THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MSc Scheme-61030

Subject : COMP5121 Data Mining and Data Warehousing Applications

Group : 101,102, 103, 104, 1888

Session : 2007 / 2008 Semester I

Date : 17 December 2007 Time : 18:30-20:30

Time Allowed : 2 Hours Subject Lecturer : Korris Chung

This question paper has ___7___ pages (cover included).

Instructions to Candidates:

Open-book examination.
Answer **ALL** questions.
Show your steps, interpret the questions logically and write down any assumption(s) you made.
Standard non-programmable calculator is allowed.

Do not turn this page until you are told to do so!

COMP5121 Data Mining & Data Warehousing Applications

2007/2008 Fall Examination

1. As a data mining consultant, you are asked to analyze the following medical database records where the last column "Disease" is the class attribute.

Patient ID	Blood Pressure Level	Sex	Risk Level	Marital Status	Disease
9100123	80-120	Male	High	Married	No
9303034	160-200	Female	Medium	Single	Yes
9210126	80-120	Male	Medium	Married	Yes
9142020	120-160	Male	Low	Single	No
9910111	160-200	Female	High	Single	Yes
9576732	80-120	Male	Low	Married	No
9910115	160-200	Female	Low	Single	Yes
9210120	120-160	Female	Medium	Single	Yes
9576737	120-160	Male	Low	Married	No

a) Suppose your clients would like to know the difference between the frequent itemsets of database records with Disease=Yes and the frequent itemsets of database records with Disease=No.
   i) Mine all frequent 3-itemsets of the database records with Disease=Yes for minimum support=50%.
   ii) Mine all frequent 3-itemsets of the database records with Disease=No for minimum support=50%.

   Note here that you are NOT required to use the Apriori algorithm (or other association mining algorithms) to produce the answer and the class attribute value is NOT considered in the frequent itemset mining process.

   (12 marks)

b) Your clients have commented that the frequent itemsets obtained in part (a) are not sufficient. They would like to know the patterns/itemsets with great differences in their occurrence between the two data partitions, namely, partition D1 consisting of records with Disease=Yes and partition D2 consisting of records with Disease=No. For example, the support of itemset "Sex=Male" in D1 is equal to 20% while that in D2 is equal to 100%. Such an itemset has great difference (20% vs. 100%) in its occurrence between partitions D1 and D2. Propose a mining algorithm/method to address this concern from your clients. Among the set of 2-itemsets {(80-120, Male), (120-160, Male), (160-200, Female)}, which one has the largest difference in its occurrence between partition D1 and D2? Show your calculation.

   (13 marks)

2. Suppose you are asked to provide data mining consulting services to an Internet DVD shop. After interviewing the shop's manager and the database administrator, the following information about the customer database and the movie database are collected.

Customer Database

Customer ID	Transaction Date	Movie Rent (Movie ID)
00001	02-01-2003	3997 (Spiderman II)
00001	12-11-2003	0553 (King Kong)
00001	15-12-2003	0150 (Cinderella Man)
00002	12-01-2003	1011 (Poltergeist)
00002	12-10-2004	0150 (Cinderella Man)
00002	10-06-2005	3997 (Spiderman II)
00003	16-03-2006	0001 (A Beautiful Mind)
00004	07-03-2005	1011 (Poltergeist)
00004	17-03-2006	0553 (King Kong)
00004	18-03-2006	3997 (Spiderman II)
00004	18-04-2007	3997 (Spiderman II)
00004	18-06-2007	0553 (King Kong)

Movie Database

Movie ID	Movie Name	Types
0001	A Beautiful Mind	Drama, Mystery, Romance
0150	Cinderella Man	Drama, Romance
0553	King Kong	Action, Thriller, Horror, Sci-Fiction
1011	Poltergeist	Horror, Thriller
3997	Spiderman II	Action, Crime, Sci-Fiction

a) If you are asked to cluster the movies, propose an appropriate dissimilarity measure for it and show the dissimilarity matrix values for the five movies in the database above.

(7 marks)

b) Based on your dissimilarity matrix obtained in part (a), compute the first reduced dissimilarity matrix, i.e., D_2, by using the complete linkage agglomerative clustering algorithm.

