TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN **TÀI LIỆU HƯỚNG DẪN THỰC HÀNH**
KHOA HỆ THỐNG THÔNG TIN Môn: Khai thác dữ liệu

Bài thực hành 2
TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP

I. Mục tiêu

1. Hiểu và vận dụng giải thuật Apriori để tìm các tập phổ biến.
2. Hiểu và vận dụng kỹ thuật tìm tập phổ biến tối đại
3. Hiểu và vận dụng kỹ thuật tìm luật kết hợp tính độ tin cậy của luật

II. Thời gian

1. Thực hành: 5 tiết
2. Bài tập làm thêm: 8 tiết

III. Hướng dẫn chung

Cho tập các hoá đơn O={o1, o2, o3, o4, o5} và tập các mặt hàng I={i1,i2,i3,i4,i5}

Mỗi hóa đơn mua các mặt hàng như sau:

o1={i1,i3,i4} ; o2={i1,i3,i4} ; o3={i3,i5} ; o4={i4,i5} ; o5={i2,i3,i5}

Yêu cầu:

Với min_sup = 0.4 và min_conf = 0.8

1. Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật Apriori.
2. Tìm tập phổ biến tối đại
3. Tìm và tính độ tin cậy cho các luật kết hợp được liệt kê từ tập phổ biến tối đại

Hướng dẫn:

1. Tìm tập phổ biến bằng giải thuật Apriori:

- Lập ma trận nhị phân từ các giao tác mua hàng

	i1	i2	i3	i4	i5
o1	1	0	1	1	0
o2	1	0	1	1	0
o3	0	0	1	0	1
o4	0	0	0	1	1
o5	0	1	1	0	1

- Với min_sup là 0.4 và tổng số giao dịch là 5 => tần số xuất hiện tối thiểu của phần tử để thỏa min_sup (min_sup count) là 2.
- Tập các ứng viên 1 phần tử F1={i1},{i2},{i3},{i4},{i5}

- Tần số xuất hiện của từng mặt hàng (support count)

  SP(s) = Số hóa đơn chứa s / Tổng số hóa đơn

  SP({i1}) =2/5 (phổ biến)

  SP({i2}) =1/5 (không phổ biến)

  SP({i3}) =4/5 (phổ biến)

  SP({i4}) =3/5 (phổ biến)

  SP({i5}) =3/5 (phổ biến)

- Tập phổ biến có 1 mặt hàng C1={ {i1} , {i3} , {i4}, {i5}}
- Tập ứng cử viên có 2 mặt hàng từ tập phổ biến có 1 mặt hàng:

	{i1}	{i3}	{i4}	{i5}
{i1}	ϕ	{i1,i3}	{i1,i4}	{i1,i5}
{i3}		ϕ	{i3,i4}	{i3,i5}
{i4}			ϕ	{i4,i5}
{i5}				ϕ

F2={ {i1,i3} ,{i1,i4},{i1,i5},{i3,i4},{i3,i5},{i4,i5} }

- Tính độ phổ biến của các tập trong F2

  SP({i1,i3}) =2/5 (phổ biến)

  SP({i1,i4}) =2/5 (phổ biến)

  SP({i1,i5}) =0/5 (Không phổ biến)

  SP({i3,i4}) =2/5 (phổ biến)

  SP({i3,i5}) =2/5 (phổ biến)

  SP({i4,i5}) =1/5 (Không phổ biến)

- Tập phổ biến có 2 mặt hàng C2 ={ {i1,i3} , {i1,i4} , {i3,i4} , {i3,t5} }
- Tập ứng cử viên có 3 mặt hàng từ tập phổ biến có 2 mặt hàng:

	{i1,i3}	{i1,i4}	{i3,i4}	{i3,i5}
{i1,i3}	ϕ	{i1,i3,i4}	{i1,i3,i4}	{i1,i3,i5}
{i1,i4}		ϕ	{i1,i3,i4}	ϕ
{i3,i4}			ϕ	{i3,i4,i5}
{i3,i5}				ϕ

