KHAI THÁC LUẬT KẾT HỢP VỚI APRIORI

LUẬT KẾT HỢP

Xét CSDL khảo sát tiện nghi sử dụng ở các hộ gia đình như sau:

Hộ	Tiện nghi sở hữu
1	Tivi, Máy vi tính
2	Tủ lạnh, Máy Lạnh
3	Tivi, Máy giặt, Máy lạnh
4	Tivi, Tủ lạnh, Máy lạnh
5	Tivi, Máy giặt, Máy vi tính
6	Tivi, Tủ lạnh, Máy giặt
7	Tivi, Tủ lạnh, Máy vi tính
8	Tivi, Tủ lạnh, Máy giặt, Máy lạnh, Máy vi tính

Luật kết hợp là phép kéo theo có dạng:

Tivi → Máyvitính [50%, 57%] hay
sử dụng:Tivi → sử dụng:Máyvitính [50%, 57%]

Nghĩa là: "57% hộ gia đình sử dụng Tivi thì cũng sử dụng Máy vi tính. Tivi và Máy vi tính xuất hiện chung trong 50% dòng dữ liệu"

KHAI THÁC LUẬT KẾT HỢP

Khai thác luật kết hợp được chia thành 2 gia đoạn

1. Khai thác tập phổ biến (FIs – Frequent Itemsets)
2. Khai thác luật kết hợp từ các tập phổ biến (ARs – Association Rules)

1. Khai thác tập phổ biến sử dụng thuật toán apriori

Độ phổ biến (support)

Cho CSDL giao dịch D và tập dữ liệu $X \subseteq \boldsymbol{I}$. Độ phổ biến của X trong D, kí hiệu $\sigma(X)$, được định nghĩa là số giao dịch mà X xuất hiện trong D.

Tập phổ biến

Tập $X \subseteq \boldsymbol{I}$ được gọi là phổ biến nếu $\sigma(X) \geq minSup$ (với *minSup* là giá trị do người dùng chỉ định).

Tính chất APRIORI

1. Mọi tập con của tập phổ biến đều phổ biến. Nghĩa là $\forall X \subseteq Y$, nếu $\sigma(X) \geq minSup$ thì $\sigma(Y) \geq minSup$
2. Mọi tập cha của tập không phổ biến đều không phổ biến. Nghĩa là $\forall Y \supseteq X$, nếu $\sigma(X) < minSup$ thì $\sigma(Y) < minSup$

Thuật toán APRIORI

- **Đầu vào**:CSDL giao dịch D và ngưỡng phổ biến *minSup*
- **Đầu ra**: FIs chứa tất cả các tập phổ biến của D
- **Mã giả**:

Gọi C_k: Tập các ứng viên có kích thước k
L_k: Các tập phổ biến có kích thước k
$L_1 = \{i \in I: \sigma(i) \geq minSup\}$
for ($k = 2$; L_{k-1} !=∅; k++) do
  C_k = {các ứng viên được tạo từ I_{k-1}}
  for each $t \in$ D do
    if $C_k \subseteq t$ then C_k.count++
  $L_k = \{C_k \mid C_k.\text{count} \geq minSup\}$
FIs = $\cup_k L_k$;

Cách tạo ứng viên của APRIORI

Nguyên tắc Apriori:

Nhớ lại tính chất: mọi tập con của tập phổ biến cũng phổ biến

- Giả sử ta có ***L3*** **= {***abc*, *abd*, *acd*, *ace*, *bcd***}**
- Xét việc kết để tao ra các ứng viên C4: ***L3*L3***
  - *abcd* được tạo từ *abc* và *abd*
  - *acde* được tạo từ *acd* và *ace*
- **Rút gọn:**
  - *acde* bị loại vì *ade* không có trong *L3*

⇒ ***C4* = {*abcd*}**

Ví dụ minh họa: Xét CSDL mẫu

Mã giao dịch	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

σ(A) = 4
σ(C) = 6
σ(D) = 4
σ(T) = 4
σ(W) = 5

Với *minSup* = 50% (50*6/100 = 3), ta có

Database (D)	
TID	**Nội dung**
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C1	
Danh mục	**Độ phổ biến**
A	4
C	6
D	4
T	4
W	5

L1	
Danh mục	**Độ phổ biến**
A	4
C	6
D	4
T	4
W	5

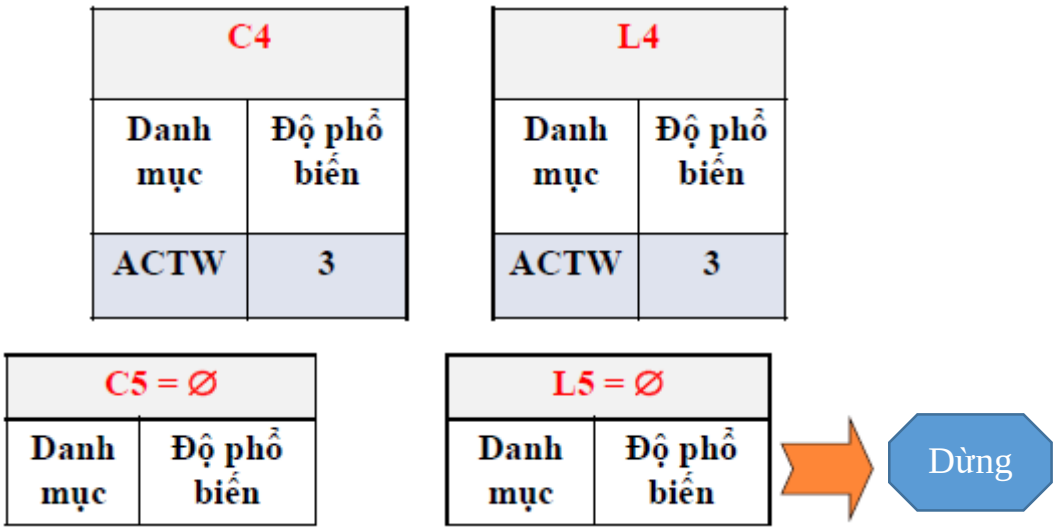
C2	
Danh mục	Độ phổ biến
AC	**4**
AD	**2**
AT	**3**
AW	**4**
CD	**4**
CT	**4**
CW	**5**
DT	**2**
DW	**3**
TW	**3**

