Informatics Research and Development Unit
Public Health Informatics & Technology Program Office
Office of Surveillance, Epidemiology, and Laboratory Services
US Centers for Disease Control and Prevention

2010 Report:

Open Source Data Mining Software Evaluation

Background

This evaluation focused on general purpose open source data mining software products. Data mining tools which were developed for a specific domain (e.g. text mining, image mining, microarray data mining, etc.) were not included in this evaluation. The data mining software products in this evaluation include: RapidMiner, Weka, Orange, Rattle, and Knime.

Given the broad potential audience for these products in the domain of public health, only those functions / features available through a graphical user interface (GUI) were evaluated. Specifically, any functions of these tools which required scripting or programming were not considered. For this reason, an additional popular open source data mining tool, known as R, was not included in this evaluation. Please note that, R provides many powerful statistical analyses and data mining functions.

Results

This summary report presents our finding from three perspectives: *general information, system features, and data mining functionality.*

Table 1. General information of selected data mining software products

	RapidMiner (YALE)	Weka	Orange	Rattle	KNIME
Edition/ Version	Community edition, version 5.0	version 3.6.2	version 2.0	version 2.5.21	Desktop edition, version 2.1.2
Company/	Rapid-I	University of	University	Togaware	KNIME.com

Organization (Country)	(Germany)	Waikato (New Zealand) / Pentaho	of Ljubljana (Slovenia)	(Australia)	GmbH (Switzerland)
Website	http://rapid-i.com/	http://www.cs.waikato.ac.nz/ml/weka/	http://www.ailab.si/orange/	http://rattle.togaware.com/	http://www.knime.org/
License	OSS, GNU AFFERO GPL, version 3	OSS, GNU GPL, version2	OSS, GNU GPL version 3	OSS, GNU GPL version 2	OSS, GNU GPL version 3
Cost	Free	Free	Free	Free	Free

Table 2. System Features

Features*	RapidMiner (YALE)	Weka	Orange	Rattle	KNIME
OS platform	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux
Documentation	4	5	3	4	5
Easy-to-learn	3	4	5	4	4
Usability	5	4	5	5	4
Support	5	5	3	2	3
Extensibility	5	5	3	3	5
Reliability	5	5	5	5	5
Installation	5	5	5	4	5
Data IO / Preprocessing	5	3	3	3	5
Data Visualization	4	3	5	5	3

*: The system feature evaluation was based on a 5-point scale, with higher scores indicating better results such as high/comprehensive/easy/simple, and lower scores for negative results such as low/none/difficult/complex.

Table 3. Data Mining Functionality

Functionality	RapidMiner (YALE)	Weka	Orange	Rattle	KNIME
Bayes Network	yes	yes	yes	no	yes
Decision Tree	yes	yes	yes	yes	yes
Neural network	yes	yes	no	no	yes
SVM	yes	yes	yes	yes	yes
Feature Selection	yes	yes	no	no	yes
Clustering	yes	yes	yes	yes	yes
Association Rules	yes	yes	yes	yes	yes
Model Information	yes	yes	yes	yes	yes
Evaluation	yes	yes	yes	yes	yes

Summary

Weka, RapidMiner and KNIME are developed in the Java software language. Rattle is a fully R-based application, and Orange is integrated with Python. Among all open-source data mining tools, Weka and RapidMiner have the biggest and most active user communities. Both of them quickly implement (and integrate) new and emerging machine learning algorithms into their systems.

Weka, as one of the best-known open-source data mining software tools, has an impressive array of data mining components, which have, in fact, been integrated into many other data mining tools including RapidMiner, Rattle, and KNIME. Weka consists of four major applications: Explorer (for exploring data), Experimenter (for performing experiments and conducting statistical tests between learning schemes), KnowledgeFlow (for incremental learning), and SimpleCLI (a command-line interface to allow direct execution of WEKA commands). For beginners, it is best for them to start with Weka Explorer, as it provides a relatively simple and easy-to-learn interface to access Weka data mining components.

RapiderMiner (formerly YALE) is built on top of Weka, and includes additional powerful data analysis functions such as data preprocessing, visualization, and additional machine learning algorithms. In addition, its user interface is more intuitive than Weka KnowledgeFlow. Users with limited knowledge in computer science and programming may find RapidMiner's learning curve to be substantial.

KNIME has one of the best built-in on-line support features, which is very helpful for new users who are in the process of building their data mining workflows. KNIME also supports running R and Python scripts. Another nice feature of KNIME is its integration of the Chemistry Development Kit with additional nodes for the processing of chemical structures, compounds, etc.

Rattle provides a graphical user interface (GUI) specifically for data mining using R. Although an understanding of R is not required to begin using Rattle for basic data mining functions, Rattle is particularly suited for users familiar with R. In addition, Rattle integrates two sophisticated tools for interactive graphical data analysis: GGobi and Latticist.

Orange has a very simple and intuitive graphical interface (GUI) for users with limited knowledge in data mining. Compared to the other data mining tools, its strength is its interactive visualization function, which enables users to set visualization parameters and choose data points or nodes directly from a graph.

Recommendations

As expected, each of the five open source data mining tools evaluated, has its unique set of pros and cons. The recommended open source data mining tool to be selected depends on the goals of your specific project, as well as you or your team's background knowledge in data mining and programming. The recommendations below are based on our personal experience, and thus, represent our own opinions.

User Characteristics	Recommendation
A tool for those users who want to develop a new software package for their specific data mining purposes. The learning curve is not very steep, the tool is quite powerful, and it has solid community support.	**Weka**
A tool for users who are willing to spend a significant amount of time to learn the art and science of data mining. In other word, this tool is for those who would consider themselves, "data miners." It has a mild learning curve, but is a very powerful tool. Like Weka, it has great community support.	**RapidMiner**
A tool for users who want a solid, but very easy-to-learn data mining tool. It's ideal for some someone who wants to do some simple exploration of an existing data set.	**Weka Explorer**
A tool for users who have minimal experience in data mining and essentially no "programming" skills - who would like to learn about data mining. It has a minimal learning curve, but has a relatively small support community.	**Orange**
An ideal tool for those who are familiar with R. R is a free software environment for statistical computing and graphics. It runs on a wide variety computer platforms.	**Rattle**
A tool ideal for users who have specific data mining needs - using either time-series data or chemistry data.	**KNIME**