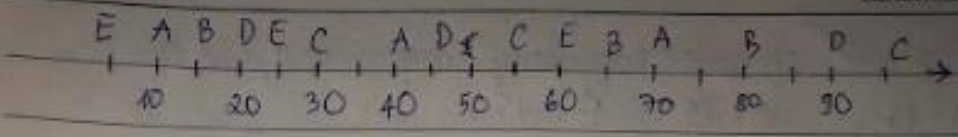
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHAI THÁC DỮ LIỆU

BÀI TẬP LAP 3

NGUYỄN THỊ NGỌC HÀ: 17520421



min_frequent = 30%, min_conf = 60%

width = 25 và step = 5

a Dãy phổ biến song song bằng thuật toán WINEPI

- Thuật toán này sử dụng cửa sổ với độ rộng cố định để trượt qua chuỗi sk ...

$$N = \frac{T_e - T_s + width}{step} - 1 =$$

$$\Rightarrow N = \frac{100 - 5 + 25}{5} - 1 = 23$$

Các cửa sổ

U_1 : -, -, -, -, E

U_2 : -, -, -, E, A

U_3 : -, -, E, A, B

U_4 : -, E, A, B, D

U_5 : E, A, B, D, E

U_6 : A, B, D, E, C

U_7 : B, D, E, C, -

U_8 : D, E, C, -, A

U_9 : E, C, -, A, D

U_{10} : C, -, A, D, -

U_{11} : -, A, D, -, C

U_{12} : A, D, -, C, E

U_{13} : D, A, -, C, E, B

U_{14} : -, C, E, B, A

U_{15} : C, E, B, A, -

U_{16} : E, B, A, -, B

U_{17} : B, A, -, B, -

U_{18} : A, -, B, -, D

U_{19} : -, B, -, D, C

U_{20} : B, -, D, C, -

U_{21} : -, D, C, -, -

U_{22} : D, C, -, -, -

U_{23} : C, -, -, -, -

Tập các episode song song ứng viên 1 phần tử

A: 15 C: 15 E: 15

B: 13 D: 15

Với số cửa sổ là 23 và min_frequent = 30%, ta có các dãy song song 1 phần tử phổ biến sau:

$L_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}\}$

- Tập ứng viên có 2 phần tử là:

$C_2 = L_1 \bowtie L_1 = \{AB: 9, AC: 8, AD: 9, AE: 11$
$BC: 7, BD: 8, BE: 9, CD: 12, CE: 8, DE: 8\}$

⇒ Các dãy phổ biến 2 phần tử thỏa min_frequent:

$L_2 = \{\{A,B\}, \{A,C\}, \{A,D\}, \{A,E\}, \{B,C\}, \{B,D\},$
$\{B,E\}, \{C,D\}, \{C,E\}, \{D,E\}\}$

- Tập ứng viên có 3 phần tử là:

$C_3 = L_2 \bowtie L_2 = \{ABC: 3, ABD: 4, ABE: 7, ACD: 6,$
$ACE: 6, ADE: 6, BCD: 5, BCE: 5, BDE: 5,$
$CDE: 6\}$

⇒ Các dãy phổ biến 3 phần tử thỏa min_frequent:

$L_3 = \{\{A,B,E\}\}$

- Tập ứng viên có 4 phần tử:

$C_4 = L_3 \bowtie L_3 = \emptyset$

Đến đây thuật toán tìm dãy phổ biến dừng lại và kq là các dãy L_1, L_2, L_3 thu được.

b. Dãy phổ biến tuần tự bằng thuật toán WINEPI

- Sử dụng dữ liệu của sổ trượt và dãy phổ biến L_1 đã làm ở câu a.
- Tập các ứng viên có 2 phần tử là:

$C_2 = L_1 \bowtie L_1 = \{$ AB:7, BA:4, AC:3, CA:5, AD:8, DA:1, AE:3, EA:9, BC:4, CB:3, BD:7, DB:1, BE:3, EB:7, CD:2, DC:10, CE:4, EC:4, DE:6, ED:3 $\}$

⇒ Các dãy phổ biến 2 phần tử thỏa min frequent:

$L_2 = \{\{A,B\}, \{A,D\}, \{E,A\}, \{B,D\}, \{E,B\}, \{D,C\}\}$

- Tập các ứng viên có 3 phần tử là:

$C_3 = L_2 \bowtie L_2 = \{$ ABD:4, EAB:3 $\}$

⇒ Không có dãy phổ biến 3 phần tử thỏa min-frequent

- Đến đây thuật toán tìm dãy phổ biến dừng lại và kết quả là các dãy L_1, L_2 thu được

c. Các luật WINEPI song song dựa trên dãy phổ biến tìm được ở câu a là:

Xét tập phổ biến ABE, tính confidence của các luật tạo ra:

A ⇒ ABE có $conf = \frac{fre(ABE)}{fre(A)} = \frac{7}{15} = 0,46 < min\_conf = 0,8$ (1)

B ⇒ ABE có $conf = \frac{fre(ABE)}{fre(B)} = \frac{7}{13} = 0,54 < min\_conf = 0,6$ (2)

E ⇒ ABE có $conf = \frac{fre(ABE)}{fre(E)} = \frac{7}{15} = 0,46 < min\_conf = 0,6$ (3)

Date:

(4) AB ⇒ ABE có $conf = \frac{fre(ABE)}{fre(AB)} = \frac{7}{9} = 0,77 > min\_conf = 0,6$

(5) BE ⇒ ABE có $conf = \frac{fre(ABE)}{fre(BE)} = \frac{7}{9} = 0,77 > min\_conf = 0,6$

(6) AE ⇒ ABE có $conf = \frac{fre(ABE)}{fre(AE)} = \frac{7}{11} = 0,64 > min\_conf = 0,6$

Như vậy, ta nhận các luật (4), (5), (6)

d. Các luật WINEPI tuần tự dựa trên dãy phổ biến tìm được

(1) A ⇒ AB có ~~conf~~ [0,30, 0,46]

(2) A ⇒ AD [0,35, 0,53]

(3) A ⇒ EA [0,39, 0,60]

(4) B ⇒ AB ~~[0,17, 0,3]~~ [0,3, 0,54]

(5) D ⇒ AD [0,35, 0,53]

(6) E ⇒ EA [0,39, 0,6]

(7) B ⇒ BD [0,3, 0,54]

(8) D ⇒ BD [0,3, 0,47]

(9) E ⇒ EB [0,3, 0,46]

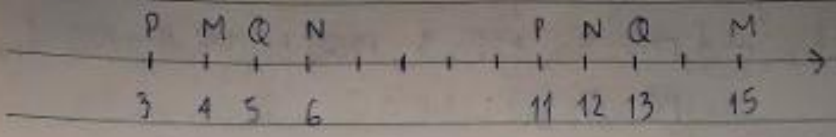
(10) B ⇒ EB [0,3, 0,54]

(11) D ⇒ DC [0,43, 0,66]

(12) C ⇒ DC [0,43, 0,66]

Vậy ta nhận các luật (3), (6), (11), (12)

2.



min_frequent = 20%, min_conf = 50%, width = 4, skip = 1

a. Các dãy phổ biến song song bằng thuật toán WINEPI

$$N = \frac{T_e - T_s + width}{step} - 1 = \frac{16 - 3 + 4}{1} - 1 = 16$$

Các cửa sổ:

U_1: -, -, -, P
U_2: -, -, P, M
U_3: -, P, M, Q
U_4: P, M, Q, N
U_5: M, Q, N, -
U_6: Q, N, -, -
U_7: N, -, -, -
U_8: -, -, -, -
U_9: -, -, -, P
U_{10}: -, -, P, N
U_{11}: -, P, N, Q
U_{12}: P, N, Q, -
U_{13}: N, Q, -, M
U_{14}: Q, -, M, -
U_{15}: -, M, -, -
U_{16}: M, -, -, -

- Tập các episode song song 1 phần

M: 8 N: 8 P: 8 Q: 8

⇒ Với 16 cửa sổ và min_frequent = 20%, ta tìm ra các dãy song song 1 phần phổ biến sau:

$L_1 = \{\{M\}, \{N\}, \{P\}, \{Q\}\}$

Date:

• Tập các ứng viên có 2 phần tử là:

$C_2 = L_1 \bowtie L_1 = \{MN:3, MP:3, MQ:5, NP:4, NQ:6, PQ:4\}$

⇒ Các dãy phổ biến 2 phần tử thỏa min_frequent:

$L_2 = \{\{M,Q\}, \{N,P\}, \{N,Q\}, \{PQ\}\}$

• Tập các ứng viên có 3 phần tử là:

$C_3 = L_2 \bowtie L_2 = \{NPQ:3\}$

⇒ Không có dãy phổ biến 3 phần tử thỏa min_frequent.

