KHAI PHÁ DỮ LIỆU

TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP

Mai Xuân Hùng

Nội dung

- Giới thiệu luật kết hợp
- Ứng dụng của luật kết hợp
- Bài toán về tập phổ biến và luật kết hợp
- Cách tìm tập phổ biến và luật kết hợp

Dạng luật kết hợp

- *Có 80% khách hàng mua bia thì sẽ mua thuốc lá*
- *Có 75 % khách hàng mùa quần tây thì sẽ mua áo sơ mi*
- *Có 87% khách hàng mua sữa hộp Minamilk thì mua trà Lipton*

Ứng dụng luật kết hợp

- Biết được xu hướng mua hàng của khách hàng
  - Có chiến lược bố trí hàng thích hợp
  - Dự tính lượng hàng nhập trong tương lai
- Phân tích dữ liệu giỏ hàng (bán hàng qua mạng)
  - Bố trí giao diện các mặt hàng.
  - Lọai bỏ, thêm mặt hàng.

Cách biểu diễn luật

- Khăn $\Rightarrow$ bia [0.5%, 60%]
- Mua:khăn $\Rightarrow$ mua:bia [0.5%, 60%]
  - Nếu mua khăn thì mua bia trong 60% trường hợp
  - Khăn và bia mua cùng 1 lúc là 0.5% dòng dữ liệu

Các thành phần trong luật

- Khăn ⇒ bia [0.5%, 60%]
  - Khăn: Vế trái
  - Bia: Mệnh đề kết quả
  - 0.5: **Support** tầng số (“trong bao nhiêu phần trăm dữ liệu thì những điều ở vế trái và vế phải cùng xảy ra")
  - 60%: **Confidence**, độ mạnh (“nếu vế trái xảy ra thì có bao nhiêu khả năng vế phải xảy ra")

Phát biểu bài toán

- Cho ngữ cảnh khai thác dữ liệu
  - ***O*** :Tập hữu hạn khác rỗng các hóa đơn.
  - ***I*** : Tập hữu hạn khác rỗng các mặt hàng.
  - ***R***: Quan hệ hai ngôi giữa ***O*** và ***I*** với o∈O và i∈I, (o,i)∈R⇔ hóa đơn ***o*** có chứa mặt hàng ***i***
  - Ngữ cảnh KTDL là bộ ba ***(O,I,R)***

Ví dụ ngữ cảnh khai thác dữ liệu

Mã hóa đơn	Mã hàng
o1	i1
o1	i2
o1	i3
o2	i2
o2	i3
o2	i4
o3	i2
o3	i3
o3	i4
o4	i1
o4	i2
o4	i3
o5	i3
o5	i4

Độ phổ biến

- Cho ngữ cảnh KTDL ***(O,I,R)*** và S ⊂ ***I***
- Độ phổ biến của S được định nghĩa là tỉ số giữa số các hóa đơn có chứa S và số lượng hoá đơn trong O
- Ký hiệu:

  $SP(S)=|\rho(S)| / |O|$

  ❖ $\rho(S)$ biểu diễn tập các hóa đơn có chung tất cả các mặt hàng trong S

Tập phổ biến

- Là những tập có độ độ phổ biến lớn hơn hoặc bằng 1 ngưỡng cho trước là minsupp

Các bước tìm tập phổ biến qua ví dụ

- Cho ngữ cảnh khai thác dữ liệu:

Mã hóa đơn	Mã hàng
o1	i1
o1	i2
o1	i3
o2	i2
o2	i3
o2	i4
o3	i2
o3	i3
o3	i4
o4	i1
o4	i2
o4	i3
o5	i3
o5	i4

❖Tìm tập phổ biến thỏa ngưỡng minsupp=0.4

Thành lập ma trận nhị phân

	i1	i2	i3	i4
o1	1	1	1	0
o2	0	1	1	1
o3	0	1	1	1
o4	1	1	1	0
o5	0	0	1	1

Tìm tập phổ biến thỏa ngưỡng

- Các tập ứng cử viên có 1 mặt hàng
  - F1={{i1},{i2},{i3},{i4}}
    - SP({i1})= 0,40 ; Phổ biến
    - SP({i2})= 0,80 ; Phổ biến
    - SP({i3})= 1,00 ; Phổ biến
    - SP({i4})= 0,60 Phổ biến
  - Tập phổ biến có 1 phần tử gồm C1={{i1},{i2},{i3},{i4}}

Tập phổ biến với mẹo Apriori

- **Bước kết hợp**: C_k được tạo bằng cách kết L_{k-1} với chính nó
- **Bước rút gọn**: Những tập kích thước (k-1) không phổ biến không thể là tập con của tập phổ biến kích thước k

