Remote Sensing of Environment 170 (2015) 115–120

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

The AmericaView classification methods accuracy comparison project: A rigorous approach for model selection

Rick L. Lawrence *, Christopher J. Moran

Spatial Sciences Center, Land Resources and Environmental Sciences Department, Montana State University, Bozeman, MT 59717, USA

ARTICLE INFO

Article history:
Received 5 May 2015
Received in revised form 25 August 2015
Accepted 21 September 2015
Available online xxxx

Keywords:
C5.0
Classification tree analysis
Classification algorithms
Logistic model trees
Multivariate adaptive regression splines
Random forest
Support vector machines

ABSTRACT

Evaluation of classification methods, whether in connection with the development of new methods or in an application setting, has been hampered by the lack of availability of adequate data and an approach for comparisons. We collected 30 mostly moderate-resolution, multispectral datasets to enable statistically rigorous comparisons of methods and have made those datasets available for other researchers. We developed a methodological approach to comparing classification methods and demonstrated the approach using six methods, C5.0, classification tree analysis, logistic model trees, multivariate adaptive regression splines, random forest, and support vector machines. We also demonstrated how these data and this approach can be used to address specific questions in addition to overall accuracy performance, including the relative effects of using derived components and ancillary data and the relative success in classifying rare classes. Most methods performed best by at least one metric with at least one dataset. Therefore, although random forest on average performed statistically significantly better than the other methods tested, we do not recommend this method as the sole option currently in remote sensing. Rather, our results suggest that remote sensing analysts should evaluate multiple methods with respect to any classification project, which can be accomplished through statistical software packages.

1. Introduction

The past two decades have seen rapid expansion of the types of classification methods used with remotely sensed imagery, especially with respect to supervised classification methods. Few methods were commonly employed in the mid-1990s, with remote sensing textbooks commonly covering parallelepiped, nearest neighbor, and maximum likelihood classifiers (e.g., Lillesand & Kiefer, 1994), while commercial image processing software rarely included other options. Machine-learning methods, in particular, have seen rapid adoption since the 1990s, perhaps starting with neural networks (e.g., Heermann & Khazenie, 1992), and then expanding into methods such as tree-based approaches (such as classification trees, Lawrence and Wright (2001), C5.0, Quinlan (1993), and random forest, Lawrence, Wood, and Sheley (2006), and support vector machines, Mountrakis, Im, and Ogole (2011)). Many of these methods have not yet been adopted within some of the most popular commercial image processing software packages, but the evidence both in published literature and anecdotally is that these methods are now in widespread use, often through add-ins to commercial software or as stand-alone programs. It is likely that we will continue to see an increasing number of new methods applied to remotely sensed data. We are aware, for example, of over 100 classification methods available in the R statistical program, most of which likely have not ever been tested in the remote sensing field, although some of the more traditional methods, such as the maximum likelihood classifier, are currently missing, complicating comparisons with such methods. Some are not appropriate or logical choices for remote sensing, but many are worth examining. The proliferation of new methods is showing no signs of abating.

The general practice when introducing new methods to the remote sensing field has been to provide very limited, if any, comparisons to other methods and to apply the new methods to one or only a few datasets. Examples from some of our publications will serve to illustrate this common approach. An early paper on classification trees (Lawrence & Wright, 2001) used a single dataset and compared no other methods. The paper introducing stochastic gradient boosting to remote sensing (Lawrence, Bunn, Powell, & Zambon, 2004) compared results to one other method, single classification trees, and used three datasets. One of the earliest papers applying random forest to remote sensing classification (Lawrence et al., 2006) compared results to two methods, single classification trees and spectral angle mapper, and used two datasets. We have used our own studies to illustrate the point (so as not to point fingers at others), but this approach of conducting very limited comparisons is common. This tendency has likely been out of necessity, rather than by choice. New methods are almost always developed or adopted from other fields in the context of the needs of a specific, often grant funded, project, thus making the collection and application to other datasets outside the bounds of the project. There historically has not been a readily available collection of datasets that could be

* Corresponding author.
E-mail address: rickl@montana.edu (R.L. Lawrence).

http://dx.doi.org/10.1016/j.rse.2015.09.008
0034-4257/

accessed for truly rigorous comparisons (this is in comparison with statistical literature where, for example, random forest was introduced using 19 datasets (Breiman, 2001)). Researchers also have been faced with determining what logical comparisons would be meaningful among the myriad possibilities available. Perhaps comparisons to maximum likelihood (the standard of the day) were logical in the 1990s, but subsequent growth in available methods presents no obvious contemporary standard, and incorporating many different methods into an analysis might not be practical.

