A REPORT

ON

RULE GENERATION USING FP TREE

BY

ANIKETH JANARDHAN REDDY	2014A7PS096H
KARTIK SETHI	2014A7PS130H
MONICA ADUSUMILLI	2014A7PS005H
VINAY DATTA	2014A7PS509H

Under the supervision of

Prof. N. L. Bhanu Murthy

Department of Computer Science and Information Systems

Submitted as a part of the academic project for the course

CS F415 Data Mining

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

LANGUAGE USED

C++

DATA PRE PROCESSING

The data has 9 attributes out of which 8 are continuous and 1 is discrete (has two values). The continuous values are discretized into 5 intervals of equal width. The width of each interval of an attribute is equal to the one fifth of the range of that attribute. Hence, a total of 42 items are made to make the data suitable for rule generation.

COMPILATION STEPS

```
$ g++ datapreprocess.cpp
```

```
$ ./a.out
```

Input: pima-indians-diabetes.data **Output:** transactions.txt

The input file contains the raw data and the output file contains the modified basket data in the form of transactions (transactions.txt) which can be fed as an input to the generatefrequentitemsets.cpp It also outputs a file valtoattr.txt which has the item number and the actual name of the item.

```
$ g++ generatefrequentitemsets.cpp
```

```
$ ./a.out
```

Input: transactions.txt **Output:** frequent_itemsets.txt and tree.txt

The input file is transactions.txt and valtoattr.txt which are the outputs of the previous file and it outputs the number frequent item sets along with the frequent item sets in the file frequent_itemsets.txt for the support mentioned in the file generatefrequentitemsets.cpp. It also outputs a file which prints the fptree called tree.txt.

```
$ g++ generaterules.cpp
```

```
$ ./a.out
```

Input: frequent_itemsets.txt **Output:** rules.txt and valtoattr.txt

The input file is the frequent_itemsets.txt and the output file is the rules.txt which generates the association rules for the confidence mentioned in the file generaterules.cpp.

SUPPORT AND CONFIDENCE VALUES, NUMBER OF RULES GENERATED

SUPPORT VALUE	CONFIDENCE VALUE	FREQUENT ITEMSETS	ASSOCIATION RULES
0.2	0.8	177	133
0.2	0.9	177	42
0.3	0.7	64	78
0.3	0.8	64	36
0.3	0.9	64	6
0.4	0.8	26	10
0.4	0.9	26	2
0.5	0.8	10	4
0.5	0.9	10	0
0.6	0.7	5	0

At a support of 0.3 and a confidence of 0.9, 6 association rules were mined and at a support of 0.4 and a confidence threshold of 0.8, 10 rules are generated. These were some of the interesting rules generated by experimenting with different values of support and confidence.

It is also observed that very low support and very high support result in too many or too less frequent item sets. Also very low confidence or very high confidence is equally bad. Hence it is important to have an optimum values of support and confidence. Low values of confidence yield more number of association rules than the frequent item sets which again is sub optimal. A confidence of 0.8-0.9 and a support of 0.3-0.4 is found as the optimal one in this case.

It is also important to optimize both support and confidence values. A low value of support with optimal confidence or low confidence with optimal support hasn’t given the best of the rules.

ASSOCIATION RULES

Some of the interesting rules generated at support of 0.4 and confidence of 0.8 are below.

Confidence = 0.82783 Rule: 0<=number-of-pregnancies<5 -> 0<=serum-insulin<5

Confidence = 0.856132 Rule: 0<=number-of-pregnancies<5 -> 21<=age<26

Confidence = 0.923706 Rule: 10<=plasma-glucose-concentration<15 -> 0<=serum-insulin<5

Confidence = 0.839674 Rule: 10<=diastolic-blood-pressure<15 -> 0<=serum-insulin<5

Confidence = 0.952663 Rule: 0<=triceps-skin-fold-thickness<5 -> 0<=serum-insulin<5

Confidence = 0.821138 Rule: 10<=body-mass-index<15 -> 0<=serum-insulin<5

Confidence = 0.872137 Rule: 0.078<=diabetes-pedigree-function<5.078 -> 0<=serum-insulin<5

Confidence = 0.898907 Rule: 0.078<=diabetes-pedigree-function<5.078, has-diabetes -> 0<=serum-insulin<5

Confidence = 0.841772 Rule: 21<=age<26 -> 0<=serum-insulin<5

Confidence = 0.878 Rule: has-diabetes -> 0<=serum-insulin<5