Evaluation of Open Source Data Mining Software Packages

Bonnie Ruefenacht[1], Greg Liknes[2], Andrew J. Lister[3], Haans Fisk[1], Dan Wendt[4]

Abstract: *Since 2001, the USDA Forest Service (USFS) has used classification and regression-tree technology to map USFS Forest Inventory and Analysis (FIA) biomass, forest type, forest type groups, and National Forest vegetation. This prior work used Cubist/See5 software for the analyses. The objective of this project, sponsored by the Remote Sensing Steering Committee (RSSC), was to evaluate other software packages, including R, SAS®, WEKA, and Orange. These software packages must work with the USFS standard remote-sensing and GIS packages such as ArcGIS and ERDAS Imagine. As part of this project, a Python script was developed that fully integrated these software packages, excluding SAS®, with ArcGIS and ERDAS Imagine. Appendix A provides the tutorial for this script. Appendix B provides a tutorial on how to write similar scripts in Python.*

Keywords: CART, Orange, WEKA, R, random forest, classification trees, regression trees, open-source software

Introduction

In 2001 the USFS investigated classification and regression-tree (CART) technology for vegetation mapping and image classification. As part of that investigation, the USFS, in conjunction with the US Geological Survey, developed a suite of CART tools to derive percent impervious and canopy cover products for the National Land Cover Dataset 2000. Since the development of these tools, which integrated RuleQuest's Cubist and See5 (http://www.rulequest.com) with ERDAS Imagine, these tools were used to create several USFS mid-level vegetation maps and several national maps, including a nationwide biomass map, forest type map, and forest type group map (Blackard et al. 2008, Ruefenacht et al. 2008). These products were well received by user groups and overall accuracies of the final products met or exceeded accuracy standards.

[1] USDA Forest Service Remote Sensing Applications Center, 2222 West 2300 South, West Valley City, UT 84119
[2] USDA Forest Service Northern Research Station, 1992 Folwell Avenue, St. Paul, MN 55108
[3] USDA Forest Service Northern Research Station, 11 Campus Blvd, Suite 200, Newtown Square, PA 19073
[4] USDA Forest Service Eastern Region – R9, 626 East Wisconsin Ave., Milwaukee, WI 53202

In: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. 2008. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008; Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1 CD.

These local, regional, and national maps not only help address the USFS mission, but also are valuable tools to aid in the understanding of the nation's forests.

New versions of Cubist and See5 have been released since the mapping tools were developed in 2001. Updating Cubist and See5 requires repurchasing Cubist and See5; the cost of Cubist and See5 has more than doubled since 2001. CART falls under the umbrella of data-mining technology, which is a popular analysis technique used in many fields. Because of this popularity, new and less expensive – or even free, open source – software packages have been and are being developed. These new software packages could broaden applicability and improve upon existing approaches.

This project investigated the feasibility of using alternative CART software to help satisfy the USFS current and future mapping needs. Specifically, the objective was to evaluate five statistical modeling software programs for cost, usability, critical mass, uniqueness, defensibility, and performance. The software packages studied were R, WEKA, Orange, SAS®, and RuleQuest's Cubist and See5. In this paper Cubist and See5 are treated as one software program because they come from the same company.

Of critical importance is the ease with which the five software packages could be integrated with the standard USFS remote sensing and GIS software packages, ERDAS Imagine and ArcGIS. As part of the evaluation, example scripts were developed using the programming language Python (http://www.python.org). Python is a versatile open source programming language that is fairly easy to learn. The example scripts were originally meant to demonstrate how to integrate the software packages being investigated with ArcGIS and ERDAS Imagine. However, these example scripts morphed into a complete software package that creates CART-formatted data sets and CART models, and applies the CART models to new data sets. A complete description and tutorial of this software package is included in Appendix A. We also developed a tutorial that demonstrates how to build similar applications using Python; this tutorial is contained in Appendix B.

