



Experiment No. 8

Name: Lokesh Patil

Branch: ETRX

Roll no: 41

Objective: Apply Apriori Algorithm to given dataset : Association Rule Mining with WEKA.

System Requirements: Weka version 3.8.6

DataSet:

Groceries.csv

	A	B	C	D	E	F	G
1	Trans_id	exista	existb	existc	existd	existe	existk
2	T1	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
3	T2	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
4	T3	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE
5	T4	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE

Results:

Exercise 1: The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) is 80% ? Trans_id Itemlist

T1 {K, A, D, B}

T2 {D, A C, E, B}

T3 {C, A, B, E}

T4 {B, A, D}

Hint: Make a tabular and binary representation of the data in order to better see the relationship between Items. First generate all item sets with minimum support of 60%. Then form rules and calculate their confidence base on the conditional probability $P(B|A) = \frac{|B \cap A|}{|A|}$. Remember to only take the item sets from the previous phase whose support is 60% or more.

Tabular Representation:

Trans id	A	B	C	D	E	K
T1	1	1	0	1	0	1
T2	1	1	1	1	1	0
T3	1	1	1	0	1	0
T4	1	1	0	1	0	0

Min Support = 0.6

Item	Frequency	Support
A	4	$4/4 = 1$
B	4	$4/4 = 1$
C	2	$2/4 = 0.5$
D	3	$3/4 = 0.75$
E	2	$2/4 = 0.5$
K	1	$1/4 = 0.25$

Considering item Sets with Support $\geq 60\%$

A = 1, B = 1, D = 0.75

Considering 2 items at a time

Pair	Frequency	Support
AB	4	$4/4 = 1$
AD	3	$3/4 = 0.75$
BD	3	$3/4 = 0.75$

Considering 3 items at a time

ABD Freq: 3 Support = $3/4 = 0.75$

Forming rules & finding confidence

A \rightarrow B	$P(B/A) = 4/4 = 1$
B \rightarrow A	$P(A/B) = 4/4 = 1$
A \rightarrow D	$P(D/A) = 3/4 = 0.75$
D \rightarrow A	$P(A/D) = 3/3 = 1$
B \rightarrow D	$P(D/B) = 3/4 = 0.75$
D \rightarrow B	$P(B/D) = 3/3 = 1$
AB \rightarrow D	$P(D/AB) = 3/4 = 0.75$
D \rightarrow AB	$P(AB/D) = 3/3 = 1$
AD \rightarrow B	$P(B/AD) = 3/3 = 1$
B \rightarrow AD	$P(AD/B) = 3/4 = 0.75$
BD \rightarrow A	$P(A/BD) = 3/3 = 1$
A \rightarrow BD	$P(BD/A) = 3/4 = 0.75$

Considering Rules with Confidence $\geq 80\%$

A \rightarrow B	100%
B \rightarrow A	100%
D \rightarrow A	100%
D \rightarrow B	100%
D \rightarrow AB	100%
AD \rightarrow B	100%
DB \rightarrow A	100%

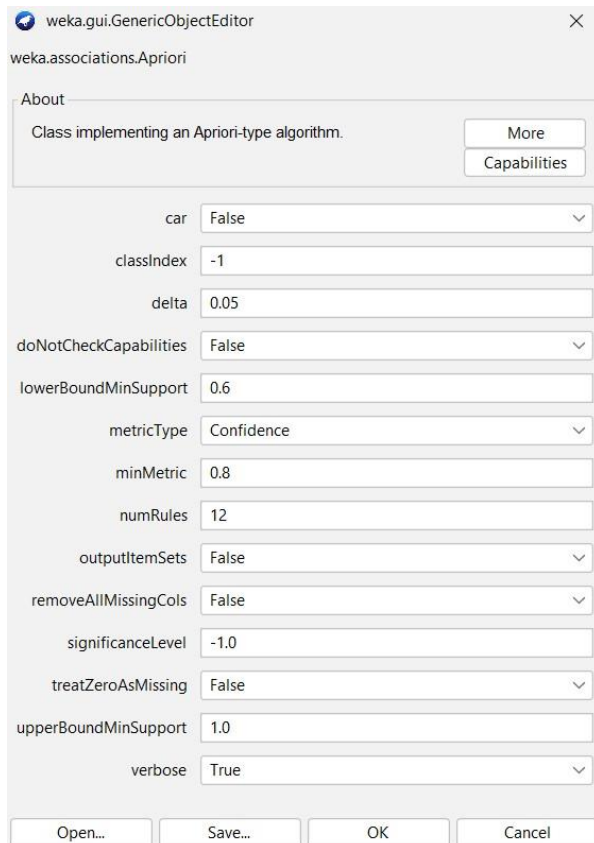
Exercise2: Input file generation and Initial experiments with Weka's association rule discovery. 1. Launch Weka and try to do the calculations you performed manually in the previous exercise. Use the apriori algorithm for generating the association rules.

