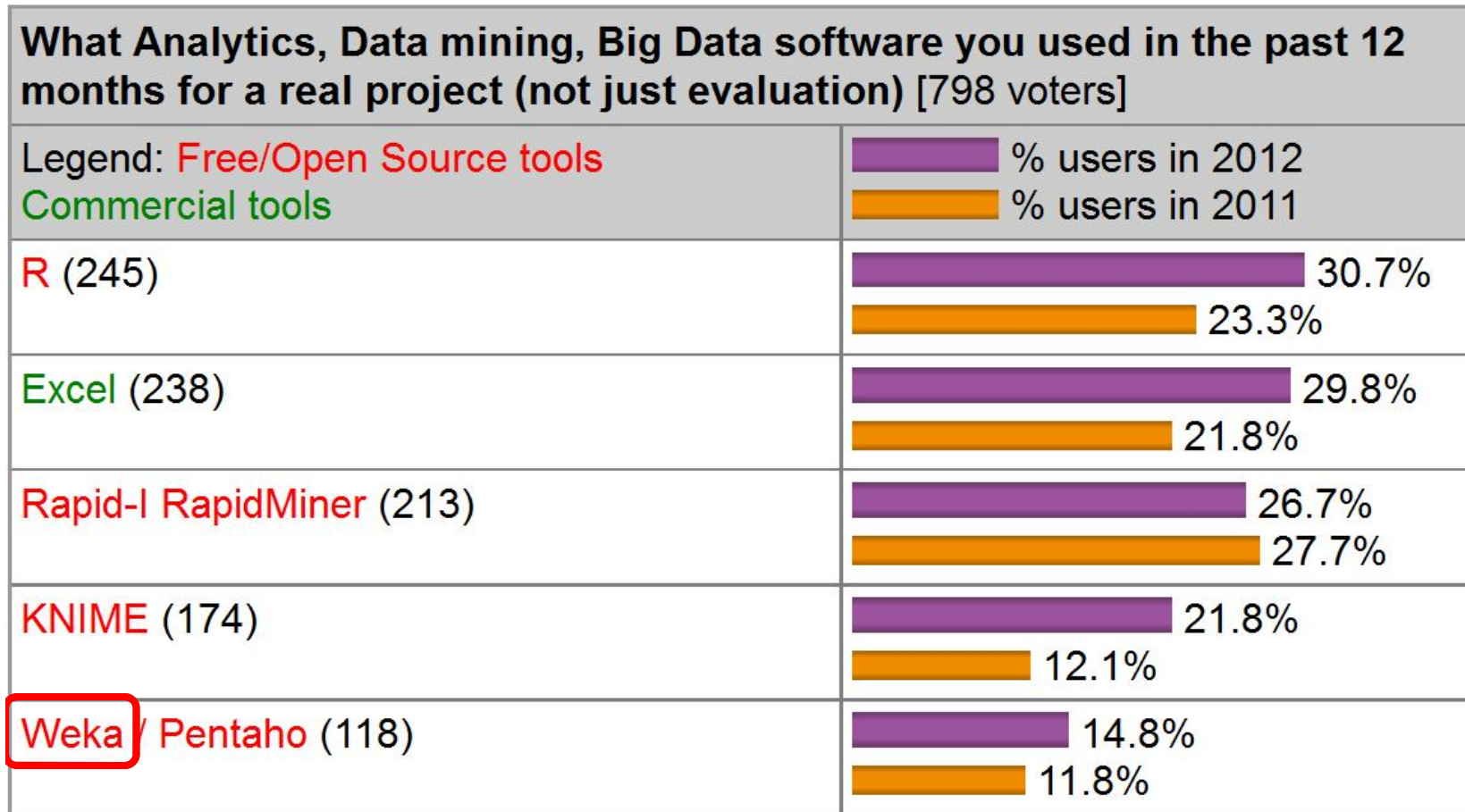


CS2220 Introduction to Computational Biology

Weka Introduction

Xiaoli Li

Most Popular Tools for Data Mining



<http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html>

WEKA – Data Mining Software in Java



Classify

Cluster

Associate



Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply

Current relation
Relation: vote
Instances: 435 Attributes: 17

Attributes
All None Invert Pattern

No.	Name
1	☒ handicapped-infants
2	☐ water-project-cost-sharing
3	☐ adoption-of-the-budget-resolution
4	☐ physician-fee-freeze
5	☐ el-salvador-aid
6	☐ religious-groups-in-schools
7	☐ anti-satellite-test-ban
8	☐ aid-to-nicaraguan-contras
9	☐ mx-missile
10	☐ immigration
11	☐ synfuels-corporation-cutback
12	☐ education-spending
13	☐ superfund-right-to-sue

Remove

Selected attribute
Name: handicapped-infants
Missing: 12 (3%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	n	236
2	y	187