Review of CART Programs

R

R (http://www.r-project.org) originated in 1993 at the University of Auckland, New Zealand. The original purpose for R was to fulfill statistical computing software needs in teaching laboratories at the University of Auckland. The creators of R were familiar with the S language and environment, which was developed by John Chambers, Rick Becker, and Allan Wilks of Bell Laboratories (now Alcatel-Lucent). S syntax and features were adopted for the development of R; R can be thought of as a dialect of S. R experienced modest growth for a

couple of years and, in 1995, R was released under the terms of the Free Software Foundation's GNU general public license (GPL). R compiles and runs on Windows, MacOS, and a wide variety of UNIX platforms and similar systems.

R is not a CART software package but rather a language and environment for statistical computing. Twelve packages are supplied with the basic R distribution. Each package includes many functions. For instance, the statistical package has 563 different statistical functions. It is possible to create new packages (and new functions) for R. CRAN (Comprehensive R Archive Network) (http://cran.r-project.org) offers 1,364 additional packages extending the basic R functionality. Within these additional packages are functions that do classification trees, regression trees, and random forests.

WEKA

WEKA (Waikato Environment for Knowledge Analysis) (http://www.cs.waikato.ac.nz/ml/weka/) was developed by the University of Waikato, New Zealand, which still actively supports the software with funding from the New Zealand government. The original purpose of WEKA was to develop machine-learning techniques and investigate their applications in the agricultural industries of New Zealand. Development of WEKA was started in 1993. WEKA was publicly released in 1996 and is available for free under GPL. WEKA can be run on Windows, MacOS, and Linux and similar systems.

WEKA is a collection of machine-learning algorithms implemented in Java. WEKA contains several data preprocessing tools, classification-tee and regression-tree algorithms, clustering algorithms, associations rule algorithms, and visualization tools. WEKA can be run using its graphical user interface (GUI) or can be implemented from customized Java code. There is also a WEKA package for R (RWeka) distributed from CRAN (http://cran.r-project.org). Rather than attempt to describe WEKA in this document, the reader is encouraged to view the Microsoft PowerPoint available from the WEKA website that demonstrates the program (http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html).

Orange

Out of the five statistical data-mining packages evaluated in this document, Orange is the newest (http://www.ailab.si/orange/) and is continuing to be developed. Orange was released from the University of Ljubljana, Slovenia in 2004 using GPL. The developers of Orange intend it to be a suite of flexible machine-learning tools where users can easily add their own machine-learning algorithms using both scripting and GUI environments. Orange has capabilities to perform classification trees, regression trees, k-nearest neighbors (k-NN), support vector machines, self-organizing maps, and Bayesian networks. New releases of Orange appear each month and new tools are continually being added. Orange also has several unique visualization tools.

The main routines and libraries of Orange are written in C++; Orange uses Python to implement the routines and access the libraries. There is a comprehensive and user-friendly tutorial on how to use Orange in the Python programming environment available on Orange's website (http://www.ailab.si/orange/doc/ofb/). Orange also has a GUI version called Orange Canvas, which allows for interactive machine-learning "visual programming".

SAS®

SAS® was originally an acronym for Statistical Analysis Software created by Jim Goodnight and North Carolina State University associates in the early 1970s. In 1976 the SAS Institute was founded to help distribute and further develop the increasingly popular software. Over time, both the "Institute" portion of the name as well as the acronym were dropped. SAS® currently has 10,658 employees and is the largest privately held software company. In 2007, the annual revenue of SAS® was $2.15 billion. SAS® is used in 109 countries and has 44,000 customer sites worldwide. SAS® users are employed by a wide variety of industries. Unlike the other software packages reviewed thus far, SAS® is not free. SAS® is purchased by contacting a distributor directly. Thus, there is no general price list available, but SAS® can cost several thousand dollars depending upon the options. The purchase of SAS® includes the software, technical support, and licenses, which are renewed regularly, incurring more costs. SAS® might be cost prohibitive for some organizations and individuals.

SAS® is not a CART software package, but it does have an extension called the Enterprise Miner™ that can perform decision-tree analysis. There is also a SAS® extension called SAS® Bridge for ESRI, which links ArcGIS with SAS®. SAS® has its own scripting language, which is not difficult to learn. SAS® does have menu and GUI systems that assist the user in performing statistical analyses.