(7 marks)

c) You are then asked to cluster the customers based on three measures, namely, the number of movie rental transactions made (which can be extracted from the customer database), the movies rent (which can be extracted from the customer database), and the types of movies rent (which can be extracted from the customer database and the movie database).

i) Prepare a task-relevant database based on the two databases shown above. You may just show the attribute values of customers 00003 and 00004.

ii) Compute the distance/dissimilarity between customer 00003 and customer 00004.

(11 marks)

3. Suppose you are asked by the Octopus Card Company to analyze the following Card Database.

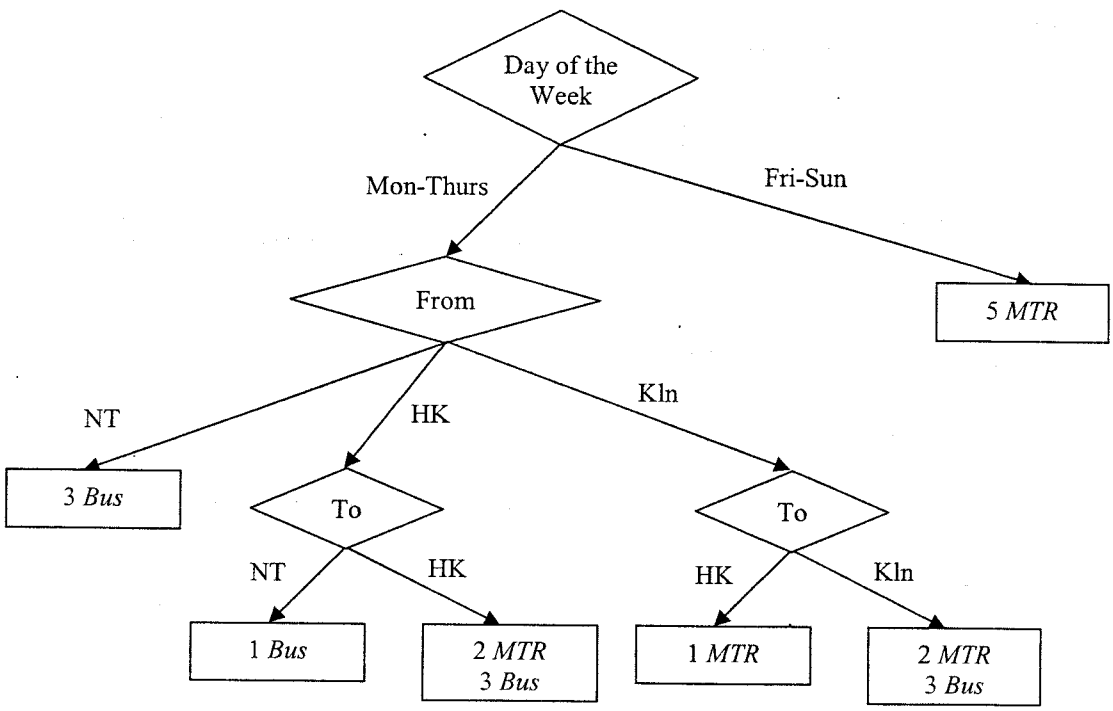
Octopus Card Database for Training

Octopus ID	*Day of the Week*	*From*	*To*	*Transportation Taken*
1	Mon-Thurs	HK	HK	MTR
2	Mon-Thurs	Kln	Kln	Bus
3	Mon-Thurs	NT	Kln	Bus
4	Fri-Sun	HK	HK	MTR
5	Mon-Thurs	Kln	Kln	MTR
6	Mon-Thurs	NT	NT	Bus
7	Mon-Thurs	HK	HK	Bus
8	Fri-Sun	Kln	Kln	MTR
9	Mon-Thurs	Kln	Kln	Bus
10	Mon-Thurs	HK	HK	Bus
11	Fri-Sun	HK	HK	MTR
12	Mon-Thurs	Kln	Kln	Bus
13	Mon-Thurs	HK	HK	Bus
14	Mon-Thurs	Kln	Kln	MTR
15	Fri-Sun	HK	NT	MTR
16	Mon-Thurs	NT	Kln	Bus
17	Mon-Thurs	HK	NT	Bus
18	Mon-Thurs	Kln	HK	MTR
19	Fri-Sun	HK	NT	MTR
20	Mon-Thurs	HK	HK	MTR

Octopus Card Database for Testing

Octopus ID	*Day of the Week*	*From*	*To*	*Transportation Taken*
21	Mon-Thurs	Kln	Kln	Bus
22	Fri-Sun	Kln	HK	MTR
23	Mon-Thurs	Kln	Kln	Bus
24	Mon-Thurs	Kln	Kln	Bus
25	Mon-Thurs	HK	HK	MTR

Your project team member tried to construct a decision tree for the first 20 records above, i.e., the training dataset, and obtained the result as follows.



a) Compute the classification rate (accuracy) of the constructed decision tree for the testing database.

