12-2017

Application of Machine Learning and Statistical Learning Methods for Prediction in a Large-Scale Vegetation Map

Carla M. Brookey
Utah State University

APPLICATION OF MACHINE LEARNING AND STATISTICAL LEARNING METHODS FOR PREDICTION IN A LARGE-SCALE VEGETATION MAP

by

Carla M Brookey

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

_____	_____
Dr. Richard Cutler, Ph.D. Major Professor	Dr. Adele Cutler, Ph.D. Committee Member
_____	_____
Dr. David Olsen, Ph.D. Committee Member	Mark R. McLellan, Ph.D. Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2017

ABSTRACT

Application of Machine Learning and Statistical Learning Methods

for Prediction in a Large-scale Vegetation Map

by

Carla M Brookey, Master of Science

Utah State University, 2017

Major Professor: Dr. Richard Cutler
Department: Mathematics and Statistics

Analyses of a large vegetation-cover dataset from Roosevelt National Forest in Colorado were carried out by Blackard (1998) and Blackard and Dean (1998; 2000). They compared classification accuracies of linear and quadratic discriminant analysis (LDA and QDA) with artificial neural networks (ANN) and obtained accuracies of 70.58% for a tuned ANN, 58.38% for LDA, and 52.76% for QDA.

Because of the development of machine learning classification methods over the last 35 years and improvements in computer hardware speed, I applied five modern machine learning algorithms to the data to determine whether significant improvements in the classification accuracy were possible with these methods. Only a tuned gradient boosting machine had a higher accuracy (71.62%) than the ANN of Blackard and Dean (1998), and the difference in accuracies was about 1%. Of the other methods, Random Forests (RF), Support Vector Machines (SVM), Classification Trees (CT), and adaboosted trees (ADA), a tuned SVM and RF had accuracies of 67.17% and 67.57%, respectively.

The partition of the data by Blackard and Dean (1998) was unusual as the training and validation datasets had equal representation of the vegetation classes, even though 85% of the data are classes 1 and 2. I decided to randomly select 60% of the data for the training data and 20% each for the validation and test data. On this partition, a single CT achieved an accuracy of 92.63% on the test data and the accuracy of RF is 83.98%. Most of the gains in accuracy were in classes 1 and 2, the largest classes which had the highest misclassification rates under the original data partition. By decreasing the size of the training data but maintaining the relative occurrences of the classes, I found that for a training dataset of the same size as that of Blackard and Dean (1998) a single CT was more accurate (73.80%) that their ANN(70.58%).

The final part of my thesis was to explore the possibility that combining several of the classifiers could result in higher predictive accuracies. In the analyses I carried out, a simple voting of five machine learning classifiers does not increase accuracy.

(36 pages)

PUBLIC ABSTRACT

Application of Machine Learning and Statistical Learning Methods for Prediction in a Large-scale Vegetation Map

Carla M Brookey

Original analyses of a large vegetation cover dataset from Roosevelt National Forest in northern Colorado were carried out by Blackard (1998) and Blackard and Dean (1998; 2000). They compared the classification accuracies of linear and quadratic discriminant analysis (LDA and QDA) with artificial neural networks (ANN) and obtained an overall classification accuracy of 70.58% for a tuned ANN compared to 58.38% for LDA and 52.76% for QDA.

Because there has been tremendous development of machine learning classification methods over the last 35 years in both computer science and statistics, as well as substantial improvements in the speed of computer hardware, I applied five modern machine learning algorithms to the data to determine whether significant improvements in the classification accuracy were possible using one or more of these methods. I found that only a tuned gradient boosting machine had a higher accuracy (71.62%) that the ANN of Blackard and Dean (1998), and the difference in accuracies was only about 1%. Of the other four methods, Random Forests (RF), Support Vector Machines (SVM), Classification Trees (CT), and adaboosted trees (ADA), a tuned SVM and RF had accuracies of 67.17% and 67.57%, respectively.

The partition of the data by Blackard and Dean (1998) was unusual in that the training and validation datasets had equal representation of the seven vegetation classes,

even though 85% of the data fell into classes 1 and 2. For the second part of my analyses I randomly selected 60% of the data for the training data and 20% for each of the validation data and test data. On this partition of the data a single classification tree achieved an accuracy of 92.63% on the test data and the accuracy of RF is 83.98%. Unsurprisingly, most of the gains in accuracy were in classes 1 and 2, the largest classes which also had the highest misclassification rates under the original partition of the data. By decreasing the size of the training data but maintaining the same relative occurrences of the vegetation classes as in the full dataset I found that even for a training dataset of the same size as that of Blackard and Dean (1998) a single classification tree was more accurate (73.80%) that the ANN of Blackard and Dean (1998) (70.58%).