Remote sensing researchers and practitioners have been left, therefore, with less rigorous bases on which to select classification methods. Options have included the perceived weight of the evidence based on many published works showing high success of certain methods, use of methods with which a researcher has had previous familiarity and success, or ease of application based on availability through a particular software program.

The goal of this project was to create and demonstrate an approach and infrastructure that will allow rigorous comparisons of classification methods for remotely sensed data. The project was bounded at this time for practical purposes to include (1) mostly multispectral, moderate spatial-resolution datasets, (2) pixel-based, supervised classification methods, and (3) classification schemes with three or more classes (because two-class problems have an additional range of methodological options). Our approach, if found useful, could readily be expanded beyond these bounds, given the availability of appropriate datasets.

The methods we selected for demonstration included four that have been widely favorably reported in the literature and, in order to demonstrate the utility of this approach for evaluating new methods, two that, to our knowledge, have been rarely or never reported as previously used for classification of remotely sensed data. We initially compared these methods based on overall accuracy. Overall accuracy, however, might not always be the only, or even primary, factor on which to base the selection of a classification method. We recognized that the approach and infrastructure we present provides the ability to rigorously compare methods based on many criteria. We therefore further demonstrated examples of how these data might be mined by conducting two other analyses. First, because many modern classification problems, in addition to using spectral band data, take advantage of ancillary data and derived components, we examined whether certain classification methods were better able to exploit these additional data by repeating our analysis excluding ancillary data and derived components and evaluating the resulting changes in overall accuracy. Second, classification of rare classes can be problematic for some classification methods (such as classification tree analysis, Chawla, Cieslak, Hall, and Joshi (2008)). We therefore also compared class accuracies among the methods for rare classes.

2. Methods

2.1. Data

We attempted to obtain a large number of datasets meeting the study's criteria in order to have sufficient statistical power to meaningfully compare methods. A number of datasets were available in-house from previously published studies (Lawrence et al., 2004; Bricklemyer, Lawrence, Miller, & Battogtokh, 2007; Savage & Lawrence, 2010). We made broadly advertised requests through several remote sensing organizations/committees with which we are involved, direct inquiries to contacts at governmental agencies, and personal requests to several remote sensing colleagues. The response was extremely limited, and personal contact indicated that, while the project was deemed highly valuable, researchers felt they did not have the time to work through their archives to obtain and provide data. The primary source of additional datasets, therefore, came from data archived on-line by the Gap Analysis Project (GAP) (Lowry et al., 2007).

The final collection of datasets used for our analyses included five in-house and 25 obtained through the GAP archive (Table 1). Most very large datasets (tens of thousands of observations) were randomly subset to 3000–5000 observations for computational efficiency. Most datasets were based on Landsat imagery and included either ancillary data (such as topographic variables), derived components (such as tasseled cap components), or both. Additional information with respect to these datasets can be found at the referenced citations.

2.2. Methods tested

Our approach was demonstrated using six selected methods. Four of these methods, classification tree analysis (CTA), C5.0 (C5), random forest (RF), and support vector machines (SVM), have been widely reported and demonstrated as successful methods for classification of remotely sensed data. One method, multivariate adaptive regression splines (MARS), has been successfully reported for mapping continuous responses with remotely sensed data (e.g., Nawar, Buddenbaum, Hill, & Kozak, 2014), but to our knowledge has not yet been widely used for classification applications (but see Quirós, Felicísimo, & Cuartero, 2009). We were not aware of any reported studies using logistic model trees (LMT) with remotely sensed data but chose to evaluate it as one of the most recent tree-based classifiers not using ensemble methods. We used, in all cases, a version of the method implemented in the R statistical package, using default parameters in order to standardize the comparisons (Table 2).

CTA, C5, RF, and SVM have been widely reported in the literature, and readers are referred to these previous studies for detailed descriptions of those methods. An overview of these methods and many others in a single volume can be found in Tso and Mather's (2009) Classification Methods for Remotely Sensed Data, Second Edition.

