Yiduo Feng

CS 699

Assignment 5

Date: 6/14/2022

**Problem 1 (25 points)**. Consider the following transactional database.

|  |  |
| --- | --- |
| TID | Items |
| 100 | 2, 3, 4, 5, 6, 8 |
| 200 | 1, 2, 3, 5, 6 |
| 300 | 1, 4, 5, 7, 8 |
| 400 | 2, 3, 4, 5, 6 |
| 500 | 1, 2, 3, 4, 5, 7 |
| 600 | 1, 3, 8 |

1. Mine all frequent itemsets using the Apriori algorithm, which we discussed in the class, with the minimum support = 50% (or 3 or more transactions). Show all candidate itemsets and frequent itemsets. You should follow the process described in the book and lecture (i.e., C1 → L1 → C2 → L2 → …). You don't need to show pruning steps. To save your time, L1 is given below:

L1:

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

C2:

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

L2:

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

C2:

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

L3:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | 2,3,4 | 2,3,5 | 2,3,6 | 2,4,5 | 2,5,6 | 3,4,5 | 3,5,6 |
| Count | 3 | 4 | 3 | 3 | 3 | 3 | 3 |

C4:

|  |  |  |
| --- | --- | --- |
| Itemset | 2,3,4,5 | 2,3,5,6 |
| Count | 3 | 3 |

L4:

|  |  |  |
| --- | --- | --- |
| Itemset | 2,3,4,5 | 2,3,5,6 |
| Count | 3 | 3 |

All Frequent Itemsets:

L = L1UL2UL3UL4

=

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | | | | | | 2 | | 3 | | | 4 | | | | 5 | | | 6 | 8 |
| 1,3 | 1,5 | 2,3 | | 2,4 | | | 2,5 | 2,6 | | 3,4 | 3,5 | | 3,6 | | 4,5 | | 5,6 | |
| 2,3,4 | | | 2,3,5 | | | 2,3,6 | | 2,4,5 | | 2,5,6 | | 3,4,5 | | 3,5,6 | |
| 2,3,4,5 | | | 2,3,5,6 | | |

Similar to the way to get frequent Itemsets, all candidate itemsets are shown in the C tables above.

1. Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset form the 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%.

First L4 is {2,3,4,5}

Conf(R1) = Conf({2,3}=>{4,5}) = 3/4 = 75%

Conf(R2) = Conf({3,4}=>{2,5}) = 3/3 = 100%

Conf(R3) = Conf({3,5}=>{2,4}) = 3/4 = 75%

Conf(R4) = Conf({4,5}=>{2,3}) =3/4 = 75%

So R2 is string rule.

**Problem 2 (25 points)**. Consider the following training dataset, which is used for classification:

|  |  |  |  |
| --- | --- | --- | --- |
| A1 | A2 | A3 | Class |
| High | On | True | Positive |
| High | On | False | Positive |
| Low | Off | True | Negative |
| High | Off | True | Negative |
| Low | On | False | Positive |
| High | Off | True | Positive |
| High | On | False | Negative |

You can generate classification rules from the above dataset using the Apriori algorithm, which we discussed in the class.

(1). Execute the Apriori algorithm on the above dataset with the minimum support = 40% or 3 transactions. You need to proceed as we discussed in the class, i.e., C1 -> L1 -> C2 > L2 -> . . . You need to show all candidate itemsets, frequent itemsets, and all rules mined from the dataset.

C1:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | High | Low | On | Off | True | False | Positive | Negative |
| Count | 5 | 2 | 4 | 3 | 4 | 3 | 4 | 3 |

L1:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | High | On | Off | True | False | Positive | Negative |
| Count | 5 | 4 | 3 | 4 | 3 | 4 | 3 |

C2:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Itemset | High,On | High,Off | | High,Ture | | High,False | | | High,Positive | | | High,Negative |
| Count | 3 | 2 | | 3 | | 2 | | | 3 | | | 2 |
| Itemset | On,True | On,False | | On,Positive | | | | On,Negative | | |
| Count | 1 | 3 | | 3 | | | | 1 | | |
| Itemset | Off,True | Off,Positive | | | Off,Negative | | | | |
| Count | 3 | 1 | | | 2 | | | | |
| Itemset | True,Positive | | True,Negative | | | |
| Count | 2 | | 2 | | | |
| Itemset | False,Positive | | False,Negative | | | |
| Count | 2 | | 1 | | | |

L2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Itemset | High,On | High,Ture | High,Positive | On,False | On,Positive | Off,True |
| Count | 3 | 3 | 3 | 3 | 3 | 3 |

C3:

|  |  |
| --- | --- |
| Itemset | High,On,Positive |
| Count | 2 |

Conf(R1) = conf(high=>on) = 3/5 = 60%

Conf(R2) = conf(high=>true) = 3/5= 60%

Conf(R3) = conf(high=>Positive) = 3/5= 60%

Conf(R4) = conf(On=>false) = 3/4= 75%

Conf(R5) = conf(On=>positive) = 3/4= 75%

Conf(R6) = conf(Off=>true) = 3/4= 75%

(2). Show only the rules that can be used for classification and calculate their confidences.