The final part of my thesis was to explore the possibility that combining several of the machine learning classifiers predictions could result in higher predictive accuracies. In the analyses I carried out, the answer seems to be that increased accuracies do not occur with a simple voting of five machine learning classifiers.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction and Previous Work

The subject of the analyses that make up my M.S. Thesis is a dataset on vegetation cover type in Roosevelt National Forest in northern Colorado taken from the UCI Data Repository (Bache & Lichman, 2013). Initial analyses of these data were carried out by Blackard (1998) and Blackard and Dean (1998; 2000) using linear dcriminant analysis (LDA), quadratic discriminant analysis (QDA), and artificial neural networks (ANN) to classify vegetation type (seven levels) using topographic, shade, and soil type variables. The intent of their work was to determine if ANN could be used to more accurately predict forest cover type than the more traditional methods. After significant tuning of the neural network, the final model of Blackard and Dean (1998), which used all 54 predictor variables, had an overall accuracy (percent correctly classified) of 70.58% compared to 58.38% for LDA and 52.76% for QDA.

Over the past 35 years there has been tremendous development of machine learning classification methods in both computer science (e.g., support vector machines) and statistics (e.g., classification and regression trees, gradient boosting machines and random forests) as well as substantial improvements in the speed of computer hardware. The initial goal of my work was to determine if other classification methods could outperform the neural network of Blackard and Dean (1998). During these analyses questions arose about the original selection of a training data by Blackard and Dean (1998) and a second piece of my thesis concerns different selections of training, validation, and test data. The application of multiple classification methods brings to mind the possibility of combining predictions from several methods, and that is another part of my research reported in this thesis.

1.2 The Data

I obtained the data for my thesis from the UC Irvine data repository (Bache and Lichman, 2013). The 581,012 observations were collected from the Rawah, Comanche Peak, Neota, and Cache la Poudre wilderness areas of the Roosevelt National Forest, in Colorado prior to 1999. These areas were chosen because there was limited human management disturbances in those areas, leaving the cover type to be determined by natural ecological processes. The data consist of 54 variables which may be broadly classified as topographic and soil type variables. Topographic variables include elevation, aspect, slope, horizontal distance to nearest surface water feature, vertical distance to nearest surface water feature, horizontal distance to nearest roadway, sunlight at 9am, at noon, at 3 pm, horizontal distance to nearest historic wildfire ignition point, wilderness area designation, and soil type. Two of these variables were then converted into a series of binary variables, the 4 wilderness areas, and 40 soil types to give the full set of 54 predictor variables as shown in Table 1.

Table 1

List and description of variables used in analyses

Name	Data Type	Measurement	Description
Elevation	Quantitative	Meters	Elevation in meters
Aspect	Quantitative	Azimuth	Aspect in degrees Azimuth
Slope	Quantitative	Degrees	Slope in degrees
Horizontal_Distance_To_Hydrology	Quantitative	Meters	Horizontal distance to nearest surface water feature
Vertical_Distance_To_Hydrology	Quantitative	Meters	Vertical distance to nearest surface water features
Hoirzontal_Distance_To_Roadways	Quantitative	Meters	Horizontal distance to nearest roadway
Hillshade_9am	Quantitative	0 to 255 index	Hillshade index at 9am, summer solstice
Hillshade_Noon	Quantitative	0 to 255 index	Hillshade index at noon, summer solstice
Hillshade_3pm	Quantitative	0 to 255 index	Hillshade index at 3pm, summer solstice
Horizontal_Distance_to_Fire_Points	Quantitative	Meters	Horizontal distance to nearest wildfire ignition points
Wilderness_Area (4 binary columns)	Quantitative	0 (absence) or 1 (presence)	Wilderness area designation
Soil_Type (40 binary columns)	Quantitative	0 (absence) or 1 (presence)	Soil type designation
Cover_Type (7 types)	Integer	1 to 7	Forest cover type designation

According to Blackard and Dean (1999), the elevation data was taken from the USGS digital elevation model. Each cell represents a unique 30x30 meter cell and the USGS digital elevation

model was used to determine aspect, slope and the measures of relative sunlight. It was also used in conjunction with USFS data concerning wildfire ignition points and hydrological data to determine several of the other variables.

It was also stated in Blackard and Dean(1998) that the cover types were determined from large scale aerial photography, which has been shown to be a reliable method for determining cover type in homogeneous stands. The soil type data and the wilderness designations came from the USFS.

The variable of interest is the cover types and were coded as shown in Table 2.

Table 2

Class codes for vegetation types

Code	Type
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonweed/Willow
5	Aspen
6	Douglas-fir
7	Krummholz (stunted windblown trees growing near the tree line on mountains)

1.3 Statistical Methods

This section contains a brief overview of the various methodologies that I used in my analyses. They are *linear discriminant analysis*, *quadratic discriminant analysis*, *classification trees*, *random forests*, *gradient boosting machines*, boosted trees using the AdaBoost algorithm, and *support vector machines*.

Linear discriminant analysis (LDA) (Fisher 1936) involves taking linear combinations of the predictor variable to create boundaries among the different classes. An assumption of LDA is that the distribution of the predictor variables is approximately multivariable normal with the same covariance matrix (but different means) for the different classes. *Quadratic discriminant analysis* (QDA) (Fisher 1936; 1938) also assumes multivariate normality of the predictor variables but allows different covariance matrices for the different classes, resulting in quadratic boundaries among the classes. For further explanation of LDA, see *A simple explanation of what is LDA classification* (Carrion, 2017).

