Birla Institute of Technology & Science-Pilani, Hyderabad Campus
1st Semester 2011-2012
Data Mining (CS / IS C415)
Test 2 (Makeup)

Type: Closed Book Time: 50 mins Max Marks:60 Date: 26/11/2011

1. Given the following database of transactions:

CustID	TransID	Items(in the form Brand-Category)
01	T100	King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread
02	T200	Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread
01	T300	Westcoast-Apple, King's-Crab, Tasty-Pie, Wonder-Bread
03	T400	Wonder-Bread,Sunset-Milk,Dairyland-Cheese

Assume that minimum support is set to 50% and minimum confidence to 80%.
a) At the level of categories, find all freqent itemsets, e.g. {Cheese, Apple}. Report these itemsets together with their support.
b) Which of the frequent itemsets under a) are closed, which of them are maximal?
c) Let k be the maximum number for which there is a frequent k-itemset. For all frequent k-itemsets, determine all corresponding association rules (having minimum confidence) and their confidence.

Solution:
a) At the level of categories, find all freqent itemsets, e.g. {Cheese, Apple}. Report these itemsets together with their support. (15 marks)
Crab (50%) Pie (50%) Apple (50%) Milk (75%) Cheese (75%) Bread (100%)
Milk, Bread (75%) Milk, Cheese (75%) Cheese, Bread (75%) Crab, Bread (50%)
Pie, Apple (50%) Pie, Bread (50%) Apple, Bread (50%)
Milk, Cheese, Bread (75%) Pie, Apple, Bread (50%)
b) Which of the frequent itemsets under a) are closed, which of them are maximal? (10 marks)
Closed frequent itemsets: Bread (100%) Crab, Bread (50%)
Milk, Cheese, Bread (75%) Pie, Apple, Bread (50%)
Maximal frequent itemsets: Crab, Bread (50%) Milk, Cheese, Bread (75%)
Pie, Apple, Bread (50%)
c) Let k be the maximum number for which there is a frequent k-itemset. For all frequent k-itemsets, determine all corresponding association rules (having minimum confidence) and their confidence. (15 marks)
association rules from frequent 3-itemset {Milk, Cheese, Bread}:
Milk, Cheese → Bread (100%)
Milk, Bread → Cheese (100%)
Bread, Cheese → Milk (100%)
Milk → Bread, Cheese (100%)
Cheese → Bread, Milk (100%)
association rules from frequent 3-itemset {Pie, Apple, Bread}:
Pie, Apple → Bread (100%)
Pie, Bread → Apple (100%)
Apple, Bread → Pie (100%)
Pie → Apple, Bread (100%)
Apple → Bread, Pie (100%)

2. Your company is trying to solve the following problem. You have some data for which you have no additional information apart from the data themselves. You want to extract as much information as possible.

You interview three data mining consultants for an opening in your firm. The three consultants propose the following approaches.
Consultant A: First we apply a decision tree. Then we prune it as much as possible so that we can identify large portions of the problem space. Then, we apply hierarchical clustering on the subspaces identified by the decision tree to find good description of the subproblems.
Consultant B: I disagree with A. We should first apply clustering and then apply decision tree using the results of the clustering process. In this way we can extract a description of the clusters.
Consultant C: They are both wrong. First apply a hierarchical clustering so that you can find some structure in the data. Then, on each cluster we just found we apply k-mean so that we can actually have a compact description of the clusters.
Which solution is the best? Why the other ones are worse?

3. There are two clusters C1 and C2 formed from a dataset. The Clustering Feature (CF) vectors of these two clusters are: CF1 = (4, 13, 51) and CF2 = (4, 34, 294). Determine the following:

a) Centroids of C1 and C2
b) Radius of C1
c) Diameter of C2
d) Define a valid inter-cluster distance and find the corresponding inter-cluster distance between C1 and C2

4. Consider a two dimensional database D with the records : R_1(2, 2), R_2(2, 4),R_3(4, 2), R_4(4, 4), R_5(3, 6), R_6(7, 6), R_7(9, 6), R_8(5, 10), R_9(8, 10), R_{10}(10, 10). The distance function is the L_1 distance (Manhattan distance).

Show the results of the k-means

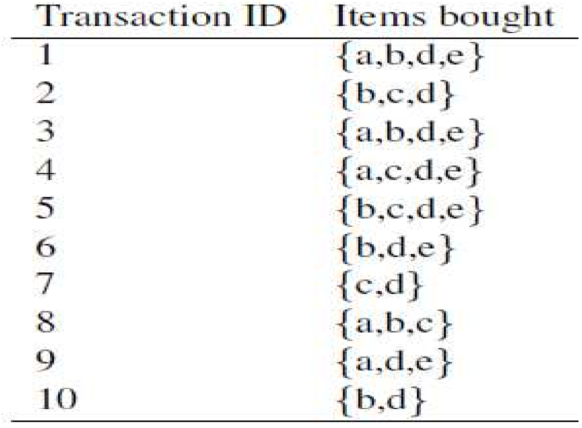
algorithm at each step, assuming that you start with two clusters (k = 2) with centers

C_1 = (6, 6) and C_2 = (9, 7).