Class: Class (Nom) Visualize All

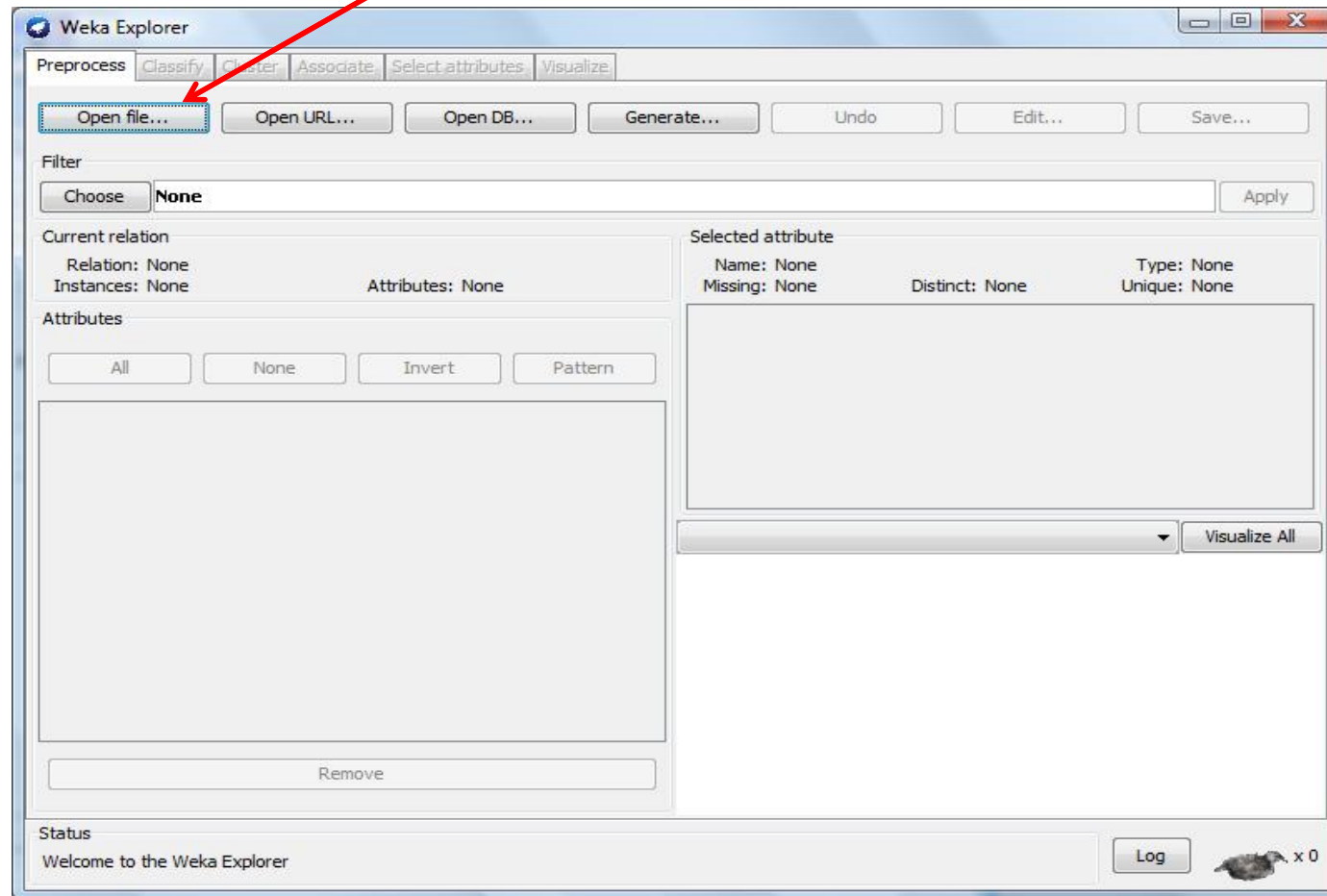
Status
OK Log x 0

<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Select the Data for Exploring



We can see and edit our data

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Open the current dataset in a Viewer for editing Apply

Current relation: weather.symbolic
Instances: 14

Attributes: All

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Remove

Status: OK

Viewer

Relation: weather.symbolic

No.	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

Selected attribute: Name: outlook
Missing: 0 (0%) Distinct: 3 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom) Visualize All

Log x 0

WEKA – Data Mining Software in Java



Example: Predict if we want to play

```
TextPad - C:\Program Files\Weka-3-6\data\weather...
File Edit Search View Tools Macros Configure Window
Help
weather.arff weather.nominal.arff
1 @relation weather
2
3 @attribute outlook {sunny, overcast, rainy}
4 @attribute temperature real
5 @attribute humidity real
6 @attribute windy {TRUE, FALSE}
7 @attribute play {yes, no}
8
9 @data
10 sunny, 85, 85, FALSE, no
11 sunny, 80, 90, TRUE, no
12 overcast, 83, 86, FALSE, yes
13 rainy, 70, 96, FALSE, yes
14 rainy, 68, 80, FALSE, yes
15 rainy, 65, 70, TRUE, no
16 overcast, 64, 65, TRUE, yes
17 sunny, 72, 95, FALSE, no
18 sunny, 69, 70, FALSE, yes
19 rainy, 75, 80, FALSE, yes
20 sunny, 75, 70, TRUE, yes
21 overcast, 72, 90, TRUE, yes
22 overcast, 81, 75, FALSE, yes
23 rainy, 71, 91, TRUE, no

TextPad - C:\Program Files\Weka-3-6\data\weather...
File Edit Search View Tools Macros Configure Window
Help
weather.arff weather.nominal.arff
1 @relation weather.symbolic
2
3 @attribute outlook {sunny, overcast, rainy}
4 @attribute temperature {hot, mild, cool}
5 @attribute humidity {high, normal}
6 @attribute windy {TRUE, FALSE}
7 @attribute play {yes, no}
8
9 @data
10 sunny, hot, high, FALSE, no
11 sunny, hot, high, TRUE, no
12 overcast, hot, high, FALSE, yes
13 rainy, mild, high, FALSE, yes
14 rainy, cool, normal, FALSE, yes
15 rainy, cool, normal, TRUE, no
16 overcast, cool, normal, TRUE, yes
17 sunny, mild, high, FALSE, no
18 sunny, cool, normal, FALSE, yes
19 rainy, mild, normal, FALSE, yes
20 sunny, mild, normal, TRUE, yes
21 overcast, mild, high, TRUE, yes
22 overcast, hot, normal, FALSE, yes
23 rainy, mild, high, TRUE, no

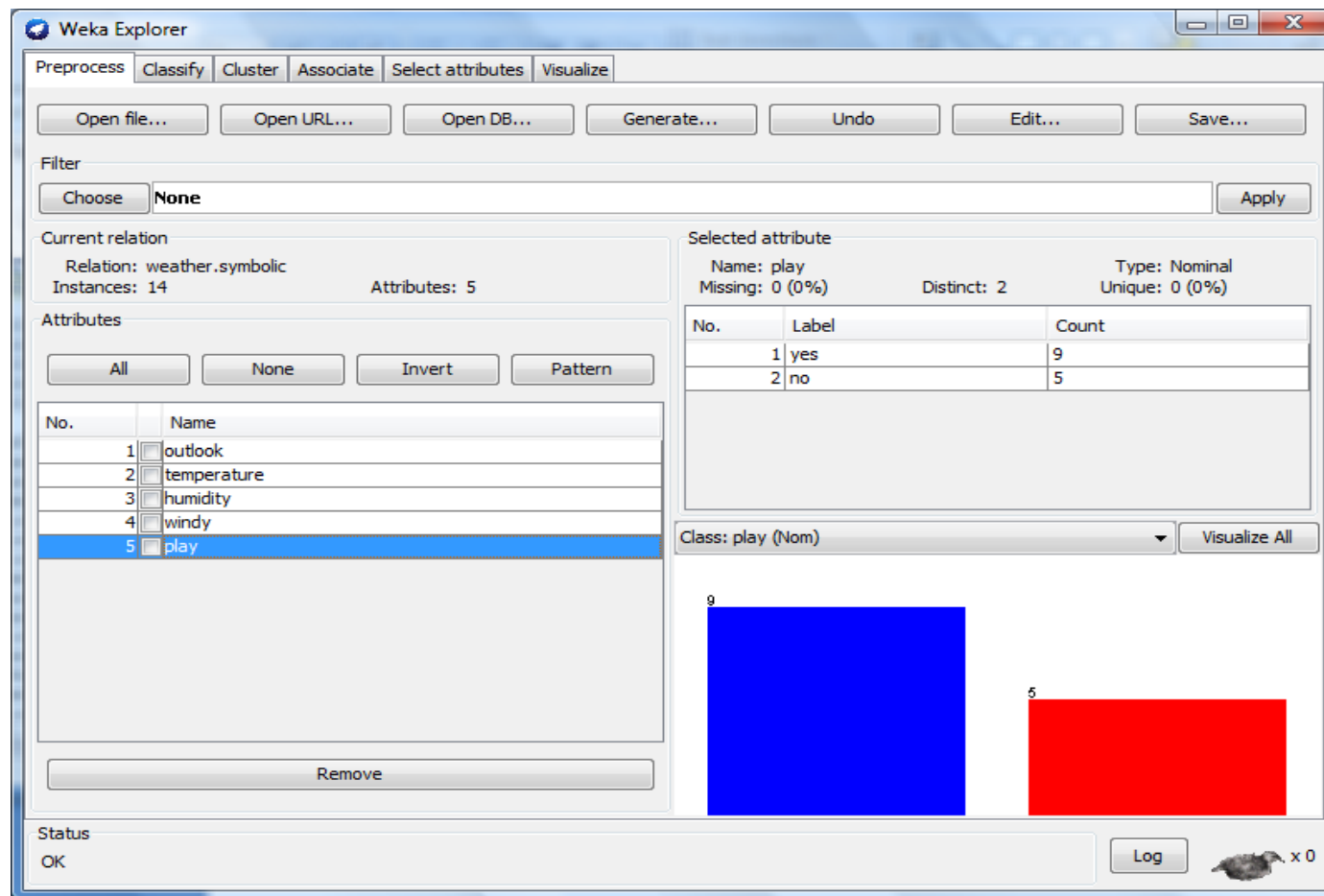
For Help, press F1 1
```