F3={ {i1,i3,i4}, {i1,i3,i5} , {i3,i4,i5} }

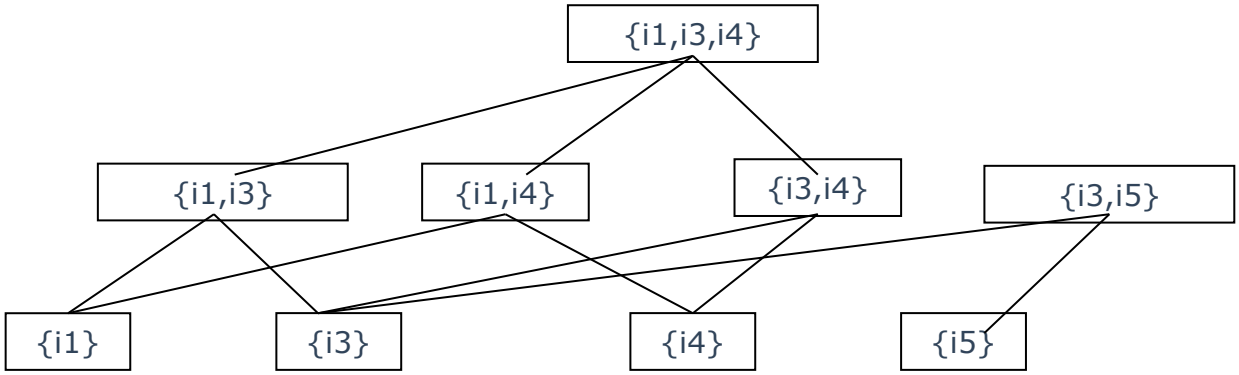
- Áp dụng giải thuật Apriori - nếu tập con không phổ biến thị tập mẹ chứa tập con đó không phổ biến. Nên tập {i1,i3,i5} không phổ biến vì tập {i1,i5} không phổ biến, tập {i3,i4,i5} không phổ biến vì tập {i4,i5} không phổ biến
- Tính độ phổ biến của các tập trong F3

  SP({i1,i3,i4}) = 2/5 (phổ biến)
- Vậy tập phổ biến thỏa ngưỡng là:

  C={ {i1}, {i3}, {i4}, {i5}, {i1,i3 } ,{i1,i4 } , {i3,i4 } ,{i3,i5 } , {i1,i3,i4} }

2. Tìm tập phổ biến tối đại từ tập phổ biến được liệt kê ở câu 1

- Từ tập phổ biến ở câu 1 ta xác định được ***dàn*** từ các tập phổ biến như sau



- Tập phổ biến tối đại là: {i3,i5} , {i1,i3,i4}

3. Tìm các luật kết hợp dựa trên các tập phổ biến tối đại

- Tạo các luật từ tập phổ biến tìm được.
- Tính confidence của các luật kết hợp
- So sánh với min_confidence để tìm các luật thỏa yêu cầu
- Xét luật F(X->Y)

  CF(X->Y) = SP(X ∪ Y) / SP(X)

  Ví dụ: Xét tập phổ biến { i1,i3,i4} có có các tập con không rỗng sau: {i1}, {i3}, {i4}, {i1, i3}, {i1, i4}, {i3, i4}

R1	{i1} -> {i3,i4}	CF(R1) = SP({i1,i3,i4})/SP({i1})= 2/5 /2/5=1
R2	{i3} -> {i1,i4}	CF(R2) = SP({i1,i3,i4})/ SP({i3})= 2/5 /4/5=1/2<0.5 loại
R3	{i4} -> {i1,i3}	CF(R3) = SP({i1,i3,i4})/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại
R4	{i3,i4} -> {i1}	CF(R4) = SP({i1,i3,i4})/ SP({i3,i4})= 2/5 /2/5=1 > 0 .8
R5	{i1,i4} -> {i3}	CF(R5) = SP({i1,i3,i4})/ SP({i1,i4})= 2/5 /2/5=1 > 0 .8
R6	{i1,i3} -> {i4}	CF(R6) = SP({i1,i3,i4})/ SP({i1,i3})= 2/5 /2/5=1 > 0 .8

Với min_conf = 0.8, dựa vào bảng trên ta có các luật kết hợp thỏa ngưỡng là:

R1, R4, R5, R6

IV. Thực hành

1. Cho bảng dữ liệu ở một cửa hàng tạp hóa có 6 giao dịch như sau

Transaction ID	Items
T1	Xúc xích, Bánh bao, Sốt cà chua
T2	Xúc xích, Bánh bao
T3	Xúc xích, Coca, Khoai tây chiên
T4	Khoai tây chiên,Coca
T5	Khoai tây chiên, Sốt cà chua
T6	Xúc xích, Coca, Khoai tây chiên

Với min_sup = 0.3334 và min_conf = 0.6, sinh viên thực hiện lại các yêu cầu trên.

