# CSC 466 Lab 2 Report: Optimizing Association Rules in Data Mining with the Apriori Algorithm

Martin Solomon Hsu, mshsu@calpoly.edu
Lana Mai Huynh, lmhuynh@calpoly.edu

### Abstract

In this lab, we mined association rules to discover skyline frequent itemsets and uncover multiple relationships that exist in the bakery and fantasy bingo dataset. We utilized Apriori and genRules algorithms to implement our discovery algorithm. Additionally, we applied a “guess and check” method to determine the optimal minimum support and confidence levels of our datasets. In the end, the associations that we found from our algorithms can be used to make strategic decisions in that domain.

## I. Introduction

Association rules can help us uncover hidden relationships between variables in large datasets. We can use association rules in different applications such as market basket analysis, customer segmentation, and fraud detection. Because of its critical application in the real world, in this lab, we explore methods to find association rules by utilizing an Apriori algorithm. One way to determine the strength of an association is to optimize the support and confidence of such association rules. In this lab report, we discover methods to find association rules and optimize the minimum support and confidence for multiple datasets.

## II. Discussion

### Description of Procedures

We started out with a relatively low *minConf* of 0.7. Then, we wanted to optimize the *minSup* first, so we started out with a relatively high value of 0.1. For every iteration, we would decrease the *minSup* until we reached a number of association rules that felt sufficient — in the case of the *Extended Bakery* dataset, we thought ten rules was a good number. Once we achieved the sweet spot number of association rules, we used the minimum support in our association rules as the *minSup* and the minimum confidence level in our association rules as our *minConf*. Additionally, we utilized visualizations to verify our values. In figure 1, we set our support to 0.02, and we notice that the subsequent itemsets are higher than that threshold. Then, we graph the confidence of such itemsets using the support of 0.02 to see the distribution. We used the graph to see if there was a clear jump in the confidence to use as a threshold. However, there was not a distinct value that we could use, so we ended up using 1 as our *minConf*.

*Figure 1. Support of different itemsets for 1k BAKERY dataset, minSup = 0.02*

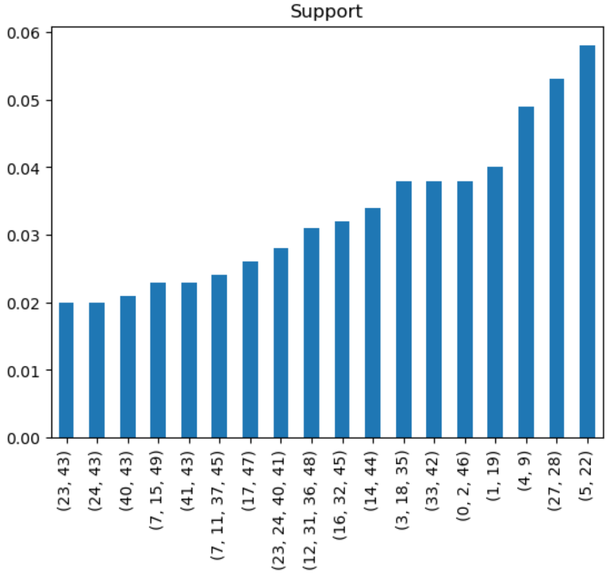

*Figure 2. Confidence of association rules given a support of 0.02 for 1k BAKERY dataset*

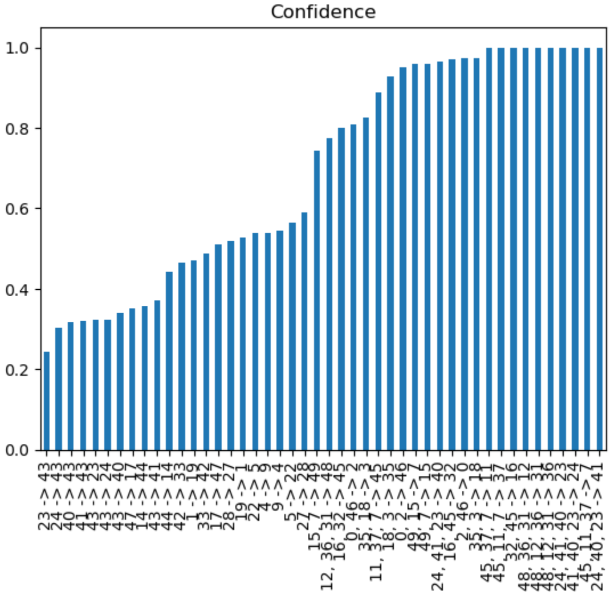

*Table 1. minSup and min Conf parameters for each Extended Bakery dataset*

| Dataset | *minSup* | *minConf* | Number of skyline association rules induced |
|---|---|---|---|
| 1,000 purchases | 0.024 | 1 | 10 |
| 5,000 purchases | 0.021 | 0.99 | 10 |
| 20,000 purchases | 0.020 | 0.98 | 10 |
| 75,000 purchases | 0.020 | 0.98 | 10 |

