

# Lab 9

Association rule learning

Alex Becheru

[irlab.becheru.net](http://irlab.becheru.net)

[irlab@becheru.net](mailto:irlab@becheru.net)

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

# Example

database:

alpha beta epsilon

alpha beta theta

alpha beta epsilon

alpha beta theta

rules:

- 100% alpha & beta
- 50% alpha, beta & epsilon
- 50% alpha, beta & theta

Association rules are intended to identify strong rules discovered in databases using different **measures of interestingness**

# Support measure

$\text{supp}(X)$ , where  $X$  is an item in the data set

$$\text{supp}(\text{alpha}, \text{beta}, \text{epsilon}) = 2/4 = 0.5$$

# Confidence of a rule

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X, Y) / \text{supp}(X)$$

$\text{conf}(X \Rightarrow Y) =$  nr of items where both X and Y appear / nr of items where X appears

$$\text{conf}(\text{alpha}, \text{beta} \Rightarrow \text{epsilon}) = 2/4 = 0.5$$

# Lift of a rule

$$\text{lift}(X \Rightarrow Y) = \text{supp}(X, Y) / \text{supp}(X) * \text{supp}(Y)$$

The ratio of the observed support to that expected if  $X$  and  $Y$  were independent.



# Conviction of a rule

$$\text{conv}(X \Rightarrow Y) = (1 - \text{supp}(Y)) / (1 - \text{conf}(X \Rightarrow Y))$$

The frequency that the rule makes an incorrect prediction if  $X$  and  $Y$  were independent

# Apriori algorithm

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time.

# Apriori treshold

A frequent item is an item that appears in the database more frequent than **a threshold**.

**(given by us)**

# Example Apriori

Itemsets:

Threshold =3

1,2,3,4

1,2,4

1,2

2,3,4

2,3

3,4

2,4

## Items of 1 element

Itemsets:

1,2,3,4

1,2,4

1,2

2,3,4

2,3

3,4

2,4

Item	Support
1	3
2	6
3	4
4	5

## Items of 2 elements

Itemsets:

1,2,3,4

1,2,4

1,2

2,3,4

2,3

3,4

2,4

Item	Support
1,2	3
1,3	1
1,4	2
2,3	3
2,4	4
3,4	3

## Items of 3 elements

Itemsets:

1,2,3,4

1,2,4

1,2

2,3,4

2,3

3,4

2,4

Item	
2,3,4	2

# Association rules in Weka

The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Choose' button is highlighted with a red box. The command line below it reads: `Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1`. The 'Result list' on the left shows three entries, with '10:52:57 - Apriori' selected. The 'Associator output' pane displays the following text:

```
Minimum metric <confidence>: 0.9
Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 6
Size of set of large itemsets L(3): 2

Best rules found:

1. int-discolor=none 581 ==> sclerotia=absent 581   conf:(1)
2. mycelium=absent int-discolor=none 575 ==> sclerotia=absent 575   conf:(1)
3. leaves=abnorm sclerotia=absent 548 ==> mycelium=absent 547   conf:(1)
4. sclerotia=absent 625 ==> mycelium=absent 619   conf:(0.99)
5. int-discolor=none 581 ==> mycelium=absent 575   conf:(0.99)
6. int-discolor=none sclerotia=absent 581 ==> mycelium=absent 575   conf:(0.9)
7. int-discolor=none 581 ==> mycelium=absent sclerotia=absent 575   conf:(0.9)
8. leaf-malf=absent 554 ==> mycelium=absent 548   conf:(0.99)
9. mycelium=absent 639 ==> sclerotia=absent 619   conf:(0.97)
10. leaves=abnorm mycelium=absent 567 ==> sclerotia=absent 547   conf:(0.96)
```

Associate panel in Explorer

Choose algorithm



# Apriori output

Minimum metric <confidence>: 0.9

Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 2

Best rules found:

1. int-discolor=none 581 ==> sclerotia=absent 581 conf:(1)
2. mycelium=absent int-discolor=none 575 ==> sclerotia=absent 575 conf:(1)
3. leaves=abnorm sclerotia=absent 548 ==> mycelium=absent 547 conf:(1)
4. sclerotia=absent 625 ==> mycelium=absent 619 conf:(0.99)
5. int-discolor=none 581 ==> mycelium=absent 575 conf:(0.99)
6. int-discolor=none sclerotia=absent 581 ==> mycelium=absent 575 conf:(0.9)
7. int-discolor=none 581 ==> mycelium=absent sclerotia=absent 575 conf:(0.9)
8. leaf-malf=absent 554 ==> mycelium=absent 548 conf:(0.99)
9. mycelium=absent 639 ==> sclerotia=absent 619 conf:(0.97)
10. leaves=abnorm mycelium=absent 567 ==> sclerotia=absent 547 conf:(0.96)

Minimum confidence

Best rules discovered

# Ex.1

- open soybean.arff
- use the apriori algorithm to discover new rules

Questions?

**T H A N K**

Five colorful speech bubble tags are hanging from black strings. The tags are orange, pink, red, light green, and blue, and they contain the letters T, H, A, N, and K respectively. The letters are in a bold, sans-serif font, with the 'K' being white and the others being a lighter shade of the tag's color.

**Y O U**

Three colorful speech bubble tags are hanging from black strings. The tags are orange, red, and light green, and they contain the letters Y, O, and U respectively. The letters are in a bold, sans-serif font, with the 'O' being white and the others being a lighter shade of the tag's color.