Classification trees (CT) (Breiman, Friedman, Olshen, & Stone, 1984) work by recursively dividing the data into smaller and smaller subsets ("nodes") that are increasingly pure with respect to the classification variable as measured by the Gini index. At each step in the process a node, a variable, and a cutoff value are chosen so as to maximize the reduction in the Gini index. The process stops when no further partitioning can reduce the value of the Gini index. Such a tree is said to be *fully grown* and the final groups of the data are *terminal nodes* or *leaves*. The number of terminal nodes may be as large as the size of the dataset. Fully grown trees tend to *overfit* data in the sense that the lower branches and leaves are modeling noise in the data rather than structure. Such trees generally have lower predictive accuracy and so methods for "pruning" trees have been developed, the most widely used of which is the 1-SE rule of Breiman et al. (1984). This method penalizes the accuracy of the tree on the training data by multiplying the number of terminal nodes in the tree by a parameter, called the *cost complexity* parameter (*cp*), and then selecting the optimal value of *cp* (and hence the optimal predictive tree) by finding the minimum cross validated prediction error among different values of *cp*. For further

information see *Accurate decision trees for mining high-speed data* (Gama, Rocha, & Medas, 2003).

Adaboost (ADA) (Freund, 1995; Freund & Schapire, 1997) is an ensemble classifier that is usually implemented using classification trees. The algorithm begins by fitting a very simple tree—perhaps with only two terminal nodes—to the data. Observations that are misclassified are upweighted and a new tree is fit to the data. The process is repeated many times, and the eventual predictions come from weighted voting of the many fitted trees with the weights of the individual trees being inversely proportional to their misclassification rates. For further information see *A decision-theoretic generalization of on-line learning and an application to boosting* (Greund & Schapire, 1997).

Gradient Boosting Machines (GBM) (Friedman, 2001) is also an ensemble classifier that works sequentially. The algorithm begins with a tree being fit to the data and a misclassification rate computed. Residuals are computed, and a tree fit to the residuals. The process is repeated many times and the predictions of the different fitted trees voted. In many applications GBM. In many applications fully tuned GBM's have been found to be among the most accurate classifiers currently available, but the devil is in the details: tuning a GBM is a time-consuming process.

Random forests (Breiman, 2001) is another ensemble classifier but works "in parallel" rather than sequentially. Many subsets of the original data are drawn. For each subset the observations that are in the original data but not in the subset are said to be *out-of-bag* (OOB). Fully grown classification trees are fit to each subset with the restriction that only a random sample of predictor variables is made available for partitioning at each node of the tree. This ensures that the fitted trees are quite different and hence will accurately prediction different observations among the original dataset. Predictions made for each tree for all observations that

are out-of-bag for the dataset to which the tree is fit, and combined (by voting) to give a single prediction for that observation. For further information on the use of Random Forests in ecology see *Random Forests for Use in Ecology* (Cutler, et al., 2007).

The default number of trees to fit in a random forest is 500 in the randomForest package in R. Due to computational limitations with some of my analyses I was not able to fit 500 trees. However, as the graph below suggests the accuracy of the predictions is very insensitive to the number of trees fit. Note that the accuracies for 50—200 trees differ only in the third decimal place.

