Chương 3 : Luật kết hợp dùng trong KPDL

Trịnh Anh Phúc [1]

[1]Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

Ngày 8 tháng 5 năm 2014

Giới thiệu

1. Các khái niệm cơ bản
   - Mô hình
   - Luật kết hợp và các phép đo
   - Mục tiêu và các tình huống đặc biệt của luật kết hợp
2. Giải thuật Apriori
   - Bước 1 : tìm tập thường xuyên
   - Bước 2 : sinh luật kết hợp từ tập thường xuyên
3. Các vấn đề liên quan đến luật kết hợp
   - Biểu diễn dữ liệu khi dùng luật kết hợp
   - Luật kết hợp với dữ liệu được phân lớp
4. Chương trình Weka

Các khái niệm cơ bản

Lịch sử hình thành

- Được đề nghị bởi Agrawal et al. (1993) [1]
- 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)

[1] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.

Các khái niệm cơ bản

Mô hình luật kết hợp

- Tập các món hàng

$$I = \{i_1, i_2, ..., i_m\}$$

- Một giao dịch (mua bán) :
  - là một tập các món hàng $t \subseteq I$
- Một cơ sở dữ liệu giao dịch T sẽ là một tập nhiều giao dịch

$$T = \{t_1, t_2, ..., t_n\}$$

Các khái niệm cơ bản

Cơ sở dữ liệu giao dịch tại một siêu thị

- 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 khái niệm 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

Cơ sở dữ liệu giao dịch : tập các văn bản

Mỗi văn bản được biểu diễn như một "túi" các từ khóa

- d_1 : sinh viên, giáo viên, trường học
- d_2 : sinh viên, trường học, thực tập
- ...
- d_n : sinh viên, giờ học, thể dục

Các khái niệm cơ bản

Các khái niệm tương ứng

- Món hàng chính là từ khóa trong văn bản
- Tập I gồm tất cả các từ khóa
- Một giao dịch t tương ứng một văn bản xuất hiện
- Tập dữ liệu giao dịch T là tập các văn bản

Các khái niệm cơ bản

Luật kết hợp

Một luật kết hợp là một sự suy dẫn có dạng

$$X \rightarrow Y \text{ trong đó } X, Y \subset I \text{ và } X \cap Y = \emptyset$$

Ngoài ra, để chi tiết hóa hơn hai vế X và Y ta có

- Một tập các món hàng (an itemset), e.g. $X = \{$sữa, bánh mỳ, ngũ cốc$\}$
- Một tập của k-món hàng (k-itemset), e.g. X là tập của 3-món hàng

Các khái niệm cơ bản

Các phép đo đối với 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

$$conf(X \rightarrow Y) = Pr(Y|X) = \frac{Pr(X, Y)}{Pr(X)} = \frac{sup(X \cup Y)}{sup(X)}$$

- Một luật kết hợp là một mẫu chỉ ra khi xuất hiện X sẽ xuất hiện Y với giá trị xác suất xác định trong tập T

Các khái niệm cơ bản

Mục tiêu

Tìm tất cả các luật thỏa mãn

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

trong đó $X, Y \subset I$ với mức hỗ trợ *minsup* và tin tưởng *minconf* do người dùng cho trước

Các tính huống đặc biệt

- Đầy đủ (completeness) : tìm tất cả các luật
- Không có đích (no target on right-hand-side) : $Y = \emptyset$

Các khái niệm cơ bản

t1: Thịt bò, Thịt gà, Sữa
t2: Thịt bò, Phô mát
t3: Phô mát, Ủng
t4: Thịt bò, Thịt gà, Phô mát
t5: Thịt bò, Thịt gà, Quần Áo, Phô mát, Sữa
t6: Thịt gà, Quần Áo, Sữa
t7: Thịt gà, Sữa , Quần Áo

- Cho $minsup = 30\%$ và $minconf = 80\%$
- Tập các món hàng xuất hiện tần suất cao {Thịt gà, Quần Áo, Sữa} có $[sup = 3/7 \geq 30\%]$

- Tập các luật kết hợp tương ứng tập các món hàng trên là
  - Quần Áo $\rightarrow$ Sữa, Thịt gà [sup $= 3/7$, conf $= 3/3 \geq 80\%$]
  - Quần Áo, Sữa$\rightarrow$ Thịt gà [sup $= 3/7$, conf $= 3/3 \geq 80\%$]
  - ...
  - Quần Áo, Thịt gà $\rightarrow$ Sữa [sup $= 3/7$, conf $= 3/3 \geq 80\%$]

