A Comparison of Classification Methods for Forest Cover Type

Alvin Au
Jared Eccles
André Haynes
Timothy Thatcher
Yicheng Zhang

November 30, 2012

Abstract

This study evaluates the performance of seven classification techniques on the problem of predicting forest cover type from cartographic data. The data was obtained from the UCI Machine Learning Repository. Techniques are evaluated based on the metric of correct classification rate for each cover type.

Contents

1 Introduction

The goal of this study is to predict the dominant forest cover type (CT) from cartographic data, for forested areas in Roosevelt National Forest in Colorado USA. These areas have experienced relatively little direct human management disturbances, thus the current composition of CT within them are primarily a result of natural ecological processes, rather than the product of active forest management.

Observations correspond to 30m $\times$ 30m grids of the forest. For each grid there is one response variable for the dominant CT, and twelve predictors that detail the features of the area. These include:

1. **Ele** - Elevation (m)
2. **Asp** - Aspect (azimuth from true north)
3. **Slo** - Slope (degrees)
4. **HDtH** - Horizontal distance to nearest surface water feature (m)
5. **VDtH** - Vertical distance to nearest surface water feature (m)
6. **HDtR** - Horizontal distance to nearest roadway (m)
7. **Hs9** - Measure of incident sunlight at 09:00 h on the summer solstice
8. **Hs12** - Measure of incident sunlight at noon on the summer solstice
9. **Hs15** - Measure of incident sunlight at 15:00 h on the summer solstice
10. **HDtFP** - Horizontal distance to nearest historic wildfire ignition point (m)
11. **WA** - Wilderness area designation (categorical)
12. **ST2** - Modified version of the soil type variable (categorical)

There are seven CT used in this study. They are classified as Lodgepole Pine (1), Spruce/Fir (2), Ponderosa Pine (3), Douglas-fir (4), Aspen (5), Cottonwoood/Willow (6), and Krummholz (7). These represent the primary dominant tree species currently found in the Roosevelt National Forest.

This study will investigate the performance of classification techniques, based on the correct classification rate at each CT level.

2 Methodology and Data Exploration

2.1 Data Exploration

Preliminary data exploration was conducted on the full dataset of 581,012 observations. The frequency plot of CT, in Figure 1 indicates types 1 and 2 dominate the data, and CT 4 is a rare class. This issue of the rare class will have implications for the sampling design.

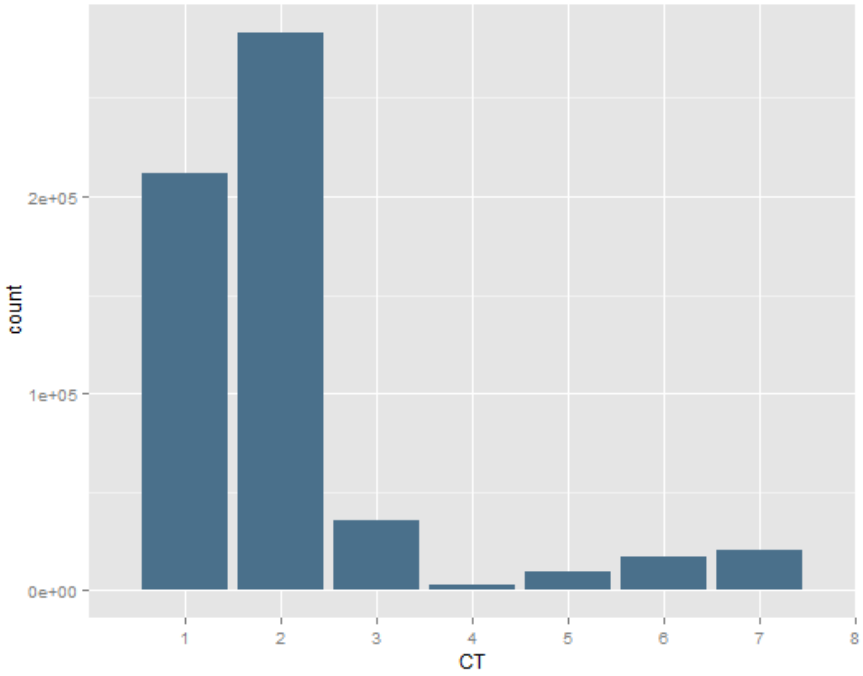

Figure 1: Histogram of the cover type frequencies.

Figure 6 in Appendix A shows the distribution of CT for each predictor. The Ele, Asp, and HDtR predictors show the highest degree of linear separability for CT. Therefore, subspaces spanned by these predictors may provide the best discrimination rules for separating the CT.

2.2 Sampling Design and Methodology

The continuous predictors were scaled to the interval $[0, 1]$ before analysis. This scaling does not represent any information loss and was done to boost the performance of some algorithms for the techniques used.

