Weka
-
Decision Trees
Weka

• **Waikato Environment for Knowledge Analysis**
• Is a collection of advanced
  – machine learning algorithms and
  – data preprocessing tools.

• Available at [http://www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)
• Book: *Data Mining, Practical Machine Learning Tools and Techniques*
  (2nd Ed)(Part II)
  (3rd Ed)(part III)
Data Mining Process

1. Problem Definition in terms of Data Mining and Business Goals
2. Data Acquisition and preparation (preprocess)
3. Building Models
4. Evaluation of Models
5. Deployment
Data Mining Process (cont’d)

• No single machine learning scheme is appropriate to all data mining problems.
  – Real data sets vary significantly.
  – Learning algorithms must match the structure of the domain.
  – Data mining is an experimental science.

• Weka supports the whole process of experimental data mining.
  – Preparing the input data (e.g. discretization, dealing with missing values, and so on)
  – Statistically evaluating learning schemes
  – Visualizing the input data and result of learning
Weka GUIs

• **Explorer**
  suitable for small data files, it loads the whole data into main memory.

• **Knowledge Flow**
  for very large datasets

• **Experimenter**
  enables Weka users to compare *automatically* a variety of learning techniques

• **Command Line Interface**
Explorer GUI
Methods in Weka

– Data preprocessing tools
– Classification
– Clustering
– Association rule mining
– Attribute selection
– Visualization

6 main tabs at the top of Explorer window for basic operations, all of them inactive at the beginning except preprocess
Preparing Data

• Weka accepts data in .arff File Format.
• An ARFF file includes independent, unordered instances and do not involve relationships among instances.
• Data from spreadsheets or databases have to be transformed to ARFF format.

1. Export data to a CSV file (Comma-Separated-Value) then
2. Convert to ARFF by adding keywords and saving as a raw text file in a text editor.
Preparing Data (cont’d)

• ARFF data file is a text file which must have the following format:

  @relation  dataset’s Name

  @attribute  att1Name  {v1,v2,...}
  @attribute  att2Name  numeric
  @attribute  att3Name  string

  ...
  @attribute  classAttName?

  @data
  \( a_1, a_2, a_3, ... \)
  \( b_1, b_2, b_3, ... \)
  ...

.arff Data File

```%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

- Some sample data files are provided by Weka (Weka-3-6\data).
- The Weka Explorer interface is also launched automatically when you double-click on an .arff file.
ARFF Data File Format

• Lines starting with % are comments.
• Each record (data point) is in one line, with attribute values separated by comma.
• If a value is missing, it is represented by a single question mark.
• Attribute types can be: nominal, numeric, string, date.
• Date values has to be in the format
  – yyyy-MM-ddTHH:mm:ss (2004-04-03T12:00:00)
• Strings are usually converted into a numeric form such as a word vector.
• Numeric attributes are usually normalized (standardized) to lie in one specific range, e.g. 0 and 1.
Open a data file

• In “Preprocess” panel, click “Open file” button, and choose an ARFF file from “data” folder. Weka-3-6/data
Now the other tabs are active.

   – If you specify a “CSV” file, it will be automatically converted into ARFF file.

   Select for example weather.nominal.arff .
Preprocess Panel
Preprocess Panel (cont’d)

• **Current relation** pane: specification of data file

• **Attributes** and **Remove** button to possibly remove some attributes from the experiment.

• **Statistics** about selected attribute (highlighted attribute.)
  – If you select a numeric attribute, *Min, Max, Mean*, and *standard deviation* are shown.

• **Histogram** shows the distribution of the class as the function of the selected attribute.

• **Edit** button to edit input data on a separate window
  – In the Edit window, right click on the caption of column or values opens a list of various available editions on data.
  –Undo, and Save buttons have the known function
Build a Decision Tree

- Switch to “Classify” tab
- Select “J48” algorithm (an implementation of C4.5) by
  - Clicking “Choose” button
  - Selecting classifiers >> trees >> J48 from the Weka tree
- Invoke classifier by clicking “start” button
- Clicking the line in front of the “choose” button, opens classifier’s Object Editor, in which any parameter can be changed.
- Weka keeps the results of different classifiers in the “Result List” pane.
Classifier
Object Editor
Decision Tree
Classifier Output

Classifier output includes:

• Summary of the data set
• 10-fold cross validation is the default “test (evaluation) mode”
• A pruned decision tree in textual format
• A colon (:) introduces the class label assigned to the leaf, followed by number of instances reached that leaf
• Number of leaves and nodes in the decision tree
• Estimates of the tree’s predictive performance
• Confusion matrix at the end
• Some other statistics