Konrad Rauscher
Introduction to Data Mining
Project 2
3/21/16

Status:
Completed

Approx. Time Spent on Project:

45 hours

Things I wish I had been told prior to being given assignment:
Nothing in particular pertaining to the concepts of this class.

Design:
I chose the FP-Tree implementation because the runtime for item set generation increases linearly, depending on the number of transactions and items.

My code executes as follows:
1) The transactions and products files are both loaded as lists of lists: transactions_prelim, and products respectively
2) The function frequentizeAndOrder is then called, returning a list of lists which is then assigned to the variable transactions. This function does the following:
    a. particularly non-frequently-occuring products (as defined by having a frequency of less than .002 % are removed from transactions. With the given data set, this reduces the number of products that are kept track of from 100 to 93. Not only does this decrease runtime, but makes intuitive sense, as products that are particularly seldom bought would be of less interest to grocery store owners.
    b. Then, the placement of products for all transactions are re-ordered according to the frequency by which each product occurs, with the most-frequent occurring product being placed first, 2nd most-frequent 2nd, etc.
3) The function createItemSets is called, with transactions, a min support value, and a min confidence value being passed as parameters. This functions returns a list of of frequent item sets in the form of: *[Set], support, confidence, K.* This functions does the following:
    a. creates an empty FPTree
    b. this FP Tree is then populated, one transaction at a time.
    c. a dictionary of the supports for all item sets is created for soon-to-occur confidence calculations of item sets that pass the minimum support threshold.
    d. a recursive call is then made to find_with_suffix, passing the populated FP Tree and an empty list, yielding frequent item sets that meet the minimum support threshold
4) Next, the largest K value of the frequent item sets is found and then used to create a list in which the frequency of lists with each K value are recorded. The number of frequent item sets that occur for each of the occurring K values (excluding 1) and the largest K that occurs in the set of generated frequent item sets are then printed to the screen.
5) Finally association rules are generated by:
    a. iterating through the list of frequent item sets and for each item set:
        i. generate a list of all subsets by calculating the combinations of the current item set
        ii. iterate through each possible subset and if the confidence of a subset in relation to the item set it is a member of is above the confidence threshold, then a rule is generated
    b. the rules are then sorted on support and then printed to the screen

BE SURE TO INCLUDE THIS DOCUMENT IN PROJECT FOLDER

BE SURE TO UPLOAD CORRECT FOLDER

MAKE PDF

Report 1:

Min-support	Min-confidence	Number of frequent item sets (report for each K)	Value of the Largest K	Number of Rules	Total Run Time (sec.)
0.2	0.75	K = 2, 69 K = 3, 3	3	11	45.4
0.4	0.75	K = 2, 1 K = 3, 0	2	1	16.6
0.5	0.75	NA	NA	0	16.4
0.2	0.6	K = 2, 69 K = 3, 3	3	65	42.1
0.4	0.6	K = 2, 1 K = 3, 0	2	2	14.6
0.5	0.6	NA	NA	0	15.4
0.2	0.5	K = 2, 69 K = 3, 3	3	83	42.3
0.4	0.5	K = 2, 1 K = 3, 0	2	2	14.9
0.15	0.75	K = 2, 350 K = 3, 51 K = 4, 1	4	71	53.7
0.15	0.6	K = 2, 350 K = 3, 51 K = 4, 1	4	183	51.3
0.15	0.5	K = 2, 350 K = 3, 51 K = 4, 1	4	246	52.1
0.05	0.1	K = 2, 3203 K = 3, 12433 K = 4, 2238 K = 5, 61	5	112196	396.8

Report 2:
The top 15 rules (based on confidence & correlation)