Conf(R1) = conf(high=>Positive) = 3/5 = 60%

Conf(R1) = conf(low=>Positive) = 1/2 = 50%

Conf(R1) = conf(On=>Positive) = 3/4 = 75%

Conf(R1) = conf(Off=>Positive) = 1/3= 33.3%

Conf(R1) = conf(True=>Positive) = 2/4 = 50%

Conf(R1) = conf(False=>Positive) =2/3 = 66.7%

Conf(R1) = conf(high=>Nagative) = 2/5 = 40%

Conf(R1) = conf(low=> Nagative) = 1/2 = 50%

Conf(R1) = conf(On=> Nagative) = 1/4= 25%

Conf(R1) = conf(Off=> Nagative) = 2/3 = 66.7%

Conf(R1) = conf(True=> Nagative) = 2/4 = 50%

Conf(R1) = conf(False=> Nagative) = 1/3 = 33.3%

You must run the Apriori algorithm yourself as we discussed in the class (i.e., you should not use Weka, JMP Pro or any other software to run an Apriori algorithm on the given dataset). You need to show all intermediate steps.

**Problem 3 (25 points)**. Consider the following contingency table.

|  |  |  |
| --- | --- | --- |
|  | *C* (buys coffee = Yes) | *C* (buys coffee = No) |
| *T* (buys tea = Yes) | 142 | 862 |
| *T* (buys tea = No) | 186 | 1859 |

Compute the *lift*, *all-confidence*, *cosine*, *Kulczynski* and *imbalance ratio* measures, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated. You must show all calculations.

142+186 = 328

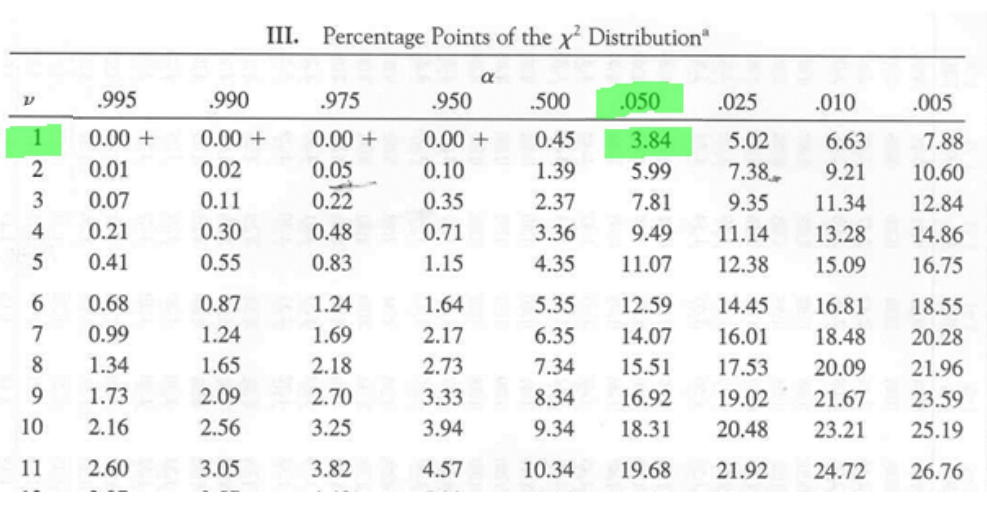
862+1859 = 2721

142+862 = 1004

186+1859 = 2045

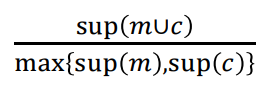
2045+1004 = 3049

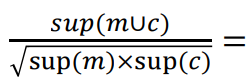
According to the formula we learn last week:

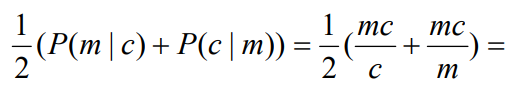


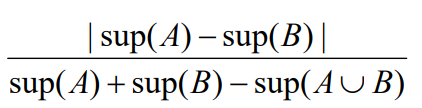
X^2 = 17.8745 < X^2\_0.05 = 3.84 so there is a correlation.

lift = P(AUB)/P(A)P(B) = (142/3049)/((328/3049)\*(1004/3049)) = 1.314735(positive)

all\_conf =  =142/1004 = 0.1414343(negative)

