

# An Introduction to Weka with Demos

Tomasz Oliwa

Ph.D., Computer Science, The University of Georgia, USA

Dipl.-Inform., University of Koblenz-Landau, Germany

Lecture in CSCI/ARTI 8950 Machine Learning

Slides at: <http://www.cs.uga.edu/~tomasz/weka2013fall/>

September 5, 2013



The University of Georgia

# Weka Software

The Weka machine learning/data mining suite:

- Rich collection of preprocessing, (un)/supervised learning and evaluation algorithms
- Accessible GUI
- Java, Free software (GPL):

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



# Attribute-Relation File Format (ARFF)

```
@relation weather
```

```
@attribute outlook sunny, overcast, rainy
```

```
@attribute temperature real
```

```
@attribute humidity real
```

```
@attribute windy TRUE, FALSE
```

```
@attribute play yes, no
```

```
@data
```

```
sunny,85,85,FALSE,no
```

```
sunny,80,90,TRUE,no
```

```
overcast,83,86,FALSE,yes
```

# Weka Modes

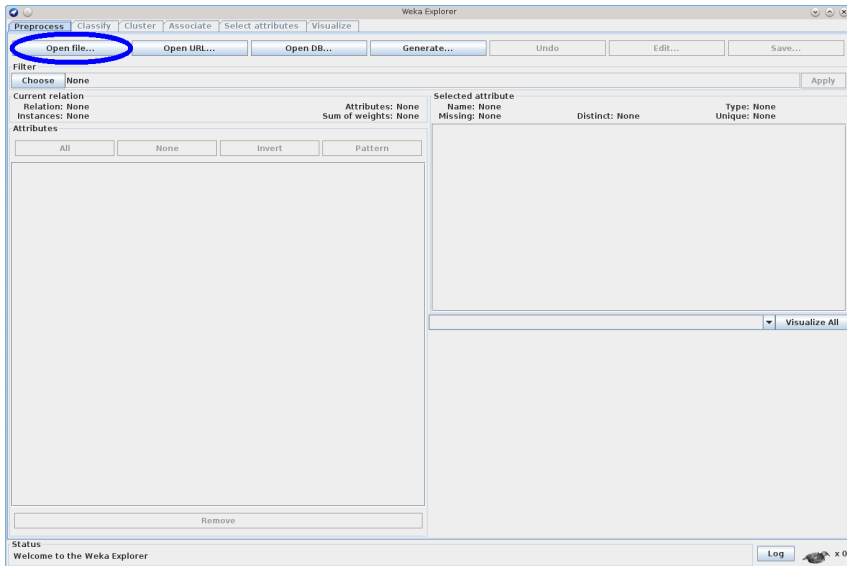
- 1 **Explorer:** Direct application and visualizations
- 2 **Experimenter:** Run & compare algorithms
- 3 **Knowledge Flow:** Visual composition of workflow
- 4 **Simple CLI:** Command line interface
- 5 **Source code:** Weka classes in Java code