Actual Weka data format

WEKA – Data Mining Software in Java



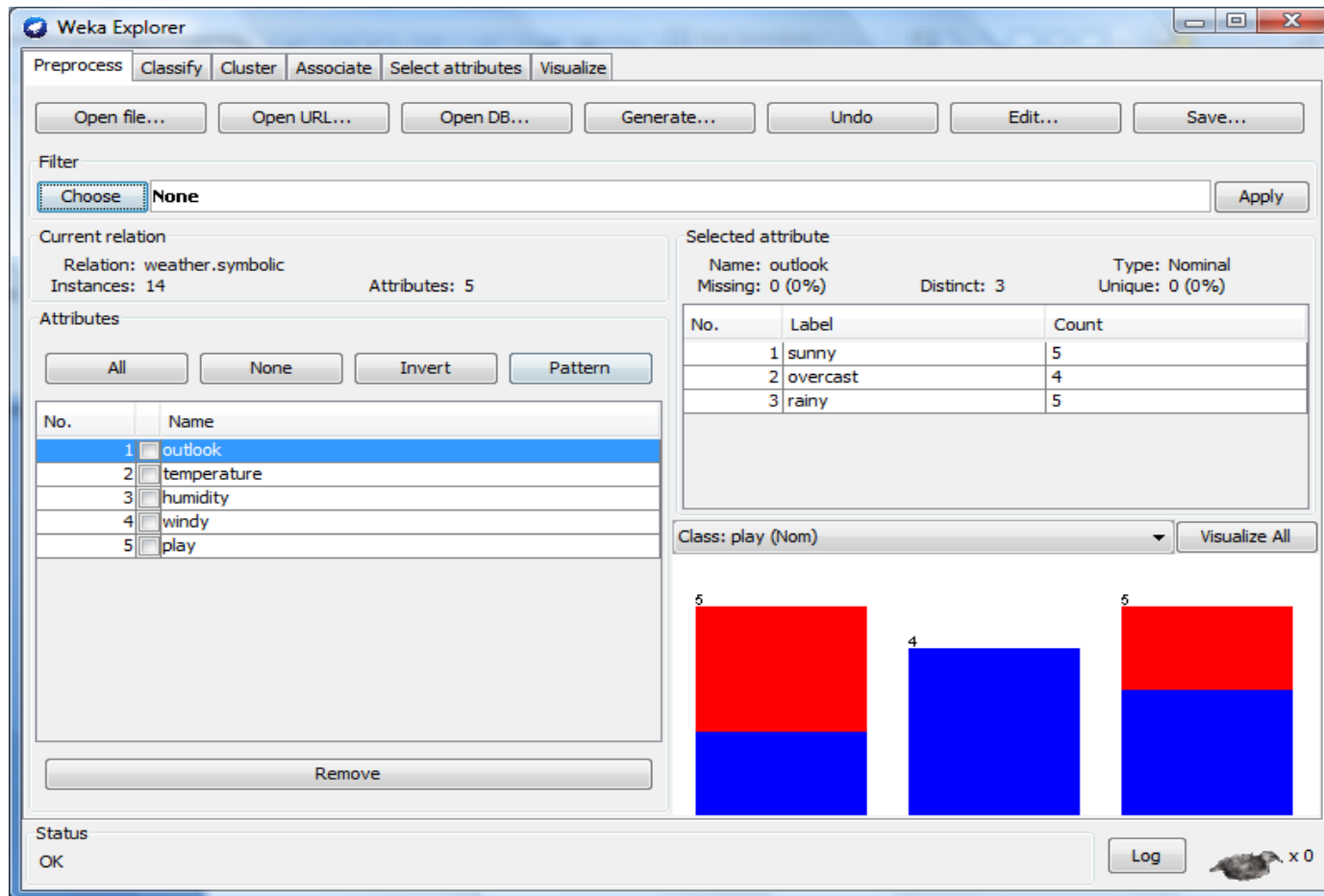
Class label: To **Play** or **Not to Play**?