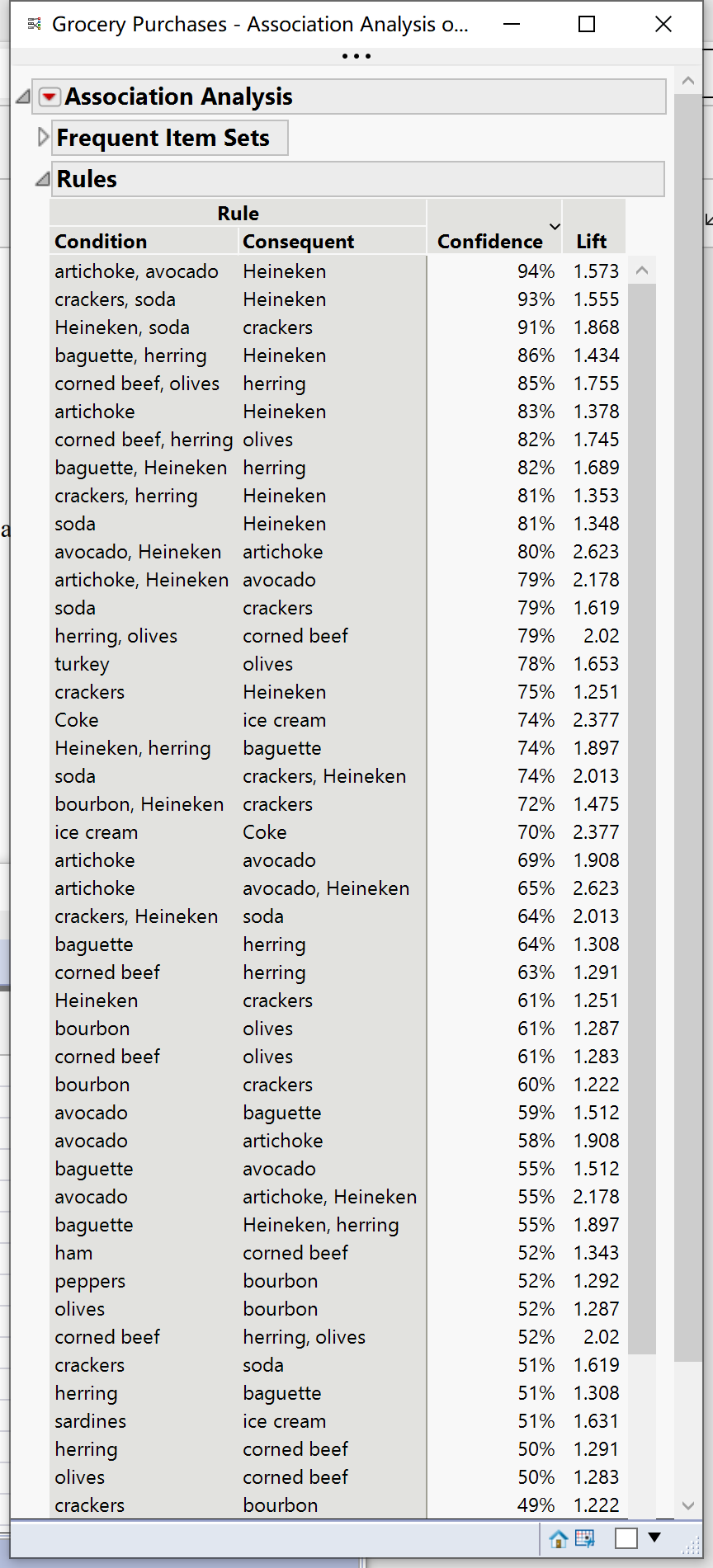
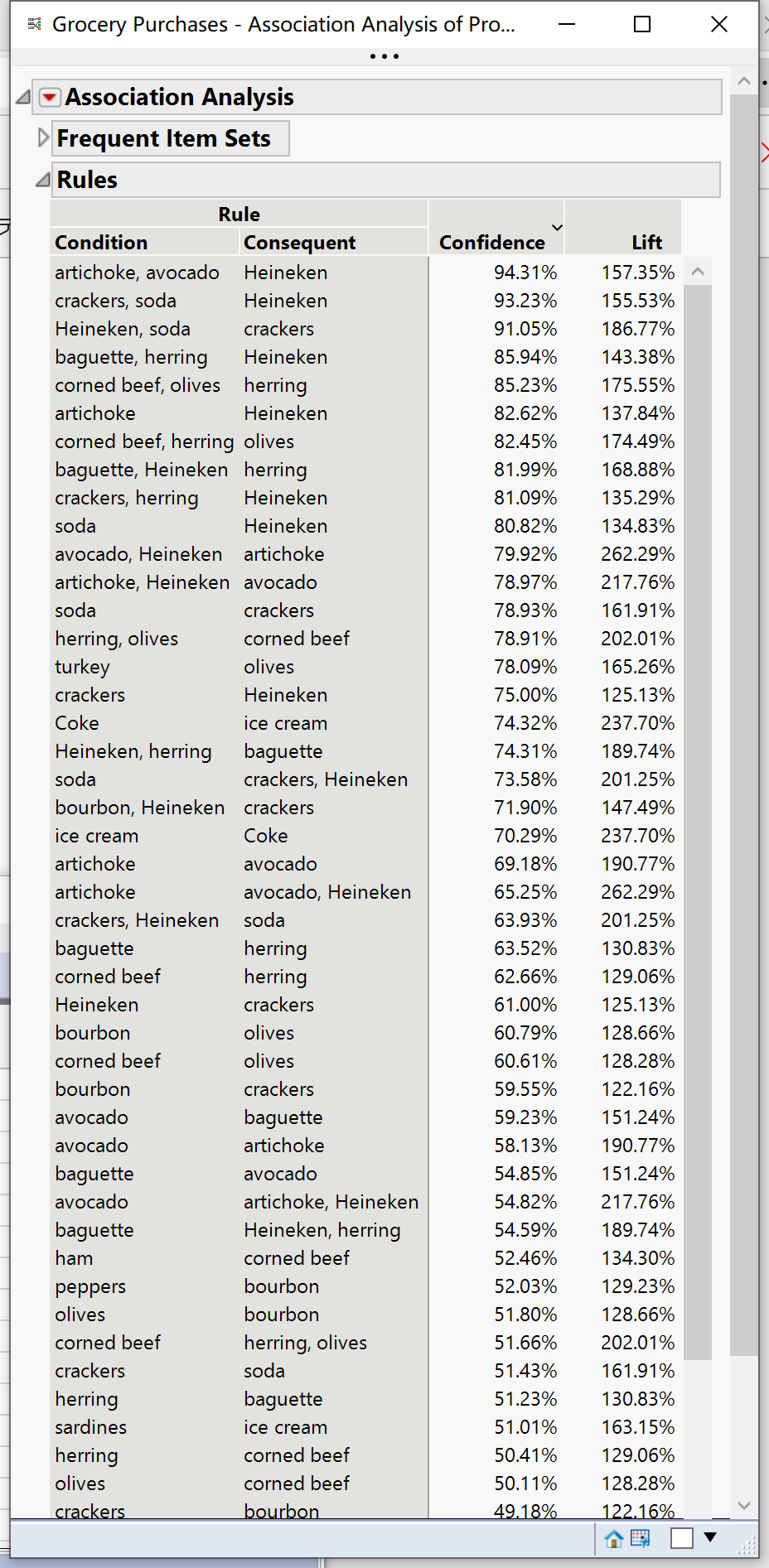
cosine =  142/sqrt(328\*1004) = 0.2474484(negative)

