Machine Learning & Data Mining

Introduction to Weka

Rafael Ramirez
rafael.ramirez@upf.edu
55.316

What is WEKA?

- Waikato Environment for Knowledge Analysis
  - It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand.
  - Weka is also a bird found only in New Zealand.

Main Features

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

Main GUI

- Three graphical user interfaces:
- The Explorer” (exploratory data analysis)
- “The Experimenter” (experimental environment)
- “The Knowledge Flow” (new process model inspired interface)

Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV
- Data can also be read from a URL or from an SQL database
- Pre-processing tools in WEKA are called "filters"
- WEKA contains filters for: Discretization, normalization, resampling, attribute selection, transforming and combining attributes

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
...
```

Numeric Attribute
Nominal Attribute

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
- …

TEST OPTIONS

- **Use training set:**
  train and classify with all the data
- **Supplied test set:**
  train with a set of data and classify with a provided new set
- **Cross-validation:**
  splits the data in n folds, train with n-1 folds and test with the other, then switch the fold and repeat for each one.
- **Percentage split:**
  It defines a percentage that will build the classifier and the remaining part will be tested

Explorer: clustering data

- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:

  k-Means, EM, …
- Clusters can be visualized and compared to “true” clusters (if given)

Explorer: finding associations

- WEKA contains an implementation of the Apriori algorithm for learning association rules
- Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
- milk, butter $\Rightarrow$ bread, eggs
- Apriori can compute all rules that have a given minimum support and exceed a given confidence

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: data visualization

- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
- To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- "Jitter" option to deal with nominal attributes
- (and to detect "hidden" data points)
- "Zoom-in" function