COMP5121 Data Mining & Data Warehousing Applications

Assignment #1 (Suggested answers)

Instructions:
- Answer all three questions.
- Interpret the questions logically, show your steps and write down your assumption(s) when necessary.
- You may use this word file to prepare your answers and submit it to L@PU before the due date.

1. In a survey, the following three questions have been asked:
   Q.1 Do you want to own a balance wheel?
   Q.2 Do you have a driver license?
   Q.3 Do you like selfie?
   After computing the corresponding statistics for 10000 participants, the following data are given.
   - ☐ Among 5000 participants who have a driver license,
     - 3250 want to own a balance wheel
     - 3750 like selfie
     - 2500 both want to own a balance wheel and like selfie
   - ☐ Among another 5000 participants who DON'T have a driver license,
     - 2750 want to own a balance wheel
     - 4000 like selfie
     - 2250 both want to own a balance wheel and like selfie

   a) List ALL strong association rules having the form {item1, item2→have a driver license} with support≥5% and confidence≥50%.
   b) Compute the interest (lift ratio) of the strong association rules found in part (a).
   c) Which of the rule(s) found in part (a) is/are most interesting? Justify your answer.
   d) Which of the rule(s) found in part (b) is/are most interesting? Justify your answer.

   Ans.
   a)

	Have a driver license			Not Have a driver license		
	BW	Not BW	sum(row)	BW	Not BW	sum(row)
Selfie	2500	1250	3750	2250	1750	4000
Not Selfie	750	500	1250	500	500	1000
sum(col.)	3250	1750	5000	2750	2250	5000

   R1: Own BW, like selfie ➔ Have a driver license
   [support=0.25, confidence=2500/4750 (0.53)]
   R2: Own BW, Not like selfie ➔ Have a driver license
   [support=0.075, confidence=750/1250 (0.6)]
   R3: Not Own BW, like selfie ➔ Have a driver license
   [support=0.125, confidence=1250/3000 (0.42)]
   R4: Not Own BW, Not like selfie ➔ Have a driver license
   [support=0.05, confidence=500/1000 (0.5)]
   All except R3 are strong!

   b) Lift Ratio: R1 = 1.05
   R2 = 1.2
   R4 = 1

c) Concerning the support, R1 should be taken a look first because it has much higher value, i.e. statistically more significant, and the confidence is not low. Concerning the confidence, R2 deserves to have a look because of its relatively higher rule's strength.

d) Concerning the support, the same comment of part (c) applies. Concerning the lift ratio, R2 has Lift Ratio most deviated from 1 and hence it is most interesting to end users looking for positive correlation information. R4 could be interesting because of the involved independence indication.

2. Consider the following stock transactions for association analysis.

Table I Stock Transaction Data

Stock	Transactions made by 10 selected investors today									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
HSBC	Buy	Buy		Buy	Buy		Buy	Buy		Buy
BoEA	Sell	Sell	Buy			Sell	Sell	Buy		
China_Mobile	Buy		Buy	Buy	Sell	Buy			Buy	Buy
China_Petroleum			Buy		Sell	Buy			Sell	Buy

That is today investor #1 buys HSBC and China_Mobile but sells BoEA, investor #2 buys HSBC but sells BoEA, investor #3 …, and investor #10 buys HSBC, China_Mobile and China_Petroleum.

a) Find all frequent itemsets using Aprior algorithm by setting min_sup=20%.

Suggested Answer:

The answer below is not unique. Other logical answers, corresponding to different handling of "no action (empty box)", are acceptable.

Table I can be represented as the following transactional form:

Transaction #	Items involved
1	Buy_HSBC, Sell_BoEA, Buy_CM
2	Buy_HSBC, Sell_BoEA
3	Sell_BoEA, BuyCM, Buy_CP
4	Buy_HSBC, Buy_CM
5	Buy_HSBC, Sell_CM, Sell_CP
6	Sell_BoEA, Buy_CM, Buy_CP
7	Buy_HSBC, Sell_BoEA
8	Buy_HSBC, Buy_BoEA
9	Buy_CM, Sell_CP
10	Sell_HSBC, Buy_CM, Buy_CP

So, there exist 8 items, namely, BHS (Buy_HSBC), SHS (Sell_HSBC), BEA (Buy_BoEA), SEA (Sell_BoEA), BCM (Buy_CM), SCM (Sell_CM), BCP (Buy_CP), SCP (Sell_CP). For min _sup=20% (i.e., ≥ 2 transactions), we have

1-itemset	Count
BHS	7
SHS	0
BEA	2
SEA	4
BCM	6
SCM	1
BCP	3
SCP	2

2-itemset	Count
BHS, BEA	1
BHS, SEA	3
BHS, BCM	3
BHS, BCP	1
BHS, SCP	1
BEA, BCM	1
BEA, BCP	1
BEA, SCP	0
SEA, BCM	2
SEA, BCP	1
SEA, SCP	0
BCM, BCP	3
BCM, SCP	1

3-itemset	Count
BHS, SEA, BCM	1

The itemsets (in black) with count ≥ 2 are frequent itemsets.

b) List ANY 3 strong rules (if exist) with 2 items on the LHS and 1 item on the RHS for min_conf=70%.