(3 marks)

b) Classify the following two records with missing attribute values (denoted as unknown below).

Octopus ID	*Day of the Week*	*From*	*To*	*Transportation Taken*
100	Unknown	NT	HK	?
101	Mon-Thurs	Unknown	HK	?

(4 marks)

c) Another project team member commented that the decision tree above is too deep. Prune the tree as much as possible so that the classification rate (accuracy) of the pruned tree for the testing database can be maintained, i.e., as high as that obtained in part (a).

(8 marks)

d) If some more database records are available and added to the training database, describe the scenarios/conditions for

i) reconstruction of the decision tree is NOT necessary;
ii) partial reconstruction of the decision tree is necessary;
iii) full reconstruction of the decision tree is necessary.

(10 marks)

4. Your R&D team has been assigned a project to carry out opinion mining on the on-line forum postings. After pre-processing the collected text documents, the following sample database is given.

Database for Opinion Mining

Document ID	A1:Number of Sentences with Positive Opinion	A2:Number of Sentences with Negative Opinion	A3:Happy Smile Code Used	Overall Opinion
F1-D1	*Less*	*Less*	*No*	*Negative*
F1-D2	*Less*	*Intermediate*	*No*	*Neural*
F1-D3	*Intermediate*	*More*	*Yes*	*Positive*
F2-D1	*More*	*Intermediate*	*Yes*	*Positive*
F2-D2	*Intermediate*	*Less*	*Yes*	*Negative*
F2-D3	*Less*	*Less*	*No*	*Neural*
F2-D4	*More*	*More*	*Yes*	*Positive*

a) You are asked to carry out an association analysis of the data above. By setting the minimum support to 25% and minimum confidence to 0%, mine all STRONG association rules satisfying:

* $\rightarrow$ Overall Opinion=Positive

or * $\rightarrow$ Overall Opinion=Neutral

or * $\rightarrow$ Overall Opinion=Negative

where * denotes a wild card. Note here that you are NOT required to apply the Apriori algorithm to generate the solution.

(14 marks)

b) You are further asked to use the association rules mined in part (a) to predict the overall opinion of an unseen document. Describe how you solve this problem (by proposing an associative classification approach) and show how the following record is classified.

Document ID	Number of Sentences with Positive Opinion	Number of Sentences with Negative Opinion	Happy Smile Code Used	Overall Opinion
F1-D1	*More*	*Less*	*Yes*	?

(11 marks)

5. a) Propose a method to normalize the following two time series subsequences so that they can be compared effectively.

Time	T1	T2	T3	T4	T5
Subsequence 1	130	135	130	125	130
Subsequence 2	5	5.5	6	5	5.5

(6 marks)

b) Propose a method to fill the missing value of the following time series subsequence.

Time	T1	T2	T3	T4	T5
Subsequence 1	135	140	?	130	130

(3 marks)

c) A bus company is starting a data warehouse project for storing and analyzing its bus route data. There are four dimension tables designed and their attributes are as follows:

- Routine: *routine_key, station_list_key, bus_type_key, frequency, start_time, end_time, fare*
- Station List: *station_list_key, begin_station, 1st_station, 2nd_station, ..., last_station*
- Bus Type: *bus_type_key, capacity, IsAirCon, year in service*
- Time: *time_key, day, day_of_week, month, quarter, year*

The foreign keys of the Bus Route fact table include *routine_key* (linking to routine dimension table) and *time_key* (linking to time dimension table). The facts being recorded by this fact table are *number of passages*, *income*, *fuel cost*, *operation cost*, and *profit/loss*. The routine dimension table has been normalized so that it is linked to station list dimension table via *station_list_key* and linked to bus type dimension table via *Bus_type_key*.

i) Design a snowflake schema based on the given information.

(8 marks)

ii) What kinds of analysis can be obtained from such a data warehouse? For example, which route(s) and what time have generated the highest profit? List some typical analysis results accordingly.

(8 marks)

- E N D -