WEKA

Prepared by:
Mohd Shamrie Sainin
http://staf.uum.edu.my/shamrie
AISIG
Dept. of Computer Science,
Faculty of Information Technology
UUM

Content

- Introduction
- Data Preparation
- Using Weka

Introduction

- Weka is an open source software that developed by Universiti Waikato New Zealand. Weka means "Waikato Environment for Knowledge Analysis".
- Weka is a tool for knowledge discovery and uses algorithms from Machine Learning and entirely developed using Java
- Some Weka function are:
  - Implementation of "state-of-the-art learning" through command line execution of the dataset
  - Tools for data preprocessing including transformation such as discretization, normalization, replacing missing values and etc.
  - Data analysis through implementation of various algorithms, testing and visualization.

Data preparation

- Data is commonly saved as databases or in spreadsheet.
- However, Weka need an input data in the format of ARFF.
- Why ARFF?
  - According to Prof. Ian H Witten, he is actually not sure why but can be understood as ‘attribute file format’
  - Therefore, ,an input file to Weka is in the extension of (*.arff)

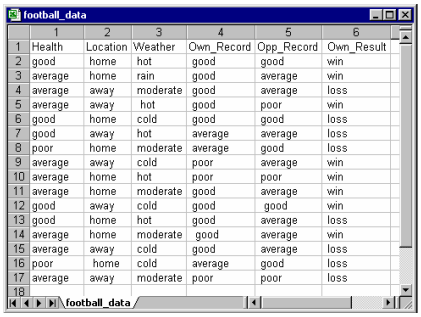
Data preparation

- Before we can use the data using any algorithm, must convert to ARFF.
- Most of spreadsheet and database program (e.g. Access/ Excel) allow us to convert or export data into a file in comma-separated format.
- In this example, we will learn how to spreadsheet and word to convert data into ARFF format.

Data from database: football_data

Health	Location	Weather	Own_Record	Opp_Record	Own_Result
good	home	hot	good	good	win
average	home	rain	good	average	win
average	away	moderate	good	average	loss
average	away	hot	good	poor	win
good	home	cold	good	good	loss
good	away	hot	average	average	loss
poor	home	moderate	average	good	loss
average	away	cold	poor	average	win
average	home	hot	poor	poor	win
average	home	moderate	good	average	win
good	away	cold	good	good	win
good	home	hot	good	average	loss
average	home	moderate	good	average	win
average	away	cold	good	average	loss
poor	home	cold	average	good	loss
average	away	moderate	poor	poor	loss

Data into Excel: football_data



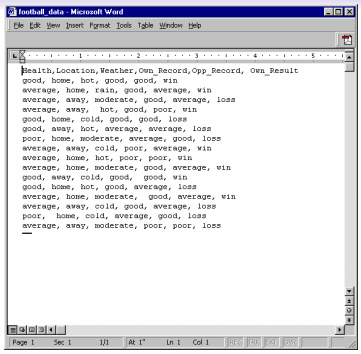
football_data

	1	2	3	4	5	6
1	Health	Location	Weather	Own_Record	Opp_Record	Own_Result
2	good	home	hot	good	good	win
3	average	home	rain	good	average	win
4	average	away	moderate	good	average	loss
5	average	away	hot	good	poor	win
6	good	home	cold	good	good	loss
7	good	away	hot	average	average	loss
8	poor	home	moderate	average	good	loss
9	average	away	cold	poor	average	win
10	average	home	hot	poor	poor	win
11	average	home	moderate	good	average	win
12	good	away	cold	good	good	win
13	good	home	hot	good	average	loss
14	average	home	moderate	good	average	win
15	average	away	cold	good	average	loss
16	poor	home	cold	average	good	loss
17	average	away	moderate	poor	poor	loss
18						

Data preparation

- Save Excel file as a CSV (comma separated) file as '**football_data.csv**'
- Close Excel and open 'football_data.csv' using Microsoft Word.

Data in Word



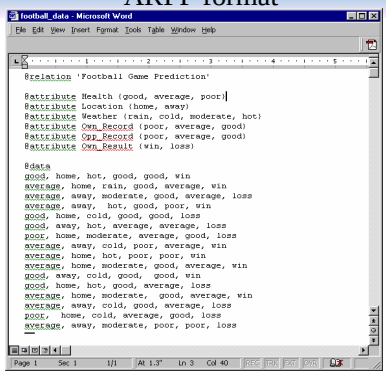
football_data - Microsoft Word

```
Health,Location,Weather,Own_Record,Opp_Record, Own_Result
good, home, hot, good, good, win
average, home, rain, good, average, win
average, away, moderate, good, average, loss
average, away,  hot, good, poor, win
good, home, cold, good, good, loss
good, away, hot, average, average, loss
poor, home, moderate, average, good, loss
average, away, cold, poor, average, win
average, home, hot, poor, poor, win
average, home, moderate, good, average, win
good, away, cold, good,  good, win
good, home, hot, good, average, loss
average, home, moderate,  good, average, win
average, away, cold, good, average, loss
poor,  home, cold, average, good, loss
average, away, moderate, poor, poor, loss
```


Format

- The heading for the ARFF format must include:
  - @relation 'TITLE'
  - @attribute ATT_NAME {VALUES}, or
  - @attribute ATT_NAME real, and
  - @data
- TITLE is the title of the data
- ATT_NAME is attribute or column name
- VALUES is possible values per attribute (Nominal)
- Real is for numeric or continuous data
- @data is a section for records/data

