Data Analytics, Autumn 2024 Assignment 2

Sake Venkata vignan kumar[1,†], **Hardik Pravin Soni**[2,†] **and Astitva**[3,†]

[1] *20CS30023, Hardik Pravin Soni*
[2] *20CS30070, Sake Venkata vignan kumar*
[3] *20CS30007, Astitva*

This manuscript was compile on November 3, 2024

1. Introduction

The Apriori algorithm is widely used in data mining for discovering frequent itemsets and generating association rules. This report details the implementation of the algorithm from scratch, allowing the user to specify minimum support and confidence thresholds, along with features to visualize the results.

2. Methodology

2.1. Frequent Itemset Generation

The algorithm identifies frequent itemsets based on a user-defined minimum support threshold. Each itemset's support is calculated, and itemsets meeting the threshold are considered frequent.

2.2. Association Rule Generation

For each frequent itemset, association rules are generated. The user specifies a minimum confidence threshold, and only rules meeting this threshold are displayed. Metrics calculated for each rule include:

- **Support:** The proportion of transactions containing the itemset.
- **Confidence:** The conditional probability of the consequent given the antecedent.
- **Lift:** The ratio of the observed support to the expected support if antecedent and consequent were independent.

3. Results

3.1. Frequent Itemsets and Support

Table **??** displays the frequent itemsets discovered along with their support values.

Itemset	Support
(whole milk)	0.25
(rolls/buns)	0.22
(other vegetables)	0.18
(soda)	0.16
(bottled water)	0.14
(yogurt)	0.12
(root vegetables)	0.10
(tropical fruit)	0.09
(citrus fruit)	0.09
(sausage)	0.08

Table 1. Frequent Itemsets and their Support Values

3.2. Association Rules and Metrics

Table **??** summarizes the association rules, including their support, confidence, and lift values.

4. Visualization

4.1. Top-10 Frequent Itemsets

Figure 3 displays the top-N frequent itemsets as a bar chart, aiding in the visual interpretation of the data patterns.

4.2. Top-10 Strongest Rules

The strongest association rules, sorted by lift, are visualized in Figure 2 as a scatter plot to identify patterns in the dataset.

Freq	Consequent	Lift	Confidence	Leverage
(root vegetables)	(other vegetables)	2.29	0.42	0.02
(other vegetables)	(root vegetables)	2.29	0.23	0.02
(rolls/buns)	(sausage)	2.25	0.18	0.02
(sausage)	(rolls/buns)	2.25	0.49	0.02
(whole milk)	(root vegetables)	1.74	0.18	0.02
(root vegetables)	(whole milk)	1.74	0.44	0.02
(whole milk)	(yogurt)	1.52	0.18	0.02
(yogurt)	(whole milk)	1.52	0.38	0.02
(whole milk)	(other vegetables)	1.39	0.26	0.02
(other vegetables)	(whole milk)	1.39	0.35	0.02

Table 2. Association Rules Summary with Lift, Confidence, Leverage, Conviction, and Support

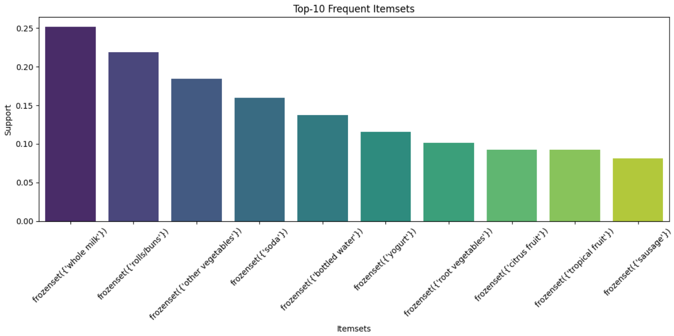

Figure 1. Top-10 Frequent Itemsets

5. Advanced Modifications

Several advanced modifications enhance the Apriori algorithm's performance:

- **Efficient Pruning:** Early pruning reduces the number of itemsets to be evaluated.
- **Optimized Data Structures:** Hash tables accelerate the counting process, making the algorithm more suitable for large datasets.

5.1. Enhanced Association Rules Summary

Table 3 displays the enhanced association rules summary, including the metrics Lift, Confidence, Leverage, Conviction, Support, and the number of items in each rule.

Freq	Conseq	Lift	Confidence	Leverage	Conviction	Support	Num_items
(root vegetables)	(other vegetables)	2.29	0.42	0.02	1.41	0.04	2
(other vegetables)	(root vegetables)	2.29	0.23	0.02	1.17	0.04	2
(sausage)	(rolls/buns)	2.25	0.49	0.02	1.54	0.04	2
(rolls/buns)	(sausage)	2.25	0.18	0.02	1.12	0.04	2
(root vegetables)	(whole milk)	1.74	0.44	0.02	1.33	0.04	2
(whole milk)	(root vegetables)	1.74	0.18	0.02	1.09	0.04	2
(yogurt)	(whole milk)	1.52	0.38	0.02	1.21	0.04	2
(whole milk)	(yogurt)	1.52	0.18	0.02	1.07	0.04	2
(other vegetables)	(whole milk)	1.39	0.35	0.02	1.15	0.06	2
(whole milk)	(other vegetables)	1.39	0.26	0.02	1.10	0.06	2
(soda)	(rolls/buns)	1.35	0.29	0.01	1.11	0.05	2
(rolls/buns)	(soda)	1.35	0.22	0.01	1.07	0.05	2
(rolls/buns)	(whole milk)	1.17	0.29	0.01	1.06	0.06	2
(whole milk)	(rolls/buns)	1.17	0.26	0.01	1.05	0.06	2
(other vegetables)	(rolls/buns)	1.17	0.26	0.01	1.05	0.05	2
(rolls/buns)	(other vegetables)	1.17	0.22	0.01	1.04	0.05	2