Cubist and See5

The first version of See5 was published in 1979 and was called ID3 (Quinlan 1979). In 1993, an improved version of ID3 was released and renamed C4.5 (Quinlan 1993). These programs were released as computer code, which anyone could compile and run. In 1997, the author of C4.5, John Ross Quinlan, started the RuleQuest company (http://www.rulequest.com). C4.5 has since been renamed to See5 and the proprietary program is sold by RuleQuest. RuleQuest also sells Cubist. The difference between Cubist and See5 is Cubist is used for the analysis of continuous variables and See5 is used for the analysis of discrete variables.

Cubist and See5 have simple and easy-to-use GUIs. Basically, the programs only do one thing: create classification trees (See5) or regression rule sets (Cubist). RuleQuest does offer computer code that allows for the classification trees or regression rule sets to be applied to non-training datasets. This computer code is written in C++ and can be easily integrated into a company's existing software.

Evaluation of CART Programs

The five statistical software packages reviewed above were evaluated according to cost, usability, critical mass, uniqueness, defensibility, and performance. Table 1 shows a summary of the evaluations. These evaluations will be discussed below.

Cost

Since R, WEKA, and Orange are all available for free under GPL, they received the highest ranking. As of this writing, the combination package of Cubist and See5 costs $1,475 for a single-seat license. Thus, Cubist and See5 received a middle ranking because they are not free nor too expensive. A limited number of SAS® licenses are available for USFS employees to use at no additional cost; see http://fsweb.rmrs.fs.fed.us/statistics/statsoftware/sas/ for more information. Even though a USFS employee can use SAS® for "free", the software was purchased and is expensive. Because of the expenses of SAS®, it received the lowest ranking.

Usability

There were four main questions related to the usability of the programs.

1. How easy is the interface to use and understand?
2. Are there a variety of models and options available?
3. How easy to use is the software's programming language?
4. How easily does the software integrate with other programs?

These questions will be answered below for each of the software packages. The highest ranking for usability is given to software programs that are easy to understand, have a variety of analysis options available, and are easy to enhance and integrate with other programs.

SAS®: The Enterprise Guide for SAS® has a user-friendly GUI system that allows for the building of graphical models. GUIs also exist for other SAS® modules, but unlike WEKA and Orange there is no universal GUI for SAS®, which resulted in a low ranking for interface ease (table 1).

SAS® is used in a wide variety of industries (http://www.sas.com) from aerospace to utilities and everything in between. To accommodate the diversity of industries, SAS® incorporates an extremely wide variety of statistical models and programs. SAS® is primarily driven by its own programming language, which means that a new user will require some training to become efficient. With

advanced programming skills and also with the help of the technical support that comes with the purchase of SAS®, it is possible to create and integrate components of SAS® into other software programs.

In summary, SAS® received a low ranking for ease of interface use because of the lack of a general GUI; a high ranking for the variety of models and options available, and moderate rankings for ease of programmability and ease of integration with other software (table 1).

R: R, like SAS®, is used by numerous industries and thus has a wide variety of models and options. Because R is open source software, new models are continuously being developed. R is driven by its own scripting language, which does require some training and/or experience to become efficient. GUIs do exist for R, but they are created by users for specific purposes. One of the products of this project was a GUI for R that does CART analyses and is loosely coupled with ArcGIS and ERDAS Imagine. Appendix B gives examples of how this was done. With this knowledge, users with intermediate programming skills should be able to create their own GUIs for their statistical remote sensing/GIS analyses.

