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 29 September 2015

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