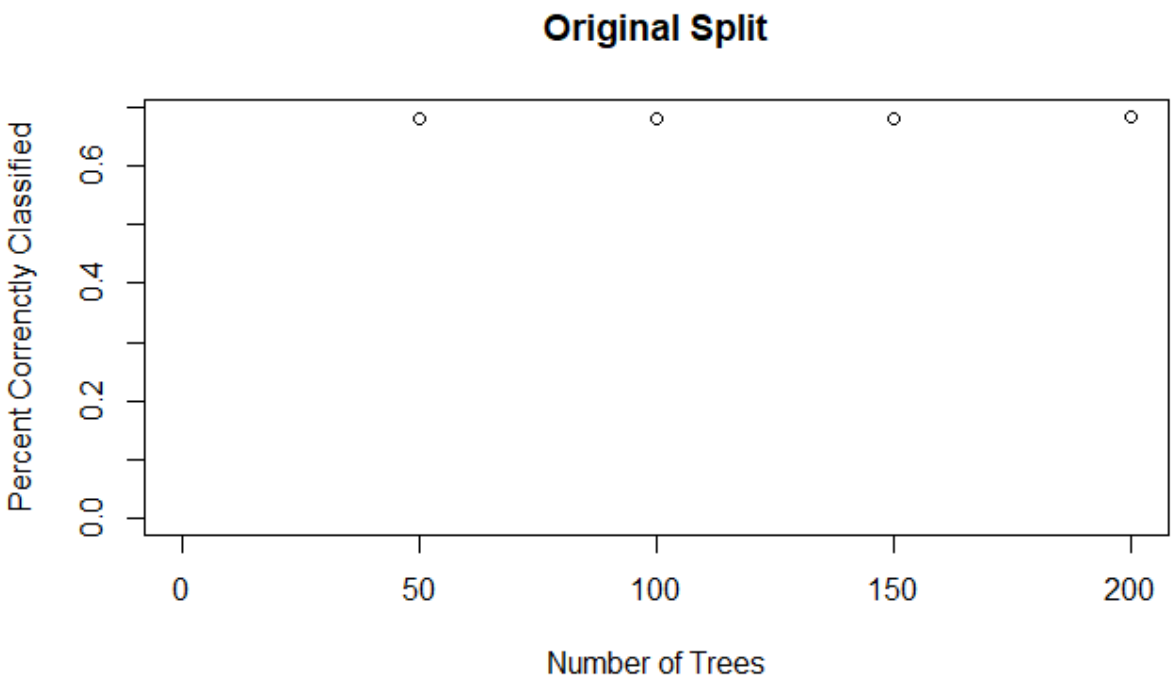

Figure 1

Number of trees used by Random Forests vs percent correctly classified on test set

Support vector machines (SVM) (Cortes & Vapnik, 1995; Vapnik, 1995) are a completely different, non-tree based classification tree methodology. SVMs may be formulated as a constrained optimization and are related to logistic regression for two-group classification. Geometrically SVMs involve projecting the data into a higher dimensional space (the *feature space*) and using linear separators of the classes, then projecting back down to the original dimension of the data (the *input space*) and obtaining highly non-linear separators among the classes.

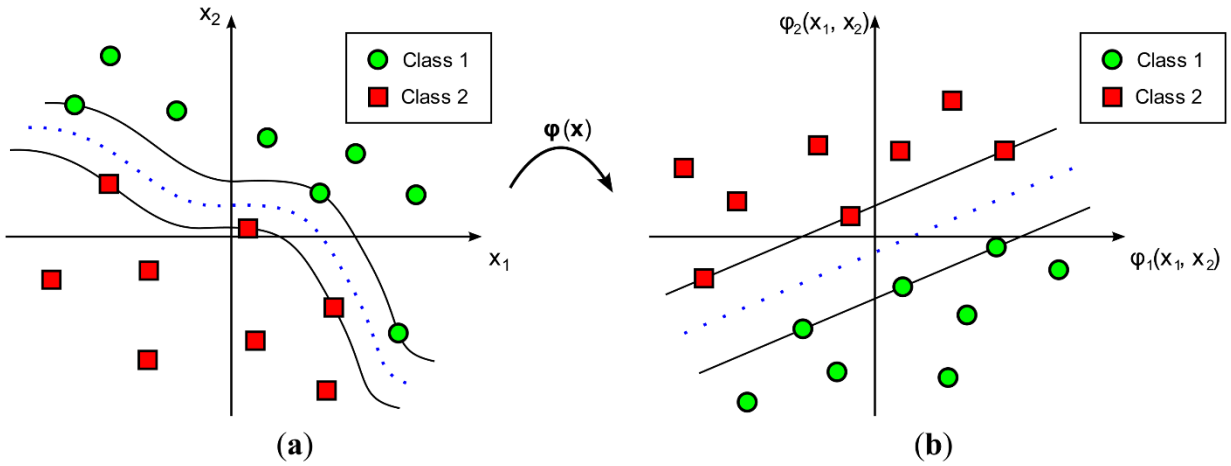

Figure 2

Visual representation of SVM taken from www.mdpi.com

More details about all these methods may be found in the original papers and in Hastie, Tibshirani and Friedman (2001).

All calculations were carried out in R (R Core Team) using the packages `MASS` (Venables & Ripley, 2002), `lda` (Chang, 2015), `rpart` (Therneau, Atkinson, & Ripley, 2015), `randomForest` (Liaw & Wiener, 2002), `gbm` (Ridgeway & with contributions from others, 2015), `caret` (Kuhn, et al., 2016), `e1071` (Meyer, Dimitriadou, Hornik, Weingessel, &

Leisch), `ada` (Culp, Johnson, & Michailidis, 2016), and `adabag` (Alfaro, Gamez, & Garcia, 2013).

1.4 Organization of Thesis

In Chapter 2 I report the results from applying the methods described above to the original division of the dataset into training, validation and test pieces. I compare the accuracies obtained to those of Blackard and Dean (1998). In Chapter 3 I explore different divisions of the data into training, validation and test components and compare the predictive accuracies of the various methods to each other and to the original results in Blackard and Dean (1998). In Chapter 4, I explore the possible increases in accuracy that might be obtained by combining the predictions from several classification methods. Chapter 5 contains an overall summary of my results and a discussion of possible future analyses of the cover type data.