Table 1: Evaluation of Programs

	R	WEKA	Orange	SAS	Cubist/See5
Cost	●	●	●	●	○
Useability					
Interface Difficulty	●	◒	○	◒	●
Variety of Models/Options	●	◒	○	●	●
Language/Programability Difficulty	○	●	●	○	◒
Integration with Other Software	●	●	●	○	◒
Critical Mass					
Number of Users	●	◒	●	●	○
Tech Support	○	○	○	●	◒
Longetivity	●	●	●	◒	○
Uniqueness					
Algorithms	●	●	●	●	●
Interface Abilities	●	●	●	●	●
Defensibility					
Peer-review Publications	●	◒	●	○	◒
Performance					
Speed	●	●	●	●	●
Accuracy	●	●	●	●	●
Stability	●	◒	●	●	●
Overall Ranking	●	◒	●	◒	○

● = High Favorable Rank
◒ = One Below High Rank
○ = Middle Rank
◒ = One Above Low Rank
● = Low Favorable Rank

In summary, R received a low ranking for interface ease because of the lack of a general GUI, and high rankings for the variety of models and options available and the ease of integration of R with other software packages (table 1). Since some training is required to master the programming language, R received a moderate ranking for ease of programmability.

WEKA: WEKA does have a comprehensive GUI with many models and options available. WEKA's GUI is easy to use. However, because of the number and complexity of options and algorithms, users need a good understanding of modeling techniques. From this standpoint, WEKA can be difficult for beginners to use.

WEKA is written in Java, which a user will need to know to expand the functionality of WEKA or to integrate WEKA into other software packages. Basic functions can be performed using WEKA's GUI without knowing JAVA. Learning Java requires advanced programming skills. All the functionality of WEKA is included in the RWeka package available in R. WEKA can be integrated with other software packages using R and the tips in Appendix B, which shows how to create GUIs in R and how to loosely couple R with ArcGIS and ERDAS Imagine.

In summary, WEKA received the second highest ranking (table 1) for ease of interface use. Even though WEKA's GUI is easy to use, users need to have some intermediate-to-advanced training on how to use the models effectively. WEKA does not have quite the wide variety of models and options as SAS® or R, but it does have a wide variety of data-mining models. WEKA received the second highest ranking with regards to the number of models and options available. Even though WEKA can be expanded and used within R, the native language of WEKA, Java, requires advanced programming skills to integrate WEKA with other software programs without using R. Thus, WEKA received the lowest ranking for ease of programmability, and for integration with other software.

Orange: Orange does not have an interface system, but Orange does have a modeling environment. A considerable amount of Orange's functionality is available within the modeling environment, but the modeling environment is mainly used for visualization purposes. Users familiar with modeling environments should find Orange easy to use.

The full functionality of Orange can be accessed through the Python scripting language, which is one of the easiest programming languages to learn. The Orange website (http://www.ailab.si/orange/) has an excellent tutorial on how to integrate Orange with Python. Because Orange and ArcGIS both use Python as a scripting language, integration of these two software packages is seamless.

In summary, Orange received the middle ranking for ease of interface use because Orange's easy-to-use interface (or modeling environment) is not comprehensive (table 1). The number of models and options available in Orange lags behind not only SAS® and R but WEKA as well. Orange received the middle ranking for the variety of models and options available. Because Orange employs the easy-to-learn Python scripting language, Orange received high rankings for ease of programmability programmability, as well as software integration ability.

Cubist and See5: Cubist and See5 only allows for two types of analyses: classification trees and regression trees. Because there are only two models available with few options, Cubist and See5 have the easiest and simplest GUI out of the five software programs evaluated. Cubist and See5 are proprietary software programs meaning the functionality cannot be extended. However, there is C code available on http://www.rulequest.com that allows the use of Cubist and See5 results within other software packages that can work with the C language.

In summary, Cubist and See5 received a high ranking for ease of interface use (table 1). Cubist and See5 received low rankings for all of the other items under the usability heading. Cubist and See5 have only two models with few options, is impossible to extend without advanced programming skills, and is difficult to integrate with other programs.

Critical Mass

Critical mass concerns how widespread is the use of the software, what kind of technical support is available, and what is the expected useful lifespan of the software.

The amount of use of the software packages reviewed in this document was based upon how many people the authors know who use the software, what kind of interest is shown in the software on the internet, and how many people write in to the forums and support pages available on the internet.