Due to the imbalanced nature of the data for CT, a uniform sampling specification was used. The training data was constructed by sampling 1620 observations from each CT - 1620 observations correspond to 60% of the rarest class. The validation data set was comprised of 540 observations from each class. The validation data was used in conjunction with the training data to tune models. The remainder of the data set was allocated as test data, and was used for final predictions. The performance of the models was assessed by the mean correct classification rate and mean miss-classification matrix based on 100 test set samples generated via SRSWOR.

CT 1 and 2 dominated the test data compared to the other CT. By sampling relatively few observations for training and validation, important but subtle information on the distribution of these CT with given predictors could be missed. Thus, this could introduce bias to misclassification of 1 and 2 in the techniques used. However, because the data is imbalanced, it was believed this sampling design would produce better overall performance across the techniques. Previous studies of this dataset have also used this design with some success [2]. The wilderness area (WA) categorical

predictor was dropped because the information conveyed wouldn't be relevant to the classification techniques. The soil type predictor (ST1) of 40 levels was reduced to an 11 level predictor based on climate and geographical information related to the original levels.

3 Classification Methods

3.1 Bayes Classifier

The Naive Bayes Classifier (NBC) applies Bayes Theorem, and predictor distributional assumptions to classify a set of observed predictors $\vec{P} = \{P_i\}_{i=1}^{n}$, according to the following classification rule:

$$\text{classify}(\vec{P}) = \arg\max_c Pr(CT = c) \prod_{i=1}^{n} Pr(\vec{P}_i = \vec{p}_i \mid CT = c)$$

NBC assumes that the conditional distributions of the predictors given the response, are independent. It also imposes assumptions on the distribution of $(\vec{P}_i \mid CT = c)$, and the prior $Pr(CT = c)$.

The NBC was applied using (i) all predictors of CT, and (ii) on a reduced set of predictors which were most uncorrelated, and whose distributions were easiest to specify given the knowledge of the data. The reduced set corresponded to predictors Ele, Slo, and Hs15. These predictors were assumed to be normally distributed, since their densities appeared to be unimodal, symmetric and bell shaped (see Figure 6 in Appendix A).

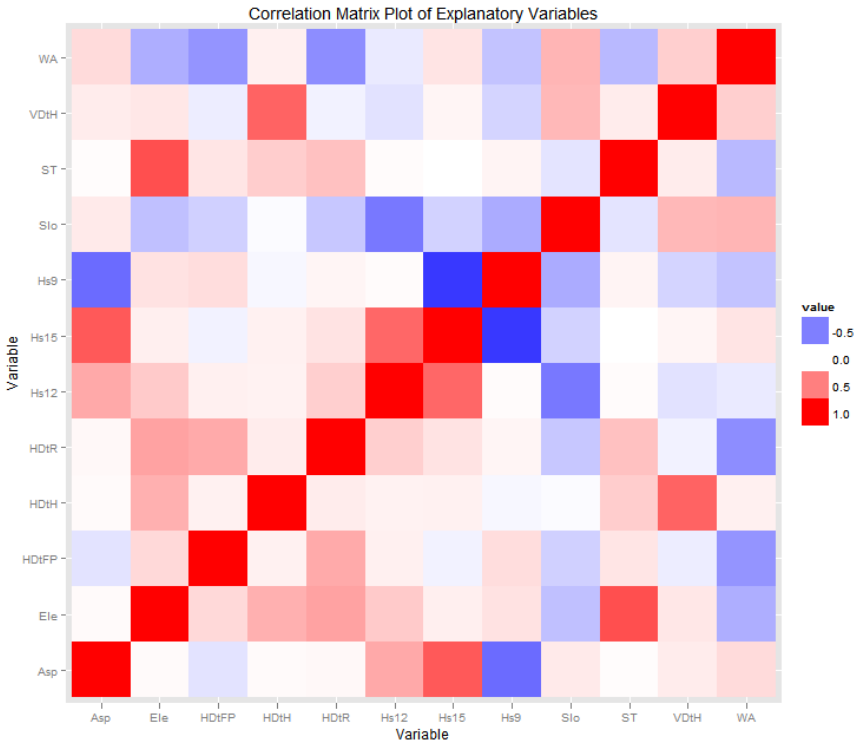

Figure 2: Heat map of the correlation matrix for the continuous predictors.

To gauge the predictive performance of the NBC, the model was assessed using two different priors:

1. The uniform prior, where $Pr(CT = c)$ is assigned equal weight for each CT class
2. The proportional prior, where $Pr(CT = c)$ is assigned the true proportion of class CT in the full dataset

The following is a table of the average correct classification rate for NBC, under the prior assumptions:

Prior Specification	**Reduced**	**Full**
Uniform Prior	0.46	0.45
Proportionally weighted Prior	0.65	0.61

Table 1: Naive Bayes classifier accuracy (%) under different prior assumptions.