• Đến đây, thuật toán tìm dãy phổ biến dừng lại và kết quả là các dãy L_1, L_2 thu được.

b. Các dãy phổ biến tuần tự bằng thuật toán WINEPI

• Sử dụng dữ liệu cửa sổ trượt và dãy phổ biến L_1 đã làm ở câu A.

• Tập các ứng viên có 2 phần tử là.

$C_2 = L_1 \bowtie L_1 = \{MN:2, NM:0, MP:0, PM:3, MQ:3, QM:2, NP:0, PN:4, NQ:3, QN:3, PQ:4, QP:0\}$

⇒ Các dãy phổ biến 2 phần tử thỏa min_frequent:

$L_2 = \{\{PN\}, \{P,Q\}\}$

• Tập các ứng viên có 3 phần tử

$C_3 = L_2 \bowtie L_2 = \emptyset$

⇒ Không có dãy phổ biến 3 phần tử

• Đến đây thuật toán tìm dãy phổ biến dừng lại và kết quả là các dãy L_1, L_2 thu được.

c. Các luật WINEPI song song dựa trên dãy phổ biến tìm được ở câu a.

M ⇒ MQ [0,31 ; 0,63] (1)

Q ⇒ MQ [0,31, 0,63] (2)

N ⇒ NP [0,25, 0,5] (3)

P ⇒ NP [0,25, 0,5] (4)

N ⇒ NQ [0,38, 0,75] (5)

Q ⇒ NQ [0,38, 0,75] (6)

P ⇒ PQ [0,25, 0,5] (7)

Q ⇒ PQ [0,25, 0,5] (8)

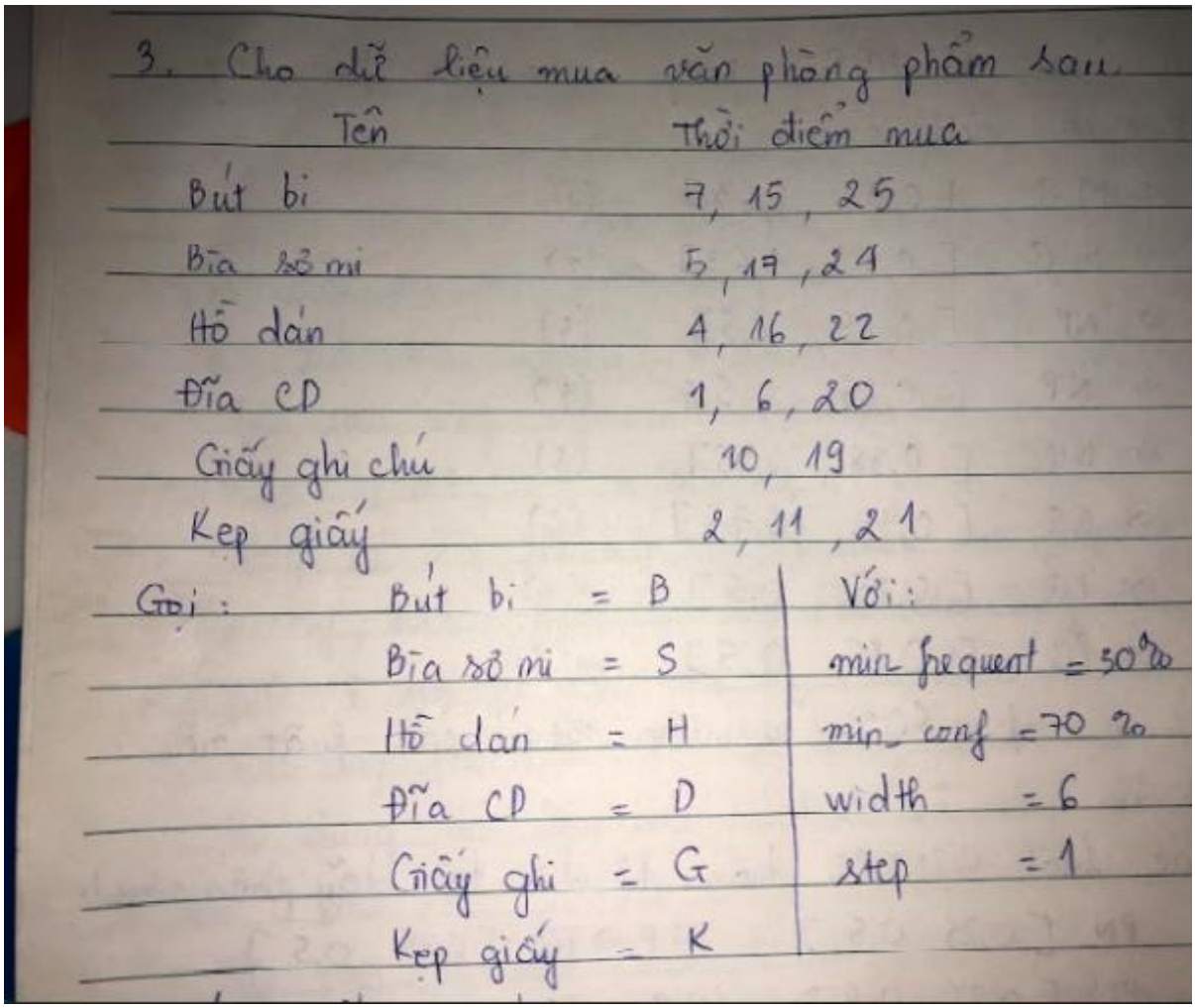
Với min_conf = 50%, ta nhận tất cả các luật nêu trên.