WEKA – Data Mining Software in Java



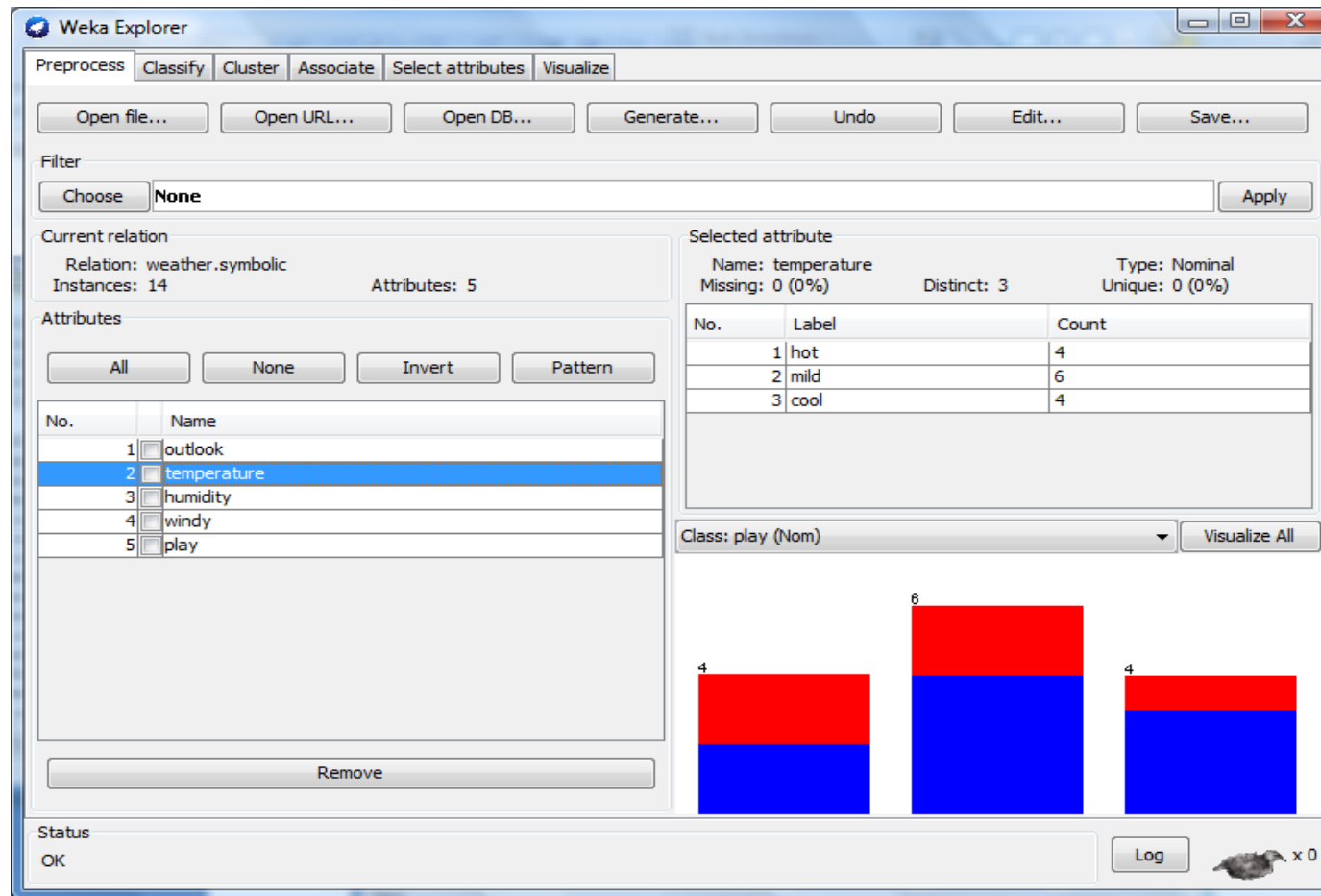
Outlook: To Play or Not to Play?



WEKA – Data Mining Software in Java



Temperature: To **Play** or **Not to Play**?



<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Humidity: To **Play** or **Not to Play**?

The screenshot shows the Weka Explorer window with the 'weather.symbolic' dataset loaded. The 'humidity' attribute is selected, and the 'play' class is set. The interface displays a table of attributes and a visualization of the data.

Current relation: weather.symbolic
Instances: 14
Attributes: 5

Attributes:

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Selected attribute:

Name: humidity
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	high	7
2	normal	7

Class: play (Nom)

Visualization: Two bar charts showing the distribution of the 'play' class for 'humidity'. The first chart shows a red bar for 'high' (7) and a blue bar for 'normal' (7). The second chart shows a red bar for 'high' (7) and a blue bar for 'normal' (7).

<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Windy: To **Play** or **Not to Play**?

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply

Current relation
Relation: weather.symbolic
Instances: 14 Attributes: 5

Attributes
All None Invert Pattern

No.	Name
1	☐ outlook
2	☐ temperature
3	☐ humidity
4	☒ windy
5	☐ play

Remove

Selected attribute
Name: windy
Missing: 0 (0%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	TRUE	6
2	FALSE	8

Class: play (Nom) Visualize All

Windy	Play	Count
TRUE	TRUE	6
TRUE	FALSE	0
FALSE	TRUE	8
FALSE	FALSE	0

Status: OK Log x 0

WEKA – Data Mining Software in Java



Classification: To **Play** or **Not to Play?**

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose ZeroR

Test options

☒ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click for options)

16:52:54 - rules.ZeroR

Classifier output

ZeroR predicts class value: yes

Time taken to build model: 0 seconds

=== Predictions on training set ===

inst#	actual	predicted	error	probability distribution
1	2:no	1:yes	+	*0.625 0.375
2	2:no	1:yes	+	*0.625 0.375
3	1:yes	1:yes		*0.625 0.375
4	1:yes	1:yes		*0.625 0.375
5	1:yes	1:yes		*0.625 0.375
6	2:no	1:yes	+	*0.625 0.375
7	1:yes	1:yes		*0.625 0.375
8	2:no	1:yes	+	*0.625 0.375
9	1:yes	1:yes		*0.625 0.375
10	1:yes	1:yes		*0.625 0.375
11	1:yes	1:yes		*0.625 0.375
12	1:yes	1:yes		*0.625 0.375
13	1:yes	1:yes		*0.625 0.375
14	2:no	1:yes	+	*0.625 0.375

=== Evaluation on training set ===

Status OK

Log x 0

WEKA – Data Mining Software in Java



J48 Decision Tree: To **Play** or **Not to Play**?

