CSE-601
DATA MINING AND BIOINFORMATICS

Dimensionality Reduction & Association Analysis - Apriori Algorithm

By:
Naina Nigam: 50208030
Surabhi Singh: 50208675
Vanshika Nigam: 50208031

Part 2: Association Analysis

Apriori Algorithm:

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.

Key Concepts :

- **Frequent Itemsets:** The sets of item which has minimum support (denoted by Li for ith-Itemset).
- **Apriori Property:** Any subset of frequent itemset must be frequent.
- **Join Operation:** To find L k , a set of candidate k-itemsets is generated by joining Lk-1 with itself.

Association Rule :

- An implication expression of the form X → Y, where X and Y are itemsets. The strength of an association rule can be measured in terms of support and confidence.
- **Support:** Fraction of transaction that contains both X and Y.
- **Confidence:** Measures how often items in Y appear in transactions that contain X
- Given a set of transactions T, the goal of association rule mining is to find all rules having
    - support ≥ *minsup* threshold
    - confidence ≥ *minconf* threshold

Mining Association Rule :

- **Two-step approach:**
    1. **Frequent Itemset Generation**
        a. Generate all itemsets whose support ≥ minsup
    2. **Rule Generation**
        a. Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

RESULTS :

Results after implementing the Apriori algorithm to find all frequent itemsets:

1. **For support value at 30% -**
   - 1.1. Number of Length - 1 Frequent Itemset: 196
   - 1.2. Number of Length - 2 Frequent Itemset: 5340
   - 1.3. Number of Length - 3 Frequent Itemset: 5287
   - 1.4. Number of Length - 4 Frequent Itemset: 1518
   - 1.5. Number of Length - 5 Frequent Itemset: 438
   - 1.6. Number of Length - 6 Frequent Itemset: 88
   - 1.7. Number of Length - 7 Frequent Itemset: 11
   - 1.8. Number of Length - 8 Frequent Itemset: 1
2. **For support value at 40% -**
   - 2.1. Number of Length - 1 Frequent Itemset: 167
   - 2.2. Number of Length - 2 Frequent Itemset: 753
   - 2.3. Number of Length - 3 Frequent Itemset: 149
   - 2.4. Number of Length - 4 Frequent Itemset: 7
   - 2.5. Number of Length - 5 Frequent Itemset: 1
3. **For support value at 50% -**
   - 3.1. Number of Length - 1 Frequent Itemset: 109
   - 3.2. Number of Length - 2 Frequent Itemset: 63
   - 3.3. Number of Length - 3 Frequent Itemset: 2
4. **For support value at 60% -**
   - 4.1. Number of Length - 1 Frequent Itemset: 34
   - 4.2. Number of Length - 2 Frequent Itemset: 2
5. **For support value at 70% -**
   - 5.1. Number of Length - 1 Frequent Itemset: 7

Generation of association rules based on the templates. (SUPPORT = 50%, CONFIDENCE = 70%):

Template 1 Results :

1. "RULE", "ANY", ['G59_Up'] : Count = 26
2. "RULE", "NONE", ['G59_Up'] : Count = 91
3. "RULE", 1, ['G59_Up', 'G10_Down'] : Count=39
4. "BODY", "ANY", ['G59_Up'] : Count=9
5. "BODY", "NONE", ['G59_Up'] : Count =108
6. "BODY", 1, ['G59_Up', 'G10_Down'] : Count=17
7. "HEAD", "ANY", ['G59_Up'] : Count=17
8. "HEAD", "NONE", ['G59_Up'] : Count=100
9. "HEAD", 1, ['G59_Up', 'G10_Down'] : Count=24

Template 2 Results :

1. "RULE", 3 : Count=9
2. "BODY", 2 : Count=6
3. "HEAD", 1 : Count =117

Template 3 Results :

1. "1or1", "BODY", "ANY", ['G10_Down'], "HEAD", 1, ['G59_Up'] : Count=24
2. "1and1", "BODY", "ANY", ['G10_Down'], "HEAD", 1, ['G59_Up'] :Count=1
3. "1or2", "BODY", "ANY", ['G10_Down'], "HEAD", 2 : Count=11
4. "1and2", "BODY", "ANY", ['G10_Down'], "HEAD", 2 : Count=0
5. "2or2", "BODY", 1, "HEAD", 2 : Count=117
6. "2and2", "BODY", 1, "HEAD", 2: Count=3