ARFF format



football_data - Microsoft Word

```
@relation 'Football Game Prediction'

@attribute Health {good, average, poor}
@attribute Location {home, away}
@attribute Weather {rain, cold, moderate, hot}
@attribute Own_Record {poor, average, good}
@attribute Opp_Record {poor, average, good}
@attribute Own_Result {win, loss}

@data
good, home, hot, good, good, win
average, home, rain, good, average, win
average, away, moderate, good, average, loss
average, away,  hot, good, poor, win
good, home, cold, good, good, loss
good, away, hot, average, average, loss
poor, home, moderate, average, good, loss
average, away, cold, poor, average, win
average, home, hot, poor, poor, win
average, home, moderate, good, average, win
good, away, cold, good,  good, win
good, home, hot, good, average, loss
average, home, moderate,  good, average, win
average, away, cold, good, average, loss
poor,  home, cold, average, good, loss
average, away, moderate, poor, poor, loss
```

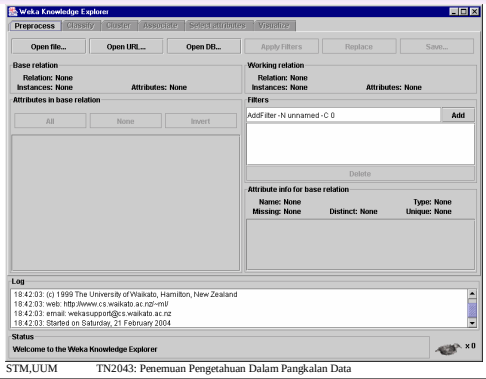

Getting distinct values from attribute

- Use database (Access) to return distinct values for certain field using SQL, or
- Use simple programming to output distinct values.
- Manually observe distinct values.

Using Weka

- From Start menu, find menu for weka-3-2 and click to launch Weka.
- The following is a main GUI for Weka. Choose 'Explorer' to start Weka application.

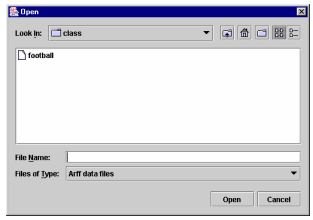
Explorer GUI



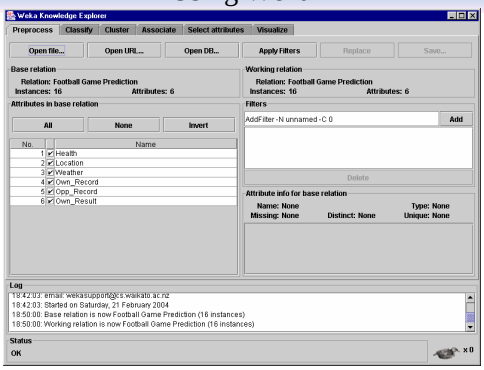
Getting started

- From that frame, the first tab is for preprocessing‘:
  - Data file input
  - Data file information
  - Data preprocessing (Filters)
- To start using Weka algorithm, open ‘football_data.arff’ and all information about football_data will be displayed.

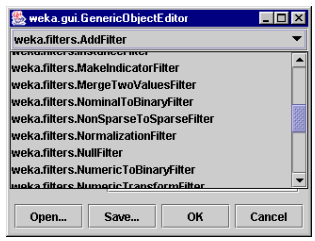
Using Weka



Using Weka

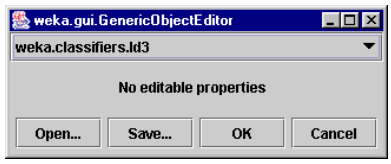


Preprocessing / Filters



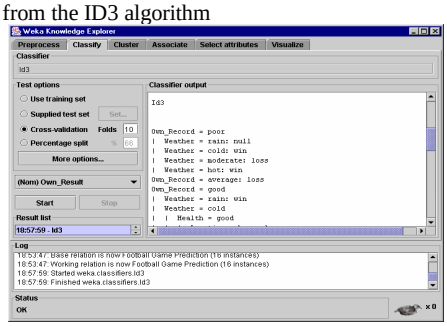
Using Weka: Classifier

- Next, select 'Classify' tab and click the box below the Classifier, and the following window will be displayed. Select 'weka.classifier.id3':



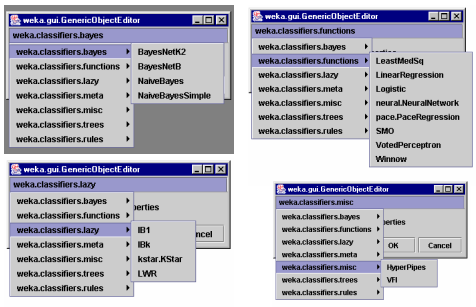
Using Weka

- Next, click Start Button to generate the output from the ID3 algorithm

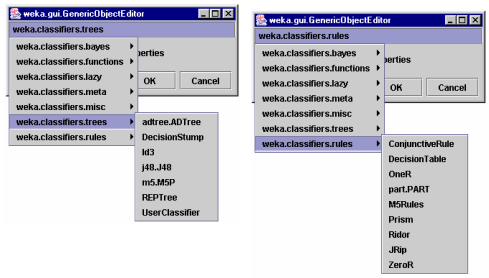


Other Functions

- Version 3.3



Using Weka



End

Prepared: Mohd Shamrie Sainin

TN2043: Penemuan Pengetahuan Dalam Pangkalan Data.

Rujukan: Witten, I.H., Frank, E. 1999. Data Mining: Practical Machine Learning Tools and Techiques with Java Implementations.