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected. In the 'Classifier' section, 'J48 -C 0.25 -M 2' is chosen. The 'Test options' section has 'Use training set' selected. The 'Test mode: evaluate on training data' is indicated. The 'Result list' shows '17:10:39 - trees.J48'. The 'Classifier output' pane displays the following information:

```
Relation: weather.symbolic
Instances: 14
Attributes: 5
  outlook
  temperature
  humidity
  windy
  play

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
-----
outlook = sunny
|  humidity = high: no (3.0)
|  humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|  windy = TRUE: no (2.0)
|  windy = FALSE: yes (3.0)

Number of Leaves :    5
```

A blue box labeled 'Actual tree structure' points to the J48 pruned tree output.

WEKA – Data Mining Software in Java



Decision Tree: To **Blue** or **Not to Play**?

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☒ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click for options)

17:10:39 - trees.J48

Classifier output

=== Predictions on training set ===

inst#	actual	predicted	error	probability distribution
1	2:no	2:no	0	*1
2	2:no	2:no	0	*1
3	1:yes	1:yes	*1	0
4	1:yes	1:yes	*1	0
5	1:yes	1:yes	*1	0
6	2:no	2:no	0	*1
7	1:yes	1:yes	*1	0
8	2:no	2:no	0	*1
9	1:yes	1:yes	*1	0
10	1:yes	1:yes	*1	0
11	1:yes	1:yes	*1	0
12	1:yes	1:yes	*1	0
13	1:yes	1:yes	*1	0
14	2:no	2:no	0	*1

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	14	100	%
Incorrectly Classified Instances	0	0	%

Status

OK

Log x 0

Training set has been used as test set

WEKA – Data Mining Software in Java



Decision Tree: To **Blue** or **Not to Play**?

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 14

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click for options)

17:10:39 - trees.J48

17:17:37 - trees.J48

17:18:48 - trees.J48

Classifier output

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: 14-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

outlook = sunny

| humidity = high: no (3.0)

| humidity = normal: yes (2.0)

outlook = overcast: yes (4.0)

outlook = rainy

| windy = TRUE: no (2.0)

| windy = FALSE: yes (3.0)

Number of Leaves : 5

Status

OK

Log x 0

Leave one out
CV
Tree structure

WEKA – Data Mining Software in Java



Decision Tree: To **Blue** or **Not to Play**?

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.35 -M 1

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 14

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click for options)

- 17:43:51 - trees.J48
- 17:43:59 - trees.J48
- 17:44:06 - trees.J48
- 17:44:22 - trees.J48
- 17:44:31 - trees.J48
- 17:44:36 - trees.J48
- 17:45:07 - trees.J48
- 17:45:13 - trees.J48
- 17:45:23 - trees.J48
- 17:45:29 - trees.J48
- 17:45:37 - trees.J48
- 17:48:12 - trees.J48

Classifier output

=== Predictions on test data ===

inst#	actual	predicted	error	probability distribution
1	2:no	1:yes	+	*1 0
1	2:no	2:no	0	*1
1	2:no	2:no	0	*1
1	2:no	2:no	0	*1
1	2:no	2:no	0	*1
1	1:yes	1:yes	*1	0
1	1:yes	1:yes	*1	0
1	1:yes	1:yes	*1	0
1	1:yes	1:yes	*1	0
1	1:yes	1:yes	*1	0
1	1:yes	1:yes	*1	0
1	1:yes	2:no	+	0 *1
1	1:yes	1:yes	*1	0

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	12	85.7143 %
Incorrectly Classified Instances	2	14.2857 %

Status

OK

Log x 0

Change the parameter of J48

WEKA – Data Mining Software in Java



Classify **Cluster** Associate



Weka Explorer

Preprocess **Classify** **Cluster** Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply

Current relation
Relation: vote
Instances: 435 Attributes: 17

Attributes
All None Invert Pattern

No.	Name
1	☒ handicapped-infants
2	☐ water-project-cost-sharing
3	☐ adoption-of-the-budget-resolution
4	☐ physician-fee-freeze
5	☐ el-salvador-aid
6	☐ religious-groups-in-schools
7	☐ anti-satellite-test-ban
8	☐ aid-to-nicaraguan-contras
9	☐ mx-missile
10	☐ immigration
11	☐ synfuels-corporation-cutback
12	☐ education-spending
13	☐ superfund-right-to-sue

Remove

Selected attribute
Name: handicapped-infants
Missing: 12 (3%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

