COMP4433 Data Mining and Data Warehousing

FAQ on Association Rule Mining I (Suggested Answers)

1. A database has four transactions. Let min_sup=60% and min_conf=80%.

TID	Date	Items_bought
100	10/15/99	{K,A,D,B}
200	10/15/99	{D,A,C,E,B}
300	10/19/99	{C,A,B,E}
400	10/22/99	{B,A,D}

a) Find all frequent itemsets using Aprior algorithm.

b) List all of the strong association rules (with support s and confidence c) matching the following metarule (form), where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

$\forall x \in$ transaction, buys(X, $item_1$) $\wedge$ buys(X, $item_2$) $\Rightarrow$ buys(X, $item_3$) [s,c]

Suggested Answer: (Please try it first!)

a) min _sup=60% (i.e., ≥ 3 transactions)

1-itemset	Count	2-itemset	Count	3-itemset	Count
A	4	**A-B**	4	**A-B-D**	3
B	4	**A-D**	3		
~~C~~	~~2~~	**B-D**	3		
D	3				
~~E~~	~~2~~				
~~K~~	~~1~~				

The frequent 2-itemsets and 3-itemsets are bolded.

b)

Rule	**Confidence**
A,B⇒D	3/4
A,D⇒B	3/3
B,D⇒A	3/3

All except the first one are strong rules for *min_conf*=80%.

D ⇒ AB

2. Suppose that you are working for a company that provides local courier services. To improve the operations for more speedy delivery, your boss would like to know if there is any association among the originating locations of the parcel to be delivered, the destinations, the type of the parcel and their weight. Given some of the data collected yesterday below, you decide to use the Apriori algorithm to mine the data for interesting association rules.

Parcel ID	Origin	Destination	Type	Weight
1	HK	HK	Parcel	Light
2	Kln	Kln	Letter	Light
3	NT	Kln	Letter	Light
4	HK	HK	Parcel	Heavy
5	Kln	Kln	Parcel	Light
6	NT	NT	Letter	Light
7	HK	HK	Letter	Light
8	Kln	Kln	Parcel	Heavy
9	Kln	Kln	Letter	Light
10	HK	HK	Letter	Light
11	HK	HK	Parcel	Heavy
12	Kln	Kln	Letter	Light
13	HK	HK	Letter	Light
14	Kln	Kln	Parcel	Light
15	HK	NT	Parcel	Heavy
16	NT	Kln	Letter	Light
17	HK	NT	Letter	Light
18	Kln	HK	Parcel	Light
19	HK	NT	Parcel	Heavy
20	HK	HK	Parcel	Light
21	Kln	Kln	Letter	Light
22	Kln	HK	Parcel	Heavy
23	Kln	Kln	Letter	Light
24	Kln	Kln	Letter	Light
25	HK	HK	Parcel	Light

a) By setting the support to 25% and confidence to 50%, use the Apriori algorithm to find all interesting association rules in the above Table.

b) Assume that a user sets the lift ratio to 1.75, which rules you discovered for a) above are still interesting?

Suggested Answer: (Please try it first!)

a) min _sup=25% (i.e., ≥ 7 records) and min_conf=50%.

The itemsets (in black) with count ≥ 7 records below are frequent itemsets.

1-itemset	Count
OHK	**11**
OKLN	**11**
ONT	3
DHK	**10**
DKLN	**11**
DNT	4
P	**12**
L	**13**
Li	**19**
H	6

2-itemset	Count
OHK, DHK	**8**
OHK, DKLN	0
OHK, P	**7**
OHK, L	4
OHK, Li	**7**
OKLN, DHK	2
OKLN, DKLN	**9**
OKLN, P	5
OKLN, L	6
OKLN, Li	**9**
DHK, P	**7**
DHK, L	3
DHK, Li	**7**
DKLN, P	3
DKLN, L	**8**
DKLN, Li	**10**
P, Li	6
L, Li	**13**

3-itemset	Count
OHK, DHK, P	5
OHK, DHK, Li	6
~~OHK, P, Li~~	
OKLN, DKLN, Li	8
~~DHK, P, Li~~	
DKLN, L, Li	8

To determine the interesting association rules for frequent 3-itemsets, i.e., {OKLN, DKLN, Li} & {DKLN, L, Li}, we have the following candidates:

{OKLN, DKLN, Li}		{DKLN, L, Li}	
OKLN, DKLN → Li	8/9	DKLN, L → Li	8/8
OKLN, Li → DKLN	8/9	DKLN, Li → L	8/10
DKLN, Li → OKLN	8/10	L, Li → DKLN	8/13
OKLN → DKLN, Li	8/11	DKLN → L, Li	8/11
DKLN → OKLN, Li	8/11	L → Li, DKLN	8/13
Li → OKLN, DKLN	8/19	Li → DKLN, L	8/19

The rules NOT highlighted in red are interesting ones w.r.t min_sup=25% and min_conf=50%. The same procedure can be repeated for frequent 2-itemsets but is omitted here.