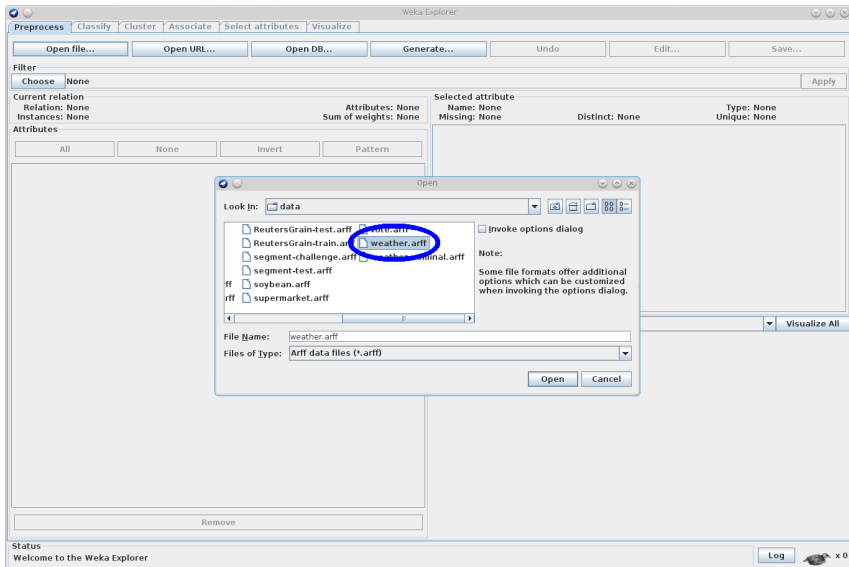
# Weka GUI Chooser



# Preprocessing



# Preprocessing - Load Data



# Preprocessing - Data Overview

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  
Instances: 14

Attributes: 5  
Sum of weights: 14

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

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Attributes

All None Invert Pattern

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

Remove

Status  
OK

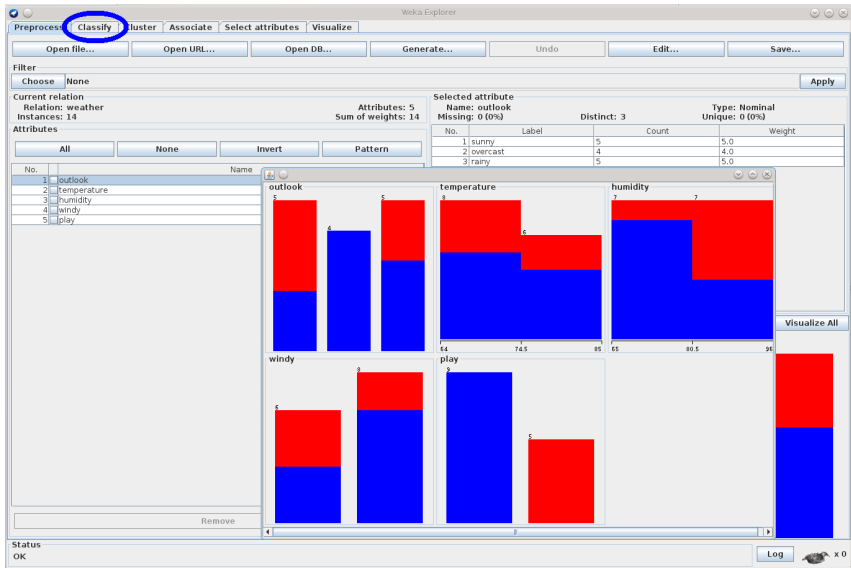
Class: play (Nom) Visualize All

Outlook	Blue (Count)	Red (Count)	Total (Count)
sunny	5	0	5
overcast	4	0	4
rainy	5	0	5

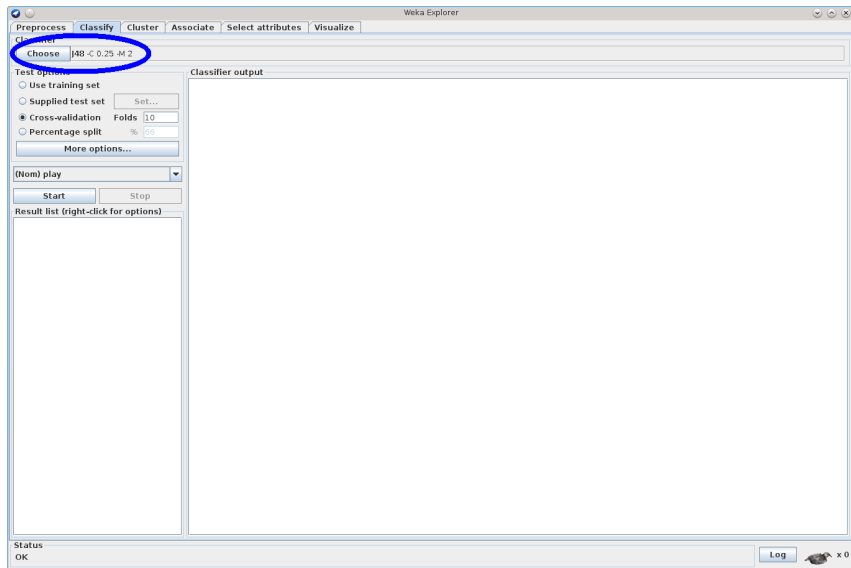
Log x 0



# Preprocessing - Visualize All



# Classification



# Classification Results

**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
- ☒ Cross-validation Folds 10
- ☐ Percentage split %

More options...

(Nom) play

Start

Result list (right-click for options)

02:15:16 - trees.J48

**Classifier output**

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

J48 pruned tree

```
outlook = sunny
| humidity <= 75: yes (2.0)
| humidity > 75: no (3.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
```

Number of Leaves : 5  
Size of the tree : 8

Time taken to build model: 0.01 seconds

==== Stratified cross-validation ====

==== SUMMARY ====

Correctly Classified Instances	9	64.2857 %
Incorrectly Classified Instances	5	35.7143 %
Kappa statistic	0.186	
Mean absolute error	0.2857	
Root mean squared error	0.4818	
Relative absolute error	60 %	
Root relative squared error	97.6586 %	
Coverage of cases (0.95 level)	92.8571 %	
Mean rel. region size (0.95 level)	64.2857 %	
Total Number of Instances	14	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.778	0.6	0.7	0.778	0.737	0.189	0.789	0.647	yes
	0.4	0.222	0.5	0.4	0.444	0.189	0.789	0.738	no
Weighted Avg.	0.643	0.465	0.629	0.643	0.632	0.189	0.789	0.808	

