Introduction to Weka
Lab 1 - SEng 474 / CSC 578D
Data Mining

Yudi Santoso

Credit: Cheng Chen, Maryam Shoaran

WEKA

(Waikato Environment for Knowledge Analysis)

- A software for Data Mining / Machine Learning.
- A collection of
  - machine learning algorithms,
  - data preprocessing tools,
  - small datasets.
- Written in Java.
- It is free! Available for Windows, Mac, and Linux.
  - So install on your own computer.

Weka supports the whole process of experimental data mining:

- Preprocessing the input data
- Statistically evaluating learning schemes
- Visualizing data and the results of learning

Fire Up “Weka” in the Lab
(using Windows 10)

Start > Weka 3.8.2 > Weka 3.8

Also check the documentation:
… > Weka 3.8.2 > Documentation

Weka 3.8.2

Explorer

Suitable for exploring features, and small datasets.

Experimenter

Comparing variety of learning techniques using statistical tests.

Knowledge Flow

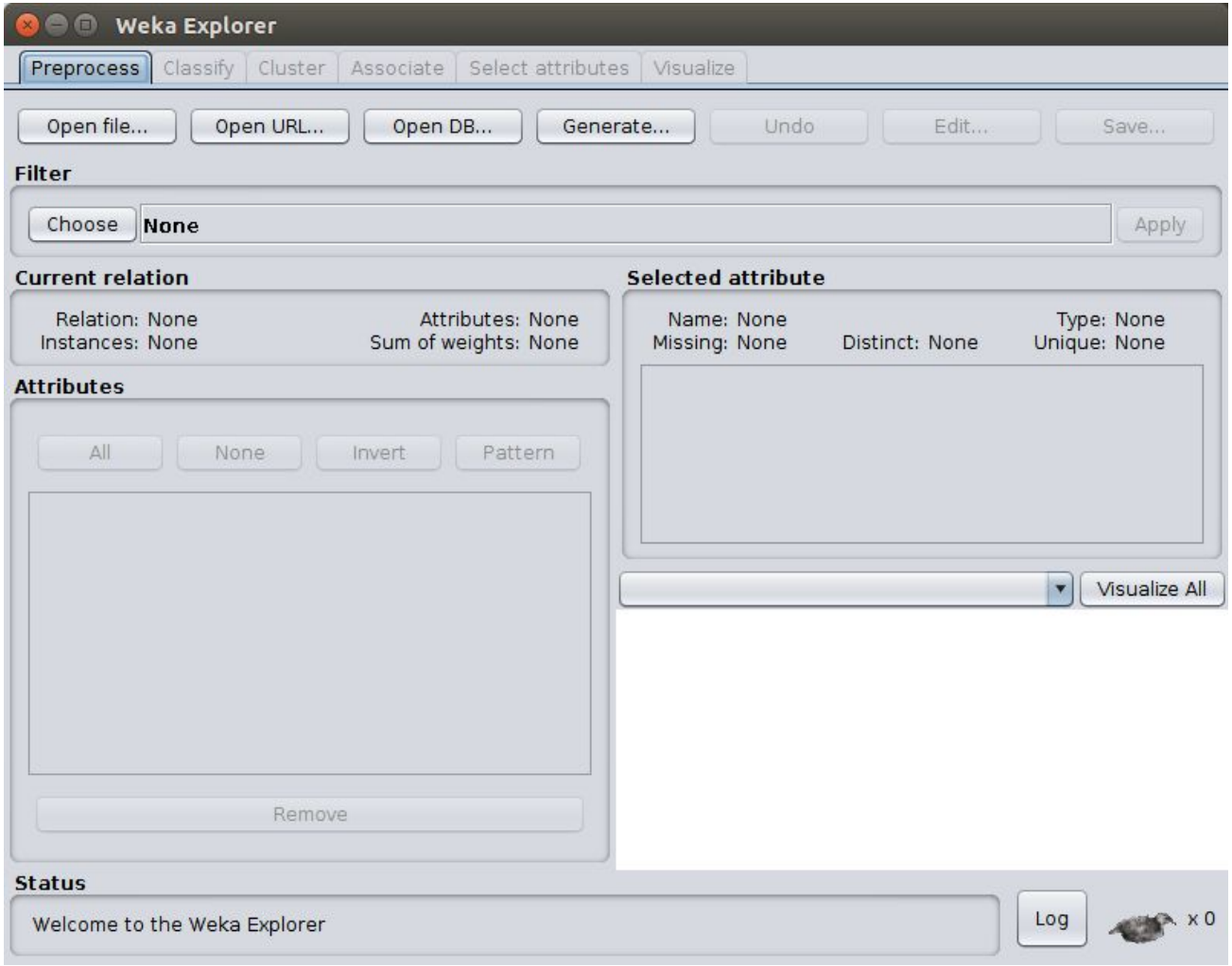
Graphical explorer, can deal with large datasets by incremental learning.