Kulczynski =  (1/2)\*(142/328+142/1004) = 0.2871805(negative)

IR = = |1004-328|/(1004+328-142) =0.5680672(unbalanced)

So they are negatively correlated.

**Problem 4 (25 points).** You will perform association analysis using JMP Pro. There is a section in *Predictive and Specialized Modeling.pdf* documentation that shows how to perform association analysis. You may want to read this section before starting the assignment. Follow the instructions in *JMP-association-analysis-assignment.pdf* file.





Crackers = 7007\*48.75% = 3415.912



heineken = 7007\*59.94% =4199.996



soda = 7007\*31.77% =2226.124



{Crackers,heineken} = 7007\*36.56% =2561.759



{Crackers, soda} = 7007\*25.07% =1756.655



{ heineken,soda} = 7007\*25.67% =1798.697



{Crackers,heineken,soda} = 7007\*23.38% =1638.237

Conf(R1) = conf(crackers=>{Heineken,soda}) = 1638.237/3415.912 = 0.4795899

Conf(R2) = conf(Heineken=>{crackers,soda}) = 1638.237/4199.996 = 0.3900568

Conf(R3) = conf(soda=>{Heineken,crackers}) = 1638.237/2226.124 = 0.7359145

Conf(R4) = conf({Heineken,crackers}=>soda) = 1638.237/2561.759 = 0.6394969

Conf(R5) = conf({crackers,soda}=>heineken) = 1638.237/1756.655 = 0.9325889

Conf(R6) = conf({Heineken,soda}=>crackers) = 1638.237/1798.697 =0.910791

If min conf is 70%, R3,R5,R6 are strong rules.