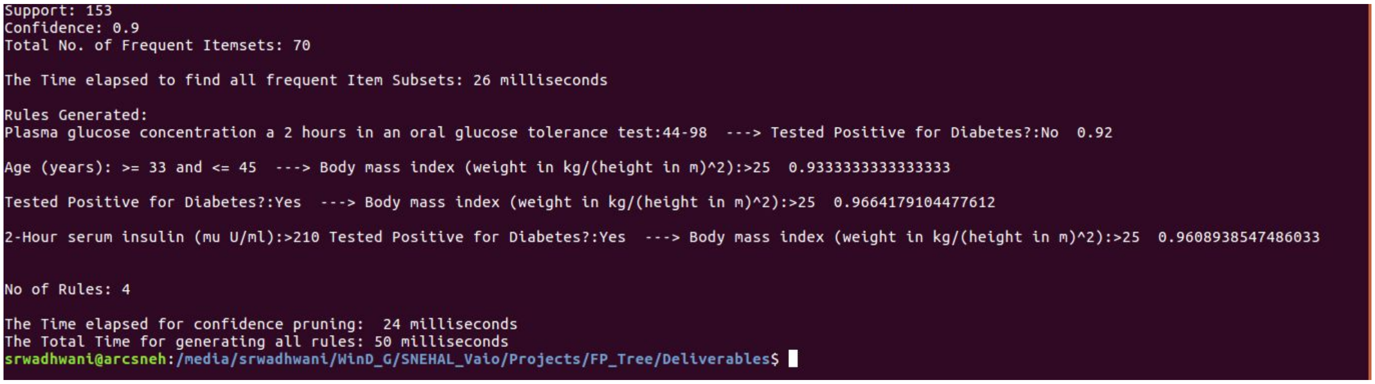
Data Mining - CS F415

Rule Mining using FP Trees

Assignment - 2



```
Support: 153
Confidence: 0.9
Total No. of Frequent Itemsets: 70

The Time elapsed to find all frequent Item Subsets: 26 milliseconds

Rules Generated:
Plasma glucose concentration a 2 hours in an oral glucose tolerance test:44-98  ---> Tested Positive for Diabetes?:No  0.92

Age (years): >= 33 and <= 45  ---> Body mass index (weight in kg/(height in m)^2):>25  0.9333333333333333

Tested Positive for Diabetes?:Yes  ---> Body mass index (weight in kg/(height in m)^2):>25  0.9664179104477612

2-Hour serum insulin (mu U/ml):>210 Tested Positive for Diabetes?:Yes  ---> Body mass index (weight in kg/(height in m)^2):>25  0.9608938547486033


No of Rules: 4

The Time elapsed for confidence pruning:  24 milliseconds
The Total Time for generating all rules: 50 milliseconds
srwadhwani@arcsneh:/media/srwadhwani/WinD_G/SNEHAL_Vaio/Projects/FP_Tree/Deliverables$ 
```

Team Members

NAME	ID Number
G V Sandeep	2014A7PS106H
Malla Naga Poojitha Reddy	2014A7PS019H
Snehal Wadhwani	2014A7PS430H

Language Chosen

Java (Development Environment - Eclipse)

Pre-processing

- Since the the attributes are continuous, the first step in pre-processing was to discretise the continuous data. The following classification principles were used to form subclasses of each of the attributes :-

Attribute	**Classification Principle**	**No. of Sub-Classes**
Number of times pregnant	Equal Frequency	6
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Equal Frequency	5
Diastolic blood pressure (mm Hg)	Domain Knowledge	4
Triceps skin fold thickness (mm)	Equal Frequency	5
2-Hour serum insulin (mu U/ml)	Equal Frequency	5
Body mass index (weight in kg/(height in m)^2)	Domain Knowledge	3
Diabetes pedigree function	Equal Width	5
Age (years)	Equal Width	5

- 0s in attributes apart from No. of pregnancies and the class to identify whether the person was tested positive for diabetes indicate missing values (This is due to the fact that attributes like Blood Pressure clearly cannot be 0 unless the data is missing). The second step of preprocessing involved handling this missing data. The mode of the attribute (ie. the class which occurs the most) replaces the missing value.
- Finally, the 9-attribute data was expanded to effectively a 40-attribute itemset (based on the number of sub-classes for each class). But, at a time since only one subclass can be present in the basket, each transaction is exactly a row of 9 items.
- The support count of each of the attributes was counted (ie. one frequent itemsets generated).

- The attributes within the transactions were sorted in descending order of their support counts, so that the FP tree is not very broad. For eg. After this preprocessing, the number of children of the root node is only 4, ie. all transactions begin with only one of those 4 attributes.

Compilation Steps

- javac FP_Growth.java
- Java FP_Growth

The program then runs for the pre-specified Minimum Support and Minimum Confidence threshold values.

Support and confidence value at which interesting rules are generated

Support : 0.2

Confidence : 0.9

No of Rules Generated : 4

1. Tested Positive for Diabetes? : Yes ---> Body mass index (weight in kg/(height in m)^2) : >25 0.9664179104477612
2. 2-Hour serum insulin (mu U/ml):>210 Tested Positive for Diabetes? : Yes ---> Body mass index (weight in kg/(height in m)^2) : >25 0.9608938547486033
3. Age (years): >= 33 and <= 45 ---> Body mass index (weight in kg/(height in m)^2) : >25 0.9333333333333333
4. Plasma glucose concentration a 2 hours in an oral glucose tolerance test : 44-98 ---> Tested Positive for Diabetes? : No 0.92

Support : 0.15

Confidence : 0.9

No of Rules Generated : 13

1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:127-150 ---> Body mass index (weight in kg/(height in m)^2):>25 0.9054054054054054
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:150+ ---> Body mass index (weight in kg/(height in m)^2):>25 0.9448275862068966
3. Diastolic blood pressure (mm Hg):>80 and <=90 ---> Body mass index (weight in kg/(height in m)^2):>25 0.952755905511811
4. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:44-98 ---> Tested Positive for Diabetes?:No 0.92
5. Age (years): >= 33 and <= 45 ---> Body mass index (weight in kg/(height in m)^2):>25 0.9333333333333333
6. Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9664179104477612
7. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:44-98 Diabetes pedigree function: >= 1.0148000000000001 and <= 1.4832000000000003 ---> Tested Positive for Diabetes?:No 0.936
8. Plasma glucose concentration a 2 hours in an oral glucose tolerance test:44-98 Age (years): >= 21 and <= 33 ---> Tested Positive for Diabetes?:No 0.9185185185185185
9. Diastolic blood pressure (mm Hg):>60 and <=80 Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9605263157894737
10. Triceps skin fold thickness (mm):>38 Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9492753623188406
11. 2-Hour serum insulin (mu U/ml):>210 Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9608938547486033
12. Diabetes pedigree function: >= 1.0148000000000001 and <= 1.4832000000000003 Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9556962025316456
13. Age (years): >= 21 and <= 33 Tested Positive for Diabetes?:Yes ---> Body mass index (weight in kg/(height in m)^2):>25 0.9596774193548387