CHAPTER 2

NEW CLASSIFICATION ANALYSES ON THE ORIGINAL PARTITION OF THE DATA

2.1 Methods

For all the analyses in this chapter I used the training, validation and test datasets used by Blackard and Dean (1998). The training data was obtained by randomly selecting 1,620 (58.97%) of the 2,747 observations in the class with the fewest observations (4 = Cottonwood/Willow) and randomly selecting an equal number of observations from each of the other six vegetation classes. The validation dataset was obtained in similar fashion, by selecting 540 observations from each of the seven vegetation classes. All the remaining data, 565,892 observations, were used as the test data. I note that the test data is very much larger than the training and validation datasets. Also, the training and validation datasets have equal representation from all the vegetation classes whereas for the dataset as a whole more than 85% of the data is in classes 1 (Spruce/Fir) and 2 (Lodgepole pine).

The variable Aspect is measured in degrees azimuth and hence is on a circular scale with the largest value, 359, being almost the same direction (north) as the smallest value, 0. Accordingly I generated new variables, Northness and Eastness, by taking the cosine and sine of Aspect, respectively. In all subsequent analyses I used Northness and Eastness rather than the original variable Aspect.

I fit LDA, QDA, CT, RF, ADA, GBM and SVM to the training data using the validation data for tuning parameters where possible. GBM and SVM perform poorly using the default parameters settings in R so tuning is very important. I used the caret and tune.svm packages in R to tune these methodologies and this greatly improved their accuracy.

GBM gave tuned parameters on a multinomial distribution of 200 trees, with an interaction depth of 22 and shrinkage of 0.1. SVM tuned on a radial kernel with cost equal to 85 and gamma equal to $1/43$.

2.2 Results

Column 1 of Table 3 contains the classification accuracies for LDA and QDA using the variable Aspect. These results perfectly match those of Blackard and Dean (1998). The second column contains the accuracies for LDA and QDA with Northness and Eastness instead of Aspect. The results are very similar to those from using the variable Aspect. LDA actually does very slightly worse with Northness and Eastness whereas QDA does very slightly better.

Table 3

Comparison of LDA and QDA results using Aspect and the transformed variables of Northness and Eastness

Method	Test Set Percent Correctly Classified (using Aspect)	Test Set Percent Correctly Classified (using Northness and Eastness)
LDA	58.38%	58.31%
QDA	52.76%	52.95%

Table 4 contains a summary of the classification accuracies for all the methods under consideration on the training ("resubstitution accuracies"), validation and test data. The results of Blackard and Dean (1998) for ANNs are included for purposes of comparison. I note that only tuned GBM produced a higher accuracy than the value obtained by Blackard and Dean (1998) for ANNs, and only by a little over 1%. Random forests also had a relatively high accuracy of 67.57% on the test data. CTs and the tuned version of SVM had an accuracy

between those of LDA/QDA and random forests. The ADA boost method performed particularly poorly, with an accuracy even lower than that of LDA/QDA.

Table 4

Comparison of all methodologies and their resulting accuracies for training, validation, and test data sets

Method	Training Data	Validation Data	Test Data
ANN	–	–	70.58%
LDA	64.78%	65.43%	58.31%
QDA	65.68%	66.14%	52.95%
Classification Trees	87.48%	78.73%	63.22%
Random Forests	80.70%	80.90%	67.57%
GBM	68.47%	68.18%	49.20%
Tuned GBM	99.88%	84.63%	71.62%
ADA Boost	66.53%	65.93%	46.20%
SVM	74.30%	73.73%	61.22%
Tuned SVM	90.41%	79.84%	67.17%

Table 5 contains the confusion matrix for tuned GBM with error rates by class. There is significant misclassification in classes 1 (Spruce/Fir) and 2 (Lodgepole pine) and because 85% of the data is in these two classes, this dominates the overall correct classification rate and misclassification rate. The classification accuracies for classes 4, 5, and 7 are particularly high, all over 90% and two of them over 95%. Classes 3 and 6 have classification accuracies over 80% which is still much higher than the overall correct classification rate. The results of these first analyses suggests that with a training dataset that has equal representation from the seven classes it is not possible to get a correct classification rate significantly higher than 70%. Part of the problem here is the unusual partition of the dataset into training, validation, and test components with equal representation of the seven vegetation categories in both the training and

validation datasets, even though most of the data is in classes 1 and 2. This observation motivates the analyses of chapter 3 of my thesis.

Table 5

Confusion matrix of tuned GBM using Northness and Eastness

		Predicted Class							
True Class		1	2	3	4	5	6	7	% correctly classified
	1	154,515	34,580	133	0	4,050	564	15,838	73.69%
	2	54,517	185,335	6,284	58	24,506	8,657	1,784	65.92%
	3	0	256	27462	1367	434	4075	0	81.18%
	4	0	0	9	570	0	8	0	97.10%
	5	19	268	111	0	6,851	84	0	93.43%
	6	1	134	1,786	312	120	12,854	0	84.53%
	7	602	20	5	0	16	0	17,707	96.50%

CHAPTER 3

ANALYSES USING A NEW 60-20-20 PARTITION

3.1 Methods

Following the analysis completed on the original partition of the data, I decided to rerun the classification methods on a different partition of the data that reflected the different numbers of observations in the vegetation classes. Blackard and Dean (1998) used 60% of the smallest class with equal numbers from each of the other classes for their training set and 20% of the smallest class with equal numbers from all other classes as their validation set, with all remaining data being used as part of the test set, so the vast majority of the data was in the test set. I chose a simple 60-20-20 random partition of the whole dataset, which gave roughly matching *proportions* of observations in the individual classes relative to their proportion as part of the whole data set.