Performance of the NBC was observed to be most optimal on the reduced predictor set. However, these results indicate that correctly specifying the prior has a more measurable impact on performance than satisfying the independence assumptions. In practice, the NBC is robust to violations of the conditional independence assumption [4].

3.2 Discriminant Analysis

3.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is one of the more classic statistical approaches to classification. The technique finds the linear combination of predictors that best separates classes, leading to linear decision boundaries. The key assumptions in this classification method is that the classes have the same covariance matrix for predictors and all predictors arise from a Gaussian distribution. In practice, these strong assumptions are rarely met.

The final model for LDA was built on a subset of the original predictors. The final model consisted of:

- Ele, Asp, HDtFP, HDtH, Hs12, Hs15, Slo, and ST2

The first four predictors appeared to be the closest to the Gaussian distribution (unimodal, symmetric, and bell shaped) based on Epanechnikov kernel density estimates shown in figure 6 in the appendix. Although the assumption of normality was violated, the predictors Hs12, Hs15, and Slo were also included because they had a positive effect on predictive accuracy of the LDA model. Similarly, the categorical predictor ST2 was included because it contained geological information that affects vegetation growth. As stated in Blackard's paper, categorical variables are often helpful to include in practice [2]. Based on the data set, class weights generated from the class proportions were included to improve predictive accuracy.

Predicted Class	**True Class** 1	2	3	4	5	6	7
1 Spruce/Fir	21.81	9.53	0.00	0.00	0.08	0.00	0.60
2 Lod. Pine	11.12	36.86	0.21	0.00	1.16	0.39	0.04
3 Pon. Pine	0.03	0.96	3.72	0.12	0.09	0.84	0.00
4 Willow	0.00	0.01	0.26	0.06	0.00	0.08	0.00
5 Aspen	0.01	0.34	0.01	0.00	0.02	0.01	0.00
6 Doug. Fir	0.03	1.49	1.78	0.02	0.02	1.44	0.00
7 Krummholz	3.90	0.26	0.00	0.00	0.00	0.00	2.68

Table 2: Average miss-classification matrix (%) for LDA classification.

The miss-classification matrix shown in table 2 indicates that most miss-classifications happen between CT 1 and CT 2. Class 1 is also often mistakenly predicted to be class 7. According to table 3, CT 2 and 7 were most accurately predicted ($> 75\%$) which were closely followed by class 1 and 3 with prediction rates above 60%. The remaining classes, 4,5, and 6, had relatively low accuracies for prediction. They are also the rarest classes and only make up approximately 5% of the data set. When CT proportions were factored in, the overall accuracy of LDA increased to approximately 66.6%.

Cover Type	1	2	3	4	5	6	7	**Overall**
Estimate	59.11	74.55	58.99	32.21	1.77	52.01	80.8	66.6

Table 3: Classification rate estimations by cover type (%) for LDA classification.

3.2.2 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is an extension of linear discriminant analysis. This classification method does not rely on the assumption of equal covariance matrices between classes, which leads to quadratic decision boundaries. The final QDA used all the continuous predictors with the exception of VDtH, since it did not lead to improved predictive accuracy. The categorical predictor ST was dropped due to numerical instability in the QDA solving method. Class weights were also used in the QDA model.

Predicted Class	**True Class** 1	2	3	4	5	6	7
1 Spruce/Fir	27.22	13.46	0.00	0.00	0.05	0	2.28
2 Lod. Pine	8.56	32.23	0.82	0.00	1.01	0.5	0.08
3 Pon. Pine	0.09	1.57	3.46	0.04	0.04	0.75	0.00
4 Willow	0.00	0.00	0.12	0.11	0.00	0.04	0.00
5 Aspen	0.14	0.65	0.00	0.00	0.25	0.00	0.00
6 Doug. Fir	0.06	1.44	1.59	0.04	0.04	1.46	0.00
7 Krummholz	0.84	0.08	0.00	0.00	0.00	0.00	0.95

Table 4: Average miss-classification matrix (%) for QDA classification.

The miss-classification matrix shown in table 4 indicates that most miss-classifications occur between CT 1 and CT 2 - the same as LDA. Contrary to LDA, CT 7 is often

mistakenly predicted to be CT 1. According to table 5, CT 1 and 2 were most accurately predicted (> 65%), closely followed by CT 3 and 4 with prediction rates of approximately 57%. CT 6 was similar in accuracy to LDA. CT 5 was more accurate with QDA, although it was still quite low. CT 7 lost a great deal of accuracy QDA.

In summary, when class proportions were factored in, the overall accuracy of QDA was approximately 65.6%. This was similar to LDA. However, there was disparity between the two methods with respect to predictive accuracies by CT tended to differ between the two methods.

Cover Type	1	2	3	4	5	6	7	**Overall**
Estimate	73.76	65.19	57.73	57.59	17.8	52.99	28.6	65.68

Table 5: Classification rate estimations by cover type (%) for QDA classification.

