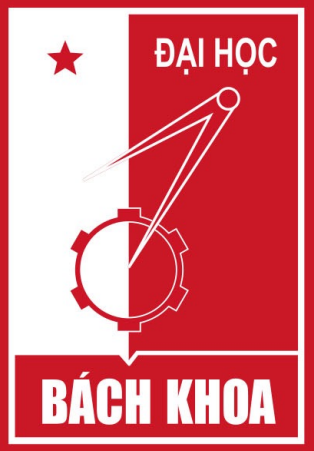

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nhập môn
Học máy và Khai phá dữ liệu
(IT3190)

Nội dung môn học

- Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- Lecture 9: Học dựa trên xác suất
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- **Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp**
- Lecture 13: Thảo luận ứng dụng trong thực tế

CÁC KHÁI NIỆM CƠ BẢN

- Lịch sử hình thành
  - Được đề nghị bởi Agrawal et al. (1993)
  - Sau đó được cộng đồng KPDL liên tục nghiên cứu trong nhiều năm
  - Giả thiết các dữ liệu đều ở dạng phân loại (rời rạc, có ý nghĩa)
  - Khởi đầu dùng với mục đích Phân tích giỏ hàng (Market Basket Analysis)

CÁC KHÁI NIỆM CƠ BẢN

- Mô hình luật kết hợp
  - Tập các món hàng $I=\{i_1, i_2, ..., i_m\}$
  - Một giao dịch t tập con I
  - Cơ sở dữ liệu giao dịch $T=\{t_1, t_2, ..., t_n\}$

CÁC KHÁI NIỆM CƠ BẢN

- Cơ sở dữ liệu giao dịch T
  - t_1= {bánh mỳ, pho mát, sữa }
  - t_2={táo, trứng, muối, sữa chua}
  - …
  - t_n={bánh bích quy, trứng, sữa}

CÁC KHÁI NIỆM CƠ BẢN

- Các thuật ngữ tương ứng
  - Món hàng (item) được để trong giỏ hàng.
  - Tập I gồm tất cả các món hàng bán trong siêu thị.
  - Một giao dịch (transaction) gồm các món hàng sẽ phải thanh toán nằm trong giỏ, thông thưởng mỗi giao dịch có một số hiệu ID (transaction ID).
  - Tập dữ liệu giao dịch T gồm có các giao dịch

CÁC KHÁI NIỆM CƠ BẢN

- Các kết hợp (association rule) : luật kết hợp là một sự suy dẫn có dạng

$$X \rightarrow Y$$

trong đó $X, Y \subset I$ còn $X \cap Y = \emptyset$.

CÁC KHÁI NIỆM CƠ BẢN

- Thuật ngữ liên quan đến luật kết hợp
  - Một tập các món hàng (an itemset)
  - Một tập của k-món hàng (k-itemset) món hàng có k món

CÁC KHÁI NIỆM CƠ BẢN

- Các phép đo dùng cho luật kết hợp
  - Hỗ trợ (support) : luật được hỗ trợ, ký hiệu sup, bao nhiêu phần trăm trong cơ sở dữ liệu T

$$sup(X \rightarrow Y) = Pr(X,Y)$$

  - Tin cậy (confidence) : luật được tin cậy, ký hiệu conf, bao nhiêu phần trăm khi có X đồng thời với Y

$$conf(X \rightarrow Y) = Pr(Y|X)$$

CÁC KHÁI NIỆM CƠ BẢN

- Bài toán khai phá luật kết hợp
  - **Đầu vào** : Tập các giao dịch T cùng *minsup*, *minconf*
  - **Đầu ra** : mọi X,Y thuộc I thỏa mãn

    $sup(X \rightarrow Y) \geq minsup$, $conf(X \rightarrow Y) \geq minconf$

MỤC LỤC

- Các khái niệm cơ bản
- Giải thuật Apriori
- Các vấn đề luật kết hợp

GIẢI THUẬT APRIORI

- Bài toán khai phá luật kết hợp
  - **Đầu vào** : Tập các giao dịch T cùng *minsup*, *minconf*
  - **Đầu ra** : $\forall X,Y \subset I$ thỏa mãn

    $sup(X \rightarrow Y) \geq minsup$, $conf(X \rightarrow Y) \geq minconf$
  - **Giải thuật Apriori** : gồm 2 bước chính
    1. Tìm tập thường xuyên (frequent itemset) ≥ *minsup*
    2. Dùng tập trên để sinh ra các luật kết hợp (generate-association-rules) ≥ *minconf*

GIẢI THUẬT APRIORI

- **Bước 1** :Tìm tập thường xuyên ≥ *minsup*
  - Một tập thường xuyên là một tập các món hàng có độ hỗ trợ ≥ minsup
  - Thuộc tính apriori : mọi tập con của tập thường xuyên cũng là tập thường xuyên
- **Ý tưởng** :
  - Khởi tạo, tìm tập thường xuyên kích thước 1 : F_1
  - Giải thuật lặp k=2,3, …
    - C_k = sinh các ƯCV tập thường xuyên kích thước k biết tập F_{k-1}
    - F_k = tập thường xuyên thực sự với $F_k \subseteq C_k$

GIẢI THUẬT APRIORI

- Dữ liệu giao dịch T với minsup = 50%

TID	Món hàng
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E

- Quét T ⇒ C_1={{A} : 2, {B} : 3, {C } :3, {D} : 1, {E } : 3}
- F_1={{A} : 2, {B} : 3, {C } : 3, {E } : 3} ⇒
- C_2={{AB}, {AC}, {AE}, {BC}, {BE}, {CE}}

