Name: Pradyun Reddy Bollepally
Njit-id:pb573
Email:pb573@njit.edu
Professor : Yasser-abduallah

Midterm Project Report-1

Using Apriori Algorithms in Retail Data Mining

Abstract : In this study, I investigate correlations within retail transactions using the Apriori Algorithm, a key data mining approach. I evaluate the algorithm's efficacy and efficiency by putting it into practice and using a variety of data mining ideas, techniques, and concepts. I build and implement a custom model for extracting insightful information from transaction data by using unique data mining techniques.

Introduction: Finding hidden patterns and relationships in huge datasets can be effectively accomplished through data mining.
Our project focuses on the basic association rule mining technique known as the Apriori Algorithm and how it might be used in a retail setting. We'll go over the fundamental ideas and guidelines of data mining that we use in our work.

In this application, we were able to identify common itemsets and association rules by applying the Apriori algorithm to a bespoke dataset linked to a retail establishment. Initializing dictionaries for candidate and frequent itemsets was one of the process's key tasks.

- Opening CSV files to load the itemsets and dataset.
- Preprocessing the dataset to guarantee item uniqueness and order.
- Gathering feedback from users to determine confidence and minimum support criteria.
- Creating candidate itemsets iteratively and updating frequently occurring itemsets with the Apriori algorithm, which uses a brute force method by taking into account every potential item combination.

Principles:
Features: The Apriori Algorithm is centered on the discovery of frequent itemsets, or collections of items that commonly appear together in transactions. These item sets offer perceptions into the preferences and purchasing habits of customers.
Support and Confidence: Support and confidence are two important data mining measures. Whereas confidence evaluates the possibility that things will be bought together, support gauges the frequency with which an item or item set is purchased. Our study is guided by these metrics.

Association Rules: I can tell which products are frequently bought together by identifying strong association rules. These guidelines are essential for maximizing sales tactics like suggestions.
Project Workflow: The application of the Apriori and other stages are part of our project's structured workflow.
Preprocessing and Data Loading: To start, we load transaction data from a dataset of retail stores. Every transaction is made up of a customer's list of purchases. We preprocess the dataset, removing duplicate items and arranging them in a predetermined order, to guarantee data accuracy.
Calculating the Minimum Level of Confidence and Support: Data mining heavily relies on user input. To weed out less important patterns, we gather the user's preferences for minimum support and confidence levels.
Iteration Through Candidate Itemsets: The Apriori Algorithm is used iteratively, producing candidate itemsets of ever larger sizes. First, we have a single item set (size K = 1), and then we move on to K = 2, K = 3, and so forth. Using a "brute force" approach, all feasible item-set combinations are generated iteratively.

Support Count Calculation: We count the number of transactions that contain each candidate itemset to determine its support. While some itemsets are destroyed, those that fulfill the minimal support criterion are kept.

Calculation of Confidence: We assess the degree of association between things by calculating the confidence of association rules.
Carefully comparing the support values for individual products and itemsets is necessary in this stage.

Generation of Association Rules: Association rules that meet the minimal standards for confidence and support are extracted. These guidelines provide insightful information about what products are frequently bought together.

Result : Support, confidence, and the ensuing association rules are examples of performance metrics used to assess the efficacy and efficiency of the project. In order to evaluate the dependability of our unique Apriori Algorithm implementation, we also compare it with the Apriori library.

Conclusion: In conclusion, our study demonstrates the application of concepts, policies, and methods related to data mining.We successfully employed the Apriori Algorithm to derive meaningful association rules using retail transaction data. The ability to modify algorithms, apply an iterative "brute force" approach, and adhere to user-specified criteria demonstrate how successful data mining is in finding important patterns that aid in retail industry decision-making.

Screenshots:

Itemsets:

```
[5] # Define product lists for different stores: Walmart, Target, and Costco
    walmart_items = [
        "Headphones", "Smart TV", "Tablet", "Fitness Tracker",
        "Bluetooth Speaker", "E-Reader", "USB Cable", "Smart Light Bulb",
        "External SSD", "Video Game Console"
    ]

    target_items = [
        "Bedding", "Kitchen Appliances", "Cookware",
        "Bath Towels", "Wall Art", "Laundry Hamper",
        "Storage Bins", "Garden Supplies", "Air Fryer",
        "Coffee Machine"
    ]

    costco_items = [
        "Television", "Gaming Console", "Laptop", "Wi-Fi Router",
        "Noise-Canceling Headphones", "Smartwatch", "Portable Speaker",
        "SSD Drive", "Home Theater", "Video Doorbell"
    ]
```

