CSCI 5523 Project 2 - Association Analysis

Due April 8, 2019

Software and Data

- Download the Apriori software from Christian Borgelt's web page at http://www.borgelt.net/apriori.html (Linux and Windows executables are available).
- It's a Command Line Interface tool so it doesn't work the way most utilities, such as Weka(Graphical User Interface) do. Please use Command Prompt or Powershell on Windows or Terminal on UNIX systems to run this tool. https://www.computerhope.com/issues/chusedos.htm is a good resource if you're new to command line interfaces.
- Look at the usage and different options that Apriori provides (run "apriori" in the command line to see them). This will help you create the commands you need to run for each question.
- Data sets:
  • mushroom
  • T10I4D100K
  • Teams and Teams_labels
  • la1

Submission Guidelines

- Prepare a report (*no more than 6 pages, no smaller than 11pt font, 1-column format*) addressing the questions asked in the following problems. **Submit the hard copy of your report in class on April 8, 2019.**
- You need to submit a ZIP containing all the outfiles on Canvas the same day by midnight.
  Generating an outfile is easy, you just need to specify the filename to which you want Apriori to write its sets/rules to as below. Some of them can be huge.
  - usage: apriori [options] infile [outfile]
- Your project will be evaluated based **only** on what you write in your report.
- Please **provide the answer to each specific question we asked separately**; do not compose a big paragraph that includes everything.
- If you are a **UNITE** student, you should submit your project report via UNITE as homework.

Problem 1:

The purpose of this problem is to make you familiar with the Apriori algorithm. You will use it on real and synthetic data sets to find frequent itemsets.

a. Run the Apriori algorithm to generate all frequent itemsets from the ' T10I4D100K ' dataset at the support thresholds of 0.0075%, 0.0275% and 0.045%, and report the number of frequent itemsets so produced. You can use a command similar to the one provided below to generate frequent itemsets:

```
apriori -ts  -s0.0075 T10I4D100K <your_outfile_name_here>
```

Compare the performance of the algorithm in terms of the time taken to produce the results at different thresholds and comment on the possible reason(s) for this difference in performance. You may estimate the amount of time spent by adding up the time displayed by the program when it is executed. Include all the times displayed by the program.

b. Run Apriori (using the -ts option) on the 'mushroom' data set to generate frequent itemsets of sizes 3 through 15 at support thresholds of 4%, 7% and 14%. **In a single figure**, for each threshold, plot the number of frequent sets (on Y-axis) obtained vs size of itemsets (on X-axis) from 3 to 15. Comment on the general trends illustrated by the plots, and comment on the reason(s) for these trends. Also comment on how the plots differ for the three thresholds. Justify your comments with proper reasoning. (You may find the –Z option useful for this problem).

c. Use the Apriori algorithm to generate closed frequent (using the -tc option) and maximal (using the -tm option) frequent itemsets from ‘mushroom’ and ‘T10I4D100K’ datasets. Use a support threshold of 7.5% for the 'mushroom' data set and 0.025% for the 'T10I4D100K' data set. Compare the total number of closed frequent and maximal frequent itemsets obtained for each dataset individually. How do these numbers compare with the number of frequent itemsets(using the -ts option) obtained from these datasets using the same threshold? What relationship among closed frequent, maximal frequent and frequent itemsets is revealed by this comparison? Why do you think this is so? NOTE: In the Apriori program, “closed” is used to refer to “closed frequent”.

Problem 2:

This question uses the data set 'Teams' that is based on Team A's performance against two teams B and C. The data set is in the file Teams.dat and contains the following items:
1: “Team_B” represents the scenario that Team A plays against Team B
2: “Team_C” represents the scenario that Team A plays against Team C.
3: “Home” represents the scenario that the game is played at Team A's home
4: “Away” represents the scenario that the game is played away from Team A's home.
5: “Won_by_A” represents the scenario that the game is won by Team A.

Every observation is analogous to transaction dataset. For example, an observation {1, 4, 5} indicates that Team A played a match against Team B away from Team A' home and the match was won by Team A.

Generate rules with 'Won_by_A' as a consequent by using the following command line , and then use these rules to answer the questions below:

```
apriori –tr –s0 –c0 Teams Q2_part1.txt
```

Refer to http://www.borgelt.net/doc/apriori/apriori.html#appearin for more details on how to generate rules with 'Won_by_A' as the consequent.

1. Find the probability of Team A winning against Team B (P[Won_By_A|Team_B]). State which rule you used to find this conditional probability. Also find the probability of Team A winning against Team C (P[Won_By_A|Team_C]). State which rule you used to find this conditional probability. Compare these probabilities.
2. From the rule set, compare the probabilities of Team A winning against Team B and C at a home venue (P[Won_by_A|Team_B,Home] and P[Won_by_A|Team_C,Home] respectively). State the rules that you based your comparison upon.
3. From the rule set, compare the probabilities of Team A winning against Team B and C when the games are played away from A's home (Pr[Won_by_A|Team_B,Away] and P[Won_by_A|Team_C,Away] respectively). State the rules that you based your comparison upon.
4. Based on the results in (2) and (3), which team is more likely to lose against Team A? Is this consistent with what you observe from the results in (1)?

Problem 3:

You are given the transaction data set 'la1'. Each of the transactions is a document and the items are words. This data set was created from a set of documents that appeared as articles in the *Los Angeles Times*.

a) What is the support of each individual item except 1 (1 has 100% support)?
(The Apriori algorithm can give this to you if you set the parameters (-m, -n, and -s) correctly).

b) Use Apriori to find frequent itemsets of size 2 or larger. First, try setting the support threshold to 15% and 40% respectively. What itemsets do you find?

c) Now lower the support threshold until you find frequent itemsets of size 2 or larger such that every item appears in at least one such itemset of size 2 or more.

(i) What is the highest support threshold that accomplishes this?
(ii) How could you use the results in (a) to help determine the starting support threshold for answering the above question (question (i) in part c)?
(iii) What are the itemsets obtained for the support threshold you listed for part (i)?

d) Now run the *hyperclique* program on your data set. You can download the Linux version (hclinux) or Windows version (hcwin.exe) from **Canvas**. Note that this program has options similar to those of Apriori. Use the command line:

```
hcwin -th -s0 -c30 la1 la1.out
```

or

```
hclinux -th -s0 -c30 la1 la1.out
```

What patterns do you find? Note that in these programs, the rest of the parameters are the same as the Apriori algorithm, but the "-th" option specifies that hypercliques should be found and the "-c30" option represents an h-confidence threshold of 30%.

e) The 6 items in the data set are 'headline', 'writer', 'hong', 'kong', 'puerto', 'rico'. Based on this information and your knowledge of frequent itemset patterns and hyperclique patterns, comment on differences between the patterns of (c) and (d) in the following three ways:

(i) in terms of the overlap across patterns (For instance, patterns {a,b} and {a,c} have overlap because of common item {a}, whereas patterns {a,b} and {c,d} have no overlap).
(ii) the difference in frequencies among items within a pattern, and
(iii) the coherence of patterns in terms of the phrases they represented.