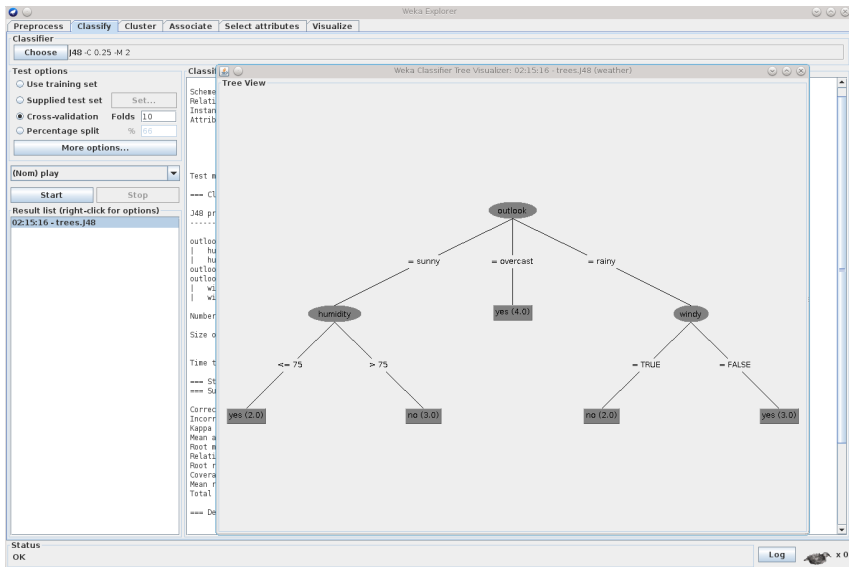
==== Confusion Matrix ====

```
a b <- classified as
7 2 | a = yes
3 2 | b = no
```

Status: OK

Log

# Classification Results - DT



# Clustering Results

Weka Explorer

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

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

Cluster mode

- ☒ Use training set
- ☐ Supplied test set
- ☐ Percentage split
- ☐ Classes to clusters evaluation

Test mode: evaluate on training data

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

02:34:35 - SimpleKMeans

Clusterer output

Relation: weather  
Instances: 14  
Attributes: outlook, temperature, humidity, windy, play

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

Number of iterations: 4  
Within cluster sum of squared errors: 6.804971358412714  
Missing values globally replaced with mean/mode

Cluster centroids

Attribute	Full Data (14)	Cluster# 0 (3)	Cluster# 1 (2)	Cluster# 2 (4)	Cluster# 3 (3)	Cluster# 4 (2)
outlook	sunny	rainy	overcast	sunny	sunny	overcast
temperature	73.5714	71	68	72	76.3333	82
humidity	81.6429	85.3333	77.5	86.5	75	80.5
windy	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
play	yes	yes	yes	no	yes	yes

Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

Cluster	Instances
0	3 ( 21%)
1	2 ( 14%)
2	4 ( 29%)
3	3 ( 21%)
4	2 ( 14%)

Status OK

Log x0

# Attribute Selection Results

The screenshot shows the Weka Explorer application window. The 'Select attributes' tab is active, and the 'CfsSubsetEval' evaluator is selected. The search method is 'BestFirst -D 1 -N 5'. The attribute selection output is displayed in the main pane, showing the selected attributes: 'outlook', 'temperature', 'humidity', 'windy', and 'play'. The 'play' attribute is circled in blue. The status bar at the bottom indicates 'OK'.

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator  
Choose CfsSubsetEval

Search Method  
Choose BestFirst -D 1 -N 5

Attribute Selection Mode  
☒ Use full training set  
☐ Cross-validation Folds: 10 Seed: 1

(Nom) play  
Start Stop

Result list (right-click for options)  
02:34:57 - BestFirst + CfsSubsetEval

Attribute selection output

```
==== Run information ====
Evaluator: weka.attributeSelection.CfsSubsetEval
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: weather
Instances: 14
Attributes: 5
    outlook
    temperature
    humidity
    windy
    play
Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ====

Search Method:
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 11
  Merit of best subset found: 0.196

Attribute Subset Evaluator (supervised, Class (nominal): 5 play):
  CFS Subset Evaluator
  Including locally predictive attributes

Selected attributes: 1.4 : 2
    outlook
    windy
```

Status  
OK

Log x 0

# Demos

Demo time:

- Preprocessing
- Classification
- Regression
- Clustering
- Feature Selection
- Experimenter
- Source code import

# Take Home Message

Weka is a powerful machine learning suite:

- **Explorer:** Intuitive suite with result visualizations
- **Experimenter:** Batch perform statistical tests on results
- **Source code:** Directly import Weka classes in your programs



# Questions

- Thanks for your attention!
- Do you have any questions?

# References and Tutorials

- Weka: <http://www.cs.waikato.ac.nz/~ml/weka/>
- Weka Wiki: <http://weka.wikispaces.com/>

## IBM Weka Tutorials:

- Introduction and Regression:  
<http://www.ibm.com/developerworks/opensource/library/os-weka1/index.html>
- Classification and clustering:  
<http://www.ibm.com/developerworks/opensource/library/os-weka2/index.html>
- Nearest Neighbor and server-side library:  
<http://www.ibm.com/developerworks/opensource/library/os-weka3/index.html>