SAS® and R offer significantly more algorithms and options than the other software packages reviewed in this document making SAS® and R attractive to a wide audience. WEKA was created a little over ten years ago; it seems to have a small number of users. Cubist and See5 are powerful CART packages and is widely used. Orange was created less than five years ago and seems to have the smallest user base.

Out of the five software packages reviewed in this document, SAS® is the only program that comes with technical support. The reason for this is SAS® costs money whereas all the other programs, except for Cubist and See5, are open source. Cubist and See5 do have informal technical support, which consists of e-mailing the author of the program, J.R. Quinlan, when problems occur.

Responses are typically very prompt. All of the open source programs do have user forums where questions and problems can be posted.

SAS® is a well-known and well-respected statistical software package that generates a lot of revenue. SAS® is currently 32 years old and could exist for many more years. However, as attractive open source statistical software packages, such as R, become more popular, the user base of costly software such as SAS will likely shrink, thereby threatening its lifespan. It is not known what will happen when the author of Cubist and See5, J.R. Quinlan, is no longer available. Will Cubist and See5 stop being available as well? The other programs, R, WEKA, and Orange, are projected to have long life spans because of the user community interest and involvement in the open source software packages.

Uniqueness

Uniqueness addresses the questions of how unique are the algorithms contained within the software and does the software have unique options or capabilities?

SAS® is a proprietary software package and therefore contains algorithms unique to it. SAS® also contains general statistical algorithms that are also available in many statistical software packages or spreadsheet and database programs such as R and Microsoft Office Excel.

Cubist and See5 are also proprietary software packages and therefore their algorithms are unique to it. However, older versions of See5, also known as ID3 or C4.5, are public domain. These older versions are available in WEKA and Orange.

R, WEKA, and Orange are open-source software packages utilizing public domain algorithms. Many of the same algorithms are contained within R, WEKA, and Orange. For instance, all of these three software packages have the capability to perform a random-forest analysis. Both WEKA and Orange have the capability to do a C4.5 analysis. Also, R contains a package called RWeka that makes all the WEKA algorithms available in R.

WEKA and Orange also have unique visualization capabilities that allow users to discern “randomness” of their data and easily spot outliers that have the potential to heavily skew analyses. These visualization tools are important features that the other three software packages, R, SAS®, and RuleQuest’s Cubist and See5, do not have.

Defensibility

The criterion used to judge defensibility is how often the citation for the particular software package appears in peer-reviewed publications. Using the Institute for Scientific information citation database (http://isiknowledge.com), cited reference searches were performed on the following citations or authors:

- Cubist and See5 author: J.R. Quinlan
- Orange citation:
- Demsar J.; Zupan, B.; Leban, G. 2004. Orange: From experimental machine learning to interactive data mining, White Paper (http://www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.
- WEKA citation:
- Witten, I.H.; Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques. 2nd Edition*, Morgan Kaufmann, San Francisco.
- R citation:
- R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- SAS author: SAS Institute

The cited reference searches were limited to the past five years (2004-2008) or, in the case of WEKA, from the year of publication of the cited work to 2008. R received the most citations at 5,433 and SAS® was cited 1,001 times. J.R. Quinlan's work appeared in 2,664 articles with 44 of those articles dealing with remote sensing or GIS issues. WEKA was cited in 462 articles with four of those articles dealing with remote sensing or GIS issues. Orange was cited in 59 articles and none of them dealt with remote sensing or GIS issues.

The number of citations is influenced by the length of time the software has been available. Various versions of Cubist and See5 have been available since the 1980s. Even though WEKA's citation year is 2005, the software package has been available since 1996. Orange software package has only been available since 2004. Because SAS® is 20 years older than R, SAS® was expected to receive more cited references. The much greater number of citations for R over SAS® demonstrates R's popularity as the statistical package of choice for many users.

Because of the unevenness in publication dates, it is unfair to judge the quality of a software package by these citation numbers. There are two conclusions to be drawn from these numbers, however. First, R is the statistical package of choice for many users. Secondly, after almost 30 years, Cubist and See5 are industry standards for data mining and have been used successfully for remote sensing and GIS projects.