LMTs are a refinement of CTA or decision trees (Landwehr, Hall, & Frank, 2005). CTA uses a single variable at each tree node to build a model. LMT, in contrast, builds a logistic regression model at each node to determine the node's binary split. Each logistic regression is built from all input variables using a stepwise variable selection approach based on model Akaike information criterion (AIC) score. This approach gives LMT the theoretical advantage of better designed splits at each node within a tree model.

MARS (Friedman, 1991), implemented in the “earth” package in R, has been used in very limited remote sensing classification applications (Quirós et al., 2009). MARS is similar to CTA in that it is a recursive partitioning algorithm. MARS, however, incorporates a multi-stage regression that uses spline functions. MARS is based on regression functions, but methods have been developed to adapt it to classification problems. A reader interested in expanded detail on the functioning of MARS is referred to the citations above.

2.3. Analysis

Training data in each case consisted of 75% of the total dataset (except for dataset #4, which was 50%). Validation data consisting of a randomly selected 25% of each dataset (except for dataset #4) were extracted, retained for accuracy assessment, and not used in model building. A function was created in the R statistical programming language for each method tested. The applicable function used the training data for each dataset sequentially to build a model for that dataset, generate accuracy statistics based on the withheld validation data, and compile the accuracy statistics for all datasets into a single spreadsheet for each method. Overall accuracies were compared pairwise between methods using a Wilcoxon's paired signed rank test with a Bonferroni correction for multiple comparisons (Demsar, 2006).

The comparative ability of each method to utilize ancillary data and derived components was evaluated by removing these components from each dataset and repeating the previous analysis using only spectral band data. Changes in accuracy compared to analyses using all data

Table 1
Information with respect to 30 datasets used for the analysis. Source A was in-house while Source B was the GAP Analysis Project. Landsat data is from the TM and ETM+ sensors, and in many cases included multiple dates and/or excluded thermal data. Locations are general, and more specific locations can be found in the publications referencing these data. Ancillary data TC = tasseled cap components and topo = topographic layers, such as elevation, slope, and aspect.

Dataset	Source	Location	Number of bands/sensor	Ancillary data	Number of classes	Number of training observations
1	A	Greater Yellowstone Ecosystem	18/Landsat	18 (TC, topo)	6	561
2	A	Yellowstone National Park	14/Landsat	37 (TC, topo)	21	5248
3	A	Virginia City, Montana	128/Probe-1	None	6	999
4	A	Northeast Montana	14/Landsat	21 (TC)	6	35950
5	A	Sierra Nevada Mountains, California	4/IKONOS	1 (topo)	4	2780
6	B	West Utah	18/Landsat	9 (TC, topo, landforms)	22	3999
7	B	Central Utah	6/Landsat	4 (TC, topo)	3	4000
8	B	Southeast Utah	18/Landsat	9 (TC, topo, landforms)	28	4000
9	B	East Central Utah	18/Landsat	9 (TC, topo, landforms)	23	4000
10	B	Northwest New Mexico	18/Landsat	11 (TC, topo, landforms)	19	4000
11	B	Southwest New Mexico	18/Landsat	11 (TC, topo, landforms)	26	5000
12	B	Northwest Nevada	12/Landsat	12 (TC, topo, landforms)	18	5000
13	B	West Central Nevada	12/Landsat	17 (TC, topo)	15	3499
14	B	Northwest Arizona	18/Landsat	9 (TC, topo, landforms)	20	4998
15	B	Northeast Arizona	18/Landsat	9 (TC, topo, landforms)	19	4000
16	B	North Central Arizona	18/Landsat	9 (TC, topo, landforms)	23	6010
17	B	Central Arizona	18/Landsat	9 (TC, topo, landforms)	19	6008
18	B	Southwest Arizona	18/Landsat	9 (TC, topo, landforms)	14	6004
19	B	Northwest Colorado	6/Landsat	16 (TC, topo)	65	6018
20	B	Central Colorado	6/Landsat	16 (TC, topo)	53	6019
21	B	East Colorado	6/Landsat	19 (TC, topo)	44	6012
22	B	West Central New Mexico	18/Landsat	11 (TC, topo, landforms)	34	1844
23	B	North Central New Mexico	18/Landsat	8 (TC, topo, landforms)	40	5426
24	B	South Central New Mexico	18/Landsat	8 (TC, topo, landforms)	45	6011
25	B	East New Mexico	18/Landsat	11 (TC, topo, landforms)	39	6009
26	B	North East Nevada	12/Landsat	9 (TC, topo)	32	6010
27	B	East Nevada	12/Landsat	17 (TC, topo)	34	6015
28	B	South Nevada	12/Landsat	9 (TC, topo)	27	6009
29	B	South West Nevada	12/Landsat	18 (TC, topo, landforms)	31	6013
30	B	North West Nevada	18/Landsat	9 (TC, topo, landforms)	13	5998

were calculated, and the relative effects of these changes were analyzed using the same Wilcoxon's test.