Workbench

All the tools together.

Simple CLI

Command line interface for direct execution of Weka commands (Java).



Weka Explorer

Weka Sample Datasets

Weka comes with a collection of small datasets.
They are in:

```
C:\Program Files\Weka-3-8\data\
```

Copy this folder to Desktop using File Explorer.
Then, in Weka Explorer (Preprocess tab):

Open file … > Look In > Engineering Home Drive (M:) > Desktop > data

Open the `weather.nominal.arff`

The ARFF data file format

```
% 1. Title: …
% 2. Sources: …
% …

@relation  <relation-name>

@attribute <attribute-name> <datatype>
@attribute …
@ …

@data
a1,b1,c1,...
a2,b2,c2,...
...
```

Easy to convert to CSV

ARFF = Attribute-Relation File Format

The <datatype>

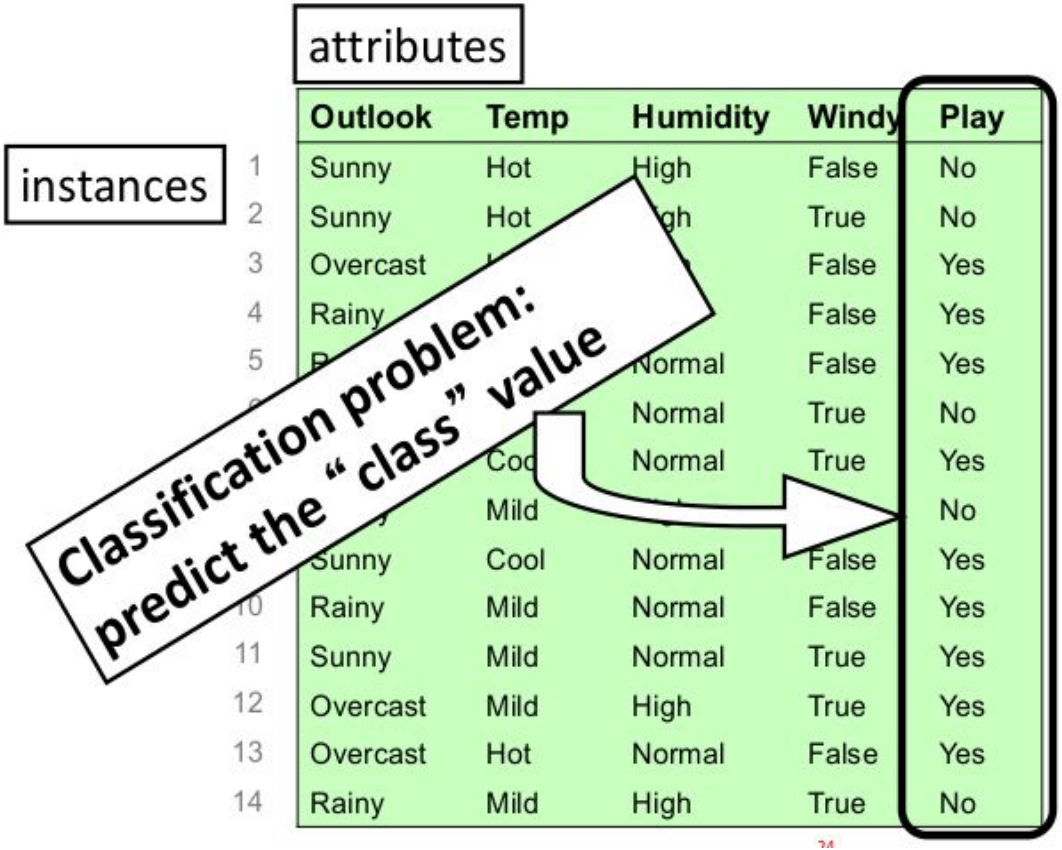
- numeric (also REAL, ..)
- <nominal-specification>
  - E.g., {blue, green, yellow}
- string
- date [<date-format>]
  - Default:
    “yyyy-MM-dd’T’HH:mm:ss”

Getting to know your data

- Relation name
- Number of attributes
- Number of instances
- Ranges of attribute values
- What is the class attribute?
- Histogram

