu sao cho kết quả phân cụm 3-means hội tụ với cụm rỗng. (ii) Kết quả chia cụm với cụm rỗng có thể là kết quả tối ưu toàn cục theo RSS?**



## Bài tập 19.6

---

**Hãy chứng minh  $RSS_{\min}(K)$  là hàm đơn điệu giảm đối với biến  $K$ .**