Suggested Answer:

Since there is no frequent 3-itemset, no strong rule with the required form can be found.

3. Your R&D team has been assigned a project to carry out price movement classification on the stock data. After pre-processing the collected numeric data, the following database is given:

Stock Price Movement Database

Stock	Price Movement from 13 Oct – 24 Oct, 2011 (only trading days Mon-Fri applied)									
	10 Sep	11 Sep	12 Sep	13 Sep	14 Sep	17 Sep	18 Sep	19 Sep	20 Sep	21 Sep
HSBCC	*Up*	*Up*	*Level*	*Down*	*Level*	*Up*	*Up*	*Down*	*Level*	*Up*

where the movement labels *Up*, *Down* & *Level* denote the stock price going up, down and level respectively in the corresponding trading day. In order to classify next trading day's price movement, the stock data above is extracted as follows.

Extracted Stock Price Movement Database for Classification

Today is	Price Movement of HSBCC for			
	2 Trading Day before (2TDB)	1 Trading Day before (1TDB)	Today (TD)	Next Trading Day (NTD)
12 Sep	*Up*	*Up*	*Level*	*Down*
13 Sep	*Up*	*Level*	*Down*	*Level*
14 Sep	*Level*	*Down*	*Level*	*Up*
17 Sep	*Down*	*Level*	*Up*	*Up*
18 Sep	*Level*	*Up*	*Up*	*Down*
19 Sep	*Up*	*Up*	*Down*	*Level*
20 Sep	*Up*	*Down*	*Level*	*Up*

Suppose you are asked to adopt the decision tree classifier to classify the given stock data with respect to the class attribute NTD (next trading day). Show how the first two rows (i.e. Today is 12 Sep & 13 Sep respectively) are classified when all the seven data records above are used for training.

Useful information:

$\log_2 x = \log_{10} x / \log_{10} 2 \cong \log_{10} x / 0.30103$

$$I(c_1, c_2, c_3) = -\frac{c_1}{c_1+c_2+c_3}\log_2\frac{c_1}{c_1+c_2+c_3} - \frac{c_2}{c_1+c_2+c_3}\log_2\frac{c_2}{c_1+c_2+c_3} - \frac{c_3}{c_1+c_2+c_3}\log_2\frac{c_3}{c_1+c_2+c_3}$$

$I(1,0,0) = I(2,0,0) = I(0,2,0) = 0$

$I(1,1,0) = I(1,0,1) = 1$

$I(1,2,0) = I(0,1,2) \cong 0.918$

$I(1,2,1) = 1.5$

$I(2,3,2) \cong 1.557$

Determining the root attribute:

I(3,2,2)=1.557

Entropy for 2TDB

2TDB	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	$I(\#_{Up}, \#_{Level}, \#_{Down})$
Up	1	2	1	1.5
Level	1	0	1	1
Down	1	0	0	0

Entropy(2TDB)=(4/7)*I(1,2,1) + (2/7)*I(1,0,1) + (1/7)*I(1,0,0)≈1.143

Information_Gain(2TDB)=1.557-1.143=0.414

Entropy for 1TDB

1TDB	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	$I(\#_{Up}, \#_{Level}, \#_{Down})$
Up	0	1	2	0.918
Level	1	1	0	1
Down	2	0	0	0

Entropy(1TDB)=(3/7)*I(0,1,2) + (2/7)*I(1,1,0) + (2/7)*I(2,0,0)≈0.679

Information_Gain(1TDB)=1.557-0.679=0.878

Entropy for TD

TD	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	$I(\#_{Up}, \#_{Level}, \#_{Down})$
Up	1	0	1	1
Level	2	0	1	0.918
Down	0	2	0	0

Entropy(TD)=(2/7)*I(1,0,1) + (3/7)*I(2,0,1) + (2/7)*I(0,2,0)≈ 0.679

Information_Gain(TD)=1.557-0.679=0.878

Hence, either TD or 1TDB can be chosen as the root node.

Let TD be the root node.

The “Down” branch will terminate.

For the “Up” branch, we have I(1,1)=1

Obviously, any of 2TDB and 1TDB can be used to differentiate the 2 samples for NTD=Up and NTD=Down

For the “Level” branch, we have I(2,0,1)=0.918

By inspection, we can easily conclude that we should choose 1TDB as the intermediate node because when 1TDB=”Down”, the two sample’s NTD=”Up” and when 1TDB=”Up”, the only one sample’s NTD=”Down”.

Hence, we have 5 rules generated. One version is:

R1: IF TD=”Down” THEN NTD=”Level”

R2: IF TD=”Up” AND 1TDB=“Level” THEN NTD=”Up”

R3: IF TD=”Up” AND 1TDB=“Up” THEN NTD=”Down”

R4: IF TD=”Level” AND 1TDB=“Down” THEN NTD=”Up”

R5: IF TD=”Level” AND 1TDB=“Up” THEN NTD=”Down”

To classify the first two rows for Today is 12 Sept. & 13 Sept., we have

12 Sept.: matched with R5

Hence, the predicted class is “Down” which is correct.

13 Sept.: matched with R1

Hence, the predicted class is “Level” which is correct.