Did you succeed? Are the results the same as in your calculations?

→ Yes

What kind of file did you use as input?

→ CSV



The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.associations.Apriori' class. The 'About' tab is active, displaying the text 'Class implementing an Apriori-type algorithm.' with 'More' and 'Capabilities' buttons. Below this, various settings are listed in a table-like format:

Property	Value
car	False
classIndex	-1
delta	0.05
doNotCheckCapabilities	False
lowerBoundMinSupport	0.6
metricType	Confidence
minMetric	0.8
numRules	12
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
treatZeroAsMissing	False
upperBoundMinSupport	1.0
verbose	True

At the bottom of the window are four buttons: 'Open...', 'Save...', 'OK', and 'Cancel'.

Associator output

```

=== Run information ===

Scheme:      weka.associations.Apriori -N 12 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.6 -S -1.0 -V -c -1
Relation:    data
Instances:   4
Attributes:  7
              Trans_id
              exista
              existb
              existc
              existd
              existe
              existk

=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.85 (3 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 5

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

Best rules found:

1. existb=TRUE 4 ==> exista=TRUE 4    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. exista=TRUE 4 ==> existb=TRUE 4    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. existd=TRUE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. existk=FALSE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. existd=TRUE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. existk=FALSE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. existb=TRUE existd=TRUE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. exista=TRUE existd=TRUE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. existd=TRUE 3 ==> exista=TRUE existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. existb=TRUE existk=FALSE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
11. exista=TRUE existk=FALSE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
12. existk=FALSE 3 ==> exista=TRUE existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

```

Exercise 3: Mining Association Rule with WEKA Explorer – Weather dataset Task 1. Run Apriori on this data with default settings. Comment on the rules that are generated. Several of them are quite similar. How are their support and confidence values related?

Task 2. It is interesting to see that none of the rules in the default output involve Class = republican. Why do you think that is?

```

=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -V -c -1
Relation:    weather.symbolic
Instances:   14
Attributes:  5
             outlook
             temperature
             humidity
             windy
             play

=== Associator model (full training set) ===

```

```

Apriori
=====

```

```

Minimum support: 0.25 (4 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 15

```

```

Generated sets of large itemsets:

```

```

Size of set of large itemsets L(1): 12

```

```

Size of set of large itemsets L(2): 26

```

```

Size of set of large itemsets L(3): 4

```

```

Best rules found:

```

```

Best rules found:

```

1. outlook=overcast 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
2. temperature=cool 4 ==> humidity=normal 4 <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
4. outlook=sunny play=no 3 ==> humidity=high 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
5. outlook=sunny humidity=high 3 ==> play=no 3 <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
8. temperature=cool play=yes 3 ==> humidity=normal 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
9. humidity=normal 7 ==> play=yes 6 <conf:(0.86)> lift:(1.33) lev:(0.11) [1] conv:(1.25)
10. play=no 5 ==> humidity=high 4 <conf:(0.8)> lift:(1.6) lev:(0.11) [1] conv:(1.25)

Changing the parameters:

With minSupport:0.2 and confidence:0.5


```

=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -V -c -1
Relation:    weather.symbolic
Instances:   14
Attributes:  5
              outlook
              temperature
              humidity
              windy
              play

=== Associator model (full training set) ===

```

```

Apriori
=====

```

```

Minimum support: 0.3 (4 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 14

```

```

Generated sets of large itemsets:

```

```

Size of set of large itemsets L(1): 12

```

```

Size of set of large itemsets L(2): 9

```

```

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

```

```

Best rules found:

```

```

Best rules found:

```

1. outlook=overcast 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
2. temperature=cool 4 ==> humidity=normal 4 <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
4. humidity=normal 7 ==> play=yes 6 <conf:(0.86)> lift:(1.33) lev:(0.11) [1] conv:(1.25)
5. play=no 5 ==> humidity=high 4 <conf:(0.8)> lift:(1.6) lev:(0.11) [1] conv:(1.25)
6. windy=FALSE 8 ==> play=yes 6 <conf:(0.75)> lift:(1.17) lev:(0.06) [0] conv:(0.95)
7. play=yes 9 ==> humidity=normal 6 <conf:(0.67)> lift:(1.33) lev:(0.11) [1] conv:(1.13)
8. play=yes 9 ==> windy=FALSE 6 <conf:(0.67)> lift:(1.17) lev:(0.06) [0] conv:(0.96)
9. temperature=mild 6 ==> humidity=high 4 <conf:(0.67)> lift:(1.33) lev:(0.07) [1] conv:(1)
10. temperature=mild 6 ==> play=yes 4 <conf:(0.67)> lift:(1.04) lev:(0.01) [0] conv:(0.71)