In doing this, I became aware of the fact that the partition used by Blackard and Dean (1998) has variables for which there is no variation within the training and validation sets. The variables Soil_Type7, Soil_Type15, and Soil_Type16 all had to be removed due to being consistent within either the training or validation set. The new 60-20-20 partition did not have any variables that were constant within their set.

The methodologies and process used to complete these analyses were the same as when working on partition the original partition of the data by Blackard and Dean (1998).

Due to processor limitations on the device used for computation, tuning of GBM and SVM on the new partition of the data has not been completed.

3.2 Results

The final results for these sets is included in the Table 6 below with the accuracies of the training ("resubstitution accuracies"), validation, and test sets all listed.

Table 6

Comparison of all methods with the accuracies for training, validation, and test data sets

	Percent Correctly Classified		
Methods	Training Set	Validation Set	Test Set
LDA	67.98%	68.30%	68.04%
QDA	66.02%	66.50%	66.20%
Classification Tree	99.00%	92.50%	92.63%
Random Forests	83.55%	83.98%	83.98%
GBM	67.10%	67.20%	67.05%
SVM	78.95%	78.96%	78.63%
Ada Boost	69.65%	69.71%	69.56%

Comparing the test set accuracies of this new 60-20-20 partition to the results on the original partition used by Blackard and Dean (Blackard & Dean, 2000) we get the following table which shows a dramatic increase in accuracy.

Table 7

Comparison of methods' accuracies on the test set between the original partition and the new 60-20-20 partition

Method	Original Partition	60-20-20 Partition	Increase from Original to 60-20-20 Partition
ANN	70.58%	–	–
LDA	58.31%	68.04%	9.73%
QDA	52.95%	66.20%	13.26%
Classification Tree	63.22%	92.63%	29.41%
Random Forests	67.57%	83.98%	16.41%
GBM	49.20%	67.07%	17.87%
Tuned GBM	71.62%	–	–
SVM	61.22%	78.64%	17.41%
Tuned SVM	78.64%	–	–
Ada Boost	46.20%	70.67%%	24.47%

The smallest gain was in LDA and that alone was nearly a 10% increase in accuracy by using a straight 60-20-20 partition over the equal numbers of each class for the training and validation sets used in the original analysis of the data. By taking a simple random sample from the data, the accuracy of the more traditional methods increased to a level comparable with the Artificial Neural Network created by Blackard and Dean.

A single classification tree did spectacularly well, increasing its accuracy by more than 20%. Using the 1-SE rule I determined to use a *cp* value of 0.000039, which is very small, but performed incredibly well with an overall accuracy of 92.63% and much higher accuracies on vegetation classes 1 and 2 than with the original partition of the data. The confusion matrix shows that even these good results still have the biggest issue differentiating between classes 1 and 2. The confusion matrix is below.

Table 8

Confusion matrix on test data of classification tree

		Predicted Class							
		1	2	3	4	5	6	7	% correctly classified
True Class	1	39,335	2,789	2	0	44	6	208	92.81%
	2	2,672	53,360	193	3	267	111	33	94.21%
	3	3	178	6,498	54	24	304	0	92.03%
	4	0	3	88	463	0	17	0	81.09%
	5	55	343	31	0	1,531	8	0	77.79%
	6	12	167	326	25	4	2,971	0	84.76%
	7	228	38	0	0	0	0	3,808	93.47%

3.3 Classification Tree Partition Reduction

Due to the single classification tree giving unexpectedly accurate results, particularly in comparison to other tree-based classifiers that typically outperform single trees, I carried out additional analyses determine how much of a reduction in size of the training set would be required to reach the same level of accuracy as the other methodologies. To do this, rpart was run on randomly generated partitions with training sets equal to 50%, 40%, 30%, 20%, 10%, 2% and 1.9%. The final two were chosen to surround the overall percentage of the partition chosen by Blackard and Dean (2000)for their original analysis using LDA, QDA, and ANN. The results for these trees are given in Table 9 below.

