Name: Tuheena Singh

**Lab #4: Association Rule Mining**

Submit **a scan copy of answered this Lab #4** to the Lab #4 Drop Box by Sunday, 10/25 @11:59PM.



**Introduction:**

In this lab we will explore association rule mining algorithms in Weka and evaluate them. The tasks include understanding the dataset and implementation of the Apriori algorithm in Weka. It is assumed that you have a basic knowledge of the Apriori algorithm.

To describe briefly, this algorithm is a divide and conquer approach. The implementation works on the concept of maximum support level, which is also referred as confidence level. The algorithm works iteratively and identifies a rule based on the confidence level. For rule generation in subsequent iterations, the dataset that is not covered in the previous rules will be used. For detailed description and pseudo code, refer to study material.

**Dataset used: *supermarket*** dataset (Check data folder under Weka installation)

It is a real world data set collected over a short duration of time at a supermarket. The attributes are aggregated to the department level. The value “t” indicates that customers shopping cart contained at least one product from that department.



**Part 1: Using Apriori**

Open Weka and navigate to Explorer window.

**Q1.1** In the Preprocess tab and select the *supermarket* dataset. Navigate to Associate tab and select and choose ‘Apriori’ in associations folder. Run Apriori using the default settings of the options. The last 10 lines of the output below ‘Best rules found’ are the 10 best rules generated. The confidence of rule 10 is 0.91. How was this confidence value computed? Write down the proportion as division.

877/969 =0.905 ~ 0.91

Q1.2 How many instances is the support of rule 7?

**701** “baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701” The number following the consequent rule is the value that Supports the entire rule.

Q1.3 Based on rule 10:

LHS (frozen foods=t fruit=t total=high) ==> RHS (bread and cake=t), are LHS and RHS independent? Justify your answer using a proper metric value.

Metric Used: Lift

Metric Value: 1.26

Justification: Lift is the metric that defines the importance of the rule. The value of Lift suggests if the items in the given rule are more prone to occurrence or least prone to occurance in combination with each other. If the value of Lift >1 then they are more likely to occur and if <1 then less likely to occur together. Incase of rule 10 the value if 1.26 which is greater than 1 and thus the items are more likely to occur together. LHS and RHS rule are not independent of each other. When the value of Lift is approximately 1 (or ideally 1) then the LHS and RHS rule are said to be statistically independent from each other.

Q1.4 Using the available Apriori algorithm options, design a method to find the two rules with the highest support count. Your answer should include which options you modified, the values you modified the options to, the resultant rules, and their support count.

The question asked for highest support count and did not say anything about the high metric value which is Confidence by default. Hence I lowered the confidence value and after couple of tries I was able to find the highest support value that works with the lowest confidence value producing rules with high support count. I increased the **“lower bound support”** to **0.5** and changed the **“minimum metric”** value to **0.4**. Additionally I also changed the **“num rules”** to **100** instead of **default 10** just to be sure that I am not missing any rules which are not displayed in the output.

The result had four rules with support values of **0.505** and **0.502**

1. milk-cream=t 2939 ==> bread and cake=t 2337 <conf:(0.8)> lift:(1.1) lev:(0.05) [221] conv:(1.37) **[Support : 2337/4627= 0.505 ]**

2. fruit=t 2962 ==> bread and cake=t 2325 <conf:(0.78)> lift:(1.09) lev:(0.04) [193] conv:(1.3) **[Support: 2325/4627= 0.502]**

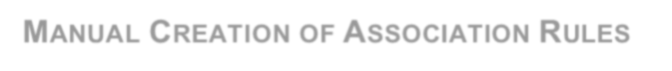
3. bread and cake=t 3330 ==> milk-cream=t 2337 <conf:(0.7)> lift:(1.1) lev:(0.05) [221] conv:(1.22)

4. bread and cake=t 3330 ==> fruit=t 2325 <conf:(0.7)> lift:(1.09) lev:(0.04) [193] conv:(1.19)

Q1.5 What does ‘best rules’ mean? What criterion is used to decide what the best rules are? (Hint: Study the descriptions of the parameters for Apriori by pressing the ‘More’ button in the window that allows you to change the options for ‘Apriori’.

The confidence value of the rule greater than the user defined threshold value are selected as the best rules. Also the rules need to fulfill the criterion for minimum support value which is 0.1 by default. Lets say the minimum metric value is 0.9(default) and the metric chosen to imply this metric value is 0.9, then all the rules whose metric value is greater than the given threshold are selected as the best rules . Also in Weka you can select how many rules you want to display in output. By default it is 10. The rules are sorted according to the confidence value.