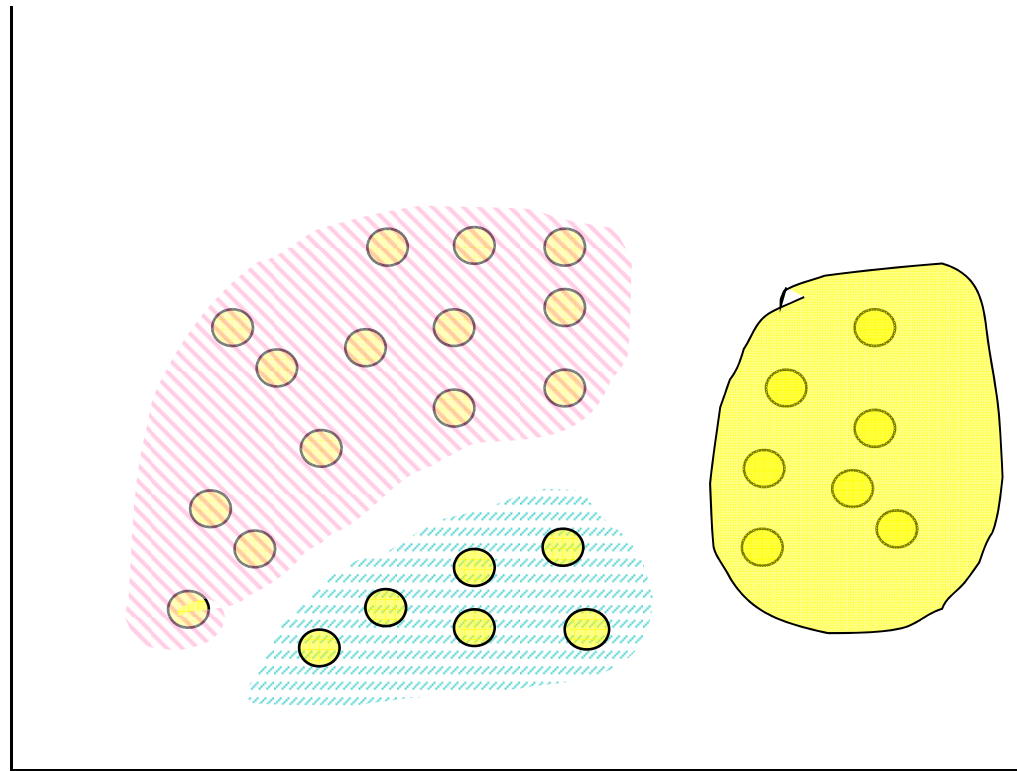
No.	Label	Count
1	n	236
2	y	187

Class: Class (Nom) Visualize All

Status
OK Log x 0

Pattern Discovery: Clustering

Find “natural” grouping of instances
given un-labeled data



WEKA – Data Mining Software in Java



Example: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

```
TextPad - E:\recsys\data mining\HDB RI - 21 June 2012\iris.arff
File Edit Search View Tools Macros Configure Window Help
iris.arff vote.arff
40
41 6. Number of Attributes: 4 numeric, predictive attributes and the class
42
43 7. Attribute Information:
44 1. sepal length in cm
45 2. sepal width in cm
46 3. petal length in cm
47 4. petal width in cm
48 5. class:
49 -- Iris Setosa
50 -- Iris Versicolour
51 -- Iris Virginica
52
53 8. Missing Attribute Values: None
54
55 Summary Statistics:
56      Min  Max  Mean  SD  Class Correlation
57 sepal length: 4.3  7.9  5.84  0.83  0.7826
58 sepal width: 2.0  4.4  3.05  0.43  -0.4194
59 petal length: 1.0  6.9  3.76  1.76  0.9490 (high!)
60 petal width: 0.1  2.5  1.20  0.76  0.9565 (high!)
61
62 9. Class Distribution: 33.3% for each of 3 classes.
63
File: iris.arff, 7486 bytes, 225 lines, UNIX, ANSI 1 1 Read Ovr Block Sync Rec Caps
```

WEKA – Data Mining Software in Java



Example: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Attributes: All None Invert Pattern

No.	Name
1	☐ sepallength
2	☐ sepalwidth
3	☐ petallength
4	☐ petalwidth
5	☒ class

Remove

Selected attribute: Name: class Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	Iris-setosa	50
2	Iris-versicolor	50
3	Iris-virginica	50

Class: class (Nom) Visualize All

Status: OK Log x 0

WEKA – Data Mining Software in Java



Clustering: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

Weka Explorer

Preprocess | Classify | **Cluster** | Associate | Select attributes | Visualize

Clusterer
Choose **SimpleKMeans** -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -S 10

Cluster mode
☐ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☒ Classes to clusters evaluation
(Nom) class
☒ Store clusters for visualization
Ignore attributes
Start Stop

Result list (right-click for options)
21:28:50 - SimpleKMeans

Clusterer output

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	Cluster# 1 (50)	Cluster# 2 (39)
sepal.length	5.8433	5.8885	5.006	6.8462
sepal.width	3.054	2.7377	3.418	3.0821
petal.length	3.7587	4.3967	1.464	5.7026
petal.width	1.1987	1.418	0.244	2.0795

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	61 (41%)
1	50 (33%)
2	39 (26%)

Status
OK

Log x 0

WEKA – Data Mining Software in Java



Clustering: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

The screenshot shows the WEKA Explorer window with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'SimpleKMeans' with parameters '-N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -S 10'. The 'Cluster mode' section has 'Classes to clusters evaluation' selected. The 'Clusterer output' pane shows the following results:

```
=== Model and evaluation on training set ===  
  
Clustered Instances  
0      61 ( 41%)  
1      50 ( 33%)  
2      39 ( 26%)  
  
Class attribute: class  
Classes to Clusters:  
0 1 2 <-- assigned to cluster  
0 50 0 | Iris-setosa  
47 0 3 | Iris-versicolor  
14 0 36 | Iris-virginica  
  
Cluster 0 <-- Iris-versicolor  
Cluster 1 <-- Iris-setosa  
Cluster 2 <-- Iris-virginica  
  
Incorrectly clustered instances : 17.0 11.3333 %
```