Các khái niệm cơ bản

Một vài lưu ý đối với dữ liệu giao dịch

- Đây là một cách nhìn đơn giản về giao dịch mua bán trong siêu thị
- Có một loạt các thông tin quan trọng không xuất hiện
  - số lượng mỗi món hàng
  - tiền phải trả tương ứng
  - số tiền thuế giá trị gia tăng
  - số hiệu nhân viên thanh toán
  - ngày giờ thực hiện giao dịch mua bán

Dành cho trả lời câu hỏi

Giải thuật Apriori

Đầu vào

- Tập dữ liệu giao dịch T
- Giá trị *minsup*
- Giá trị *minconf*

Đầu ra

- Tập tất cả các luật kết hợp $X \rightarrow Y$ sao cho
  - $sup(X \rightarrow Y) \geq minsup$
  - $conf(X \rightarrow Y) \geq minconf$

Giải thuật Apriori

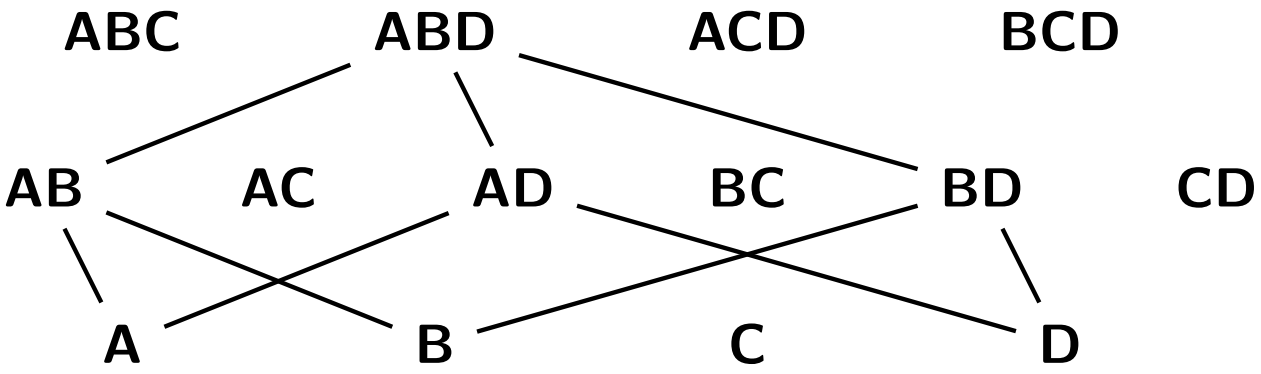
Giải thuật Apriori gồm hai bước chính

- **Bước 1** : tìm tất cả tập các món hàng có độ hỗ trợ lớn hơn bằng *minsup* hay tập thường xuyên (frequent itemsets)
- **Bước 2** : dùng tập trên để sinh ra các luật kết hợp (generate-association-rules) có độ tin cậy lớn hơn bằng *minconf*

Giải thuật Apriori

Bước 1 : tìm tập thường xuyên

- Một tập thường xuyên là một tập các món hàng có độ hỗ trợ $\geq$ *minsup*
- Thuộc tính apriori : mọi tập con của tập thường xuyên cũng là tập thường xuyên



Giải thuật Apriori

Bước 1 : tìm tập thường xuyên (tiếp)

Giải thuật lặp theo bước k = 2,3...

- Khởi tạo, tìm tập thường xuyên kích thước 1 : F_1
- Lặp với k = 2,3,....
  - C_k = 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

1. Quét T $\Rightarrow C_1 = \{\{A\} : 2, \{B\} : 3, \{C\} : 3, \{D\} : 1, \{E\} : 3\}$

   $F_1 = \{\{A\} : 2, \{B\} : 3, \{C\} : 3, \{E\} : 3\} \Rightarrow$
   $C_2 = \{\{AB\}, \{AC\}, \{AE\}, \{BC\}, \{BE\}, \{CE\}\}$

2. Quét T $\Rightarrow C_2 = \{\{AB\} : 1, \{AC\} : 2, \{AE\} : 1, \{BC\} : 2, \{BE\} : 3, \{CE\} : 2\}$

   $F_2 = \{\{AC\} : 2, \{BC\} : 2, \{BE\} : 3, \{CE\} : 2\}$
   $C_3 = \{BCE\}$

3. Quét T $\Rightarrow C_3 = \{\{BCE\} : 2\} \Rightarrow F_3 = \{BCE\}$

Giải thuật Apriori

Lưu ý khi biểu diễn dữ liệu giao dịch

- Các món hàng trong cùng một giao dịch nên được xếp theo thứ tự al pha bét (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** $(k \leftarrow 2; F_{k-1} \neq \emptyset;$k++$)$ **do**
3. $\quad C_k \leftarrow$ **candidate-gen**(F_{k-1}) // Hàm sinh ƯCV
4. $\quad$ **for** mỗi giao dịch $t \in T$ **do**
5. $\quad\quad$ **for** mỗi ƯCV $c \in C_k$ **do**
6. $\quad\quad\quad$ **if**(c chứa trong t) **then** $c.count \leftarrow c.count + 1$ **endif**
7. $\quad\quad$ **endfor**
8. $\quad$ **endfor**
9. $\quad F_k \leftarrow \{c | c \in C_k, c.count/n \geq minsup\}$
10. **endfor**
11. **return** $F \leftarrow \cup_k F_k$