3.3 Multinomial Logistic Regression

Similar to LDA and QDA, multinomial logistic regression (MLR) is another traditional statistical method of classification. Like LDA, MLR produces linear decision boundaries, but the boundaries are based on the maximum expected membership probability among classes I.E. the prediction is the class with the highest probability of membership for a set of predictor values. Contrary to LDA, MLR makes no assumption about the prior distribution, where as LDA implicitly assumes that the priors are Bernoulli distributed. Due to the similar linear decision boundaries and weaker assumptions, it would be expected that MLR would be a more robust method and perform similarly to LDA [3].

The `multinom` function in the `nnet` package was used to fit a multiple logistic regression model for CT. This required scaling of the data for efficient computation and stability with `nnet`'s internal methods. The predictorswere chosen based on highest predictive accuracy using repeated models and validation sets:

- Ele, Asp, HDtH, VDtH, HDtR, Hs12, Hs15, HDtFP, and ST2

Predicted Class	**True Class** 1	2	3	4	5	6	7
1 Spruce/Fir	25.16	9.14	0.00	0.00	0.01	0.00	1.83
2 Lod. Pine	10.59	39.31	0.63	0.00	1.32	0.77	0.03
3 Pon. Pine	0.00	0.81	4.87	0.11	0.04	1.66	0.00
4 Willow	0.00	0.01	0.2	0.08	0.00	0.04	0.00
5 Aspen	0.00	0.03	0.00	0.00	0.00	0.01	0.00
6 Doug. Fir	0.00	0.11	0.29	0.01	0.00	0.28	0.00
7 Krummholz	1.14	0.04	0.00	0.00	0.00	0.00	1.46

Table 6: Average miss-classification matrix (%) for multinomial logistic regression based classification.

The miss-classification matrix in table 6 indicates that CT 1 and 2 are still commonly miss-classified between each other. MLR also miss-classifies class 1 as CT 7 and vice versa. Prediction accuracies are high (> 65%) for CT 1, 2, and 3, but the remaining

classes have a less than 50% prediction rate. Overall, the MLR had an accuracy above 70%, which was an improvement from LDA and QDA.

Cover Type	1	2	3	4	5	6	7	**Overall**
Estimate	68.19	79.49	81.32	39.26	0.25	10.17	43.95	71.16

Table 7: Classification rate estimations by cover type (%) for multinomial logistic regression classification.

3.4 Support Vector Machines

Support vector machines (SVMs) are one of the more modern techniques used in classification. The method produces a series of hyperplanes that partition the data by mapping it into a high dimensional space via kernel functions. Since much of the CT data is not linearly separable, a non-linear kernel, the radial basis function, was used for building SVM model. This produced non-linear decision boundaries.

The R package `e1071` contained the function `svm` for support vector machine classification. One of the advantages of support vector machines is that the method has a sparse solution regardless of the input dimension. This means that dimensionality reduction techniques often do not need to be applied before building SVM models, since computational complexity is dependent on the number of support vectors [1]. SVM's predictive performance was tested against each of the kernels offered in the `svm` function: linear, polynomial, radial basis, and sigmoid. Of the four, the radial basis kernel performed the best in validation tests. However, each instance of `svm` was costly in terms of time and was especially computationally intensive for prediction. This made extensive tuning of the SVM parameters infeasible due to time constraints. After testing *cost* parameter values between 1 and 1000 and several *gamma* parameter levels for more effective values, the default parameters were found to be the most accurate. However, these results are not unexpected since the `svm` function is optimized for the radial basis kernel [7], even if the kernel is suboptimal for the data.

Predicted Class	**True Class** 1	2	3	4	5	6	7
1 Spruce/Fir	26.45	8.70	0.00	0.00	0.07	0.01	0.95
2 Lod. Pine	9.49	39.41	0.46	0.00	1.06	0.60	0.02
3 Pon Pine	0.01	0.75	5.11	0.07	0.05	1.27	0.00
4 Willow	0.00	0.00	0.11	0.12	0.00	0.04	0.00
5 Aspen	0.01	0.13	0.00	0.00	0.20	0.00	0.00
6 Doug. Fir	0.02	0.37	0.30	0.01	0.01	0.85	0.00
7 Krummholz	0.92	0.08	0.00	0.00	0.00	0.00	2.35

Table 8: Average miss-classification matrix (%) for support vector machine classification.

Table 8 again shows that CT 1 and 2 are difficult to classify. However, SVM performed well relative to the other classification methods in this respect. In fact, SVM acquired the highest overall accuracy among all methods. Table 9 indicates that the first four CT, as well as CT 7, have a high prediction accuracy with most classes

over 70%. However, CT 5 and 6 are still troublesome, with accuracies of 30% and below.