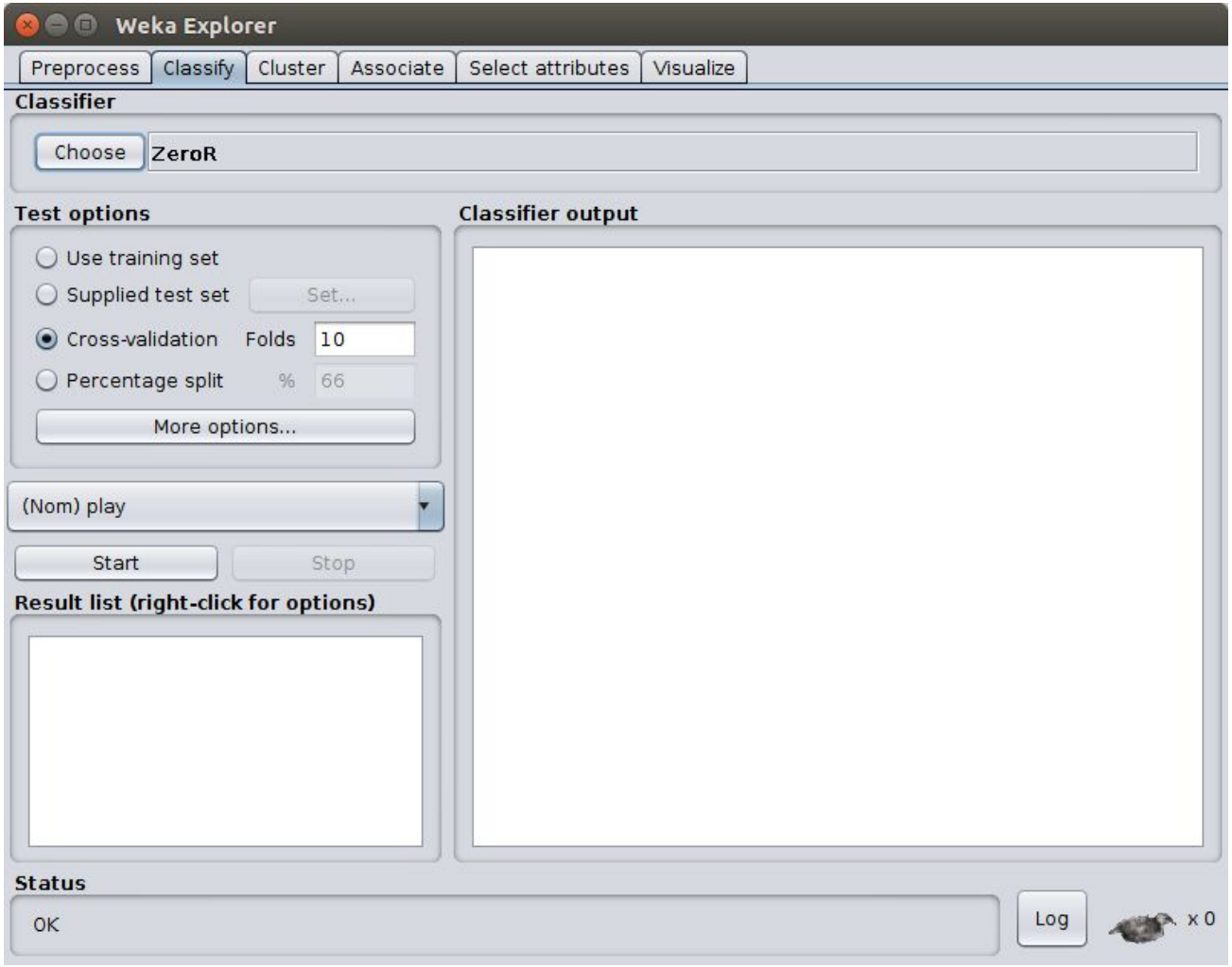
+ Open the arff file in an editor (notepad++ or WordPad) to get more info.