Rule	Support	Confidence
Puff'sPlusFacialTissue, SkippyPeanutButter, --> Avocado	0.65675	0.93227513227
SkippyPeanutButter, --> Avocado	0.65675	0.80526735834
Dole6-PackPineappleJuice, Puff'sPlusFacialTissue, --> Avocado	0.65675	0.932
RealLemonPureLemonJuice, --> Dole6-PackPineappleJuice	0.512	0.932
SkippyPeanutButter, --> Puff'sPlusFacialTissue	0.48275	0.92007104795
Avocado, SkippyPeanutButter, --> Puff'sPlusFacialTissue,	0.48275	0.87314172448
GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels	0.444	0.97668393782
Avocado, GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels	0.444	0.97902097902
Bounty8-PackWhitePaperTowels, --> GeneralMillsMultiBranChex,	0.386	0.84909909909
Avocado, Bounty8-PackWhitePaperTowels, --> GeneralMillsMultiBranChex,	0.386	0.85440278988
Puff'sPlusFacialTissue, --> Avocado	0.65675	0.93837389953
Lettuce, SkippyPeanutButter, --> Avocado,	0.65675	0.81100917431
BigelowApple&CinnamonHerbTea, --> Avocado	0.65675	0.65812542144
KraftMacaroni&Cheese, --> Avocado	0.65675	0.64610866373

Analysis:
In creating my list the 15 top rules, I had to make a decision as to which metric I would prioritize for determining the value/interestingness of a rule. I chose to prioritize Confidence over Support because it is a measure of correlation between the items in the rule, rather than support, which is more of a measure of the proportion of frequency that the rule occurs within the entire set of transactions.

One interesting observation regarding the generation of rules from varying support and confidence thresholds is that identical support min-support thresholds correspond to the same number of frequent item sets that are generated, regardless of the confidence threshold used. Rule generation, however, does vary. This makes sense because the Min-support is being used for frequent item set generation and min-confidence is being used for association rule generation!

Another observation is that a min-support threshold of .5 generated no frequent item sets. This is because there are no item sets (of size greater than 1) that exist in more than half of all the transactions in the given data set.

I found one of the best rule generation settings to be with min-Support = .15 and min-Confidence = .75. Not only did these settings provide an adequately large set of rules, 71, (so

as not to miss interesting rules) which was also not overwhelmingly large (which involves a long run time), but I think these settings are also great because they make intuitive sense, in that they generate rules that are particularly strong, meaning that they are defined by being highly correlative (over a support of .75). It makes intuitive sense that one would want to find rules that express strong relationships between objects. Further, with a min-Support = .15, the rules that are generated are relevant in that the item sets that they describe occur concurrently relatively frequently.

The top rules that I found to be most interesting are:

Avocado, GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels
Interesting because: this rule had the highest confidence (0.979020979021) of all the rules generated. A personal theory of mine for the particularly high correlation between these two sets is that both GeneralMillsMultiBranChex and Bounty8-PackWhitePaperTowels are bought in high frequency by the same subset of customers: families, and are thus frequently bought together.

RealLemonPureLemonJuice, --> Dole6-PackPineappleJuice
Interesting because: These are both types of citrus juice, and it makes intuitive sense that a costumer who enjoyed one type of citrus would also enjoy and thus purchase other types of citrus as well.

SkippyPeanutButter, --> Puff'sPlusFacialTissue
Interesting because: this rules makes intuitive sense from personal experience. My face usually become more greasy and susceptible to acne when I eat greasy foods like peanut butter, which results in a greater desire for me to clean my face, often with facial tissues. It is quite possible that the same dynamic is occurring to some extent here.

Addressing questions raised from intriguing observations:
Why are there only unique support values out of 15 association rules and why do rules such as 8 and 9 share almost the same support and confidence?
One explanation is that there are several rules that are subsets or reciprocals of other rules also in my top list of 15. For example, because Avocado, GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels and Avocado, Bounty8-PackWhitePaperTowels, --> GeneralMillsMultiBranChex, are both variations of the same co-occurance and are both in my list of 15 rules, it makes sense that they would share the same support and confidence values. Further, because GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels is a subset of Avocado, GeneralMillsMultiBranChex, --> Bounty8-PackWhitePaperTowels, it makes sense that they would share the same support and confidence values.

Why are avocados such a frequent item in these top 15 association rules (are included in10 of 15)? From this observation, It must be that avocados are a product that are purchased by a relatively very high proportion of costumers. Interesting inferences from this observation could be that perhaps this data set of transactions was recorded either during the summer, at a store location neighboring a large latin population, or in California.