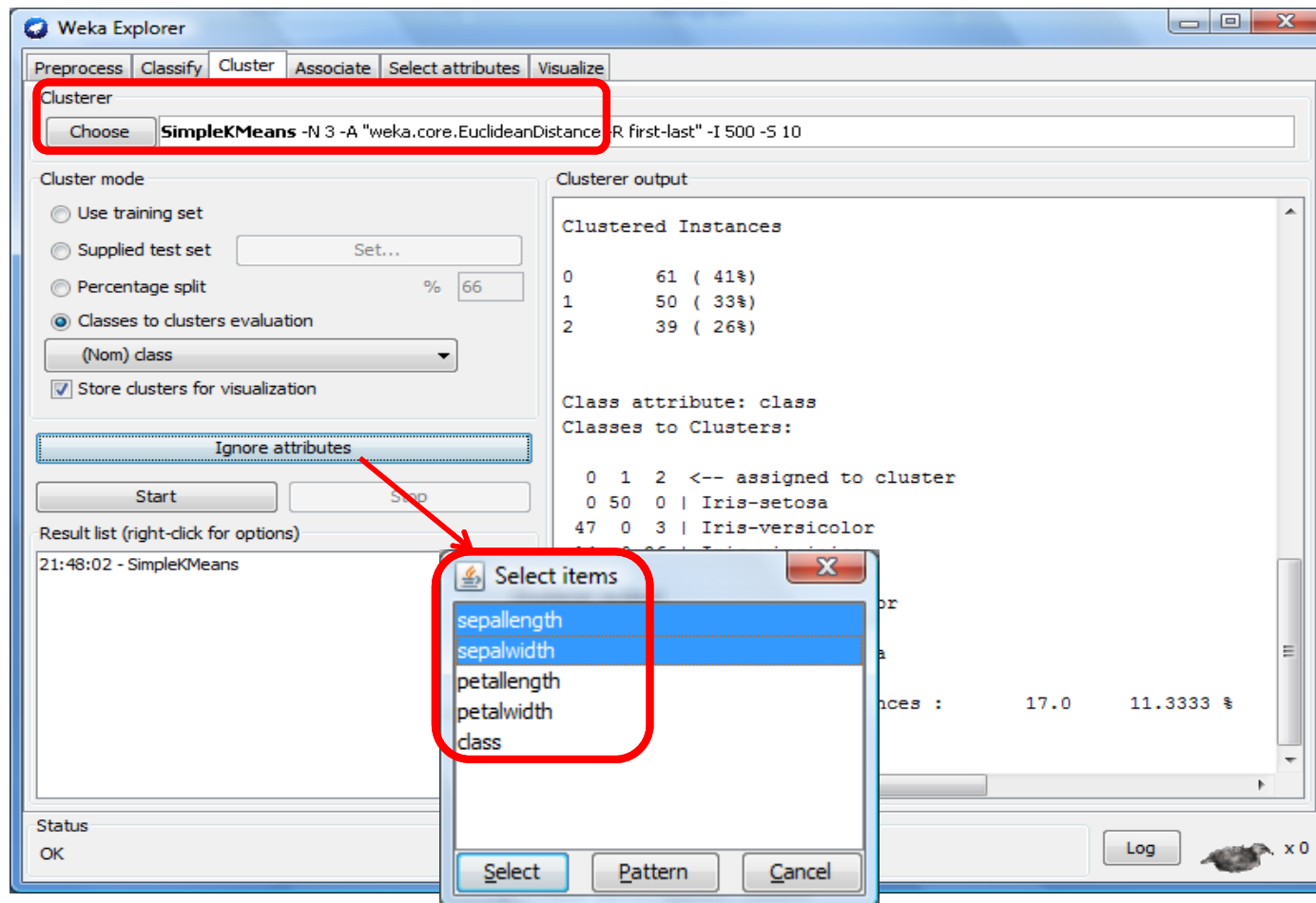
The 'Result list' on the left shows '21:36:53 - SimpleKMeans' as the selected result.

<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Clustering: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}



<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Clustering: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

Weka Explorer

Preprocess | Classify | **Cluster** | Associate | Select attributes | Visualize

Clusterer
Choose **SimpleKMeans** -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -S 10

Cluster mode
☐ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☒ Classes to clusters evaluation
(Nom) class
☒ Store clusters for visualization
Ignore attributes
Start Stop

Result list (right-click for options)
21:48:02 - SimpleKMeans
22:01:51 - SimpleKMeans

Clusterer output

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (52)	Cluster# 1 (50)	Cluster# 2 (48)
petallength	3.7587	4.2962	1.464	5.5667
petalwidth	1.1987	1.325	0.244	2.0563

Time taken to build model (full training data) : 0 seconds
=== Model and evaluation on training set ===

Clustered Instances

0	52 (35%)
1	50 (33%)
2	48 (32%)

Status
OK Log x 0

<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Clustering: 3 Types of IRIS Plant
{Setosa, Versicolor & Virginica}

Weka Explorer

Preprocess | Classify | **Cluster** | Associate | Select attributes | Visualize

Clusterer
Choose **SimpleKMeans** -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -S 10

Cluster mode
☐ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☒ Classes to clusters evaluation
(Nom) class
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)
21:48:02 - SimpleKMeans
22:01:51 - SimpleKMeans

Clusterer output

=== Model and evaluation on training set ===

Clustered Instances

0	52	(35%)
1	50	(33%)
2	48	(32%)

Class attribute: class
Classes to Clusters:

0	1	2	<-- assigned to cluster
0	50	0	Iris-setosa
48	0	2	Iris-versicolor
4	0	46	Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances : 6.0 4 %

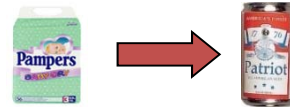
Status
OK

Log x 0

<http://www.cs.waikato.ac.nz/~ml/weka/>

Association Rules (Unsupervised Learning)

Finding groups of items that tend to occur together



WEKA – Data Mining Software in Java



Classify

Cluster

Associate



Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply

Current relation
Relation: vote
Instances: 435 Attributes: 17

Attributes
All None Invert Pattern

No.	Name
1	☒ handicapped-infants
2	☐ water-project-cost-sharing
3	☐ adoption-of-the-budget-resolution
4	☐ physician-fee-freeze
5	☐ el-salvador-aid
6	☐ religious-groups-in-schools
7	☐ anti-satellite-test-ban
8	☐ aid-to-nicaraguan-contras
9	☐ mx-missile
10	☐ immigration
11	☐ synfuels-corporation-cutback
12	☐ education-spending
13	☐ superfund-right-to-sue

Remove

Selected attribute
Name: handicapped-infants
Missing: 12 (3%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	n	236
2	y	187

Class: Class (Nom) Visualize All

Status
OK Log x 0

WEKA – Data Mining Software in Java



Example: Supermarket Purchases

```
TextPad - E:\recsys\data mining\HDB RI - 21 June 2012\supermarket.arff
File Edit Search View Tools Macros Configure Window Help
[Icons]
supermarket.arff
13 @attribute 'baby needs' { t}
14 @attribute 'bread and cake' { t}
15 @attribute 'baking needs' { t}
16 @attribute 'coupons' { t}
17 @attribute 'juice-sat-cord-ms' { t}
18 @attribute 'tea' { t}
19 @attribute 'biscuits' { t}
20 @attribute 'canned fish-meat' { t}
21 @attribute 'canned fruit' { t}
22 @attribute 'canned vegetables' { t}
23 @attribute 'breakfast food' { t}
24 @attribute 'cigs-tobacco pkts' { t}
25 @attribute 'cigarette cartons' { t}
26 @attribute 'cleaners-polishers' { t}
27 @attribute 'coffee' { t}
28 @attribute 'sauces-gravy-pkle' { t}
29 @attribute 'confectionary' { t}
30 @attribute 'puddings-deserts' { t}
31 @attribute 'dishcloths-scour' { t}
32 @attribute 'deod-disinfectant' { t}
33 @attribute 'frozen foods' { t}
34 @attribute 'razor blades' { t}
35 @attribute 'fuels-garden aids' { t}
36 @attribute 'spices' { t}
For Help, press F1 1 1 Read Ovr Block Sync Rec Caps
```

