**Association Rule Mining**

**Mauro Travieso – ID: 986965**

**Statement 1.** When considering the following transactional database.

|  |  |
| --- | --- |
| **TID** | **Items** |
| 100 | 1, 2, 3, 5, 6, 8 |
| 200 | 1, 2, 4, 5, 6 |
| 300 | 1, 2, 3, 4, 5, 6, 8 |
| 400 | 1, 2, 3, 5, 7 |
| 500 | 1, 2, 3, 8 |

**(1)** When mining all frequent itemsets using Apriori algorithm, it is shown all candidate itemsets and frequent itemsets, following the process described including the pruning steps ( C1 → L1 → C2 → L2→ ...). **Minimum support = 60% (or 3 or more transactions)**.

C1:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Count | 5 | 5 | 4 | 2<3 | 4 | 3 | 1<3 | 3 |

L1:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Itemset | 1 | 2 | 3 | 5 | 6 | 8 |
| Count | 5 | 5 | 4 | 4 | 3 | 3 |

C2:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | {1,2} | {1,3} | {1,5} | {1,6} | {1,8} | {2,3} | {2,5} | {2,6} | {2,8} | {3,5} | {3,6} | {3,8} | {5,6} | {5,8} | {6,8} |
| Count | 5 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 2 |

L2:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | {1,2} | {1,3} | {1,5} | {1,6} | {1,8} | {2,3} | {2,5} | {2,6} | {2,8} | {3,5} | {3,8} | {5,6} |
| Count | 5 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 3 | 3 | 3 |

C3:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | {1,2,3} | {1,2,5} | {1,2,6} | {1,2,8} | {1,3,5} | {1,3,6} | {1,3,8} | {1,5,6} | {1,5,8} | {1,6,8} | {2,3,5} | {2,3,6} | {2,3,8} | {2,5,6} | {2,5,8} | {2,6,8} | {3,5,8} |
| Count | 4 | 4 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 2 | 2 |

L3:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | {1,2,3} | {1,2,5} | {1,2,6} | {1,2,8} | {1,3,5} | {1,3,8} | {1,5,6} | {2,3,5} | {2,3,8} | {2,5,6} |
| Count | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

C4:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | {1,2,3,5} | {1,2,3,6} | {1,2,3,6} | {1,2,3,8} | {1,2,5,6} | {1,2,5,8} | {1,2,6,8} | {1,3,5,8} | {2,3,5,8} |
| Count | 3 |  | 2 | 3 | 3 | 2 | 2 | 2 | 2 |

L4:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Itemset | {1,2,3,5} | {1,2,3,6} | {1,2,3,8} | {1,2,5,6} |
| Count | 3 |  | 3 | 3 |

C5:

|  |  |
| --- | --- |
| Itemset | {1,2,3,5,8} |
| Count | 2 |

L5:

|  |  |
| --- | --- |
| Itemset | Ø |
| Count |  |

**(2)** Sorting all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset form this sorted list of frequent 4-itemsets and mine all strong rules from this itemset that have the format {W, X} => {Y, Z}, where W, X, Y, and Z are individual items. **Assume that** **minimum confidence = 80%**.

From L4 (frequent 4-itemset), the first frequent 4-itemset is: **{1,2,3,5}**

**Format:** {W,X} => {Y,Z}

Confidence = support(X→Y)/support(X), then using the format:

support(X→Y) = minimum support = 60% (or **3** or more transactions)

|  |  |  |
| --- | --- | --- |
| **Rule** | **X -> Y, where Y = I – X** | **Confidence = support({W,X}→{Y,Z})/support({W,X})** |
| R1: | {1,2} -> {3,5} | Confidence = **3**/*5 ← from L2: {1,2}* => 60% |
| R2: | {1,3} -> {2,5} | Confidence = **3**/*4 ← from L2: {1,3} => 75%* |
| R3: | {1,5} -> {2,3} | Confidence = **3**/*4 ← from L2: {1,5} => 75%* |
| R4: | {2,3} -> {1,5} | Confidence = **3**/*4 ← from L2: {2,3} => 75%* |
| R5: | {2,5} -> {1,3} | Confidence = **3**/*4 ← from L2: {2,5} => 75%* |
| R6: | {3,5} -> {1,2} | Confidence = **3**/*3 ← from L2: {3,5} => 100%* |

Because the minimum confidence to select the rule is 80%, **R6** is a strong rule.

**Statement 2.** Consider the following contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | C (buys coffee = Yes) | (buys coffee = No) | Total |
| T (buys tea = Yes) | 362 | 823 | 1185 |
| (buys tea = No) | 49 | 1527 | 1576 |
| Total | 411 | 2350 | 2761 |

To compute the *lift*{T->C} and to determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated.

Total transactions = 2761

support( {T} ) = 1185/2761 = 0.43 => 43%

support( {T,C} ) = 362/2761 = 0.13 => 13%

confidence( T -> C ) = support( {T -> C} )/support( {T} ) = 0.13/0.43 = 0.30 => 30%

**Coffee purchase rate:**

- with tea: confidence( T -> C ) = support( {T -> C} )/support( {T} ) = 0.3 => 30%

- without tea: support( {C} ) = 411/2761 = 0.15 => 15%

correlation: ***lift{T→C}* =** confidence( T -> C )/ support( {C} ) =

**=**30%/15% = **2 > 1** => **positive correlation**.

**Conclusion:** Coffee is more bought when buying Tea. The rule indicates positive correlation.

Buy Coffee with Tea = 30% > buy Coffee alone = 15%