Table 9

Comparison of the accuracies of a single classification tree (using the 1-SE rule to choose *cp*) for various sized training, validation, and test data sets

Partition Percentages (Training-Validation-Test)	Training Percentage Correctly Classified	Validation Percentage Correctly Classified	Test Percentage Correctly Classified
60-20-20	99.00%	92.50%	92.63%
50-25-25	97.24%	92.03%	91.90%
40-30-30	97.20%	91.17%	90.99%
30-35-35	95.95%	89.79%	89.67%
20-40-40	95.12%	87.81%	87.74%
10-45-45	93.27%	84.06%	83.90%
2-49-49	95.28%	75.16%	75.07%
1.9-49-49.1	79.54%	73.63%	73.80%

As can be seen from the table, and by recalling the results of the ANN model created by Blackard and Dean (1998), there's a high chance the high accuracy achieved by ANN in comparison to other statistical methods may have been due in part to the choice of training data. A single classification tree is outperforming the tuned ANN with equally small training sets (the original training set was just over 1.9% of the total dataset).

I also looked at the influence of the *cp* value on the results of the classification tree. Starting with the original *cp* value, and doubling it until the accuracy on the test set was comparable to the results of random forests. Doing so showed that I could have needed to take the *cp* value from the one chosen ($5 * 10^{-6}$) to one 32 time larger ($1.6 * 10^{-4}$) to get results comparable to those of Random Forests as shown in the table below.

Table 10

Percent correctly classified by a single classification tree as the *cp* was doubled on the new 60-20-20 partition

cp value	Percentage Correctly Classified of Test Set
0.000005	92.91%
0.00001	92.67%
0.00002	91.29%
0.00004	89.06%
0.00008	85.90%
0.00016	81.95%

CHAPTER 4

COMBINING CLASSIFIERS

4.1 Analyses

Based on the 60-20-20 partition, I ran further analysis to determine if the various methods were misclassifying the same observations or if it was unique to the method. The results of that analysis are summarized below.

Table 11

Counts of correctly and incorrectly classified observations for 4 methods on the 60-20-20 partition

Method	Number Correct	Number Incorrect	Percent Correct
Tree	107,966	8,236	92.91%
Random Forest	97,593	18,609	83.99%
SVM	91,378	24,824	78.64%
GBM	77,931	38,271	67.07%
Ada Boost	82,117	34,085	70.67%

Table 12

Counts of how many times a given observation was misclassified by the four methods

Number of times mis-classified	Count	Percent of Total	Cumulative Percent
0	68,149	58.65%	58.65%
1	12,676	10.91%	69.56%
2	13,039	11.22%	80.78%
3	7,385	6.36%	87.14%
4	11,649	10.02%	97.16%
5	3,304	2.84%	100%

The worst-case scenario being that those misclassified 3 or more times as the same incorrect class, a straight vote of these four methods would produce accuracies of 80.78% , which is substantially less than the accuracy of the single classification tree. Should those that were

misclassed be misclassified as different classes, it would be possible to achieve up to 87.14% accuracy by voting. Given the high accuracy of a single classification tree, this would perhaps not be the best option to pursue. However, due to these results a similar analysis was completed using the results of the original partition (equal numbers for the training and validation set based on 60% and 20% of the smallest class respectively). Those results are summarized in the following two tables. Since Ada Boost returned such poor results, I decided to replace it with LDA which performed better for the purposes of this voting.

Table 13

Counts of correctly and incorrectly classified observations by four methods on the original partition

Method	Number Correct	Number Incorrect	Percent Correct
Tree	358,168	207,724	63.29%
Random Forest	382,316	183,576	67.56%
SVM - tuned	380,124	185,768	67.17%
GBM - tuned	405,294	160,598	71.16%
LDA	329,972	235,920	58.31%

Table 14

Counts of how many times a given observation was misclassified by the four methods

Number of times mis-classified	Count	Percent of Total	Cumulative Percent
0	211,465	37.37%	37.37%
1	111,488	19.70%	57.07%
2	64,282	11.36%	68.43%
3	53,001	9.37%	77.80%
4	53,752	9.50%	87.30%
5	71,904	12.71%	100%

Also, in this case, voting classifiers does not seem to help the predictive accuracy. The worst-case scenario that each time an observation was misclassified it was consistently misclassified as the same class would give an overall accuracy of 68.43%. The best we could get, should those that were misclassified be misclassified as a different class each time, would give at best an overall accuracy of 77.80%. This range indicates that a voted prediction of each observation by these classifiers would give a comparable result to that of the ANN created by Blackard and Dean (1998).

Another option for voting would be some sort of weighted votes where the weight would be inversely related to the error rate of the particular method, giving higher weight to classifications that came from a highly accurate method. This could potentially increase the overall accuracy to something slightly higher than the ANN result.

4.2 Results

It seems that voting would improve the results on the original partition of the data, however, for the new 60-20-20 partition, the single classification tree still seems the best choice.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND FUTURE WORK

In conclusion, it seems possible that a simple random sample partition would have prevented the superiority of ANN. A tuned GBM was the best performer on the type of partition used by Blackard and Dean. And with a straight partition, a single classification tree consistently performed better.

I began by replicating the results of Blackard and Dean (1998) for LDA and QDA on the cover type data and then applied a number of classification methods that have emerged from the statistics and computer science literature in the last 35 years. My results suggested that with the original partition of the data it was not possible to significantly improve on the classification accuracy obtained by Blackard and Dean (1998) using an artificial neural network. The best classification accuracy I obtained was for tuned gradient boosting machines at 71.62% compared to 70.58% for the ANN of Blackard and Dean (1998).