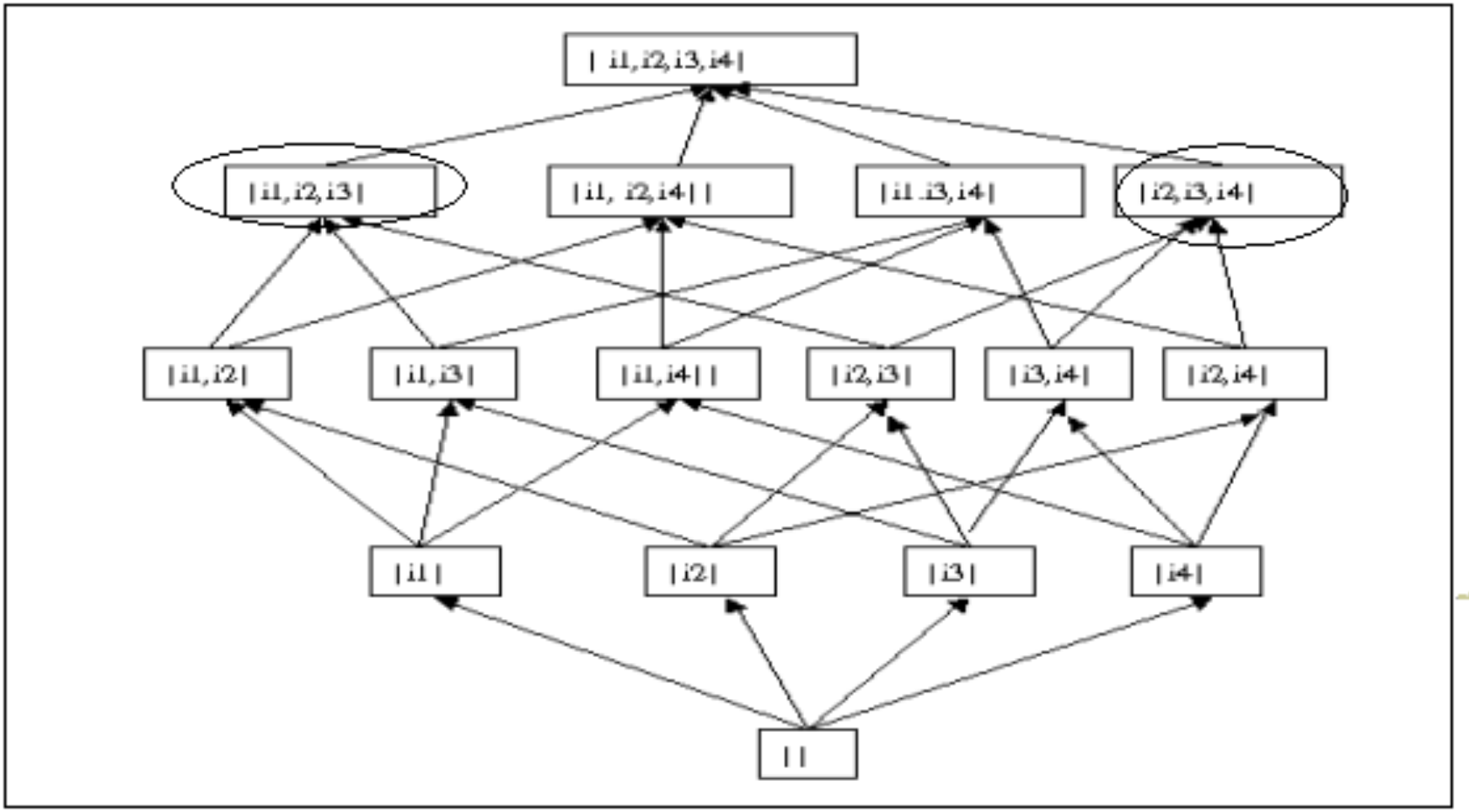
Tìm tập phổ biến thỏa ngưỡng (tt)

- Các tập ứng cử viên có 2 phần tử từ tập C1
  - F2={{i1,i2},{i1,i3},{i1,i4},{i2,i3},{i2,i4},{i3,i4}}
    - SP({i1,i2})= 0.4
    - SP({i1,i3})= 0.4
    - SP({i1,i4})= 0.0
    - SP({i2,i3})= 0.8
    - SP({i2,i4})= 0.4
    - SP({i3,i4})= 0.6
  - Các tập phổ biến có 2 phần tử
  - C2={{i1,i2}, {i1,i3}, {i2,i3}, {i2,i4}, {i3,i4}}

Tìm tập phổ biến thỏa ngưỡng (tt)