WEKA – Data Mining Software in Java



Example: Supermarket Purchases

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply

Current relation
Relation: supermarket-weka.filters.unsupervised.attribute.Remove-R...
Instances: 4627 Attributes: 99

Attributes
All None Invert Pattern

No.	Name
1	☒ baby needs
2	☐ bread and cake
3	☐ baking needs
4	☐ coupons
5	☐ juice-sat-cord-ms
6	☐ tea
7	☐ biscuits
8	☐ canned fish-meat
9	☐ canned fruit
10	☐ canned vegetables
11	☐ breakfast food
12	☐ cigs-tobacco pkts
13	☐ cigarette cartons

Remove

Selected attribute
Name: baby needs Type: Nominal
Missing: 4008 (87%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count
1	t	619

No class Visualize All

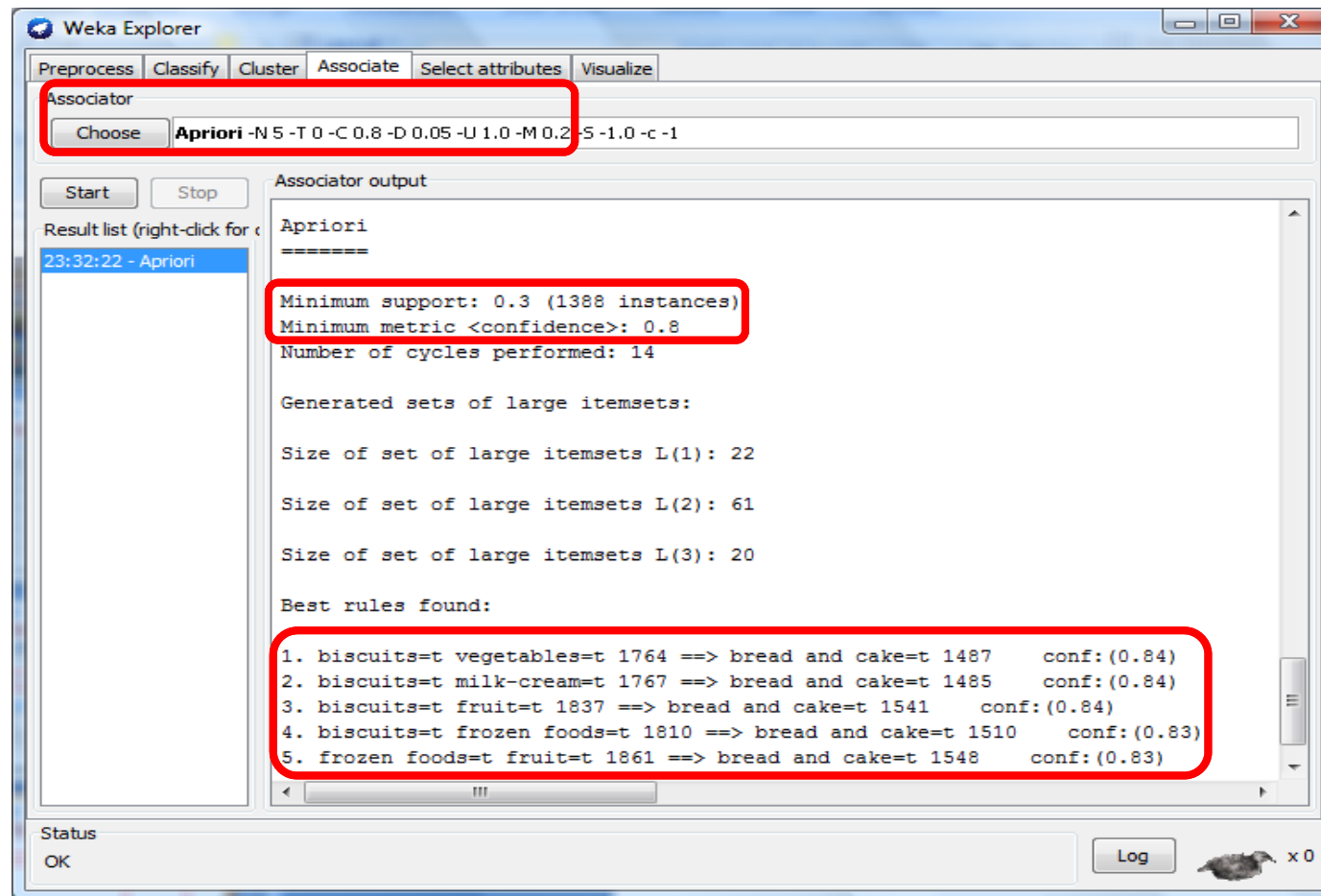
619

Status
OK Log x 0

WEKA – Data Mining Software in Java



Association Rules: Supermarket Purchases



<http://www.cs.waikato.ac.nz/~ml/weka/>

WEKA – Data Mining Software in Java



Association Rules: Supermarket Purchases

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Associator

Choose **Apriori** -N 5 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1

Start Stop

Result list (right-click for context menu)

- 23:32:22 - Apriori
- 23:34:51 - Apriori

Associator output

Apriori
=====

Minimum support: 0.5 (2314 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 2

Best rules found:

1. milk-cream=t 2939 ==> bread and cake=t 2337 conf: (0.8)
2. fruit=t 2962 ==> bread and cake=t 2325 conf: (0.78)
3. bread and cake=t 3330 ==> milk-cream=t 2337 conf: (0.7)
4. bread and cake=t 3330 ==> fruit=t 2325 conf: (0.7)

Status
OK

Log x 0

Thank You

Contact: xli@i2r.a-star.edu.sg if you have questions