5. Given the transactions given in the table 1, draw the itemset lattice induced by the items, and label each node of the lattice with the following letters:

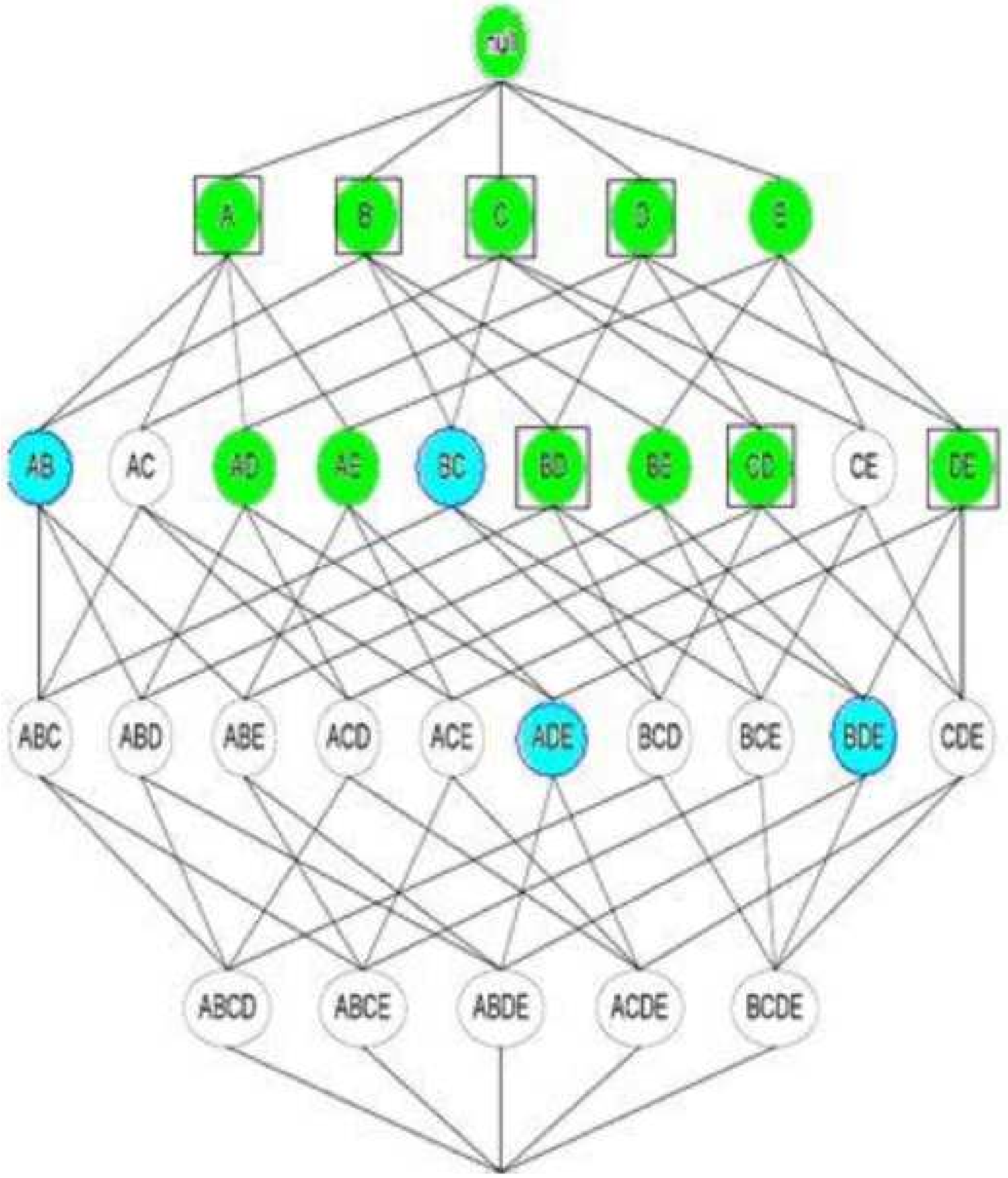
- M, if the node is a maximal frequent itemset
- C, if it is a closed frequent itemset
- N, if it is frequent but neither maximal nor closed
- I, if it is infrequent.

Assume a minsup threshold of 30%.



Transaction ID	Items bought
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,c,d,e}
6	{b,d,e}
7	{c,d}
8	{a,b,c}
9	{a,d,e}
10	{b,d}

Table 1: Marker basket transactions



Green nodes are frequent (also null), white infrequent, blue maximal and boxed ones closed itemsets.

Using the transaction data in table 1, draw the FP-tree data structure used by the FP-growth algorithm, using two different ordering for the items:
(a) Alphabetical order
(b) Descending order of support

3. Using the FP-tree structure for transaction data in table 1 using the alphabetical order for items, draw the
(a) conditional FP-tree for itemsets ending with fdg
(b) conditional FP-tree for itemsets ending with fc,dg
http://www.cs.helsinki.fi/group/bioinfo/teaching/dami_s10/solutions_ex2.pdf

An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 multiple choice questions with four possible answers each. How would you convert this data into a form suitable for association analysis?
One option would be giving unique values to each possible selection. There would be then 100x4=400 values altogether. In this case the transaction would be formed as a row with 100 items.

Question no/ Response no	1	2	3	4	5	6	7	...	100
1	C1	D2	D3	B4	A5	A6	B7	...	B100
2	C1	B2	A3	D4	C5	A6	D7	...	C100
3	A1	C2	C3	B4	C5	A6	B2	...	C100