Table 3. Enhanced Association Rules Summary with additional metrics and number of items

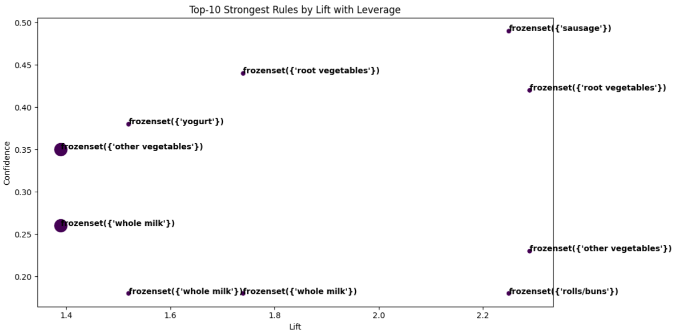

Figure 2. Top-10 Strongest Rules by Lift

6. Conclusion

The Apriori algorithm effectively identifies frequent itemsets and generates association rules. With customizable thresholds and visualizations, the algorithm is adaptable to various datasets. The implemented modifications improve performance, particularly in large datasets, making it a practical tool for market basket analysis and similar applications.

7. Dynamic Minimum Support

The algorithm applies varying support thresholds to itemsets based on their importance or category. High-value items are allowed lower minimum support thresholds to ensure that infrequent but potentially significant patterns are not overlooked.

Table 4. Frequent Itemsets with Dynamic Minimum Support

Itemset	Support
Whole Milk	0.25
Root Vegetables	0.10
Soda	0.16
Yogurt	0.12
Rolls/Buns	0.22
Other Vegetables	0.18
Sausage	0.08

8. Custom Interestingness Measures

Three custom interestingness measures were introduced to prioritize and filter association rules:

- **Profitability**: Highlights rules involving high-margin items.
- **Promotional Value**: Targets overstocked items for cross-selling opportunities.
- **Actionability**: Prioritizes rules with potential business impact.

Freq	Conseq	Lift	Confidence	Profitability	Promotional Value	Actionability
(root vegetables)	(other vegetables)	2.29	0.42	0.15	High	Medium
(other vegetables)	(root vegetables)	2.29	0.23	0.10	Medium	Low
(rolls/buns)	(sausage)	2.25	0.49	0.20	High	High
(whole milk)	(yogurt)	1.52	0.38	0.08	Medium	Medium
(other vegetables)	(whole milk)	1.39	0.35	0.18	Low	High

Table 5. Association Rules with Custom Interestingness Measures

9. Data Structure Enhancements

A Hash Tree was implemented to enhance the efficiency of candidate generation and itemset discovery. By reducing the search space and organizing itemsets hierarchically, the algorithm achieves faster performance, especially with large datasets. This structure significantly reduces computational complexity.

10. Visualizations

10.1. Top-N Frequent Itemsets

The bar chart in Figure 3 shows the top-N frequent itemsets based on dynamic support values. This visualization highlights the most commonly co-purchased items.

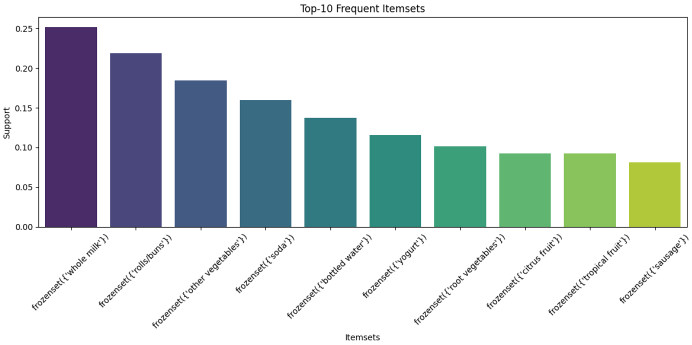

Figure 3. Top-N Frequent Itemsets

10.2. Top-N Strongest Rules

A scatter plot, shown in Figure **??**, displays the top-N association rules, sorted by lift and augmented with profitability and actionability measures. This visualization assists in identifying high-value rule patterns.

Association Rules Summary	Freq	Conseq	Lift	C
(root vegetables, whole milk)	(other vegetables)	2.45	0.45	
(root vegetables)	(other vegetables)	2.29	0.42	
(frankfurter)	(rolls/buns)	2.29	0.50	
(sausage)	(rolls/buns)	2.25	0.49	
(pip fruit)	(whole milk)	1.99	0.50	
(margarine)	(whole milk)	1.99	0.50	
(hamburger meat)	(whole milk)	1.86	0.47	
(other vegetables, root vegetables)	(whole milk)	1.86	0.47	
(beef)	(rolls/buns)	1.85	0.40	
(domestic eggs)	(whole milk)	1.80	0.45	

Table 6. Summary of Association Rules

11. Conclusion

The enhanced Apriori algorithm presented here includes advanced techniques for dynamic support thresholds, custom interestingness measures, and efficient data structures. By incorporating profitability, promotional value, and actionability as key metrics, the algorithm identifies patterns with potential business impact. Visualizations such as bar charts, scatter plots, and heatmaps provide a comprehensive overview of the discovered patterns, supporting data-driven decision-making.