L2	
Danh mục	Độ phổ biến
AC	**4**
AT	**3**
AW	**4**
CD	**4**
CT	**4**
CW	**5**
DW	**3**
TW	**3**

C3	
Danh mục	**Độ phổ biến**
ACT	**3**
ACW	**4**
ATW	**3**
CDW	**3**
CTW	**3**

L3	
Danh mục	**Độ phổ biến**
ACT	**3**
ACW	**4**
ATW	**3**
CDW	**3**
CTW	**3**

Lưu ý: CDT không có trong C_3 vì DT không có trong L_2



Kết luận:

Các tập phổ biến thõa ngưỡng minsupp = 50%

L1: {A, C, D, T, W}
L2: {AC, AT, AW, CD, CT, CW, DW, TW}
L3: {ACT, ACW, ATW, CDW, CTW}
L4: {ACTW}

2. *Khai thác luật kết hợp từ các tập phổ biến*

Định nghĩa:

Luật kết hợp là biểu thức có dạng $X \rightarrow Y - X$ (q,p) (X, Y là các tập phổ biến) trong đó $X,Y \neq \emptyset$, $X \subset Y$ và $p = \sigma(Y)/\sigma(X) \geq minConf$ gọi là độ tin cậy của luật còn $q = \sigma(Y) \geq minSup$ được gọi là độ phổ biến của luật.

Như vậy: luật kết hợp là luật sinh ra giữa các tập phổ biến $X, Y \in FI$ trong đó $X \subset Y$.

Ví dụ

Các luật kết hợp thõa ngưỡng minsupp = 50% và minconf = 80% trong ví dụ 2

- {ACTW} – Tập phổ biến thuộc L4
  Các tập con khác rỗng của tập phổ biến {ACT, ACW,ATW, CTW}, {AC, AT, AW, CT, CW, TW}, {A, C, T, W}
  Các luật có thể

ACT => W	conf = 3/3 = 100%	CT => AW	conf = 3/4 = 75%(loại)
ACW => T	conf = 3/4 = 75%(loại)	CW => AT	conf = 3/5 = 60%(loại)
ATW => C	conf = 3/3 = 100%	TW => AC	conf = 3/3 = 100%
CTW => A	conf = 3/3 = 100%	A => CTW	conf = 3/4 = 75%(loại)
AC => TW	conf = 3/4 = 75%(loại)	C => ATW	conf = 3/4 = 75%(loại)
AT => CW	conf = 3/3 = 100%	T => ACW	conf = 3/4 = 75%(loại)
AW => CT	conf = 3/4 = 75%(loại)	W => ACT	conf = 3/60 = 100%

- {ACT} – Tập phổ biến thuộc L3
  Các tập con khác rỗng của tập phổ biến: {AC}, {AT}, {CT}, {A}, {C}, {T}

  Các luật có thể:

  AC => T conf = 3/4 = 75% (loại)

  AT => C conf = 3/3 = 66% (loại)

  CT => B conf = 3/4 = 75% (loại)

  A => CT conf = 3/4 = 75% (loại)

  C => AT conf = 3/6 = 50% (loại)

  T => AC conf = 3/4 = 75% (loại)

Làm tương tự với tất cả các tập phổ biến trong L3, L2 đề tìm ra tất cả các luật thõa ngưỡng minconf = 80%

Kết luận: các luật kết hợp thõa ngưỡng minsupp = 50% và minconf = 80% là

ACT =>W, ATW => C….

BÀI TẬP

Bài 1: Cho CSDL giao dịch bên dưới

1. Sử dụng thuật toán Apriori để tìm các tập phổ biến với minsupp = 22 %

3. Tìm tât cả các các luật kết hợp thõa minsupp=22% và

a. Minconf = 50 %

b. Minconf = 70%

TID	Items
100	M1, M2, M5
200	M2, M4
300	M2, M3
400	M1,M2, M4
500	M1, M3
600	M2, M3
700	M1, M3
800	M1, M2, M3, M5
900	M1, M2, M3

Bài 2:

Cho CSDL giao dịch sau và **minsupp = 60%, minconf = 70%**

a) Hãy sử dụng thuật toán **Apriori** để tìm tất cả các tập phổ biến .

b) Tìm các luật kết hợp được xây dựng từ các tập phổ biến tối đại thỏa mãn các ngưỡng minsupp, minconf đã cho

TID	Items
100	K, D, A, B, C, F
200	A, H, C, D
300	C, I , D, E, G, F
400	B, C, H, A, I, D, F ,G
500	F, C, K, E, G

Bài 3:

Cho CSDL sau

a. Hãy sử dụng thuật toán **Apriori** để tìm ***tất cả*** các tập phổ biến thỏa mãn ngưỡng **minsupp=60%.**

b. Tìm các luật kết hợp được xây dựng từ **tập phổ biến**, thỏa mãn ngưỡng **minconf =85%**.

TID	A	B	C	D	E	F	G	H	K	M	N
10		1			1	1	1	1			
20				1		1		1	1		
30	1		1	1		1		1		1	1
40	1	1			1			1	1	1	1
50	1	1					1	1		1	1