Our analysis of the ability of each method to classify rare classes began with a definition of "rare classes". We set as an arbitrary standard for this study any class where the size of the class's training dataset was less than one percent of the total training dataset. Many datasets, especially those from the GAP project, had a large number of classes (mean of 25 and range of 3–65), and this threshold produced an appropriate relative amount of rare classes (303 of 747 total classes). We then compared the class accuracies for these rare classes as a group using the same Wilcoxon's test. Producer's and user's accuracies were tested separately.

Table 2
Summary of methods tested and key parameters applied for each method. The default parameters in the R statistical package were used in each case to standardize comparisons.

Method	Parameters
Random forest	ntree = 500 mtry = square root of number of variables sampsize = two thirds of dataset
C5.0	trials = 10
Logistic model trees	None
Support vector machines	kernel = radial degree = 3 gamma = number of variables cost = 1 nu = 0.5
Multivariate adaptive regression splines	pmethod = backward ncross = 1 nfold = 0 varmod.conv = 1 varmod.clamp = 0.1 varmod.minspan = −3
Classification tree analysis	None

3. Results

3.1. Overall accuracies

Individual dataset overall accuracies were similar to those reported for these datasets in peer-reviewed articles using these data. Mean overall accuracies were low (51–73%) primarily because of the effect of the GAP datasets (Table 3, Fig. 1). The GAP analysis involved a large number of classes (125 identified classes within the entire region), and the GAP study reported an overall mean accuracy of 61% (Lowry et al., 2007). There were statistically significant differences in all pairwise comparisons of overall accuracy among the classifiers (all p-values <0.01, except RF compared to C5, p-value = 0.05).

RF resulted in the highest mean overall accuracy, although C5 resulted in a mean overall accuracy within 1% of RF. CTA performed worst on average, trailing RF by over 22%. The superior overall accuracy of RF, however, did not mean that it was always the best classifier for individual datasets. RF had the highest overall accuracy for 18 datasets, while C5 was best for 11, and LMT was best for one. The difference between the best and second best classifier in many cases was less

Table 3
Comparisons of overall accuracy based on 30 datasets and all variables, including spectral band data, derived components, and ancillary data. Pairwise differences in mean overall accuracy were all statistically significant (p-values <0.01, except random forest compared to C5.0 p-value = 0.05).

Method	Mean overall accuracy	Number of times as best classifier
Random forest	73.19%	18
C5.0	72.35%	11
Logistic model trees	64.82%	1
Support vector machines	62.28%	0
Multivariate adaptive regression splines	58.50%	0
Classification tree analysis	50.84%	0