GIẢI THUẬT APRIORI

- Dữ liệu giao dịch T với minsup = 50%

TID	Món hàng
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E

- Quét T ⇒ C_2 = {{AB}:1, {AC}:2, {AE} :1, {BC}:2, {BE}:3, {CE}:2}
- F_2 = {{AC}:2, {BC}:2, {BE}:3, {CE}:2 }
- C3 = {BCE }

GIẢI THUẬT APRIORI

- Dữ liệu giao dịch T với minsup = 50%

TID	Món hàng
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E

- Quét T ⇒ C_3 = {{BCE} : 2} ⇒
- F_3 = {BCE}

GIẢI THUẬT APRIORI

- **Bước 1** : Lưu ý khi biểu diễn dữ liệu giao dịch
  - Các món hàng trong cùng một giao dịch nên được xếp theo thứ tự alphabét (nhỏ đến lớn)
  - Các món hàng trong một tập thường xuyên cũng được sắp xếp theo thứ tự
  - $\{i_1 , ..., i_k \}$ biểu diễn một tập thường xuyên thì tuân theo thứ tự $i_1 < ... < i_k$

GIẢI THUẬT APRIORI

- **Function** frequent-itemsets(T)

1. $C_1 \leftarrow$ init-pass(T); $F_1 \leftarrow \{f | f \in C_1, f.count/n \geq minsup\}$;
2. **for** ($k \leftarrow 2; F_{k-1} != \emptyset; k++$) **do**
3. $C_k \leftarrow$ **candidate-gen**(F_{k-1}) // Hàm sinh ƯCV
4. **foreach** giao dịch $t \in T$ **do**
5. **foreach** $c \in C_k$ **do**
6. **if**(c chứa trong t) **then** c.count $\leftarrow$ c.count + 1 **endif**
7. **endfor**
8. **endfor**
9. $F_k \leftarrow \{c | c \in C_k , c.count/n \geq minsup\}$
10. **endfor**
11. **return** $F \leftarrow \cup_k F_k$

GIẢI THUẬT APRIORI

- **Bước 1** : Hàm phụ trợ
- **Function candidate-gen**(F_{k-1})
  1. **forall** ($f_1 = \{i_1 , ..., i_{k-2} , i_{k-1} \}$, $f_2 = \{i_1,..., i_{k-2}, i'_{k-1}\} \in F_{k-1}$ với $i_{k-1} < i'_{k-1}$) **do**
  2. $c \leftarrow \{i_1 , ..., i_{k-2} , i_{k-1} , i'_{k-1}\}$ // nối f_1 và f_2
  3. $C_k \leftarrow C_k \cup \{c\}$
  4. **foreach** (tập con s kích thước k-1 của c) **do**
  5. **if**(s not in F_{k-1}) **then** $C_k \leftarrow C_k - \{c\}$ **endif** // xén bớt
  6. **endfor**
  7. **endfor**
  8. **return** C_k

GIẢI THUẬT APRIORI

- **Ví dụ** : sinh ƯCV
- F_3 = {{1, 2, 3},{1, 2, 4}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}}
- Sau khi nối
  - C_4 = {{1, 2, 3, 4},{1, 3, 4, 5}}
- Sau khi xén
  - C_4 = {{1, 2, 3, 4}} vì {1, 4, 5} không có trong F_3 nên loại bỏ {1, 3, 4, 5}

GIẢI THUẬT APRIORI

- **Bước 2** : tập thường xuyên để sinh ra các luật kết hợp ≥ *minconf*
  - **Đầu vào** : Tập các tập thường xuyên F
  - **Đầu ra** : Tập các luật kết hợp ≥ *minconf*
- **Function generate-association-rules**(F)
  1. **forall** f ∈ F do
  2. **forall** X là tập con khác rỗng của f do
  3. Y ← f – X
  4. **if**(conf(X → Y)≥minconf) **then** R←RU(X → Y) **endif**
  5. **endfor**
  6. **endfor**
  7. **return** R

GIẢI THUẬT APRIORI

- **Ví dụ** : Giả sử ta có tập thường xuyên f={2 ,3, 4} với độ hỗ trợ sup=50%
- Các tập con khác rỗng cũng độ hỗ trợ {2,3}:50%, {2,4}:50%,{3,4}:75%, {2}:75%, {3}:75%, {4}:75%
- Với minconf ≥ 50% thì
  - 2,3 → 4 conf=100%
  - 2,4 → 3 conf=100%
  - 3,4 → 2 conf=67%
  - ...

MỤC LỤC

- Các khái niệm cơ bản
- Giải thuật Apriori
- Các vấn đề luật kết hợp

Các vấn đề luật kết hợp

- Với kiểu dữ liệu giao dịch ta không có đích đến là một món hàng cụ thể trong quá trình suy dẫn theo luật
- Đưa ra mọi khả năng của luật kết hợp hay mọi món hàng, tập món hàng đều có thể là kết luận do luật suy dẫn
- Tuy nhiên, có nhiều ứng dụng người dùng quan tâm đến một vài món hàng cụ thể hay có đích đến từ đầu

Các vấn đề luật kết hợp

- Khai phá luật kết hợp với dữ liệu lớp - mining Class Association Rules (CARs) dùng để phân loại
- Khai phá luật kết hợp dùng với hệ khuyến nghị người dùng
- Khai phá luật kết hợp dùng để khai phá dữ liệu đồ thị

Tổng kết

- Tập thường xuyên và luật kết hợp dùng để giải quyết nhiều bài toán khác nhau : phân loại, tổ hợp...
- Khai phá luật kết hợp vẫn là mô hình phân tích giỏ hàng phổ biến
- Xem thêm cây tiền tố FP-tree để thay thế giải thuật Apriori

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict