CSE 601: Data Mining and Bioinformatics
Project 1 - Part B:

Apriori Algorithm

Association Rules Mining

University at Buffalo

Prashanth Seralathan #50204883
Manishkumarreddy Jarugu #50206843
Haneesh Reddy Poddutoori #50208280

Apriori Algorithm:

Apriori algorithm is used for generating the association rules on a dataset - This association rules can be used for prediction of occurrence of a particular itemset based on the occurrences of other items.

For Example: A simple association rule can be shown in the form of an implication expression,

{A, B} -> {C}

Here occurrence of {A,B} implies the occurrence of {C}.

Association rule mining tasks will involve generating of all possible rules for a given support, confidence factor. Association mining task would involve listing all possible combinations of a given itemset to figure out the combination that is most definitely possible to occur. This term “most definitely” can be interpreted in terms of **support** and **confidence**.

Support: Fraction of transactions that contains both the implication and the occurrence itself.

$$\text{Support} = \frac{Number\ of\ Transactions\ that\ contains\ (\{A,B,C\})}{Total\ Number\ of\ Transactions}$$

Confidence: Measures how strong the implication factor is, For example {A,B} -> {C} we measure the confidence as,

$$\text{Confidence} = \frac{Number\ of\ Transactions\ that\ contains\ (\{A,B,C\})}{Number\ of\ Transactions\ that\ contains(\{A,B\})}$$

The main methodology involved in Apriori algorithm is we first calculate the frequently occurring items and proceed in such a way of incrementally apply the algorithm extending them to bigger and bigger itemsets in such a way providing an efficient way for mining the association rules.

The Apriori algorithm for association rules generation can be described as follows,

1. Initially generate itemsets of length 1 and retain the frequent itemsets.
2. Generate itemsets of length one greater than the previous itemset length.
3. Validate the itemset subsets of length(n-1) whether they are part of the previous frequent itemsets.

4. Remove the itemsets that are having support lesser than the minimum support specified.
5. Repeat steps 2 to 4 until no further successful frequent itemsets can be generated.

Association Rule generation:

Each and every itemset is split into two subsets with each subset of minimum length one. Out of the two subsets, one of it is named as head and the other one is named as body.

A RULE is defined as BODY → HEAD.

Let's say there is an itemset {A,B,C}

This is split into 2 subsets with each subset of minimum length one.

The possible splits and the prospective rules that can be generated by these splits are shown in the table

Splits	Rule	Body	Head
{A} {B,C}	{A} → {B,C}	{A}	{B,C}
	{B,C} → {A}	{B,C}	{A}
{B}, {A,C}	{B} → {A,C}	{B}	{A,C}
	{A,C} → {B}	{A,C}	{B}
{C} {A,B}	{C} → {A,B}	{C}	{A,B}
	{A,B} → {C}	{A,B}	{C}

These rule are valid if and only if it has confidence greater than or equal to the confidence threshold. In general, for a itemset of size k there will be $2^k - 2$ prospective rules that is ignoring the cases where a head or body might be an empty set.

Results:

Frequent Itemsets results:

The results obtained by different support values for different length and total frequent itemsets for each support value is tabulated below

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
Total: 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
Total: 1077

Support - 50%
Number of length 1 frequent itemsets:109 Number of length 2 frequent itemsets:63 Number of length 3 frequent itemsets:2
count :174

Support - 60%
Number of length 1 frequent itemsets:34 Number of length 2 frequent itemsets:2
count :36

Support - 70%
Number of length 1 frequent itemsets:7
count :7

Number of Rules generated for given Support/Confidence:

Support/Confidence Values	Number of Rules
Support - 30%, Confidence - 70%	31759
Support - 40%, Confidence - 70%	1137
Support - 50%, Confidence - 70%	117
Support - 60%, Confidence - 70%	4
Support - 70%, Confidence - 70%	0

From the above table, we can see for a given confidence as the support increases the number of rules generated decreases.

Template 1 Results:

Query	Result Count
RULE, ANY, [G59_Up]	26
RULE, NONE, [G59_Up]	91
RULE, 1, [G59_Up,G10_Down]	39
BODY, ANY, [G59_Up]	9
BODY, NONE, [G59_Up]	108
BODY, 1, [G59_Up,G10_Down]	17
HEAD, ANY, [G59_Up]	17
HEAD, NONE, [G59_Up]	100
HEAD, 1, [G59_Up,G10_Down]	24

Template 2 Results:

Query	Result Count
RULE, 3	9
BODY, 2	6
HEAD, 1	117

Template 3 Results:

Query	Result count
"1or1", BODY, ANY, [G10_Down],HEAD, 1, [G59_Up]	24
"1and1", BODY, ANY, [G10_Down],HEAD, 1, [G59_Up]	1
"1or2", BODY, ANY, [G10_Down],HEAD, 2	11
"1and2", BODY, ANY, [G10_Down],HEAD, 2	0
"2or2", BODY, 1, HEAD, 2	117
"2and2", BODY, 1, HEAD, 2	3