Classification Problem



	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	gh	True	No
3	Overcast			False	Yes
4	Rainy			False	Yes
5	R		Normal	False	Yes
			Normal	True	No
		Coo	Normal	True	Yes
		Mild			No
	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

Ian Witten: Data Mining with Weka



Weka Classify

Using Classifier

In Weka - Classify:

> Choose > trees > J48

J48 -C 0.25 -M 2

➔ Click to edit properties

- More
  - Info about this classifier
- Capabilities
  - Info on usage (datatype)

Using Classifier

J48 > More

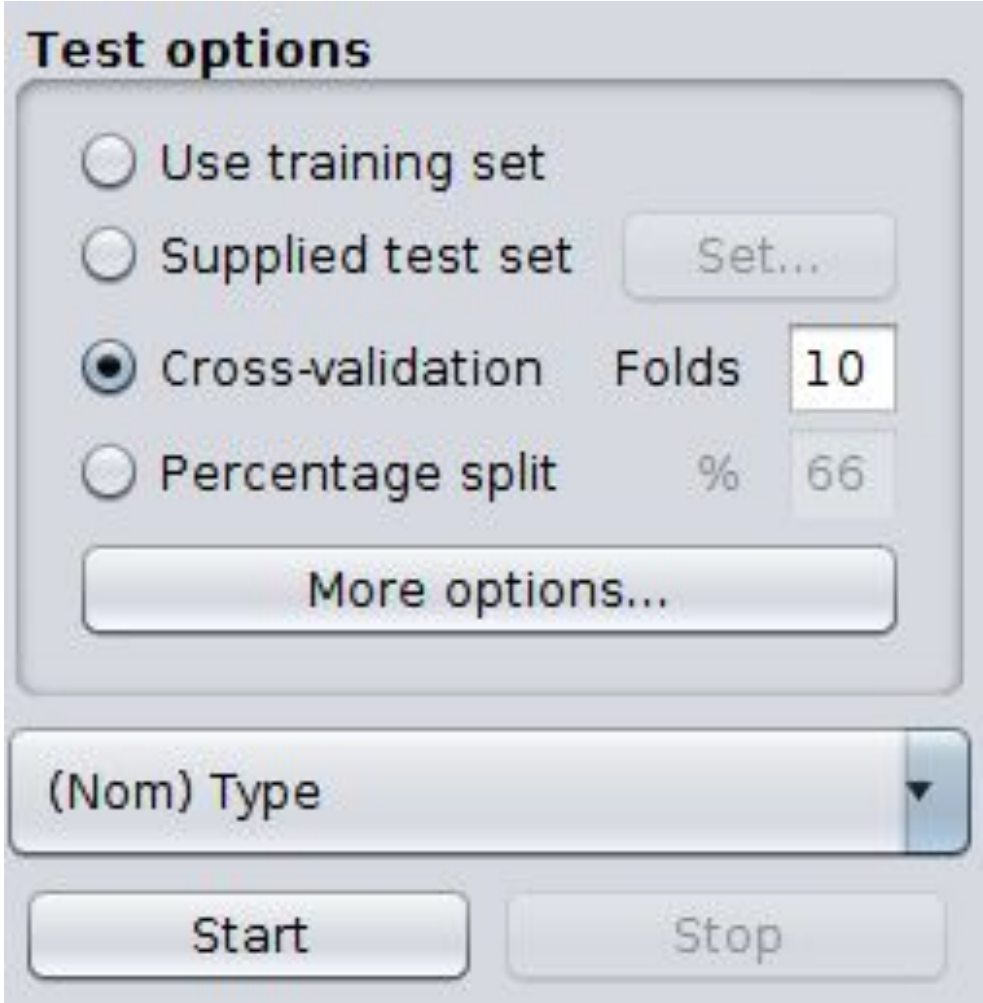
➔ Info about this classifier:

NAME
weka.classifiers.trees.J48

SYNOPSIS
Class for generating a pruned or unpruned C4.5 decision tree. For more information, see

Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Using Classifier



Classifier Output – Model

```
=== 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

Size of the tree :  8
```

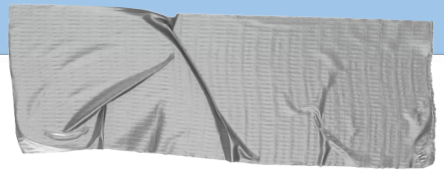
Classifier Output – Summary

```
=== Summary ===

Correctly Classified Instances           7               50      %
Incorrectly Classified Instances         7               50      %
Kappa statistic                         -0.0426
Mean absolute error                      0.4167
Root mean squared error                  0.5984
Relative absolute error                 87.5    %
Root relative squared error            121.2987 %
Total Number of Instances               14
```

```
=== Confusion Matrix ===

 a b   <-- classified as
 5 4 | a = yes
 3 2 | b = no
```



Closing

- **In this lab:**

  We have learn about Weka

- **We have seen that**

  Weka is a very useful tool for data mining.

- **What's next?**

  Experiment more!