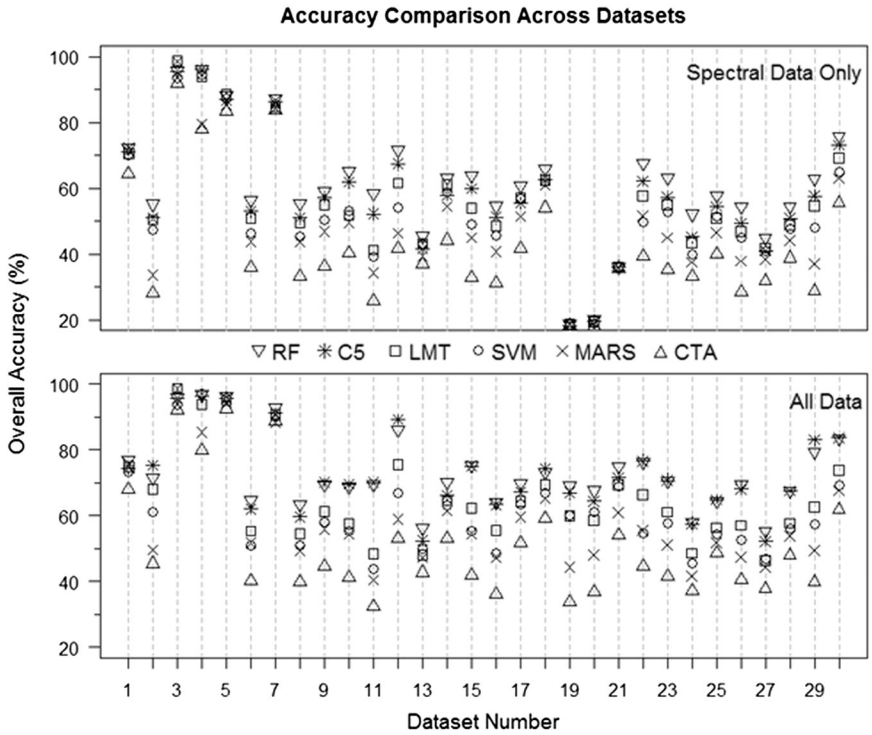

Fig. 1. Comparison of six classification methods across 30 datasets, including analysis of spectral band data, derived components, and ancillary data combined (All Data) and spectral band data only. Methods compared are random forest (RF), C5.0 (C5), logistic model trees (LMT), support vector machines (SVM), multivariate adaptive regression splines (MARS), and classification tree analysis (CTA). See Section 2.2 for method descriptions.

than 1% of overall accuracy, however, while in some cases is was substantially more (Fig. 1).

3.2. Spectral band only accuracies

The relative performance of the methods when classifications were evaluated using only spectral bands (excluding both ancillary data and derived components) was the same as with all variables included (Table 4, Fig. 1). All classifiers performed worse in terms of overall accuracy with the reduced number of variables. There were substantial differences, however, in how much worse each performed, indicating differences in how effectively each method was able to use the additional variables. C5 demonstrated nearly a 15% decrease in mean overall accuracy, indicating that on average it was best able to take advantage of the additional variables, while RF had a decrease of over 12%. The decrease for other classifiers was less than 10%, with CTA showing the smallest decrease of less than 8% on average, indicating that overall it was least effective in taking advantage of the additional variables.

There were also differences in which methods achieved the highest overall accuracy with each dataset. RF more often achieved the highest overall accuracy when using only spectral band data, while C5 only achieved the highest overall accuracy with one dataset. LMT achieved the highest overall accuracy with three datasets (as opposed to only once when all variables were used), while in one case MARS tied RF with the highest overall accuracy.

3.3. Class accuracies for rare classes

Our datasets contained a large number of rare classes, 303, primarily because of the large number of classes specified for the GAP project. All differences between methods were statistically significant (p-values < 0.01), except for the best methods for both user's and producer's accuracy, C5 and LMT for user's accuracy (p-value = 0.10) and RF and C5 for producer's accuracy (p-value = 0.09).

C5 produced the best rare class results when mean producer's and user's accuracies were averaged (Table 5). C5 had the highest mean user's accuracy (although it was not statistically significantly different than LMT at an alpha of 0.05) and the second highest mean producer's accuracy (although it was not statistically significantly different than RF at an alpha of 0.05). RF and LMT also produced relatively high class accuracies.

CTA produced the worst class accuracies, nearly completely failing to successfully differentiate any rare classes. MARS and SVM, while having higher class accuracies than CTA, also produced relatively low accuracies compared to the best methods.

There were substantial differences in how often each method produced the best class accuracies, similar to the overall accuracies. RF

Table 4
Comparisons of overall accuracy based on 30 datasets and only spectral band data, excluding derived components and ancillary data. Pairwise differences in mean overall accuracy were all statistically significant (p-values < 0.01). Random forest and multivariate adaptive regression splines tied as the best classifier in one case.

Method	Mean overall accuracy	Number of times as best classifier	Mean decrease in accuracy vs all components
Random forest	60.75%	25	12.44%
C5.0	57.47%	1	14.88%
Logistic model trees	55.52%	3	9.30%
Support vector machines	53.18%	1	9.10%
Multivariate adaptive regression splines	49.47%	1	9.03%
Classification tree analysis	42.90%	0	7.94%

Table 5
Comparisons of user's and producer's accuracies based on 303 rare classes and all variables, including spectral band data, derived components, and ancillary data. Pairwise differences in mean overall accuracy were all statistically significant (p-values < 0.01) except C5.0 and logistic model trees for user's accuracy p-value = 0.10 and random forest and C5.0 for producer's accuracy p-value = 0.09.