*Table 2. Skyline association rules for the largest Extended Bakery dataset*

| Association rule | Support | Confidence |
|---|---|---|
| Hot Coffee, Apple Pie, Almond Twist ---> Coffee Eclair | 0.02792 | 0.99525 |
| Hot Coffee, Almond Twist, Coffee Eclair ---> Apple Pie | 0.02792 | 0.99289 |
| Hot Coffee, Apple Pie, Coffee Eclair ---> Almond Twist | 0.02792 | 0.99383 |
| Cherry Soda, Apple Danish, Apple Croissant ---> Apple Tart | 0.02059 | 0.98974 |
| Cherry Soda, Apple Tart, Apple Danish ---> Apple Croissant | 0.02059 | 0.99293 |
| Cherry Soda, Apple Tart, Apple Croissant ---> Apple Danish | 0.02059 | 0.99101 |
| Lemon Lemonade, Raspberry Lemonade, Green Tea, Lemon Cookie ---> Raspberry Cookie | 0.02073 | 1 |
| Lemon Lemonade, Raspberry Lemonade, Green Tea, Raspberry Cookie ---> Lemon Cookie | 0.02073 | 0.99936 |
| Raspberry Lemonade, Green Tea, Raspberry Cookie, Lemon Cookie ---> Lemon Lemonade | 0.02073 | 1 |
| Lemon Lemonade, Green Tea, Raspberry Cookie, Lemon Cookie ---> Raspberry Lemonade | 0.02073 | 1 |

**Optimal Parameters for the *Fantasy Bingo* Dataset**

When determining the optimal *minSup* and *minConf* parameters for the *Fantasy Bingo* dataset, we wanted to find a number of skyline frequent itemsets and association rules that made sense in real world application. Originally, we tried to capture association rules with two or three authors on the left side because we thought the public would be more interested in knowing the association between groups of two or more authors. However, our list of skyline frequent item sets was quite large, so we gradually lowered our support at the cost of decreasing the cardinality in our set. In the end, we attained an optimal *minSup* = 0.15 and *minConf* = 0.4.

After finalizing the list of association rules, we explored the authors to verify the validity of our results. In general, the results were not too surprising; a comic book author would point to a fiction author. Another feature of the list that we noticed was that many of the rules were symmetric; for example, Layman, John points to Banks, Iain M, and  Banks, Iain M. points to Layman, John. Intuitively, this makes sense as people generally read the same group of authors.

*Table 3. List of skyline frequent itemsets for the Fantasy Bingo dataset*

| Size k=1 Skyline Frequent Itemsets, minSup=0.15 | |
|---|---|
| ('Aguirre, Ann') | ('Armstrong, Kelley') |
| ('Ballantine, Philippa / Ballantine, Pip') | ('Brett, Peter V.') |
| ('Butler, Octavia E.') | ('Chambers, S. J.') |
| ('Eddings, David') | ('Gleason, Patrick') |
| ('Hocking, Amanda') | ('Kirkman, Robert') |
| ('Leavy, Barbara Fass') | ('Maberry, Jonathan') |
| ('McClung, Michael') | ('Miller, Frank') |
| ("O'Donnell, Peter") | ('Rowland, Diana') |
| ('Schmitz, James H.') | ('Summers, Ella') |
| ('Van Coops, Nathan') | ('Vaughn, Carrie') |
| Size k=2 Skyline Frequent Itemsets, minSup=0.15 | |
| ('Banks, Iain M.', 'Layman, John') | ('Banks, Iain M.', 'Presley, M. D.') |
| ('Banks, Iain M.', 'Saunders, Charles R.') | ('Galland, Nicole', 'Saunders, Charles R.') |
| ('Johansen, Erika', 'Saunders, Charles R.') | ('Layman, John', 'Saunders, Charles R.') |

*Table 4. List of skyline association rules for the Fantasy Bingo dataset*

| Association rule | Support | Confidence |
|---|---|---|
| Layman, John ---> Banks, Iain M. | 0.16872 | 0.54667 |
| Banks, Iain M. ---> Layman, John | 0.16872 | 0.42268 |
| Presley, M. D. ---> Banks, Iain M. | 0.16461 | 0.58824 |
| Banks, Iain M. ---> Presley, M. D. | 0.16461 | 0.41237 |
| Saunders, Charles R. -- Banks, Iain M. | 0.19342 | 0.46078 |
| Banks, Iain M. ---> Saunders, Charles R. | 0.19342 | 0.48454 |
| Galland, Nicole ---> Saunders, Charles R. | 0.15226 | 0.56923 |
| Saunders, Charles R. ---> Johansen, Erika | 0.16872 | 0.40196 |
| Johansen, Erika ---> Saunders, Charles R. | 0.16872 | 0.46591 |
| Layman, John ---> Saunders, Charles R. | 0.16049 | 0.52000 |