d. Các luật WINEPI tuần tự dựa trên dãy phiên câu b.

(1) P ⇒ PN [0,25, 0,5] (3) P ⇒ PQ [0,25, 0,5]

(2) N ⇒ PN [0,25, 0,5] (4) Q ⇒ PQ [0,25, 0,5]

Với min_conf = 50%, ta nhận tất cả các luật nêu trên.



3. Cho dữ liệu mua văn phòng phẩm sau.

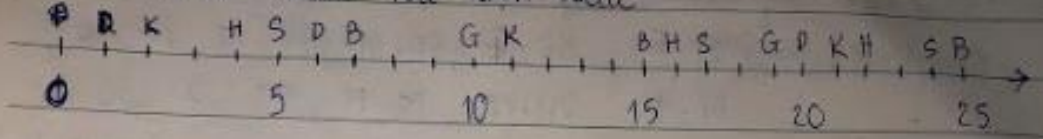
Tên	Thời điểm mua
Bút bi	7, 15, 25
Bìa sơ mi	5, 17, 24
Hồ dán	4, 16, 22
Đĩa CD	1, 6, 20
Giấy ghi chú	10, 19
Kẹp giấy	2, 11, 21

Gọi: Bút bi = B; Bìa sơ mi = S; Hồ dán = H; Đĩa CD = D; Giấy ghi = G; Kẹp giấy = K

Với: min_frequent = 50%; min_conf = 70%; width = 6; step = 1

Date:

Ta có chuỗi sự kiện sau



$$N = \frac{T_e - T_s + width}{Step} - 1 = 30$$

U_1: -,-,-,-,-,D
U_2: -,-,-,-,D,K
U_3: -,-,-,D,K,-
U_4: -,-,D,K,-,H
U_5: -,D,K,-,H,S
U_6: D,K,-,H,S,D
U_7: K,-,H,S,D,B
U_8: -,H,S,D,B,-
U_9: H,S,D,B,-,-
U_{10}: S,D,B,-,-,G
U_{11}: D,B,-,-,G,K
U_{12}: B,-,-,G,K,-
U_{13}: -,-,G,K,-,-
U_{14}: -,G,K,-,-,-
U_{15}: G,K,-,-,-,B
U_{16}: K,-,-,-,B,H
U_{17}: -,-,-,B,H,S
U_{18}: -,-,B,H,S,-
U_{19}: -,B,H,S,-,G
U_{20}: B,H,S,-,G,D
U_{21}: H,S,-,G,D,K
U_{22}: S,-,G,D,K,H
U_{23}: -,G,D,K,H,-
U_{24}: G,D,K,H,-,S
U_{25}: D,K,H,-,S,B
U_{26}: K,H,-,S,B,-
U_{27}: H,-,S,B,-,-
U_{28}: -,S,B,-,-,-
U_{29}: S,B,-,-,-,-
U_{30}: B,-,-,-,-,-

- Tập các episodes song song 1 phí:

$C_1 = \{B:18, D:17, K:18, H:18, S:18, G:12\}$

Với 30 của số và min_frequent = 30%, ta tìm ra các dãy phổ biến song song 1 phí sau:

$L_1 = \{\{B\}, \{D\}, \{K\}, \{H\}, \{S\}, \{G\}\}$

• Tập các ứng viên có 2phần tử là:

$C_2 = L_1 \bowtie L_1 = \{$ BD: 5, BK: 7, BA: M, BS: 12, BG: 6,
DK: 12, DH: 12, DS: 11, DG: 7, SG: 6
KH: 11, KS: 8, KG: 9, HS: 15, HG: 6 $\}$

⇒ Các dãy phổ biến 2phần tử thỏa min-frequent:

$L_2 = \{ \{B,H\}, \{B,S\}, \{D,K\}, \{D,H\}, \{D,S\},$
$\{K,H\}, \{K,G\}, \{H,S\} \}$

• Tập các ứng viên có 3phần tử là:

$C_3 = L_2 \bowtie L_2 = \{$ DKH: 9, BHS: 10, DHS: 10 $\}$

⇒ Các dãy phổ biến 3phần tử thỏa min-frequent:

$L_3 = \{ \{D,K,H\}, \{B,H,S\}, \{D,H,S\} \}$

• Tập các ứng viên có 4phần tử là

$C_4 = L_3 \bowtie L_3 = \emptyset$

⇒ Không có dãy phổ biến 4phần tử

• Đến đây thuật toán tìm dãy phổ biến dừng lại và kết quả là L_1, L_2, L_3 thu được.

b. Các dãy phổ biến tuần tự bằng thuật toán WINEPI

• Sử dụng dữ liệu cửa sổ trượt và dãy phổ biến L_1 đã làm ở câu a

• Tập các ứng viên có 2phần tử là:

$C_2 = L_1 \bowtie L_1 = \{$ BD: 1, DB: 6, BK: 2, KB: 5, BH: 5, HB: 6,
BS: 4, SB: 9, BG: 5, GB: 1, DK: 11, KD: 2,
DH: 7, HD: 6, DS: 4, SD: 8, DG: 2, GD: 5,
SG: 5, GS: 1, KH: 10, HK: 1, KS: 6, SK: 2,
KG: 0, GK: 9, HS: 14, SH: 1, HG: 3, GH 3 $\}$

⇒ Các dãy phổ biến 2 phần tử thỏa min-sup
$L_2 = \{\{S,B\}, \{D,K\}, \{K,H\}, \{G,K\}, \{H,S\}\}$

• Tập các ứng viên 3 phần tử
$C_3 = L_2 \bowtie L_2 = \phi$

⇒ Không có dãy phổ biến 3 phần tử

• Đến đây thuật toán tìm dãy phổ biến dừng lại và kết quả là L_1, L_2 thu được.

c. Các luật WINEPI song song dựa trên dãy phổ biến tìm được ở câu a.

(1) D ⇒ DKH [0,3, 0,53] < min-conf = 0,7
(2) K ⇒ DKH [0,3, 0,5] < min-conf = 0,7
(3) H ⇒ DKH [0,3, 0,5] < min-conf = 0,7
(4) KH ⇒ DKH [0,3, 0,82] > 0,7
(5) DH ⇒ DKH [0,3, 0,75] > 0,7
(6) KD ⇒ DKH [0,3, 0,75] > 0,7

Với min-conf = 70%, ta nhận được các luật (4), (5), (6)
Làm tương tự với tập phổ biến DHS và BHS, ta được
DH ⇒ DHS [0,33, 0,83] > 0,7
DS ⇒ DHS [0,33, 0,9] > 0,7
BH ⇒ BHS [0,33, 0,9] > 0,7
BS ⇒ BHS [0,33, 0,83] > 0,7
...

Date:

d. Các luật ~~phổ biến~~ WINEPI tuần tự thỏa min_conf dựa trên dãy phổ biến ở câu b.

(1) S ⇒ SB [0,3, 0,5] (6) K ⇒ KH [0,33, 0,56]

(2) B ⇒ SB [0,3, 0,5] (7) H ⇒ KH [0,33, 0,56]

(3) D ⇒ DK [0,37, 0,65] (8) G ⇒ GK [0,3, 0,5]

(4) K ⇒ DK [0,37, 0,61] (9) K ⇒ GK [0,3, 0,5]

(5) H ⇒ HS [0,47, 0,78] (10) S ⇒ HS [0,47, 0,78]

Với min_conf = 70%, ta nhận được các luật (5), (10).