Transactions:

```
# Transactions for each store
store_transactions = {
    "Walmart": [
        ["Headphones", "Fitness Tracker", "Bluetooth Speaker"],
        ["Smart TV", "Tablet"],
        ["Smart Light Bulb", "External SSD"],
        ["Video Game Console", "Headphones", "Fitness Tracker"],
        ["Bluetooth Speaker", "Smart TV"],
        ["Tablet", "USB Cable"],
        ["Headphones", "E-Reader"],
        ["Fitness Tracker", "Video Game Console"],
        ["Smart Light Bulb", "Bluetooth Speaker"],
        ["External SSD", "Tablet"]
    ],
    "Target": [
        ["Bedding", "Kitchen Appliances", "Bath Towels"],
        ["Cookware", "Air Fryer"],
        ["Wall Art", "Laundry Hamper"],
        ["Storage Bins", "Garden Supplies", "Coffee Machine"],
        ["Bedding", "Laundry Hamper"],
        ["Kitchen Appliances", "Bath Towels"],
        ["Air Fryer", "Storage Bins"],
        ["Coffee Machine", "Wall Art"],
        ["Garden Supplies", "Bedding"],
        ["Bath Towels", "Kitchen Appliances"]
    ],
    "Costco": [
```

```
        ],
        "Costco": [
            ["Television", "Gaming Console", "Wi-Fi Router"],
            ["Laptop", "Noise-Canceling Headphones"],
            ["Smartwatch", "Portable Speaker"],
            ["SSD Drive", "Home Theater", "Video Doorbell"],
            ["Television", "Laptop"],
            ["Wi-Fi Router", "Smartwatch"],
            ["Noise-Canceling Headphones", "Gaming Console"],
            ["Portable Speaker", "Video Doorbell"],
            ["SSD Drive", "Home Theater"],
            ["Smartwatch", "Gaming Console"]
        ]
    }
```

Support and confidence:

```
# Get user input for minimum support and confidence
min_support = float(input("Enter minimum support  "))
min_confidence = float(input("Enter minimum confidence  "))
```

```
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the fu
  and should_run_async(code)
Enter minimum support (e.g., 0.2): 0.2
Enter minimum confidence (e.g., 0.7): 0.7
```

Outputs for brute force, fp-tree, apriori:

```
--- Walmart ---
Exhaustive Search Time: 0.0003 seconds
Frequent Itemsets (Exhaustive Search): [({'Headphones'}, 0.3), ({'Smart TV'}, 0.2), ({'Tablet'}, 0.3), ({'Fitness Tracker'}, 0.3), ({'Bluetooth Speaker'}, 0.3), ({'Smart L

Association Rules (Exhaustive Search):
Rule 1: ['Video Game Console'] -> ['F', 'i', 't', 'n', 'e', 's', 's', ' ', 'T', 'r', 'a', 'c', 'k', 'e', 'r']
Confidence: 100.00%
Support: 20.00%


Association Rules (Apriori):
Rule 1: ['Video Game Console'] -> ['Fitness Tracker']
Confidence: 100.00%
Support: 20.00%

Apriori Time: 0.0114 seconds
Frequent Itemsets (Apriori):
   support                              itemsets
0      0.3                          (Headphones)
1      0.2                            (Smart TV)
2      0.3                              (Tablet)
3      0.3                     (Fitness Tracker)
4      0.3                   (Bluetooth Speaker)
5      0.2                    (Smart Light Bulb)
6      0.2                        (External SSD)
7      0.2                  (Video Game Console)
8      0.2         (Fitness Tracker, Headphones)
9      0.2  (Fitness Tracker, Video Game Console)
FP-Growth Time: 0.0102 seconds
Frequent Itemsets (FP-Growth):
   support                              itemsets
0      0.3                   (Bluetooth Speaker)
1      0.3                     (Fitness Tracker)
2      0.3                          (Headphones)
3      0.3                              (Tablet)
4      0.2                            (Smart TV)
5      0.2                        (External SSD)
6      0.2                    (Smart Light Bulb)
7      0.2                  (Video Game Console)
8      0.2         (Fitness Tracker, Headphones)
9      0.2  (Fitness Tracker, Video Game Console)

Association Rules (FP-Growth):
Rule 1: ['Video Game Console'] -> ['Fitness Tracker']
Confidence: 100.00%
Support: 20.00%


--- Target ---
Exhaustive Search Time: 0.0005 seconds
Frequent Itemsets (Exhaustive Search): [({'Bedding'}, 0.3), ({'Kitchen Appliances'}, 0.3), ({'Bath Towels'}, 0.3), ({'Wall Art'}, 0.2), ({'Laundry Hamper'}, 0.2), ({'Stora
```