Cover Type	1	2	3	4	5	6	7	**Overall**
Estimate	71.67	79.69	85.35	63.10	14.42	30.71	70.73	74.49

Table 9: Classification rate estimations by cover type (%) for support vector machine classification.

3.5 Tree Based Methods

Decision trees are well suited to many classification problems. Due to their simplicity and robustness to violations of any assumptions about the underlying distribution of the data [9]. Simple decision trees were used to classify CT under various model conditions. The following table shows the initial decision trees investigated, and the corresponding overall classification rate:

Tree Type	**Overall Classification Rate**
Simple tree, full model, no prior specified	46%
Simple tree, full model, prior specified	69%
Boosted simple tree **	46%
Bagged simple tree **	52%

Table 10: A comparison of different tree-based models.

After pruning the simple tree (with prior specified), the overall classification rate was improved to 72%. The trade-off for gaining a better average classification rate from using the priors in simple classification trees, is a decline in the classification of the rare classes: without using the priors, the mean estimate over 100 samples for classes 4,5 and 6 were 84%, 83%, and 42% respectively; with the inclusion of the priors, these estimates dropped to 13%, 9.7%, and 19.7%, but the method significantly improved prediction on the majority classes.

The random forest technique was also used, giving a average overall classification rate of 73%. This compared favourably with trees, since the classification accuracy of the rare classes was not sacrificed in favour of this high rate.

Cover Type	**Simple Tree without prior**	**Simple Tree with Prior**	**Random Forest**
1 - Spruce/Fir	58.17	69.50	75.87
2 - Lodgepole Pine	30.34	79.09	67.60
3 - Ponderosa Pine	60.70	83.41	80.99
4 - Cottonwood/Willow	84.38	13.27	97.19
5 - Aspen	83.34	9.78	95.00
6 - Douglas-fir	42.00	19.76	86.35
7 - Krummholz	87.64	62.70	95.99
Overall	45.49	72.54	73.35

Table 11: Classification rate estimations by cover type (%) for different tree methods.

3.6 Artificial Neural Networks

The motivation for classifying the response using artificial neural networks (ANN) is the fact that the data is not linearly separable. The `nnet` library in R was used to construct the ANN, implying that all ANN models that were produced contained a single hidden layer. In validation, a 20-30% increase in classification accuracy resulted from omitting the categorical variable ST. For this reason, it were omitted from the final ANN model. The remainder of the continuous predictors were scaled between 0 and 1 to allow for computation optimality in the `nnet` function.

Due to the guess-and-check nature of the tuning parameters for ANN, several thousand artificial neural networks were created with varying *decay* values ranging from 0 to 0.3 at varying node levels in the hidden layer. It was concluded that a *decay* value of 0.005 performed best on the validation set, with an average validation classification accuracy of 72%.

To identify the optimal number of nodes in the single hidden layer in the ANN, a simple guess-and-check method was required. Models were built with 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130 and 140 nodes; at each node level 100 models were built and their performance on the validation dataset was assessed. Node levels were chosen in steps of 10 nodes so that a large range of values could be covered, while minimizing required computing time.

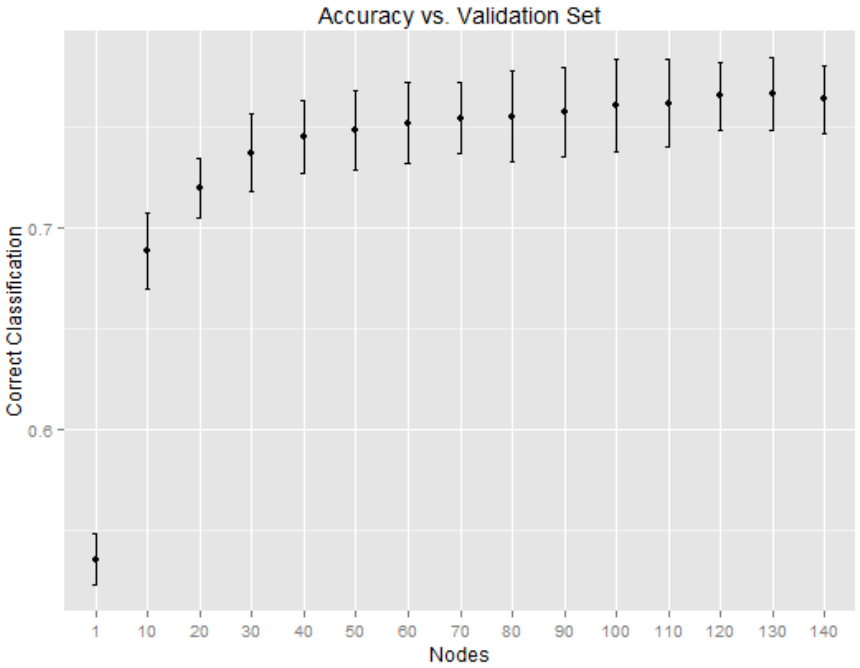

Figure 3: Accuracy versus the validation set for potential ANN models.