2. Cho bảng dữ liệu ở một cửa hàng văn phòng phẩm như sau:

TID	KÉO	COMPA	THƯỚC	TẬP TRẮNG	BÚT BI	BÚT MÀU	TẤY
T1		x		x	x		
T2	x		x	x	x		
T3		x		x	x		
T4	x	x		x	x		
T5			x				
T6					x		
T7				x			
T8							x
T9						x	x
T10						x	

Với min_sup = 0.3 và min_conf = 0.8, sinh viên thực hiện lại các yêu cầu trên.

3. CSDL về Nhân viên được cho trong bảng sau:

	Giới tính (GT)	Tuổi (T)	Năng lực làm việc (NL)	Đã lập gia đình (LGD)	Thu nhập (TN)	Thăng chức (TC)
1	Nữ	20..25	Giỏi	Rồi	Rất cao	Có
2	Nam	20..25	Khá	Chưa	Khá	Không
3	Nữ	26..30	Giỏi	Chưa	Khá	Có
4	Nữ	31..40	T.Bình	Chưa	T.Bình	Có
5	Nam	26..30	T.Bình	Rồi	Rất cao	Không
6	Nữ	26..30	Khá	Chưa	Cao	Không
7	Nữ	31..40	Khá	Chưa	T.Bình	Không
8	Nam	26..30	Khá	Rồi	Cao	Có
9	Nữ	>40	Giỏi	Rồi	T.Bình	Không
10	Nữ	26..30	Giỏi	Chưa	Khá	Có

Cho B = {Tuổi, Năng lực làm việc, Thăng chức}. Hãy tìm tất cả các luật kết hợp có vế phải chỉ gồm thuộc tính Thăng chức (TC) thỏa ngưỡng minsup=0.3 và minconf=0.8

V. Bài tập thêm

1. Trong các phương pháp tìm kiếm luật kết hợp trên ta sử dụng hai giá trị min_sup và min_conf để đánh giá các luật tìm được. Tuy nhiên trong thực tế, nếu chỉ sử dụng hai giá trị này thì mô hình vẫn có thể sinh ra một số luật phi lí. Vì thế để giới hạn vấn đề này ta có thể bổ sung thêm một giá trị để đánh giá luật kết hợp đó là tính tương quan giữa hai vế của luật.

   Sinh viên tìm hiểu hai phương pháp phân tích tính tương quan giữa hai vế của luật sử dụng giá trị **Lift** và χ^2 và sử dụng để đánh giá các luật tìm được ở phần thực hành.
2. Chọn một ngôn ngữ lập trình, cài đặt giải thuật Apriori.
3. Chọn 1 trong các kĩ thuật sau: Dùng bảng băm, giảm số lượng giao dịch trong tập giao dịch, chia nhỏ tập giao dịch và lấy mẫu trên tập giao dịch để cải tiến giải thuật Apriori đã viết trong câu 1.
4. Cho mẫu dữ liệu[1] về các giao dịch trong một tháng của một của hàng outlet, gồm 9835 giao dịch và 169 items.
   a) Sử dụng kết quả lập trình ở câu 1 và 2 tìm tất cả các luật kết hợp bằng thuật toán Apriori (sinh viên tự chọn min_sup và min_conf).
   b) Sử dụng các thư viên trong R, Python hoặc Weka, tìm tất cả các luật kết hợp từ mẫu dữ liệu.

VI. Tài liệu tham khảo

1. *Groceries Dataset*, [Michael Hahsler et al., 2006] Michael Hahsler, Kurt Hornik, and Thomas Reutterer (2006) Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization*, pages 598–605. Springer-Verlag.;