[redacted]

```
Confidence: 100.00%
Support: 20.00%


--- Target ---
Exhaustive Search Time: 0.0005 seconds
Frequent Itemsets (Exhaustive Search): [({'Bedding'}, 0.3), ({'Kitchen Appliances'}, 0.3), ({'Bath Towels'}, 0.3), ({'Wall Art'}, 0.2), ({'Laundry Hamper'}, 0.2), ({'Storage

Association Rules (Exhaustive Search):
Rule 1: ['Kitchen Appliances'] -> ['B', 'a', 't', 'h', ' ', 'T', 'o', 'w', 'e', 'l', 's']
Confidence: 100.00%
Support: 30.00%

Rule 2: ['Bath Towels'] -> ['K', 'i', 't', 'c', 'h', 'e', 'n', ' ', 'A', 'p', 'p', 'l', 'i', 'a', 'n', 'c', 'e', 's']
Confidence: 100.00%
Support: 30.00%


Association Rules (Apriori):
Rule 1: ['Bath Towels'] -> ['Kitchen Appliances']
Confidence: 100.00%
Support: 30.00%

Rule 2: ['Kitchen Appliances'] -> ['Bath Towels']
Confidence: 100.00%
Support: 30.00%

Apriori Time: 0.0071 seconds
Frequent Itemsets (Apriori):
   support                           itemsets
0      0.3                          (Bedding)
1      0.3               (Kitchen Appliances)
2      0.3                      (Bath Towels)
3      0.2                         (Wall Art)
4      0.2                   (Laundry Hamper)
5      0.2                     (Storage Bins)
6      0.2                  (Garden Supplies)
7      0.2                        (Air Fryer)
8      0.2                   (Coffee Machine)
9      0.3  (Bath Towels, Kitchen Appliances)
FP-Growth Time: 0.0064 seconds
Frequent Itemsets (FP-Growth):
   support                           itemsets
0      0.3                      (Bath Towels)
1      0.3               (Kitchen Appliances)
2      0.3                          (Bedding)
3      0.2                        (Air Fryer)
4      0.2                   (Laundry Hamper)
5      0.2                         (Wall Art)
6      0.2                   (Coffee Machine)
7      0.2                  (Garden Supplies)
8      0.2                     (Storage Bins)
```

[redacted]

```
Confidence: 100.00%
Support: 30.00%


--- Costco ---
Exhaustive Search Time: 0.0003 seconds
Frequent Itemsets (Exhaustive Search): [({'Television'}, 0.2), ({'Gaming Console'}, 0.3), ({'Laptop'}, 0.2), ({'Wi-Fi Router'}, 0.2), ({'Noise-Canceling Headphones'}, 0.2), (

Association Rules (Exhaustive Search):
Rule 1: ['Home Theater'] -> ['S', 'S', 'D', ' ', 'D', 'r', 'i', 'v', 'e']
Confidence: 100.00%
Support: 20.00%

Rule 2: ['SSD Drive'] -> ['H', 'o', 'm', 'e', ' ', 'T', 'h', 'e', 'a', 't', 'e', 'r']
Confidence: 100.00%
Support: 20.00%


Association Rules (Apriori):
Rule 1: ['SSD Drive'] -> ['Home Theater']
Confidence: 100.00%
Support: 20.00%

Rule 2: ['Home Theater'] -> ['SSD Drive']
Confidence: 100.00%
Support: 20.00%

Apriori Time: 0.0231 seconds
Frequent Itemsets (Apriori):
    support                      itemsets
0       0.2                  (Television)
1       0.3              (Gaming Console)
2       0.2                      (Laptop)
3       0.2                (Wi-Fi Router)
4       0.2  (Noise-Canceling Headphones)
5       0.3                  (Smartwatch)
6       0.2            (Portable Speaker)
7       0.2                   (SSD Drive)
8       0.2                (Home Theater)
9       0.2               (Video Doorbell)
10      0.2     (SSD Drive, Home Theater)
FP-Growth Time: 0.0066 seconds
Frequent Itemsets (FP-Growth):
    support                      itemsets
0       0.3              (Gaming Console)
1       0.2                (Wi-Fi Router)
2       0.2                  (Television)
3       0.2  (Noise-Canceling Headphones)
4       0.2                      (Laptop)
5       0.3                  (Smartwatch)
6       0.2            (Portable Speaker)
7       0.2              (Video Doorbell)
8       0.2                (Home Theater)
```

Github: https://github.com/Pradyun-reddy/Data_mining_midterm