Comparing Weka and R

Overview

Purpose and Audience

This document is intended for users in the process of evaluating and comparing data mining offerings from Pentaho with the R statistical software suite. Wherever possible, this document attempts to use publicly verifiable information and respected third-party sources.

Functional Comparison

Introduction

Weka and R are two prominent open-source software systems for analytics. Both originate from academia, but have different goals and focus. While R comes from the statistics community and is a general-purpose environment for statistical analysis, Weka's origin is in computer science, and, as such, was designed specifically for machine learning and data mining. When choosing analytical software, you need to carefully consider the goals for data mining within your organization, including potential deployment of predictive models. Pentaho Data Mining, based on Weka, is 100% Java, facilitating simple integration and deployment within a Pentaho BI solution.

Solution Breadth

Weka provides a broad selection of data mining and machine learning techniques, more so than R does. R is a general purpose statistical environment and has facilities that you would not necessarily expect to see in a data mining tool. Weka is arguably more user friendly, with familiar point-and-click graphical user interfaces, while R is driven by what is essentially a functional programming language.

Data Import

Feature	Weka	R
Text file – delimited	√	√
ARFF	√	√
C4.5 format	√	
Database	√	√
SAS file		√
SPSS file		√
Minitab		√

Data Exploration/Visualization

Feature	Weka	R
Descriptive statistics	√	√
Frequency table	√	√
Scatter plot	√	√
Scatter plot matrices	√	√
Histograms	√	√
Tree/Graph visualization	√	√
Boxplots		√
ROC curve	√	√
Precision/recall curve	√	√
Lift chart	√	√
Cost curve	√	√

Data Preparation

Feature	Weka	R
Sampling		
Oversampling/balancing	√	√
Random	√	√
Stratified	√	√
Discretization (binning)		
Equal width	√	√
Equal frequency	√	√
Supervised	√	
Reorder fields	√	√
Identifier fields	√	√
Normalization/standardization	√	√
Binarization	√	√
Derived fields	√	√
Outlier detection	√	√
Principal components	√	√
Random projections	√	√
Attribute selection	√	
Arbitrary kernels	√	

Modelling

Feature	Weka	R
Bayesian		
Naïve Bayes	√	
Naïve Bayes multinomial	√	
Complement naïve Bayes	√	
Averaged one-dependence estimators	√	
Weigted averaged one-dependence estimators	√	
Bayes nets	√	
Naïve Bayes trees	√	
Bayesian additive regression trees		√
Lazy Bayesian rules	√	
Functions		
Linear regression	√	√
Logistic regression	√	√
Isotonic regression	√	√
Least median squares regression	√	√
Pace regression	√	
Support vector machines	√	√ (via interface to third party app)
Multilayer perceptron (neural net)	√	√ (single hidden layer NN)
Radial basis function network	√	
Gaussian processes	√	√
Voted perceptron	√	
Lazy		
K-nearest neighbors	√	√
Locally weighted learning	√	
Trees		
ID3	√	
C4.5	√	
CART	√	√
Decision stumps	√	√
Random forests	√	√
Best first tree	√	
Logistic model trees	√	
M5 model tree	√	
Alternating decision trees	√	
Interactive tree construction	√	
KNN trees		√
Rules		
Decision table	√	
RIPPER	√	
Conjunctive rule	√	

Feature	Weka	R
M5 Rules	√	
PART	√	
Ripple down rules (Ridor)	√	
NNge	√	
OneR	√	
Ensmeble learning		
AdaBoost	√	√
LogitBoost	√	√
Additive regression	√	√
Bagging	√	√
Stacking	√	
Dagging	√	
Grading	√	
MultiBoost	√	
Voted classifier	√	
MetaCost	√	
Ensembles of nested dichotomies	√	
Multi instance learning methods	√	
Clustering		
EM	√	√
KMeans	√	√
XMeans	√	
COBWEB (hierarchical)	√	
OPTICS	√	
Farthest first clustering	√	
Hierarchical clustering		√
Agglomerative nesting		√
Fuzzy C-means clustering		√
Bagged clustering		√
Cluster ensembles		√
Convex clustering		√
Association rules		
Apriori	√	√ (via interface to third pary app)
Predictive Apriori	√	
Tertius	√	
Generalized sequential patterns	√	
Eclat		√ (via interface to third party app)

Evaluation

Feature	Weka	R
Prediction accuracy	√	√
Confusion matrix	√	√
AUC	√	√
Information-retrieval stats	√	√
Information-theoretic stats	√	
ROC / lift charts	√	√
Experiment facility	√	

Deployment

Feature	Weka	R
Serialized java object	√	
Java source code (limited)	√	
PMML (limited)		√