Method	Mean user's accuracy	Number of times with highest user's accuracy	Mean producer's accuracy	Number of times with highest producer's accuracy
Random forest	36.03%	196	59.77%	134
C5.0	44.03%	53	56.41%	108
Logistic model trees	41.40%	28	43.51%	58
Support vector machines	13.44%	13	25.33%	2
Multivariate adaptive regression splines	6.24%	13	13.63%	1
Classification tree analysis	0.22%	0	0.85%	0

most often produced the highest user's and producer's accuracies. C5 was second with respect to both types of class accuracies, generating the highest user's accuracies only 27% as often as RF, while generating the highest producer's accuracies almost as often as RF. LMT also often generated the highest rare class accuracies, while SVM and MARS rarely had the highest rare class accuracies. CTA never had the highest rare class accuracies.

4. Discussion and conclusions

We conducted statistically rigorous comparisons of six remote sensing classification methods using 30 datasets from a diversity of sources. Our comparisons focused initially on overall accuracy. We also demonstrated how these data could be mined to answer more specific questions by evaluating (1) how well the methods were able to exploit derived spectral and other ancillary data and (2) how the methods compared for classifying rare classes.

A simplistic conclusion from our results, that we believe should be rejected, might be to use RF, because it produced the highest overall accuracy on average. Delving only slightly deeper into the results revealed that there are many individual cases where RF did not result in the highest accuracy. RF generated the highest overall accuracy only 60% of the time with our full data sets, while C5 produced higher overall accuracy in most other cases and LMT had the highest overall accuracy once. Analyses using only spectral band data resulted in RF having the highest overall accuracy with increased frequency (over 83% of the time), but, depending on the data set, every other method had the highest accuracy at least once, except for CTA.

The addition of components other than spectral bands on average improved the performance of all methods (Table 4). The improvement for individual datasets, however, was highly variable, ranging from virtually no improvement on average to an average of 37% overall accuracy improvement. The level of improvement was also highly method dependent with some datasets, while it was much more consistent with other datasets (standard deviation of improvement among methods ranged from less than 1% to 14%).

A straightforward, but we believe similarly erroneous, analysis of how well each method was able to use components other than spectral band data also could be misleading. C5 demonstrated the greatest improvement when using the additional components; however RF still produced the highest overall accuracies. One should not conclude, therefore, that when using additional components beyond spectral band data that there is necessarily an advantage to using C5 as opposed to RF. An analyst also should not ignore the other methods simply because RF and C5 have a strong tendency to produce the highest accuracies. Each of the other methods was able to produce the highest accuracy at least once, except for CTA, and for three datasets (all GAP data with a high number of classes) the consistently high performing RF and C5 methods performed very poorly, as did all other methods (Fig. 1).

Our conclusions from our analyses of overall accuracies (e.g., that the best performing method is dataset dependent) was strongly supported by our analysis of class accuracies for rare classes. There were again substantial differences among methods as to which produced, on average, the highest producer's and user's accuracies. An analysis solely of average accuracy percentages might lead one to favor C5, as it produced the highest average accuracies counting both producer's and user's accuracy. An examination of how often each method produced the highest class accuracy, however, demonstrated the RF produced the highest class accuracy much more often. Each method had the highest producer's and user's accuracy for rare classes at least once, except for CTA.

Our results do provide strong evidence that, compared to the other methods tested, CTA is an inferior classifier as it was implemented in this study. CTA never produced the highest overall accuracy either with all components or with spectral bands only and had the worst accuracy in all but two cases. CTA was also the worst method for classifying rare classes, almost completely failing to classify the rare classes. CTA, however, often requires more user interaction to determine a parsimonious pruning level that fits an analyst's objectives, which might increase its accuracy especially for rare classes, while in this study we implemented an automated pruning algorithm to enable efficient processing and remove subjective pruning that would compromise objective comparisons. The ease of interpreting the dichotomous tree produced by CTA also might make it desirable for certain applications, including exploratory data analysis (Lawrence & Wright, 2001).

Our results also demonstrated a strong relationship between overall accuracies and number of classes (p-value for each method <0.01). The amount of variability in overall accuracy accounted for by this relationship, however, varied substantially among methods (R^2 for each method: RF = 28%; C5.0 = 48%; LMT = 33%; SVM = 29%; MARS = 48%; CTA = 46%).