The plot shows that the initial increase from a simple ANN to a more complex (higher node count) yields large gains in accuracy. However, beyond 60 nodes the marginal gains in accuracy relative to the increased computational cost were considered irrelevant. These results are in line with the argument that D.J. Hand makes regarding classification, and more specifically how the performance of a 'simple' model

will compare to a more complex tree model [5]. Due to these results, it was decided that an ANN with 60 nodes (75.2% accuracy) and 1087 weights using the logistic activation function would be used for simplicity's sake. With the final ANN model chosen (60 nodes, *decay* value = 0.005), the model was run against 100 test sets, and resulted in an estimated correct classification of 62.6% with 95% confidence interval (60.6, 64.5).

Predicted Class	**True Class**						
	1	2	3	4	5	6	7
1 Spruce/Fir	24.69	10.96	0.00	0.00	0.01	0.00	0.2
2 Lod. Pine	6.75	27.13	0.09	0.00	0.08	0.04	0.01
3 Pon. Pine	0.04	1.79	4.2	0.01	0.03	0.42	0.00
4 Willow	0.00	0.04	0.31	0.18	0.00	0.11	0.00
5 Aspen	1.31	6.98	0.17	0.00	1.09	0.06	0.00
6 Doug. Fir	0.13	2.15	1.17	0.01	0.03	2.04	0.00
7 Krummholz	4.13	0.61	0.00	0.00	0.00	0.00	3.03

Table 12: Average miss-classification matrix (%) for ANN classification.

We see that approximately 18% (6.75% + 10.96%) of the total miss-classifications between CT 1 and 2. Therefore, $18/(100-62.6) = 48.1\%$ of the total misclassifications come from confusion between the first two cover types. This is likely due to the dominance of CT 1 and CT 2 in the test set, which is much larger and unequal in proportion compared to the validation set. CT 1 and 2 are very similar in their attributes as we see in the distribution of the individual covariates in Appendix A. This can explain the frequent misclassification between these two types.

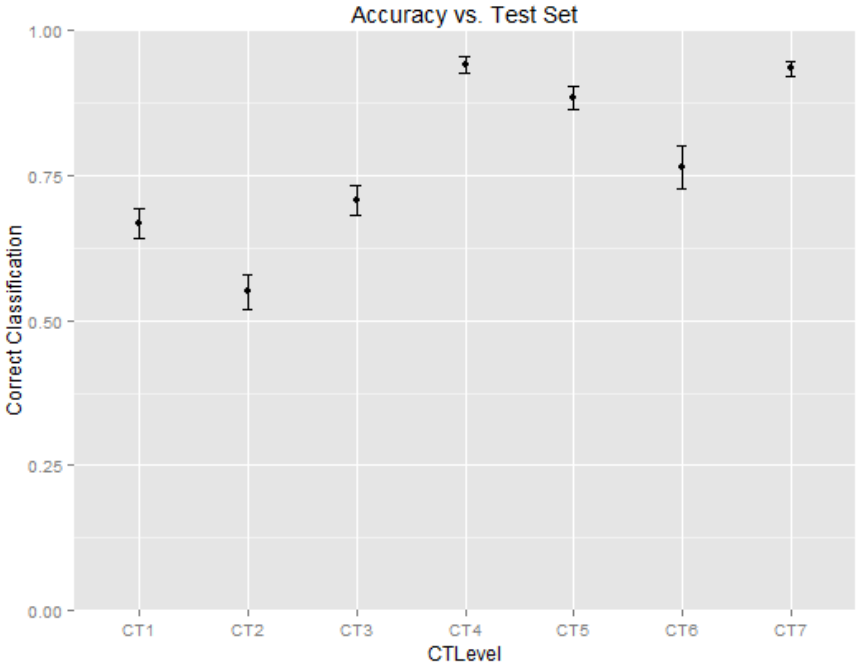

Figure 4: Prediction accuracy over the test set by cover type.

Finally, figure 4 is a plot of the models performance against each response level of CT. It reiterates the problem seen in the miss-classification matrix, that due to CT 1 and CT 2 being misclassified with each other, their respective overall classification rates drop to be the least successful of all of the levels. Conversely, the rare class CT 4 is classified the best under ANN with success rate of approximately 94%, when less than 1% of the data is of type CT 4.

4 Conclusion

We observed frequent misclassification in majority classes CT 1 and 2. As mentioned in the methodology, this could be due to the sampling design which over sampled for the minority classes. Thus, many of the models may have over fit to these minor classes and reduced overall accuracy. No one model performed well in discriminating between all seven CT levels. Overall, SVM gave the most accurate classifications, but this came at considerable computational cost for prediction. The classical statistical classification techniques like MLR and LDA performed well when considering overall speed, accuracy, and interpretability of the model. Overall, the techniques investigated in this paper outperformed the methods used in previous studies [2].

