UNIVERSITY AT BUFFALO

CSE601 – Data Mining and Bioinformatics

Project 1 – Part 2 Report (Association)

Aayush Shah (50207564)
Aniruddh Chaturvedi (50206958)
Haril Satra (50208283)
9/30/2017

Apriori Algorithm

A. Frequent Itemset Generation

1. Let k be the length of the frequent itemsets. Set k=1.
2. Generate length 1 frequent itemsets. This can be done by basically counting the occurrences of the individual items. The items whose count is greater than the minimum support will be considered as length 1 frequent itemsets.
3. Generate length (k+1) candidate itemsets from length k frequent itemsets. We can use the Apriori principle which states 'If an itemset is frequent, then all of its subsets must also be frequent' to prune the candidates. This can be done by creating length (k+1) candidates from taking union of candidates of the previous frequent itemsets whose first k-1 items are the same.
4. Count the support of each candidate in the newly generated frequent itemset generated in step 4.
5. Eliminate candidates that are infrequent, leaving only those that are frequent.
6. Repeat step 2 to step 5 until no new frequent itemset is generated.

B. Association Rule Generation

1. For each frequent itemset 'L', generate all non empty subsets of L.
2. For every non empty subset f of L, calculate the confidence as following:

   confidence = support_count(L) / support_count(f)

3. If confidence >= min_conf where min_conf is the minimum confidence threshold, output the rule " **f → (L-f)** ".

Results

Task 1 (Variable length frequent itemsets for various support count)

SUPPORT: 30%

Number of length 1 frequent itemsets: 196
Number of length 2 frequent itemsets: 5340
Number of length 3 frequent itemsets: 5287
Number of length 4 frequent itemsets: 1518
Number of length 5 frequent itemsets: 438
Number of length 6 frequent itemsets: 88
Number of length 7 frequent itemsets: 11
Number of length 8 frequent itemsets: 1
Number of length 9 frequent itemsets: 0
Number of all lengths frequent itemsets: 12879

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
Number of length 6 frequent itemsets: 0
Number of all lengths 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
Number of length 4 frequent itemsets: 0
Number of all lengths frequent itemsets:174

SUPPORT: 60%

Number of length 1 frequent itemsets: 34
Number of length 2 frequent itemsets: 2
Number of length 3 frequent itemsets: 0
Number of all lengths frequent itemsets:36

SUPPORT: 70%

Number of length 1 frequent itemsets: 7
Number of length 2 frequent itemsets: 0
Number of all lengths frequent itemsets:7

Task 2 (Answer to sample queries with Support = 50% and Confidence = 70%)

TEMPLATE 1:

(result11, cnt) = asso_rule.template1("RULE", "ANY", ["G59_Up"]) **26**
(result11, cnt) = asso_rule.template1("RULE", "NONE", ["G59_Up"]) **91**
(result13, cnt) = asso_rule.template1("RULE", 1, ["G59_Up","G10_Down"]) **39**
(result14, cnt) = asso_rule.template1("BODY", "ANY", ["G59_Up"]) **9**
(result15, cnt) = asso_rule.template1("BODY", "NONE", ["G59_Up"]) **108**
(result16, cnt) = asso_rule.template1("BODY", 1, ["G59_Up","G10_Down"]) **17**
(result17, cnt) = asso_rule.template1("HEAD", "ANY", ["G59_Up"]) **17**
(result18, cnt) = asso_rule.template1("HEAD", "NONE", ["G59_Up"]) **100**
(result19, cnt) = asso_rule.template1("HEAD", 1, ["G59_Up","G10_Down"]) **24**

TEMPLATE 2:

(result21, cnt) = asso_rule.template2("RULE", 3) **9**
(result22, cnt) = asso_rule.template2("BODY", 2) **6**
(result23, cnt) = asso_rule.template2("HEAD", 1) **117**

TEMPLATE 3:

(result31, cnt) = asso_rule.template3("1or1", "BODY", "ANY", ["G10_Down"],"HEAD", 1, ["G59_Up"]) **24**
(result32, cnt) = asso_rule.template3("1and1", "BODY", "ANY", ["G10_Down"],"HEAD", 1, ["G59_Up"]) **1**
(result33, cnt) = asso_rule.template3("1or2", "BODY", "ANY", ["G10_Down"],"HEAD", 2) **11**
(result34, cnt) = asso_rule.template3("1and2", "BODY", "ANY", ["G10_Down"],"HEAD", 2) **0**
(result35, cnt) = asso_rule.template3("2or2", "BODY", 1, "HEAD", 2) **117**
(result35, cnt) = asso_rule.template3("2and2", "BODY", 1, "HEAD", 2) **3**