WHICH CLASSIFICATION METHOD IS BEST? AN INFRASTRUCTURE FOR RIGOROUS COMPARISONS OF CLASSIFICATION ALGORITHMS

R. Lawrence [a], *

[a] Spatial Sciences Center, Land Resources and Environmental Sciences Department, Montana State University, Bozeman, Montana 59717, USA – rickl@montana.edu

THEME: Data and information systems and spatial data infrastructure

KEY WORDS: CART, C5.0, logistic model trees, multivariate adaptive regression splines, randomForest, support vector machines

ABSTRACT:

Classification methods have greatly proliferated in the past few decades. New methods have historically been introduced with comparisons to at most few other methods, while these comparisons generally have been based on one (or in rare cases a small number) of datasets. Evaluation of these new methods, therefore, has been dependent of usage over time, with no rigorous statistical comparisons conducted. The purpose of this project was to provide a common infrastructure to enable rigorous evaluation of new and existing methods. The initial focus is on moderate-spatial-resolution, multispectral data. A diverse collection of 30 datasets was collected, and six methods were compared for overall accuracy for demonstration purposes. Four methods have been widely used for classification of remotely sensed imagery: CART, C5.0, randomForest, and support vector machines. Two methods were newly introduced for this purpose: logistic model trees and multivariate adaptive regression splines. Differences among all methods were statistically significant (p-values < 0.05), with randomForest outperforming other methods by an average of 1% to 22%, while CART performed worse than the other methods by an average of 8% to 22%. These results can be further analysed for many purposes, such as which method performs best for rare classes, which performs best for certain land cover types, or which provides the most balanced class accuracies. The datasets and code for running comparisons will be made publicly available for developers of new algorithms and other analysts.

---

* Corresponding author.