### **Forest Cover-Type Prediction**

**Abstract:**

Predicting forest cover type in forests and natural reserves provides an advantage in the conservation and management of nature in the event of a disaster. The process of measuring and recording the cover types is time-consuming and costly in some situations. In these situations, predictive models provide an alternative method for obtaining suck data. In this study, we aim to compare various techniques for predicting forest cover types from cartographic variables using various classification algorithms.

**Data and Features:**

The study area is situated in the Roosevelt National Forest of Northern Colorado. The data was gathered from the US Forest Service (USFS) Region 2 Resource Information System (RIS) data. These areas have minimal human footprint so the cover types are a result of the ecological process. The observations are taken from 30m by 30m patches of forest that are classified as one of seven cover types:

* Spruce/Fi, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz

There are four wilderness areas here, Neota, Rawah, Comanche Peak and Cache la Poudre. The major tree species are found in Neota which probably has the highest mean elevational value of 4, It is typically covered by Spruce/Fur. while Rawah and Comanche Peak is covered by lodgepole pine, spruce/fir and aspen. Cache la Poudre would tend to have Ponderosa pine, Douglas-fir, and cottonwood/willow.

The various features of the dataset are:

|  |  |  |
| --- | --- | --- |
| **Name** | **Measurement** | **Description** |
| Elevation | meters | Elevation in meters |
| Aspect | azimuth | Aspect in degrees azimuth |
| Slope | degrees | Slope in degrees |
| Horizontal Distance To Hydrology | meters | Horz Dist to nearest surface water features |
| Vertical Distance To Hydrology | meters | Vert Dist to nearest surface water features |
| Horizontal Distance To Roadways | meters | Horz Dist to nearest roadway |
| Hillshade 9am | 0 to 255 index | Hillshade index at 9am, summer solstice |
| Hillshade Noon | 0 to 255 index | Hillshade index at noon, summer soltice |
| Hillshade 3pm | 0 to 255 index | Hillshade index at 3pm, summer solstice |
| Horizontal Distance To Fire Points | meters | Horz Dist to nearest wildfire ignition points |
| Wilderness Area (4 binary columns) | 0 (absence) or 1 (presence) | Wilderness area designation |
| Soil Type (40 binary columns) | 0 (absence) or 1 (presence) | Soil Type designation |
| Cover Type | Classes 1 to 7 | Forest Cover Type designation - Response Variable |

**Related Work:**

There have been many research and various methods to predict and classify forest cover types that can help in further research of forest fire susceptibility, the spread of the infestation [2], and other deforestation problems.

In our recent studies on forest cover type prediction, they have used a dataset from the UCI Machine Learning Repository where 15120 samples of 30\*30 patches of Roosevelt National Forest [1]. In these implementations of predicting forest cover type, they have used 54 cartographic features [1] and also by removing the 44 Boolean features and making them into dimensions of 10 features of the data. The features and labels include the elevation, hydrologic, soil, and sunlight and the 7 cover types. In the study, they have implemented a variety of classification algorithms such as Multi-class support vector machine and K-Means Clustering using Principal Component Analysis. Principal component analysis is a method to reduce the dimensions of the data by making the mean to zero and variance to one. This has been visualized in three dimensions for 8000 samples. When applying the data with reduced dimensions, the runtime of multi-class SVM has also been reduced but the loss of the variance will decrease the performance.

In Multi-Class SVM [1], the data is trained using the Boolean and without Boolean information where the 7 forest cover types are classified into 21 separate binary classifiers to predict the cover types of trees in the wooden area. After training the model it has been tuned with two hyper-parameters to produce better accuracy using grid search and 10 cross-validations [1]. The results of the model obtained are 81.35% training and 78.24% testing accuracy. Also by removing the Boolean features accuracies dropped to 75.21% and 72.75%.

Likewise, K Means Clustering [1] has been used for the same dataset to classify the cover types, here the data is grouped into clusters where the model is developed without the labels of the data. Each of the clusters is observed and named based on the most common cover type. This has been run for 10 times for better accuracy. The results showed that when k=7 for each of the 7 cover types performance was very poor and once the number of clusters has increased the test error reduced with 0.38 for the complete dataset and 0.55 for dataset without Boolean parameters.

The study evaluated that reducing the dimensions using PCA and transforming the data from 54 features to 10 features with Multi-Class SVM and K-Means Clustering performed worse in training and testing than using the entire dataset. Although the positive aspect would be this work reduced the overfitting and demonstrated lower generalization error.

In one of the previous studies and experimentation of using artificial neural networks and discriminant analysis [3] says that the results of the feed-forward artificial neural network model predict more accurately about the forest cover type than the traditional statistical model based on Gaussian discriminant analysis [3]. In the approach of ANN one hidden layer and backpropagation learning algorithm is used with mean squared error (MSE) function. The 54 input variables are analyzed for the reduction process to identify the variables that did not contribute to the overall predictive capability of the system. The experiment showed that 150 hidden nodes were used to minimize the MSE with the best learning rate and momentum rate of 0.05 and 0.5. Also, the classification accuracy of the prediction model was 70.58%.

The second approach in the discriminant analysis [3] is implemented based on two main assumptions. One being the data distributions of all dependent and independent variables are normal and second is the covariance matrix for different groups are equal. The classification accuracy for the discriminant analysis model was 58.38%.

The results of this study and experiments conclude that the ANN model outperforms the DA model in the prediction of forest cover type. The negative aspect implies that both models misclassify ponderosa pine, Douglas-fir, and cottonwood/willow cover types with each other. This is because of the geographic proximity of the different cover types. Also, another factor that impacts the approach of both the classification models is the amount of computational time that is required to develop the prediction.

References

[1] Crain, K., and G. Davis. "Classifying forest cover type using cartographic features." Published report (2014).

[2] D.A. Leatherman, Colorado State Forest Service entomologist (retired); 2/99. [http://www.ext.colostate.edu/pubs/insect/05528.html Revised 9/11](http://www.ext.colostate.edu/pubs/insect/05528.html%20Revised%209/11).

[3] Blackard, Jock A., and Denis J. Dean. "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables." Computers and electronics in agriculture 24, no. 3 (1999): 131-151.