- Các tập ứng cử viên có 3 phần tử từ tập C2
  - F3={{i1,i2,i3}, {i1,i2,i4}, {i2,i3,i4},{i1,i3,i4}}
    - SP({i1,i2,i3})= 0,40;
    - SP({i2,i3,i4} = 0,40;
  - Các tập phổ biến có 3 phần tử C3={{i1,i2,i3}, {i2,i3,i4}
- Các tập phổ biến thỏa ngưỡng {i1}, {i2}, {i3}, {i4}, {i1,i2},{i1,i3},{i2,i3},{i2,i4} {i3,i4} {i1,i2,i3}, {i2,i3,i4}

Định nghĩa dàn tập các mặt hàng



Tìm tập phổ biến tối đại

- FS(O,I,R,minsupp) là tập phổ biến
- M được gọi là tập phổ biến tối đại nếu không tồn tại $S \in FS(O,I,R,minsupp)$, $M \neq S$, $M \subset S$
- Trong ví dụ trên tập phổ biến tối đại là: {i1,i2,i3}, {i2,i3,i4}.

Độ tin cậy của luật

- Độ tin cậy của luật kết hợp X →Y
  - Ký hiệu CF(X →Y)
  - CF(X →Y)=SP(S)/SP(X)
  - S=X ∪ **Y**
  - Luật kết hợp hợp lệ là những luật có
    - CF >= minconf

Tìm luật kết hợp thỏa độ tin cậy minconf

- Với ngữ cảnh KTDL trong ví dụ trên, ngưỡng minsupp=0.4
  - Và xét tập phổ biến tới đại {i1,i2,i3}
  - Thì luật r1: {i1,i2}→{i3}
  - Là một luật kết hợp hợp lệ theo ngưỡng minconf=0,67

Bài tập 1

- Cho bối cảnh gồm các giao tác : o1={d1,d3,d4} ; o2={d1,d3,d4}, o3={d3,d5}; o4={d4,d5} ; o5 = {d2,d3,d5}
- Tìm các tập phổ biến tối đại minsupp=0,3
- Liệt kê 1 số luật thảo ngưỡng minconfidence =1.0

Bài tập 2

- **Cho bối cảnh khai thác dữ liệu gồm**

  o1 = {i1, i3, i4, i6}, o2 = { i1, i3, i6}

  o3 = {i3, i5, i6}, o4 = {i1, i2, i4, i5}

  o5 = {i2, i4, i6}, o6 = {i1, i2, i4, i5, i6}

  - Tìm Các tập phổ biến tối đại theo ngưỡng minsupp = 0.3
  - Các luật kết hợp từ tập phổ biến tối đại theo ngưỡng minconf = 1.0

Tìm tập phổ biến bằng thuật giải không tăng cường

- Các khái niệm
- Minh họa thuật toán

Vector biểu diễn tập mặt hàng

- Cho ma trận M là ma trận biểu diễn ngữ cảnh khai thác dữ liệu ***(O,I,R)***.
- Gọi m=|O| và cho tập mặt hàng S ⊂ I, vector biểu diễn tập mặt hàng S ký hiệu là v(S) là vector nhị phân có m thành phần, thành phần thứ i của vector v(S) có trị 1 nếu hóa đơn o_i có chứa tất cả các mặt hàng trong S, hoặc giá trị của các phần tử nằm trên dòng i và trên các cột ứng với các mặt hàng trong S đều có giá trị 1, ngược sẽ có trị 0

Ví dụ Vector biểu diễn tập mặt hàng

	i1	i2	i3	i4
o1	1	1	1	0
o2	0	1	1	1
o3	0	1	1	1
o4	1	1	1	0
o5	0	0	1	1

S = $\{i_2, i_3\}$ thì vector biểu diễn v(S) = { 1, 1, 1, 1, 0 }

Tích vector biểu diễn

- Cho **S** $\subset$ ***I*** và T $\subset$ ***I***, gọi v(S)=(s_1,...,s_m) và v(T)=(t_1,...,t_m) lần lượt là các vector biểu diễn của S và T.
- Tích vector biểu diễn của vector v(S) và v(T) ký hiệu là $\otimes$ là vector biểu diễn v(Z)=(z_1,…,z_m) với z_k=min (s_k, t_k), k=1, …,m.

Độ phổ biến của vector biểu diễn

- Cho S $\subseteq$ ***I***, độ phổ biến của vector biểu diễn v(S) được ký hiệu là SPV(v(S)) là tỉ số giữa số thành phần khác 0 trong v(S) và số thành phần của v(S) hay số dòng của ma trận biểu diễn ngữ cảnh khai thác dữ liệu.
- SPV(v(S)) = SP(S)

Ví dụ

- Tính SPV(v(S)) với S = {i_2 , i_3} thì v(S) = {1,1,1,1,0} vì:
  - S_1 = {i_2 } thì v(S_1) = {1,1,1,1,0} và S_2 = {i_3} thì v(S_2) = {1,1,1,1,1}, do vậy:
  - v(S) = v(S_1) ⊗ v(S_2) = {1,1,1,1,0} và SP(S)=SPV(v(S)) = 0,8