b) Recall from lecture notes that

$$Interest(A \rightarrow B) = \frac{P(A \wedge B)}{P(A)P(B)} = \frac{P(A \wedge B)}{P(A)} \times \frac{1}{P(B)} = conf(A \rightarrow B) \times \frac{1}{P(B)}$$

{OKLN, DKLN, Li}		{DKLN, L, Li}	
OKLN, DKLN → Li	(8/9)/(19/25)	DKLN, L → Li	(8/8)/(19/25)
OKLN, Li → DKLN	(8/9)/(11/25)	DKLN, Li → L	(8/10)/(13/25)
DKLN, Li → OKLN	(8/10)/(11/25)	L, Li → DKLN	(8/13)/(11/25)
OKLN → DKLN, Li	(8/11)/(10/25)	DKLN → L, Li	(8/11)/(13/25)
DKLN → OKLN, Li	(8/11)/(9/25)	L → Li, DKLN	(8/13)/(10/25)
Li → OKLN, DKLN	(8/19)/(9/25)	Li → DKLN, L	(8/19)/(8/25)

The rules above with interest≥1.75 are considered interesting.

3. Given the following web page content database. Let *min_sup*=60% and *min_conf*=80%.

URL	Web Page ID	Keywords Found
Jackchan.com	P100	Popstar, Actor, Movie, Hollywood
Nictsz.com	P101	Actor, Popstar, Music, Band, Movie
Faywang.com	P102	Actress, Popstar, Music, Movie, Hollywood
Allantam.com	P103	Actor, Popstar, Music, CEO, Movie
SammyChen.com	P104	Popstar, Actress, Movie, Music

a) Find all frequent 2-itemsets and 3-itemsets using Apriori algorithm.

b) How many scans of the database are required by part (a)?

c) Based on the results of part (a), generate all strong association rules (with sufficient support and confidence) matching the following metarule.
$\forall P \in$ *web page*, *contains*$(P, keyword_1) \wedge$ *contains*$(P, keyword_2) \Rightarrow$ *contains*$(P, keyword_3)$

d) Describe how the association rule mining procedure should be modified to discover rules of the following form.
$\forall P \in$ *web page*, *contains*$(P, keyword_1) \wedge$ ***NOT contains***$(P, keyword_2) \Rightarrow$ *contains*$(P, keyword_3)$

Suggested Answer: (Please try it first!)

a) min _sup=60% (i.e., ≥ 3 Web pages)

1-itemset	Count	2-itemset	Count	3-itemset	Count
Popstar	5	**Popstar-Actor**	3	**Popstar-Actor-Movie**	3
~~Actress~~	~~2~~	**Popstar-Movie**	5	**Popstar-Movie-Music**	4
Actor	3	**Popstar-Music**	4		
Movie	5	**Actor-Movie**	3		
Music	4	~~Actor-Music~~	~~2~~		
~~Band~~	~~1~~	**Movie-Music**	4		
~~Hollywood~~	~~2~~				
~~CEO~~	~~1~~				

The frequent 2-itemsets and 3-itemsets are bolded.

b) Theoretically: 4 (i.e., n+1 scans), specific to this question: 3 (because C_4 will be an empty set and there is no need to scan the database for minimum support verification in the fourth iteration of the Apriori algorithm)

c)

Rule	**Confidence**
Popstar,Actor⇒Movie	3/3
Popstar,Movie⇒Actor	3/5
Actor,Movie⇒Popstar	3/3
Popstar,Movie⇒Music	4/5
Popstar,Music⇒Movie	4/4
Movie,Music⇒Popstar	4/4

All except the second one are strong rules for *min_conf*=80%.

d) This is quite an open question. One possible solution is to append a list of keywords not found to the list of keywords found as shown below. The Apriori algorithm can then be applied to the appended database and rules following the form above can be mined accordingly.

Relational form

Web Page ID	Popstar	Actor	Actress	Movie	Music	Band	Hollywood	CEO
P100	Yes	Yes	-	Yes	-	-	Yes	-
P101	Yes	Yes	-	Yes	Yes	Yes	-	-
P102	Yes	-	Yes	Yes	Yes	-	Yes	-
P103	Yes	Yes	-	Yes	Yes	-	-	Yes
P104	Yes	-	Yes	Yes	Yes	-	-	-

➔

Transactional form

Web Page ID	Keyword Info.
P100	Popstar, Actor, ^Actress, Movie, ^Music, ^Band, Hollywood, ^CEO
P101	Popstar, Actor, ^Actress, Movie, Music, Band, ^Hollywood, ^CEO
P102	Popstar, ^Actor, Actress, Movie, Music, ^Band, Hollywood, ^CEO
P103	Popstar, Actor, ^Actress, Movie, Music, ^Band, ^Hollywood, CEO
P104	Popstar, ^Actor, Actress, Movie, Music, ^Band, ^Hollywood, ^CEO