End

Giải thuật Apriori

Thủ tục con : candidate-gen(F_{k-1})

Thủ tục trả lại tập các ƯCV từ tập thường xuyên bước trước F_{k-1}, gồm có hai bước

- bước nối (join step) : sinh ra tất cả các khả năng của tập thường xuyên C_k kích thước k
- bước xén (prune step) : loại bỏ những ƯCV trong C_k mà chúng không thể trở thành tập thường xuyên

Giải thuật Apriori

Function candidate-gen(F_{k-1})

1. $C_k \leftarrow \emptyset$
2. **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**
3. $c \leftarrow \{i_1, ..., i_{k-2}, i_{k-1}, i'_{k-1}\}$ // nối f_1 và f_2
4. $C_k \leftarrow C_k \cup \{c\}$
5. **for**(mỗi tập con s kích thước k-1 của c) **do**
6. **if**(s $\notin F_{k-1}$) **then** $C_k \leftarrow C_k - \{c\}$ **endif** // xén bớt
7. **endfor**
8. **endfor**
9. **return** C_k

End

Giải thuật Apriori

Ví dụ

- $F_3 = \{\{1, 2, 3\},\{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}\}$
- 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

Sinh luật kết hợp từ tập thường xuyên

- luật kết hợp $X \rightarrow Y$ không phải là tập thường xuyên (frequent itemsets)
- cần có một bước nữa sinh ra các luật này từ tập thường xuyên
- các luật này phải có độ tin tưởng lớn hơn bằng *minconf* được xác định trước

Giải thuật Apriori

Thủ tục sinh luật kết hợp

- **Đầu vào** : Tập các tập thường xuyên F
- **Đầu ra** : Tập các luật kết hợp
  R=$\{X \rightarrow Y | conf(X \rightarrow Y) \geq minconf\}$

Function generate-association-rules(F)

1. **forall** tập thường xuyên $f \in F$ **do**
2. **forall** X là tập con khác rỗng của f **do**
3. $Y \leftarrow f - X$
4. **if**$(conf(X \rightarrow Y) \geq minconf)$**then**$R \leftarrow R \cup (X \rightarrow Y)$ **endif**
5. **endfor**
6. **endfor**
7. **return** R

End

Giải thuật Apriori

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%
- Chúng sinh ra các luật kết hợp sau đều có độ tin tưởng $\geq$ 50%
  - 2,3 $\rightarrow$ 4 conf=100%
  - 2,4 $\rightarrow$ 3 conf=100%
  - 3,4 $\rightarrow$ 2 conf=67%
  - 2 $\rightarrow$ 3,4 conf=67%
  - 3 $\rightarrow$ 2,4 conf=67%
  - 4 $\rightarrow$ 2,3 conf=67%

Giải thuật Apriori

Tổng kết bước 2 của giải thuật : sinh luật kết hợp

- Để có được luật kết hợp $X \rightarrow Y$, ta cần $sup(X \cup Y)$ và $sup(X)$
- Mọi giá trị trên đã được tính ở bước 1 của giải thuật, không cần quét cơ sở dữ liệu giao dịch T thêm nữa
- Bước 2 này yêu cầu thời gian ít hơn rất nhiều so với bước 1 ở trước khi ta cần tìm tập thường xuyên

Giải thuật Apriori

Nhận xét chung bước 1 của giải thuật : tìm tập thường xuyên

Dường như khá tốn kém vì ta phải tìm không gian tổ hợp tất cả m món hàng $I = \{i_1, i_2, ..., i_m\}$ tại mỗi bước lặp k

- Mỗi bước lặp k chỉ tăng kích thước lên một mỗi lần
- Nó quét cơ sở dữ liệu T tại mỗi bước lặp
- Trên thực tế thì số bước lặp tối đa K phụ thuộc *minsup* thường là khá nhỏ, $O(10)$ (giải thích ?)
- Trong một vài trường hợp, ta có thể tìm các tập với thời gian tuyến tính $O(n)$
- Thời gian tính sẽ lớn lên phụ thuộc kích thước tập dữ liệu

Giải thuật Apriori

Nhận xét chung bước 2 của giải thuật : sinh luật kết hợp từ tập thường xuyên

- Rõ ràng không gian các tập kết hợp là trong không gian hàm mũ $O(2^m)$ các tập thường xuyên
- KPDL tạo nên không gian rỗng (sparseness) của dữ liệu thông qua xác định hai giá trị *minsup* và *minconf*
- Tuy vậy, số luật kết hợp tạo thành vẫn là hàng ngàn, hàng chục ngàn thậm chí triệu...