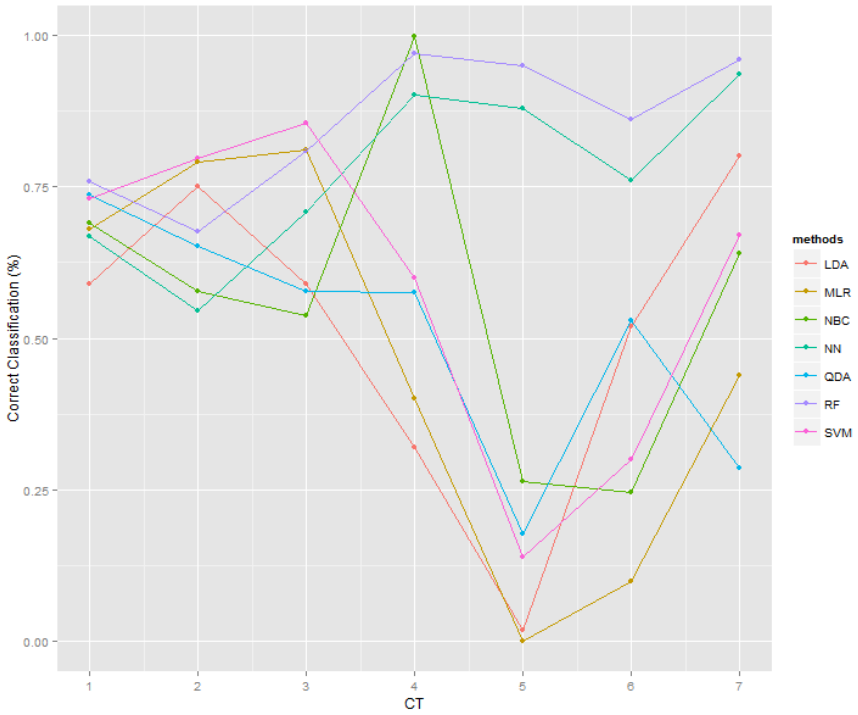

Figure 5: Comparison of correct classification by techniques used in study

A Additional Figures

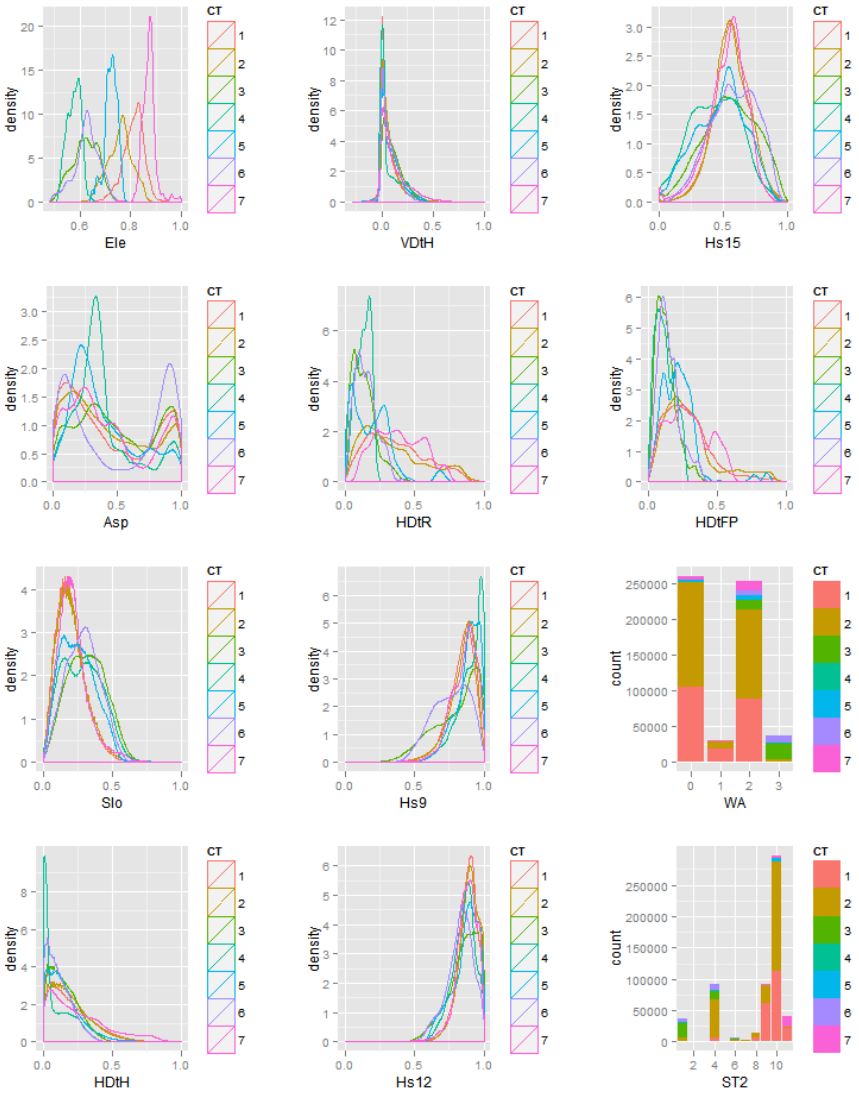

Figure 6: The distribution of each predictor by cover type.