Performance

Performance looks at the speed, stability, and accuracy of the software package.

Performance tests were run on a dataset consisting of 9,802 training points. Classification trees were produced using each software package except for SAS®. R, Cubist, and See5 took one second to create a classification tree. WEKA and Orange both took three seconds to produce a classification tree. Orange, WEKA, and R all contain the random-forest model, but Orange and WEKA could not run the random-forest model because the training dataset was too large. R performed a random-forest classification in 2 minutes 24 seconds.

The results of the above classification trees were applied to a dataset of 427,558 data points. Cubist and See5 both took 17 seconds to apply the classification-tree model to this dataset. R had the second fastest time of 35 seconds. R also took 2:08 minutes to apply the random-forest results to this dataset. For WEKA and Orange, this dataset was too large.

Orange is the least stable of the five software packages reviewed; it has difficulty handling large datasets and also has bugs and flaws. New versions of Orange are released monthly, but it has not achieved a stable, relatively bug-free version yet. WEKA is a stable program, but also does not work well with large datasets. Our tests indicate WEKA is a fast program, but some WEKA users have noted its slowness. R, Cubist, See5, and SAS® are very stable and efficient programs.

Since these software packages utilize the same or similar algorithms, all would be expected to achieve comparable accuracy. However, selection of different options will affect accuracy. On the same dataset, it is possible to achieve very high accuracy or a much lower accuracy by simply varying the options selected. Users need to investigate the literature and experiment on their own to determine which software program or model is appropriate and will achieve high accuracy for their classification problem.

Conclusion

This project evaluated five statistical modeling software programs: R, WEKA, Orange, SAS®, and RuleQuest's Cubist and See5. The overall ranking shown in table 1 indicates R came out on top in most of the categories with SAS®, RuleQuest's Cubist and See5, WEKA, and Orange following R in that order.

Each program has various strengths and weaknesses. In order to take full advantage of these programs and their strengths, users might want to employ several if not all of these software packages. For example, users might want to

take advantage of the visualization and attribute selection options in WEKA and/or Orange. Then the user might want to run a random-forest analysis in R or produce a classification tree in See5.

To aid users of these software packages, a Python script is available; it will do CART analysis using R, WEKA, Orange, Cubist, and See5. Appendix A contains a tutorial for the Python script. Also, Appendix B shows examples of Python scripts, which users can expand to meet their needs and take full advantage of R, WEKA, Orange, Cubist, and See5.

Acknowledgments

We would like to thank the USDA Forest Service Remote Sensing Steering Committee for funding this project. We would also like to thank Nancy Solomon and Paul Maus for reviewing and editing the manuscript.

References

Blackard, J.; Finco, M.; Helmer, E.; Holden, G.; Hoppus, M.; Jacobs, D.; Lister, A.; Moisen, G.; Nelson, M.; Riemann, R.; Ruefenacht, B.; Salajanu, D.; Weyermann, D.; Winterberger, K.; Brandeis, T.; Czaplewski, R.; McRoberts, R.; Patterson, P.; Tymcio, R. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and Terra MODIS-based information. Remote Sensing of the Environment. 112(4):1658-1671.

Ruefenacht, B.; Finco, M.V.; Nelson, M.D.; Czaplewski, R.; Helmer, E.H.; Blackard, J.A.; Holden, G.R.; Lister, A.J.; Salajanu, D.; Weyermann, D.; Winterberger, K. 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. Photogrammetric Engineering and Remote Sensing. 74(11): 1379-1388.

Appendices

Appendix A: Tutorial of Classification and Regression-tree (CART) Python Program

Appendix B: Image Processing in a Python Environment

Due to their length, Appendices A and B are available upon request from the author:

Bonnie Ruefenacht, PhD
USDA Forest Service
Remote Sensing Applications Center
2222 West, 2300 South
Salt Lake City, UT 84119 - 2020
bruefenacht@fs.fed.us