Exercise 4: Mining Association Rule with WEKA Explorer – Vote

```

=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -V -c -1
Relation:    vote
Instances:   435
Attributes:  17
              handicapped-infants
              water-project-cost-sharing
              adoption-of-the-budget-resolution
              physician-fee-freeze
              el-salvador-aid
              religious-groups-in-schools
              anti-satellite-test-ban
              aid-to-nicaraguan-contras
              mx-missile
              immigration
              synfuels-corporation-cutback
              education-spending
              superfund-right-to-sue
              crime
              duty-free-exports
              export-administration-act-south-africa
              Class

=== Associator model (full training set) ===

```

```
Apriori
=====
```

```
Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11
```

```
Generated sets of large itemsets:
```

```
Size of set of large itemsets L(1): 20
```

```
Size of set of large itemsets L(2): 17
```

```
Size of set of large itemsets L(3): 6
```

```
Size of set of large itemsets L(4): 1
```

```
Best rules found:
```

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 <conf:(1)> lift:(1.63) lev:(0.19) [84] conv:(84.58)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 <conf:(1)> lift:(1.63) lev:(0.18) [76] conv:(76.46)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74)
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 <conf:(1)> lift:(1.62) lev:(0.18) [77] conv:(39.01)
5. physician-fee-freeze=n 247 ==> Class=democrat 245 <conf:(0.99)> lift:(1.62) lev:(0.21) [93] conv:(31.8)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 <conf:(0.98)> lift:(1.77) lev:(0.2) [85] conv:(22.18)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 <conf:(0.98)> lift:(1.76) lev:(0.2) [88] conv:(18.46)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 <conf:(0.98)> lift:(1.72) lev:(0.19) [82] conv:(22.18)
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 <conf:(0.97)> lift:(1.57) lev:(0.17) [71] conv:(9.85)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 <conf:(0.96)> lift:(1.7) lev:(0.2) [86] conv:(10.47)

Exercise 5: Let's run Apriori on another real-world dataset.

Load data at Preprocess tab. Click the Open file button to bring up a standard dialog through which you can select a file. Choose the supermarket.arff file. To see the original dataset, click the Edit button, a viewer window opens with dataset loaded.

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1
Relation: supermarket
Instances: 4627
Attributes: 217
[list of attributes omitted]

=== Associator model (full training set) ===

Apriori

=====

Minimum support: 0.4 (1851 instances)

Minimum metric <confidence>: 0.7

Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 16

Best rules found:

1. biscuits=t 2605 ==> bread and cake=t 2083 <conf:(0.8)> lift:(1.11) lev:(0.04) [208] conv:(1.4)
2. milk-cream=t 2939 ==> bread and cake=t 2337 <conf:(0.8)> lift:(1.1) lev:(0.05) [221] conv:(1.37)
3. fruit=t 2962 ==> bread and cake=t 2325 <conf:(0.78)> lift:(1.09) lev:(0.04) [193] conv:(1.3)
4. baking needs=t 2795 ==> bread and cake=t 2191 <conf:(0.78)> lift:(1.09) lev:(0.04) [179] conv:(1.29)
5. frozen foods=t 2717 ==> bread and cake=t 2129 <conf:(0.78)> lift:(1.09) lev:(0.04) [173] conv:(1.29)
6. vegetables=t 2961 ==> bread and cake=t 2298 <conf:(0.78)> lift:(1.08) lev:(0.04) [167] conv:(1.25)
7. juice-sat-cord-ms=t 2463 ==> bread and cake=t 1869 <conf:(0.76)> lift:(1.05) lev:(0.02) [96] conv:(1.16)
8. vegetables=t 2961 ==> fruit=t 2207 <conf:(0.75)> lift:(1.16) lev:(0.07) [311] conv:(1.41)
9. fruit=t 2962 ==> vegetables=t 2207 <conf:(0.75)> lift:(1.16) lev:(0.07) [311] conv:(1.41)
10. bread and cake=t 3330 ==> milk-cream=t 2337 <conf:(0.7)> lift:(1.1) lev:(0.05) [221] conv:(1.22)