**Part 2: MANUAL CREATION OF ASSOCIATION RULES**



*Market basket Analysis:*

*The table below shows details of 5 transactions at a supermarket.*

|  |  |
| --- | --- |
| **TID** | **Items** |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Bread, Milk, Diaper, Beer, Coke |
| 4 | Bread, Diaper, Beer |
| 5 | Bread, Diaper, Coke |

Let’s make a tabular using binary representation of the above where A = Bread; B= Beer; C=Coke, D=Diaper; E=Egg; F=Milk:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TID | A | B | C | D | E | F |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 1 | 1 | 0 |
| 3 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 1 | 1 | 0 | 0 |

Q2.1 Form the item sets. Let’s start by forming the item set containing one item. Write the number of occurrences and the support of each item set. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | #  Occurrences | Support  (%) |
| A | 5 | 100% |
| B | 3 | 60% |
| C | 2 | 40% |
| D | 4 | 80% |
| E | 1 | 20% |
| F | 2 | 40% |
|  |  |  |

Q2.2 Now let’s form the item sets containing 2 items. We only take the item sets from Q2.1 whose minimum support is 60%. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | #  Occurrences | Support  (%) |
| A B | 3 | 60% |
| A D | 4 | 80% |
| B D | 3 | 60% |
|  |  |  |
|  |  |  |
|  |  |  |

Q2.3 List item sets containing 3 items from Q2.2. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | #  Occurrences | Support  (%) |
| A B D | 3 | 60% |
|  |  |  |
|  |  |  |

Q2.4 Now form the rules and calculate their confidence (c) using minsup = 60%. (Note: All rows might not be necessary)

|  |  |
| --- | --- |
| Rules | Confidence (c) |
| A -> B P(B|A)=|B∩A| / |A| = 3/5 | 60% |
| B -> A P(A|B)=|A∩B| / |B| = 3/3 | 100% |
| A -> D P(D|A)=|D∩A|/|A| = 4/5 | 80% |
| D -> A P(A|D)=|A∩D|/|D| = 4/4 | 100% |
| B -> D P(D|B)=|D∩B|/|B| = 3/3 | 100% |
| D -> B P(B|D)=|B∩D|/|D| = 3/4 | 75% |
| AB-> D P(D|AB)=|D∩AB|/|AB| = 3/3 | 100% |
| D -> AB P(AB|D)=|AB∩D|/|D| = 3/4 | 75% |
| AD -> B P(B|AD)=|B∩AD|/|AD| = 3/4 | 75% |
| B -> AD P(AD|B)=|AD∩B|/|B| = 3/3 | 100% |
| BD -> A P(A|BD)=|A∩BD|/|BD| = 3/3 | 100% |
| A -> BD P(BD|A)=|BD∩A|/|A| = 3/5 | 60% |
|  |  |
|  |  |

Q2.5 Prune the rules in Q2.4 using confidence measure of 80%. (Note: All rows might not be necessary)

|  |  |
| --- | --- |
| Rules | Confidence(c) |
| B -> A P(A|B)=|A∩B| / |B| = 3/3 | 100% |
| A -> D P(D|A)=|D∩A|/|A| = 4/5 | 80% |
| D -> A P(A|D)=|A∩D|/|D| = 4/4 | 100% |
| B -> D P(D|B)=|D∩B|/|B| = 3/3 | 100% |
| AB-> D P(D|AB)=|D∩AB|/|AB| = 3/3 | 100% |
| B -> AD P(AD|B)=|AD∩B|/|B| = 3/3 | 100% |
| BD -> A P(A|BD)=|A∩BD|/|BD| = 3/3 | 100% |
|  |  |

LAB 4 Gradesheet Name:

|  |  |  |  |
| --- | --- | --- | --- |
| Questions | Max. Points | Points Earned | Comments |
| Part 1 |  |  |  |
| Q1.1 | 10 |  |  |
| Q1.2 | 10 |  |  |
| Q1.3 | 10 |  |  |
| Q1.4 | 10 |  |  |
| Q1.5 | 10 |  |  |
| Part 2 |  |  |  |
| Q2.1 | 10 |  |  |
| Q2.2 | 10 |  |  |
| Q2.3 | 10 |  |  |
| Q2.4 | 10 |  |  |
| Q2.5 | 10 |  |  |
| TOTAL | 100 |  |  |