UNIVERSITY AT BUFFALO

CSE 601
Data Mining and Bioinformatics
Fall 2019

Project 1
Dimensionality Reduction & Association Analysis

Part 2

Amit Anilkumar Menon (amitanil@buffalo.edu)
Linus Castelino (linuscas@buffalo.edu)
Deepak Ranjan (dranjan@buffalo.edu)

Apriori Algorithm

Apriori algorithm provides a series of steps to realize association rules from a given set of records by identifying frequent item sets.

The algorithm begins with an initial set of length-1 item sets and proceeds further by merging and generating item sets of a greater length until all the frequent item sets are obtained. The non-frequent item sets are filtered based on a given minimum support value at every step of execution. This filtration is known as the pruning step in which all item sets that do not satisfy the minimum support constraint are ignored.

Once the frequent item sets are obtained, association rules are isolated using binary partitioning and are filtered based on a given minimum confidence value.

Implementation

Our implementation of the Apriori algorithm is performed in 4 stages as follows-

Stage 1: Data preprocessing

Input data provided is a gene expression dataset with dimensions 100 rows (records) x 101 columns.

Each value in column 1-100 represents a gene expression as Up or Down that indicates a binary representation of the presence of the gene expression in the patient sample. The last column indicates the disease corresponding to the patient sample.

In the preprocessing step, we rearrange the columns to begin at index 1 (instead of 0 – default in python Dataframe) and append the respective gene expression identifier to every value in the column (For example, every value in column 1 is appended with 'G1_', every value in column 2 is appended with 'G2_', and so on.)

Stage 2: Frequent item set generation

For frequent item set generation, we initially iterate through every cell in the dataset and store all the unique items as an array of sets (raw_set[]).

We then compute the support value for every item set in the array and prune the item sets that are lesser minimum support value. The item sets having support greater than or equal to the minimum support value are considered to be frequent item sets. The resulting item sets are then merged to form new item sets of length +1 than the length of the item sets in the current step. We continue this until we are unable to create any more frequent item sets.

The merging of frequent item sets of length k to form new k+1 length item sets is performed using the F_{k-1} x F_{k-1} method mentioned in the reference book Chapter 6.

Stage 3: Association rule generation

For generating association rules from the identified frequent item sets in the previous stage, we use the partitioning method.

For every frequent item set, we generate all rules having 1 item in the consequent.

From the obtained rules, we prune all rules having confidence lesser than the given minimum confidence value. We then merge the available rules to generate new rules by using intersection of the antecedents and union of the consequents. This step is repeatedly performed using recursion until no new rules can be formed.

Stage 4: Query processing

We have created 3 methods to process the user queries based on the query templates provided in class.

We accept the user query in the following formats:

Template 1 - (RULE|BODY|HEAD;ANY|NUMBER|NONE;ITEM1,ITEM2,...)

Template 2 - (RULE|BODY|HEAD;NUMBER)

Template 3 - (1or1;HEAD;ANY;G10_Down;BODY;1;G59_UP)

Note - The template parameters are semicolon delimited and the list items are comma delimited.

Implementation Output

Following is the output generated by our implementation of the Apriori algorithm for the given tasks.

Task 1 –

Find frequent item sets for support 30%, 40%, 50%, 60% and 70%.

Support is set to be 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 all lengths frequent itemsets: 12879

Support is set to be 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 all lengths frequent itemsets: 1077

Support is set to be 50%
Number of length-1 frequent itemsets: 109
Number of length-2 frequent itemsets: 63
Number of length-3 frequent itemsets: 2
Number of all lengths frequent itemsets: 174

Support is set to be 60%
Number of length-1 frequent itemsets: 34
Number of length-2 frequent itemsets: 2
Number of all lengths frequent itemsets: 36

Support is set to be 70%
Number of length-1 frequent itemsets: 7
Number of all lengths frequent itemsets: 7

Task 2 –

Association rules generation for support 50% and confidence 70% for given queries.

Number of rules generated for support 50% and confidence 70% - 117

Template 1 Queries:

Queries	Rule Count
RULE;ANY;G59_Up	26
RULE;NONE;G59_Up	91
RULE;1;G59_Up,G10_Down	39
HEAD;ANY;G59_Up	17
HEAD;NONE;G59_Up	100
HEAD;1;G59_Up,G10_Down	24
BODY;ANY;G59_Up	9
BODY;NONE;G59_Up	108
BODY;1;G59_Up,G10_Down	17

Template 2 queries:

Queries	Rule Count
RULE;3	9
HEAD;2	3
BODY;1	117

Template 3 queries:

Queries	Rule Count
1or1;HEAD;ANY;G10_Down;BODY;1;G59_Up	16
1and1;HEAD;ANY;G10_Down;BODY;1;G59_Up	0
1or2;HEAD;ANY;G10_Down;BODY;2	13
1and2;HEAD;ANY;G10_Down;BODY;2	0
2or2;HEAD;1;BODY;2	117
2and2;HEAD;1;BODY;2	6