Machine Learning with WEKA

Dieter Merkl

Die meisten Folien basieren auf einem Foliensatz von Eibe Frank,
The University of Waikato
http://prdownloads.sourceforge.net/weka/weka.ppt
WEKA: the bird

Copyright: Martin Kramer (mkramer@wxs.nl)
WEKA: the bird

- Latin name: Gallirallus australis
- Subspecies: North Island, Western, Stewart Island, Buff
- Category: Endemic (only found in New Zealand)
- Status: Threatened / Common - rare in the North Island
- Size: 53cm, Weight: 1000/700 g
- Characteristics: Flightless, rail, fast runner
- Habitat: Forest, scrub
- Nests: On the ground in short burrows, hollow logs, under plants
- Food: Worms, snails, insects, seeds, fruit, eggs, small birds, lizards, rats, mice
- Lifespan: 15 years
WEKA: the software

• Machine learning/data mining software written in Java (distributed under the GNU Public License)
• Used for research, education, and applications
• Complements “Data Mining” by Witten & Frank
WEKA: the software

• Main features:
  – Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  – Graphical user interfaces (incl. data visualization)
  – Environment for comparing learning algorithms
WEKA: versions

• There are several versions of WEKA:
  – WEKA 3.2: “GUI version” adds graphical user interfaces (book version is command-line only)
  – WEKA 3.3: “development version” with lots of improvements

• These slides are based on the current version WEKA 3.4.10
WEKA: download

Software

Projects • Software • Book • Publications • People • Related

Snapshots
Every night a snapshot of the CVS is taken, compiled and put together in ZIP files. For those who want to have the latest bugfixes, they can download these snapshots here.

Book version
Weka 3.4 is the latest stable version of Weka, and the one described in the data mining book. There are different options for downloading and installing it on your system:

- Windows
   - Click here to download a self-extracting executable that includes Java VM 1.4 (weka-3-4-10-re.exe; 24,333,382 bytes)
   - Click here to download a self-extracting executable without the Java VM (weka-3-4-10.exe; 10,301,908 bytes)
   - These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.4 (or later) on your system.

- Other platforms (Linux, etc.)
   - Click here to download a zip archive containing Weka (weka-3-4-10.zip; 10,392,250 bytes)
   - First unzip the zip file. This will create a new directory called weka-3-4-10. To run Weka, change into that directory and type:
     
     java -jar weka.jar
     
     Note that Java needs to be installed on your system for this to work. Also note, that using -jar will override your current CLASSPATH variable and only use the weka.jar.
WEKA only deals with “flat” files

@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
WEKA: the interface
WEKA: the interface
Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …
Explorer
Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, …
- “Meta”-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, …
Explorer: PART on Iris Dataset
Explorer: output
Explorer: MLP on Iris Dataset
Explorer: clustering data

- WEKA contains “clusterers” for finding groups of similar instances in a dataset.
- Implemented schemes are:
  - $k$-Means, EM, Cobweb, $X$-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given).
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution.
Explorer: k-Means on Iris Dataset
Explorer: attribute selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - An evaluation method: correlation-based, wrapper, information gain, chi-squared, …
- Very flexible: WEKA allows (almost) arbitrary combinations of these two
Explorer: Information Gain on Iris Dataset

Attribute Evaluator:
- Choose InfoGainAttributeEval

Search Method:
- Choose Ranker -T -1.7976931348623157E308 -N -1

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

(Nom) class

Attribute selection output:
- sepalwidth
- petallength
- petalwidth
- class

Evaluation mode: evaluate on all training data

Attribute selection:

- Attribute: sepalwidth
  - Information Gain: 1.418
  - Ranked: 3rd
- Attribute: petallength
  - Information Gain: 1.418
  - Ranked: 3rd
- Attribute: petalwidth
  - Information Gain: 0.698
  - Ranked: 2nd
- Attribute: sepalwidth
  - Information Gain: 0.396
  - Ranked: 1st

Selected attributes: 3,4,1,2 : 4
Explorer: attribute selection on Iris Dataset
Explorer: MLP on reduced Iris Dataset
Experimenter: performing experiments

- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing built in!
Experimenter
Conclusions: give it a go!

• WEKA is available at

http://www.cs.waikato.ac.nz/ml/weka/