**CSE 601: Data Mining and Bioinformatics**

Project 1

part2

Association Analysis

Jingsong Li (50322345)

Kaige Gao (50320916)

Junyang Li (50320301)

Brief description of Apriori algorithm and the flow of the association rule generation algorithm.

1. Steps to implement Apriori algorithm:
2. Import data from the given file, then add description prefix of each data as required.
3. Set the minimum support and confidence.
4. Generate candidate itemsets of length=1 and name it as d1.
5. Prune d1, delete unfrequent items and generate frequent length=1 itemsets.
6. Repeat followed steps until there is no new frequent itemsets:
7. generate length=k candidate itemsets based on previous length=(k-1) frequent itemsets, compare the first k-2 attributes of two length=(k-1) itemsets, if they are the same, merge them to a new length=k candidate itemset
8. count the support of each new length=k candidate itemset
9. eliminate infrequent candidate itemsets, generate length=k frequent itemsets and name it as dk
10. Flow of the association rule generation algorithm:
11. Start from d2, compare each item in dk with previous generated items
12. If item in dk contains the frequent set generated, we can say it can form a rule
13. Use the item in dk as rule, and item generated as body, the part not in item generated is head
14. Store rule, head and body into object
15. Store rules into a list

Result (This is the statistical result of the data. See anwser.txt file generated during program execution for detailed data.) :

Support = 30%

number of length-1 frequent itemsets: 196

number of length-2 frequent itemsets: 5323

number of length-3 frequent itemsets: 5251

number of length-4 frequent itemsets: 1463

number of length-5 frequent itemsets: 388

number of length-6 frequent itemsets: 61

number of length-7 frequent itemsets: 3

Total number of frequent itemsets: 12685

Support = 40%

number of length-1 frequent itemsets: 167

number of length-2 frequent itemsets: 753

number of length-3 frequent itemsets: 149

number of length-4 frequent itemsets: 7

number of length-5 frequent itemsets: 1

Total number of frequent itemsets: 1077

Support = 50%

number of length-1 frequent itemsets: 109

number of length-2 frequent itemsets: 63

number of length-3 frequent itemsets: 2

Total number of frequent itemsets: 174

Support = 60%

number of length-1 frequent itemsets: 34

number of length-2 frequent itemsets: 2

Total number of frequent itemsets: 36

Support = 70%

number of length-1 frequent itemsets: 7

Total number of frequent itemsets: 7

Support = 50% and Confidence = 70%

number of length-1 frequent itemsets: 109

number of length-2 frequent itemsets: 63

number of length-3 frequent itemsets: 2

Total number of frequent itemsets: 174

number of rules is 117

Template1

asso\_rule.template1("RULE", "ANY", ['G59\_UP']) 26

asso\_rule.template1("RULE", "NONE", ['G59\_UP']) 91

asso\_rule.template1("RULE", 1, ['G59\_UP', 'G10\_Down']) 40

asso\_rule.template1("HEAD", "ANY", ['G59\_UP']) 9

asso\_rule.template1("HEAD", "NONE", ['G59\_UP']) 108

asso\_rule.template1("HEAD", 1, ['G59\_UP', 'G10\_Down']) 17

asso\_rule.template1("BODY", "ANY", ['G59\_UP']) 17

asso\_rule.template1("BODY", "NONE", ['G59\_UP']) 100

asso\_rule.template1("BODY", 1, ['G59\_UP', 'G10\_Down']) 24

Template2

asso\_rule.template2("RULE", 3) 9

asso\_rule.template2("HEAD", 2) 6

asso\_rule.template2("BODY", 1) 117

Template3

asso\_rule.template3("1or1", "HEAD", "ANY", ['G10\_Down'], "BODY", 1, ['G59\_UP']) 24

asso\_rule.template3("1and1", "HEAD", "ANY", ['G10\_Down'], "BODY", 1, ['G59\_UP']) 1

asso\_rule.template3("1or2", "HEAD", "ANY", ['G10\_Down'], "BODY", 2) 11

asso\_rule.template3("1and2", "HEAD", "ANY", ['G10\_Down'], "BODY", 2) 0

asso\_rule.template3("2or2", "HEAD", 1, "BODY", 2) 117

asso\_rule.template3("2and2", "HEAD", 1, "BODY", 2) 3