In examining the confusion matrix from the GBM classification it became clear that most of the misclassifications were for classes 1 (Spruce/Fir) and 2 (Lodgepole Pine), which comprise over 85% of the data. The selection of the training and validation data by Blackard and Dean (1998) with equal numbers of observations of the 7 vegetation classes works well for the smaller classes, but very poorly for the two most common classes.

So, I randomly partitioned the dataset with 60% of all observations making up the training data, 20% the validation data, and the remaining 20% the test data. In the training dataset that I selected the numbers of observations in the different vegetation classes mirrored the dataset as a whole. I reran all the classification methods, with tuning where appropriate, and found much higher classification accuracies for the populous vegetation classes 1 and 2. For

some of the smaller classes the classification accuracies were not quite as high as they were with the training data selected by Blackard and Dean (1998).

In the second batch of analyses, I noticed that the overall prediction accuracy for a single classification tree was especially high, 93.67% on the test data. This is surprising because normally ensemble tree classifiers do better than a single tree. I do not have a good explanation of this result. I decided to see what the effect of reducing the size of the training data would be and found that accuracies of 90% or higher were achieved with a single tree for training datasets as small as 20% of the data. I chose a training dataset in this proportional manner that was the same size as the original training data of Blackard and Dean (1998) and found that on these data a single classification tree was a more accurate predictor of vegetation class that the ANN of Blackard and Dean (1998) using their training data.

Finally, in running different classification methods I saw that the predictions were not quite the same even for methods that had comparable classification accuracies. I decided to "vote" the results from 5 classifiers to see if increased predictive accuracy could be obtained, particularly for the original partition of the data. I found that this voting has the potential of improving the overall accuracy greatly to make it comparable to the ANN created by Blackard and Dean (1998).

Some things that I have not resolved in my thesis work and which could be the subject of future work include figuring out why a single tree does so well compared to ensembles of trees, and the effect of training dataset size on all the other classification methods. (I only explored this for classification trees). I think it would also be valuable to apply modern neural net packages to see how ANNs compare with other methods on a proportional partition of the data.

And finally, determining the most useful voting method would be of value, as either a straight vote or weighted vote based on the overall accuracy of the particular method.

REFERENCES

Alfaro, E., Gamez, M., & Garcia, N. (2013). adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, 1-35.

Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA, United States: University of California, School of Information and Computer Sciences.

Blackard, J. A. (1998). Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types. *Ph.D. dissertation*. Fort Collins, Colorado: Department of Forest Sciences, Colorado State University.

Blackard, J. A., & Dean, D. J. (1998). Comparative Accuracies of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables. *Second Southern Forestry GIS Conference* (pp. 189-198). Athens, GA: University of Georgia.

Blackard, J. A., & Dean, D. J. (2000). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 131-151.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall.

Carrion, S. (2017, March 22). *A simple explanation of what is LDA classification*. Retrieved from Stack Overflow: http://stackoverflow.com/questions/35489544/a-simple-explanation-of-what-is-lda-classification

Chang, J. (2015). lda: Collapsed Gibbs Sampling Methods for Topic Models. *R package version 1.4.2*. https://CRAN.R-project.org/package=lda.

Cortes, C., & Vapnik, V. (1995). Support-vector Networks. *Machine Learning*, 273-297.

Culp, M., Johnson, K., & Michailidis, G. (2016). ada: The R Package Ada for Stochastic Boosting. *R package version 2.0-5*. https://CRAN.R-project.org/package=ada.

Cutler, R., Edwards Jr., T., Beard, K., Cutler, A., Hess, K., Gibson, J., & Lawler, J. (2007). Random Forests for Classification in Ecology. *Ecology*, 2783-2792.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, pp. 179-188.

Fisher, R. A. (1938). The Statistical Utilization of Multiple Measurements. *Annals of Eugenics*, pp. 376-386.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 256-285.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 119-139.

Friedman, J. (2001). Greedy Function Approximation: The Gradient Boosting Machine. *Annals of Statistics*, 1189-1232.

Gama, J., Rocha, R., & Medas, P. (2003). Accurate decision trees for mining high-speed data streams. *KDD '03*, 523-528.

Greund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Sciences*, 119-139.

Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Hunt, T. (2016). caret: Classification and Regression Training. *R package version 6.0-73*. https://CRAN.R-project.org/package=caret.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 18-22.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (n.d.). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *R package version 1.6-7*. https://CRAN.R-project.org/package=e1071.

R Core Team. (n.d.). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: URL http://www.R-project.org/.

Ridgeway, G., & with contributions from others. (2015). gbm: Generalized Boosted Regression Models. *R package version 2.1.1*. http://CRAN.R-project.org/package=gbm.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. *R package version 4.1-10*. https://CRAN.R-project.org/package=rpart.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York, NY: Springer-Verlag, Inc.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition.* NY: Springer.