Parameters for all methods were standardized, in addition to the use of automated pruning for CTA, in order to allow valid comparisons among classifications. Default parameters were used in all cases, although for some methods classification accuracies can be substantially affected by the parameters selected. Optimization routines exist for some methods, while others rely on heuristic approaches. An analyst selecting methods for a single classification would be well advised to evaluate the effect of parameters on final classification accuracy.

Our results strongly indicate that, when classifying remotely sensed data, an analyst is not well served in relying on a single method. A method agnostic approach, rather, is recommended, where multiple methods are compared to evaluate the best performing method for a given dataset and analysis objective. The absence of modern classification methods from commercial software packages has resulted in many remote sensing scientists increasingly relying on more flexible approaches, such as statistical packages, to process their images. This can, in many cases, make it relatively easy to modify code to substitute classification methods and efficiently compare alternative approaches and method parameters, such as kernel configuration for SVM (Huang, Davis, & Townshend, 2002), pruning level for CTA, and ensemble parameters for RF.

We also recommend that remote sensing scientists developing or applying new classification methods should embark upon rigorous comparisons to existing widely used methods. We have encouraged

this approach by making the datasets and a sample of R code used for this study downloadable at www.americaview.org. These data are provided in comma-delimited csv file format, which might need reformatting depending on the data input requirements for the new methods, but we hope that the effort will be worthwhile in enabling the remote sensing community to have more meaningful analyses of new approaches. Finally, there are many other questions in addition to those we have addressed here that can be answered with these data. Utilizing confusion matrixes and the resulting producer's, user's, and overall accuracies (Congalton & Green, 2009) are but a few of the metrics available for assessing classification accuracy as well (Foody, 2002). We encourage those interested to download these data and embark upon further analyses.

Acknowledgments

The authors would like to acknowledge the support of the AmericaView consortium for this project and the anonymous reviewers for helpful and insightful comments. The project described in this publication was supported by Grant/Cooperative Agreement Number 08HQGR0157 from the United States Geological Survey. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the USGS.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
Bricklemyer, R. S., Lawrence, R. L., Miller, P. R., & Battogtokh, N. (2007). Monitoring and verifying agricultural practices related to soil carbon sequestration. *Agriculture, Ecosystems and Environment*, *118*, 201–210.
Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, *17*, 225–252.
Congalton, R. G., & Green, K. (2009). *Assessing the accuracy of remotely sensed data.* Boca Raton, FL: CRC Press (184 pp.).
Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning, 7*, 1–30.
Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185–201.
Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *1*, 1–67.
Heermann, P. D., & Khazenie, N. (1992). Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Transactions on Geoscience and Remote Sensing*, *30*, 81–88.
Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*, 725–749.
Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*, 161–205.
Lawrence, R. L., Bunn, A., Powell, S., & Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, *90*, 331–336.
Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*, 1137–1142.
Lawrence, R. L., Wood, S., & Sheley, R. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (random forest). *Remote Sensing of Environment*, *100*, 356–362.
Lillesand, T., & Kiefer, R. W. (1994). *Remote Sensing and Image Interpretation* (2nd ed.). Hoboken, NJ: John Wiley & Sons (721 pp.).
Lowry, J., Ramsey, R. D., Thomas, K., Schrupp, D., Sajwaj, T., Kirby, J., ... Prior-Magee, J. (2007). Mapping moderate-scale land-cover over very large geographic areas within 3 a collaborative framework: a case study of the Southwest Regional Gap Analysis Project (SWReGAP). *Remote Sensing of Environment*, *108*, 59–73.
Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*, 247–259.
Nawar, S., Buddenbaum, H., Hill, J., & Kozak, J. (2014). Modeling and mapping of soil salinity with reflectance spectroscopy and Landsat data using two quantitative methods (PLSR and MARS). *Remote Sensing*, *6*, 10813–10834.
Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann Publishers (302 pp.).
Quirós, E., Felicísimo, A. M., & Cuartero, A. (2009). Testing multivariate adaptive regression splines (MARS) as a method of land cover classification of Terra-Aster satellite images. *Sensors*, *9*, 9011–9028.
Savage, S., & Lawrence, R. (2010). Vegetation dynamics in Yellowstone's northern range: 1985–1999. *Photogrammetric Engineering and Remote Sensing*, *76*, 547–556.
Tso, B., & Mather, P. M. (2009). *Classification Methods for Remotely Sensed Data* (2nd ed.). Boca Raton, FL: CRC Press (356 pp.).