Final Exam for Data Mining (Fall of 2008)
Instructor: Kyuseok Shim

1. Consider the following market basket database. Assume we have the minimum support of 3 transactions.

Transaction ID	Transaction
1	(a, c, d, f, i, m, p)
2	(a, b, c, f, m)
3	(a, b, i)
4	(b, c, f, p)
5	(a, c, e, f, p, m)

- Find all frequent itemsets whose size is at least 3 using FP-growth algorithm ran for the above sequence data. Show the detailed steps. [25 pts.]

SOL: First, find 1-frequent items and sort them based on their frequency. Then we have the following list.

Item	Support
a	4
c	4
f	4
b	3
m	3
p	3

Using this list, remove unfrequent items and sort the remaining items for each transaction from the given data. Then, we have the following.

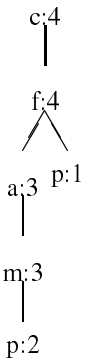
TID	Itemset
1	a,c,f,m,p
2	a,c,f,b,m
3	a,b
4	c,f,b,p
5	a,c,f,m,p

Since we want to find the frequent itemsets whose length is at least 3, remove the third transaction and decrease the frequency of a and b. Then, b becomes unfrequent item and thus remove b from the dataset. Then we have the following list and the dataset.

Item	Support
c	4
f	4
a	3
m	3
p	3

TID	Itemset
1	c,f,a,m,p
2	c,f,a,m
4	c,f,p
5	c,f,a,m,p

Using this dataset, produce FP-tree.

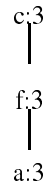


Then, generate p-conditional database and its FP-tree.

TID	Itemset
1	c,f,a,m:2
4	c,f:1

Note that a and m are pruned because they are not frequent. Since this FP-tree's height is exactly 2, we can generate one frequent itemset $\{c, f, p\}$ whose length-3 and which contains p. Then, go back to the original FP-tree.

Now, we generate m-conditional FP-tree, then we have



Since this FP-tree is a single path, we do not need any further recursive generation of FP-tree from this FP-tree and we can generate 4 frequent itemset whose length is at least 3 and which contains m, $\{c, f, a, m\}$, $\{c, f, m\}$, $\{c, a, m\}$ and $\{f, a, m\}$.

Then, we generate a-conditional FP-tree.

Since this FP-tree's height is exactly 2, we can generate one frequent itemset $\{c, f, a\}$ whose length-3 and which contains a.

Finally, for the items, c and f, we do not need FP-tree for them because the heights of their FP-tree is at most 2.

Overall, we found 6 frequent itemsets, $\{c, f, p\}$, $\{c, f, a, m\}$, $\{c, f, m\}$, $\{c, a, m\}$, $\{f, a, m\}$, and $\{c, f, a\}$.

2. Suppose that we have the following two time series and we want to minimize the sum of L_2 errors for them using only 4 Haar wavelet coefficients.

Sequence ID	Time series
1	$< 1, 2, 3, 4 >$
2	$< 1, 4, 7, 10 >$

- Show which coefficients you select and describe the reason. [10 pts.]

SOL: If we compute Haar wavelet coefficients, we get the following two collections, W1 and W2.

Wavelet ID	Normalized Haar Wavelet Coefficients
W1	$5, -2, -0.5\sqrt{2}, -0.5\sqrt{2}$
W2	$11, -6, -1.5\sqrt{2}, -1.5\sqrt{2}$

If we choose the coefficients with the largest absolute value, L_2 error is minimized by Parseval's theorem. Thus, we have to select 5 in $W1$ and 11, -6 and -1.5*$\sqrt{2}$ in $W2$.

3. Consider the following database.

Point
(1,1)
(1,2)
(2,1)
(5,5)
(5,6)
(6,6)

- For threshold $\epsilon = 2$ and branching factor $L = 2$, show the detailed steps of the pre-clustering phase of BIRCH algorithm using L_2 distance. [20 pts.]
  SOL: We scan the given dataset and find a CF-tree, i.e. pre-cluster them. First, we read (1,1) and it becomes one leaf node. Since (1,2) is within 2 from (1,1), it is absorbed and becomes (1,1.5). (2,1) is also within 2 from (1,1.5), they are merged to (4/3,4/3). Since (5,5) is out of within 2 from (4/3,4/3), it becomes a new entry. Since (5,6) is within 2 from (5,5), they are merged and they are represented as (5,5.5) and then since (6,6) is within 2 from (5,5.5), they are also merged to (16/3,17/3). Since the branching factor is 2, they all are contained in one leaf node.
- Show the detailed steps of CURE algorithm to find 2 clusters for the above data where the number of representatives $c = 2$ and the shrinking factor $\alpha = 0.5$. [25 pts.]
  SOL: First, all points are individual clusters. Then, we repeat to find a pair of the closest clusters and merge them. (1,1) and (1,2), and (5,6) and (6,6) are the closest clusters' pairs and so we merge them into 2 clusters. Then, (2,1) is merged to the cluster which consists of (1,1) and (1,2). At this time, since we have 3 points in one cluster, we have to choose 2 representatives. Since the center is (4/3, 4/3) and (1,2) or (2,1) is the farthest from the center, we arbitrary choose (1,2) and a representative is (0.5, 1) by the shrinking factor of 0.5. Then, among the remaining two points, (2,1) is chosen and another representative shrinks to (1, 0.5) since it is farther from (1,2) than (1,1). Finally, since the closest pair is the cluster of (5,5) and the cluster with (5,6) and (6,6), they become one cluster and we found 2 clusters.

4. Consider the following database. We want to compress it using Fascicle. Assume the error tolerances are 2, 10 and 0 for attributes A, B and C, respectively.

A	B	C
10	100	bad
11	101	bad
11	99	good
21	101	good
14	102	good
19	101	good

- Find 2-D fascicles using single-k algorithm where k is 2. Show the detailed steps. [10 pts.]

  SOL: First, we read the first record which becomes a 2-D tipset. Then, read the second record and compare the first record. Since all three attribute values are within error tolerances, we still have 2-D tipset. Then, if we read the third record, we find the third attribute value is out of error tolerance but we still have a 2-D tipset because the first and second attribute values are compact. The forth record eventually makes the current considering tipset to 1-D and thus the forth one should be a new 2-D tipset and the first 3 records consist of one 2-D fascicle which we found until now.

  The next 2 further records contribute the new 2-D tipset and thus it becomes another 2-D fascicle whose compact attributes are B and C.

  Now, we found two 2-D fascicles and finally maximize both fascicles by one more scan. Since the third record has the similar values which are within error tolerance from the second 2-D fascicle, the second fascicle contains the last 4 records.

- Using these 2-D fascicles, represent the given data as a compressed form. Show the detailed steps. [10 pts.]

  SOL: Now, we have two 2-D fascicles, the one consists of the first 3 records and its compact attributes are A and B. The other one is composed of the last 4 records and its compact attributes are B and C.

  By greedy selection, we first select the second fascicle since its coverage, 4×2, is larger than the other's coverage, 3×2. Then, we select the first one and so the first 2 records are covered by the first fascicle and the last 4 records are covered by the second one.

  Thus, we can represent the dataset as the following compressed form.

Fascicle ID	Values
1	A:10.5, B:100.5
2	B:100.75, C:good

Fascicles

Fascicle ID	Remaining Values
1	bad
1	bad
2	11
2	21
2	14
2	19

Compressed dataset