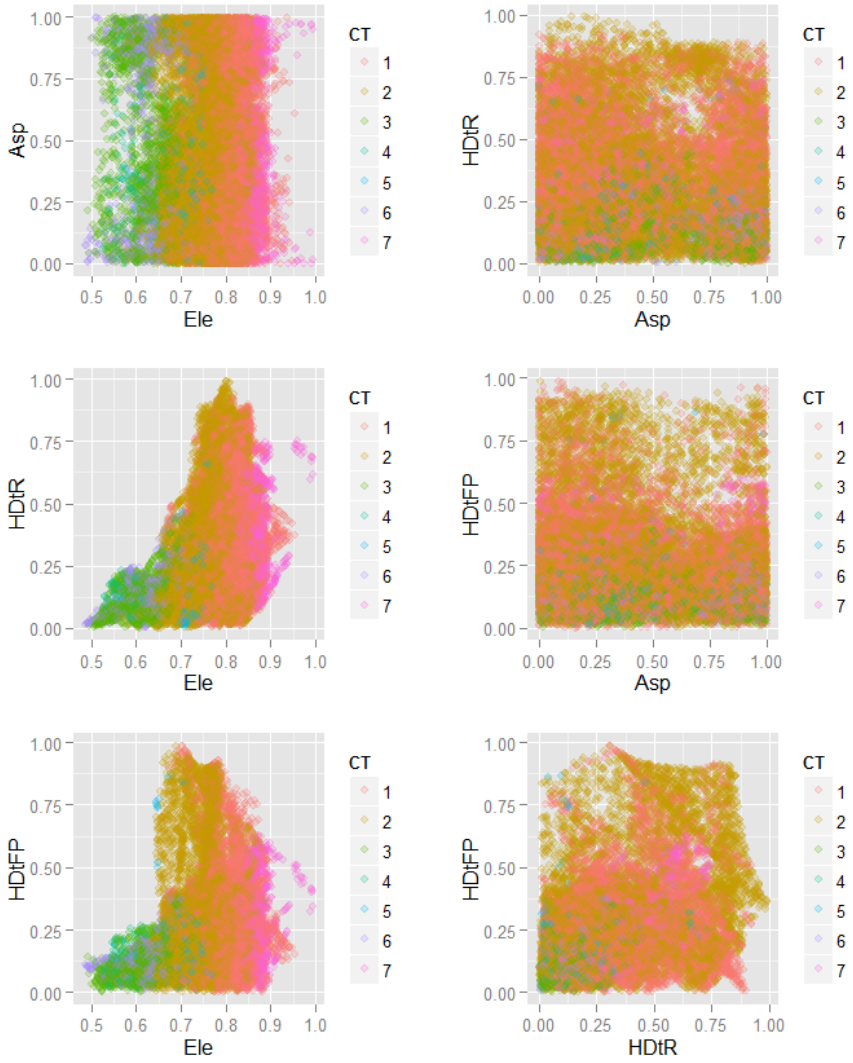

Figure 7: Pairs plot of covariates most similar to a Gaussian distribution.

References

[1] Auria, L., and Moro, R. A., 2008: Support Vector Machines (SVM) as a Technique for Solvency Analysis. *German Institute for Economic Research*, research paper no. **811**.

[2] Blackard, J. A., and Dean, D. J., 1999: Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables. *Computers and Electronics in Agriculture*, **24**, 131-151.

[3] Blas, M., Pohar, M., and Turk, S., 2004: Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. *Advances in Methodology and Statistics*, **1**, 143-161.

[4] Domingos, P., and Pazzani, M., 1997: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, **29**, 103130.

[5] Hand, D. J., 2006: Classifier Technology and the Illusion of Progress. *Statistical Science*, **21**, 1-15.

[6] Hastie, T., Tibshirani, R., and Friedman, J., 2001: The Elements of Statistical Learning. *Springer New York Inc., New York, NY, USA*, 745 pp.

[7] Meyer, D., 2012: Misc Functions of the Department of Statistics (e1071). *CRAN* [Available online at http://cran.r-project.org/web/packages/e1071/e1071.pdf]

[8] University of California, Irvine, cited 2012: Forest Cover Type Data Set. [Available online at archive.ics.uci.edu/ml/datasets/Covertype].

[9] Vaughn, B. K., and Wang, Q., 2008: Classification based on tree-structured allocation rules. *Journal of Experimental Education*, **76**, 315-340.