Luật kết hợp
Association rules

Đỗ Thanh Nghị
dtnghi@cit.ctu.edu.vn

Outline

- Giới thiệu
- Luật kết hợp
- Ứng dụng

Outline

- **Giới thiệu**
- Luật kết hợp
- Ứng dụng

Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Transaction

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

ITEMS:

A = milk
B= bread
C= cereal
D= sugar
E= eggs

Instances = Transactions

Transaction

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Attributes converted to binary flags

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Định nghĩa

- Item: *cặp thuộc tính* = *giá trị* hay *giá trị*
- Itemset I : tập của các items
  - ví dụ : I = {A,B,E} (thứ tự không quan trọng)
- Transaction: (TID, itemset)
  - TID là transaction ID

Support và Frequent Itemsets

- Support của itemset
  - sup(I) = số lượng của transactions t có chứa I
  - ví dụ : sup ({A,B,E}) = 2, sup ({B,C}) = 4
- Frequent itemset I *là* tập có support tối thiểu là minimum support
  - sup(I) >= *minsup*

Tính chất của subset

- **Mọi tập con của 1 frequent set là frequent!**
  - ví dụ : giả sử {A,B} là frequent, khi đó số lần xuất hiện của cả A,B là frequent => hiển nhiên là số lần xuất hiện của A hoặc B cũng frequent
- tất cả các giải thuật luật kết hợp đều dựa trên tính chất subset

Outline

- Giới thiệu
- **Luật kết hợp**
- Ứng dụng

Luật kết hợp

- Luật kết hợp *R* : *Itemset1* => *Itemset2*
  - *Itemset1, 2* không giao nhau và *Itemset2* không rỗng
  - ý nghĩa : nếu transaction có chứa *Itemset1* thì nó cũng chứa *Itemset2*
- ví dụ
  - A,B => E,C
  - A => B,C

Luật kết hợp

- *cho frequent set {A,B,E}, luật kết hợp có thể là*
  - A => B, E
  - A, B => E
  - A, E => B
  - B => A, E
  - B, E => A
  - E => A, B
  - __ => A,B,E (empty rule) hay true => A,B,E

khác nhau giữa luật phân lớp và luật kết hợp

luật phân lớp

- tập trung vào 1 thuộc tính target
- measures: accuracy

luật kết hợp

- nhiều thuộc tính target
- measures: support, confidence, Lift

Support và Confidence

- giả sử luật *R : I => J*
  - sup (R) = sup (I ∪ J)
    - support của itemset I ∪ J
  - conf (R) = sup(R) / sup(I) là *confidence* của luật R
- luật kết hợp có minimum support thường được cho là luật "***strong***"

Luật kết hợp

- ***cho frequent set {A,B,E}, luật kết hợp có minsup = 2 và minconf= 50%***

  A, B => E : conf=2/4 = 50%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Luật kết hợp

- ***cho frequent set {A,B,E}, luật kết hợp có minsup = 2 và minconf= 50%***

  A, B => E : conf=2/4 = 50%

  A, E => B : conf=2/2 = 100%

  B, E => A : conf=2/2 = 100%

  E => A, B : conf=2/2 = 100%

những luật không "tốt"

A =>B, E : conf=2/6 =33%< 50%

B => A, E : conf=2/7 = 28% < 50%

__ => A,B,E : conf: 2/9 = 22% < 50%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Tìm luật mạnh

- những luật có sup >= *minsup* và conf >= *minconf*
  - sup(R) >= *minsup* and conf (R) >= *minconf*
- tìm tất cả frequent itemsets

Tìm itemsets

- giải thuật Apriori (Agrawal & Srikant, 1993)
- ý tưởng : sử dụng tập 1-item để sinh ra tập 2-item, tập 2-item dùng để sinh ra tập 3-item, ...
  - nếu (A B) là frequent itemset, thì (A) và (B) phải là frequent itemsets
  - nếu X là frequent *k*-item set, thì tất cả (*k*-1)-item subsets của X cũng là frequent

  $\Rightarrow$ tính *k*-item set bằng cách merge (*k*-1)-item sets

Sinh luật kết hợp

- 2 bước :
  - xác định frequent itemsets với giải thuật Apriori
  - cho mỗi frequent itemset I
    - cho mỗi subset J của I
      - xác định tất cả các luật kết hợp : $I\text{-}J \Rightarrow J$
- ý tưởng chính : tính chất subset

ví dụ : sinh luật kết hợp từ Itemset

- Frequent itemset của tập weather :

Humidity = Normal, Windy = False, Play = Yes (4)

- 7 luật tiềm năng :

```
If Humidity = Normal and Windy = False then Play = Yes              4/4
If Humidity = Normal and Play = Yes then Windy = False              4/6
If Windy = False and Play = Yes then Humidity = Normal              4/6
If Humidity = Normal then Windy = False and Play = Yes              4/7
If Windy = False then Humidity = Normal and Play = Yes              4/8
If Play = Yes then Humidity = Normal and Windy = False              4/9
If True then Humidity = Normal and Windy = False and Play = Yes     4/12
```

Luật kết hợp cho weather

- luật có support > 1 và confidence = 100% :

	Association rule		Sup.	Conf.
1	Humidity=Normal Windy=False	⇒Play=Yes	4	100%
2	Temperature=Cool	⇒Humidity=Normal	4	100%
3	Outlook=Overcast	⇒Play=Yes	4	100%
4	Temperature=Cold Play=Yes	⇒Humidity=Normal	3	100%
...	...	...	...	...
58	Outlook=Sunny Temperature=Hot	⇒Humidity=High	2	100%

- 3 luật có support là 4, 5 luật có support bằng 3, và 50 luật có support là 2

Lọc luật kết hợp

- tập dữ liệu lớn => số luật sinh ra rất lớn mặc dù đã sử dụng Confidence và Support
- tìm cách lọc hay chọn lựa các luật hữu dụng : sử dụng các độ đo khác (tham khảo tài liệu của Howard Hamilton)
- mining luật kết hợp ☺

Outline

- Giới thiệu
- Luật kết hợp
- **Ứng dụng**

Ứng dụng

- Market basket analysis
  - Store layout, client offers
- Wal-Mart knows that customers who buy Barbie dolls have a 60% likelihood of buying one of three types of candy bars.
- What does Wal-Mart do with information like that? 'I don't have a clue,' says Wal-Mart's chief of merchandising, Lee Scott
- See - KDnuggets 98:01 for many ideas www.kdnuggets.com/news/98/n01.html
- Diapers and beer urban legend
- Finding unusual events
  - WSARE – What is Strange About Recent Events