Dành cho trả lời câu hỏi

Các vấn đề liên quan đến luật kết hợp

Dữ liệu thường có dạng bảng là chính nên ta phải chuyển về dạng dữ liệu giao dịch

- Dạng bảng

Thuộc tính 1	Thuộc tính 2	Thuộc tính 3
a	b	d
b	c	e

- Dạng giao dịch

Số hiệu TID	Món hàng
T1	a, b
T2	a, c, d, e
T3	a, d, f

Các vấn đề liên quan đến luật kết hợp

Cách chuyển đổi trực tiếp, ta dùng cặp (Thuộc tính, Giá trị) $\Rightarrow$ Món hàng

- Dạng bảng

Thuộc tính 1	Thuộc tính 2	Thuộc tính 3
a	b	d
b	c	e

- Chuyển sang dạng giao dịch

Số hiệu TID	Món hàng
T1	(TT1,a) , (TT2,b), (TT3,d)
T2	(TT1,b), (TT2,c), (TT3,e)

Các vấn đề liên quan đến luật kết hợp

Đặt vấn đề

- 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 liên quan đến luật kết hợp

Định nghĩa bài toán

Khai phá luật kết hợp với dữ liệu lớp - mining Class Association Rules (CARs) - được phát biểu như sau

- Xét T là tập dữ liệu giao dịch
- Mỗi giao dịch t sẽ được gán một nhãn lớp y
- Gọi I là tập các món hàng còn Y là tập các nhãn được gán, ta có $I \cap Y = \emptyset$
- Một luật kết hợp phân lớp (class asociation rule) là một suy dẫn $X \rightarrow y$ với $X \subseteq I$ còn $y \in Y$

Các vấn đề liên quan đến luật kết hợp

Mỗi văn bản bây giờ được gán một nhãn, cho $minsup = 20\%$ và $minconf = 60\%$

TID	$X \subseteq I$	$y \in Y$
d_1	sinh viên, giáo viên, trường học	giáo dục
d_2	sinh viên, trường học	giáo dục
d_3	giáo viên, trường học, thành phố, trò chơi	giáo dục
d_4	bóng chày, bóng rổ	thể thao
d_5	bóng chày, VĐV, người xem	thể thao
d_6	bóng rổ, HTV, trò chơi, đồng đội	thể thao
d_7	bóng rổ, đồng đội, thành phố, trò chơi	thể thao

- sinh viên, trường học $\rightarrow$ giáo dục [sup$= 2/7$, conf $= 2/2$]
- trò chơi $\rightarrow$ thể thao [sup$= 2/7$, conf $= 2/3$]

Các vấn đề liên quan đến luật kết hợp

Nhận xét về khai phá luật kết hợp phân lớp

- Khác với khai phá luật kết hợp thông thường, khai phá luật kết hợp phân lớp chi cần một bước
- Ở đây ta chỉ cần tìm tất cả các tập thường xuyên theo dạng

$$(condset, y)$$

  trong đó $condset \subseteq I$ và $y \in Y$ với $sup(condset, y) \geq minsup$
- Giải thuật Apriori có thể thay đổi để thực hiện tạo luật kết hợp phân lớp

Các vấn đề liên quan đến luật kết hợp

Tổng kết

- Luật kết hợp được nghiên cứu rất nhiều trong cộng đồng KPDL, có nhiều giải thuật và biến thể liên quan đến nó
- Có rất nhiều mô hình và phép đo dùng ứng dụng của luật kết hợp
- Ngoài ra còn có nhiều vấn đề liên quan không được trình bầy
  - Khai phá luật kết hợp nhiều độ tin tưởng (multi-confidences)
  - Khai phá luật kết hợp đa tầng (multi-level)
  - Khai phá luật kết hợp có ràng buộc
  - Khai phá luật kết hợp có tập thường xuyên lớn nhất (maximal frequent itemset)
  - Khai phá luật kết hợp với dữ liệu số (numerical)
  - Khai phá luật kết hợp với giải thuật song song
  - ...

Chương trình Weka

Yêu cầu

- Tải tập dữ liệu weather.nominal.arff
- Chạy giải thuật Apriori
- Giải thích các tham số của giải thuật