Ví dụ minh họa thuật toán

- Cho ngữ cảnh KTDL ***(O,I,R)*** trong bảng sau:

	i1	i2	i3	i4
o1	1	1	1	0
o2	0	1	1	1
o3	0	1	1	1
o4	1	1	1	0
o5	0	0	1	1

- Ngưỡng phổ biến tối thiểu *minsupp*=0,4. Các bước của thuật toán tìm tập phổ biến như sau

Tính F1={ {i1},{i2},{i3},{i4}}

- {i1}, v({i1})= (1, 0, 0, 1, 0)
- SP({i1})=SPV(v({i1}))= 2/5= 0,40 –Tập phổ biến
- {i2}, v({i2})= (1, 1, 1, 1, 0)
- SP({i2}=SPV(v({i2}))=4/5=0,80 - Tập phổ biến
- {i3}, v({i3})= (1, 1, 1, 1, 1)
- SP({i3})=SPV(v({i3}))=1,00 – Tập phổ biến
- {i4}, v({i4})= (0, 1, 1, 0, 1,)
- SP({i4})=SPV(v({i4}))= 0,60 – Tập phổ biến

F2={{i1,i2},{i1,i3},{i2,i3}, {i2, i4}, {i3, i4}}

- {i1,i2},v({i1,i2})= v({i1}) ⊗ v({i2})=(1, 0, 0, 1, 0)
- SP({i1,i2}) = SPV(v({i1,i2}))=2/5=0,40– Tập phổ biến
- {i1,i3}, v({i1,i3})= v({i1})⊗v({i3})=(1, 0, 0, 1, 0)
- SP({i1,i3}) = SPV(v({i1,i3}))=2/5=0,40–Tập phổ biến
- {i1,i4}, v({i1,i4})= v({i1})⊗v({i4})=(0, 0, 0, 0, 0)
- SP({i1,i4}) = SPV(v({i1,i4}))=0/5=0,0– Không phải tập phổ biến

Tính F2 (tt)

- {i2,i3}, v({i2,i3})= v({i2})⊗v({i3})=(1, 1, 1, 1, 0)
  - SP({i2,i3}) = SPV(v({i2,i3}))=4/5=0,80– Tập phổ biến
- {i2,i4}, v({i2,i4})= v({i2})⊗v({i4})=(0, 1, 1, 0, 0)
  - SP({i2,i4}) = SPV(v({i2,i4}))=2/5=0,40– Tập phổ biến
- {i3,i4}, v({i3,i4})=v({i3})⊗v({i4})= (0, 1, 1, 0, 1)
  - SP({i3,i4}) = SPV(v({i3,i4}))=3/5=0,80–Tập phổ biến

F3= {{i1,i2,i3} , { i2,i3,i4}}

- {i1,i2,i3 }, v({i1,i2,i3})= v({i1,i2})⊗v({i1,i3})= (1, 0, 0, 1, 0)
- SP({i1,i2,i3}) = SPV(v({i1,i2,i3}))=2/5=0,40– Tập phổ biến
- {i1,i2,i4 }, v({i1,i2,i4})= v({i1,i2})⊗v({i2,i4})= (0, 0, 0, 0, 0)
- SP({i1,i2,i4}) = SPV(v({i1,i2,i4}))=0/5=0,0– Không phải tập phổ biến

Tính F3 (tt)

- {i1,i3,i4}, v({i1,i3,i4)= v({i1,i3})⊗v({i3,i4})=(0, 0, 0, 0, 0)
  - SP({i1,i3,i4}) = SPV(v({i1,i3,i4}))=0/5=0,0– Không phải tập phổ biến
- {i2,i3,i4}, v({i2,i3,i4})= v({i2,i3})⊗ v({i3,i4})= (0, 1, 1, 0, 0)
  - SP({i2,i3,i4}) = SPV(v({i2,i3,i4}))=2/5=0,40– Tập phổ biến

F4= {{i1,i2,i3,i4} }

- {i1,i2, i3, i4},v({ i1,i2,i3,i4})= v({i1,i2,i3})⊗v({i2,i3,i4})= (0, 0, 0, 0, 0)
- SP({i1,i2,i3,i4}) = SPV(v({i1,i2,i3,i4}))=0/5=0,0– Không phải là tập phổ biến

Kết thúc thuật toán 2.2 . Kết quả

FS(O,I,R,minsupp=0.4) = F1 ∪ F2 ∪ F3

FS(O,I,R,minsupp=0,4)={ {i1},{i2},{i3},{i4}, {i1,i2}, {i1,i3},{i2,i3}, {i2, i4}, {i3, i4},{i1,i2,i3}, {i2,i3,i4} }