Forest Cover-Type Prediction

First Author#1, Second Author\*2, Third Author#3

#First-Third Department, First-Third University  
Address Including Country Name

1first.author@first-third.edu

3third.author@first-third.edu

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 data. In this study, we aim to predict forest cover types from cartographic variables using classification algorithms.

Keywords— cover-type, data mining, forest types, ExtraTreesClassifier,KNN

1. Introduction

Natural resources are vital for the existence of human life. These natural resources information are stored in the federal land management agencies for inventory management. Forest cover type is one of the basic information that is recorded in these inventories. Generally, the cover type is recorded by people or geo-sensed data. Both of these techniques may be prohibitively time confusing and costly in some situations. In these situations, predictive models provide an alternative method for obtaining data. This has so much potential for positive change, particularly in areas like environmental conservation, flora and fauna research, and geological studies. In this study, we aim to predict forest cover types from cartographic variables using classification algorithms.

1. 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 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.

1. DATA MINING METHODOLOGY

The methodology used in this project management process is CRISP -DM [4] (Cross Industry Standard Process for Data Mining). The application of using this in our research is discussed below with exploratory data analysis.

1. PROCESS FLOW

In Phase 1, Business perspective on the application of classifying forest cover types from cartographic variables are studied and analysed. The advantages of predicting forest cover type trees in wooden area helps in conservation and proper management of forest trees without any fire, infestation [2] and disasters. The business goal on this application is to understand which trees species grow predominantly in what kind of wilderness area with data collected from hill shade, slope and soil aspects for the challenges of US Forest management services. This is implemented by understanding the seven forest cover types from four different wilderness areas in Roosevelt National Forest of Northern Colorado.

The next step is the data understanding phase where all the features are analysed with verification on quality of the data and finding the outliers. There is 54 features with one target variable with 581012 instances. This is a Multi-class Classification for seven discrete categories of forest cover types. From the descriptive statistics of the dataset we could say that the features of the data is complete and numeric without any missing/null values. It includes 10 continuous variables and 44 Boolean variables with one-hot encoded columns such as soil type and wilderness area. Along with this an anomaly is identified and checked in cover type’s data frame on Vertical Distance To Hydrology column.  This feature explains the vertical distance from nearest surface water where negative values are displayed. These values show that the nearest surface water is below the sea level so the values shall remain the same.

During the Data Preparation stage, the goal is to focus on data transformation and feature selection. We analysed the different features and its correlation with each other using python libraries but there were some challenges on the measurements used for different features. In order to resolve the issues we reversed our process flow to data understanding stage to find that meters and degrees are used for measuring the input features which leads to unrelated wide spread of data’s. All the feature inference are explained in EDA.

Exploratory Data Analysis (EDA) is the first step on this workflow process to analyse the features and attribute selection. From this analysis, valuable insights can be obtained by looking at the data distribution of target variable, relationship between 54 features and target variable and links between the features. The distribution of the forest cover type is unequal, where Lodge pole Pine has the highest number of observations followed by Spruce. These 2 cover types add up to 495,141 number of observations out of 581,011 total which covers approx. 85.2% of data., figure(1).

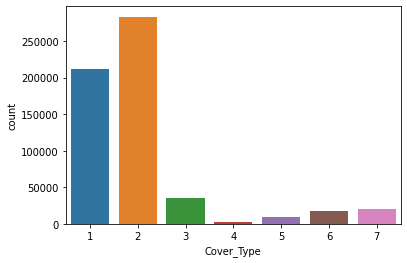


Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Univariate analysis on all the features has minimum value of 0 except Elevation and Vertical Distance\_ To Hydrology features, where Elevation has the highest minimum value and Vertical Distance To Hydrology has the lowest, being negative. With some research and using logic, negative values show that the nearest surface water is below that data point or it is below the sea level. Mean of the features vary from as low as 14 to as high as 2959, where different features take on different ranges of values. The reason some features are so widely spread and having high values and some features don't is because 5 out of 10 variables are measured in meters, includes ('Elevation', 'Horizontal Distance To Hydrology', ’Vertical Distance To Hydrology', 'Horizontal Distance To Roadways', ’Horizontal Distance To Fire Points'), so it makes sense that these have high values and ranges. Features like Aspect and Slope are measured in degrees so its maximum value can't go above 360. While Hill shades features can take on a max value of 255.

The correlation between and among the features are explored to better understand underlying relationships in the data using a correlation matrix: The features which are strongly correlated with a threshold above 0.5 can be found below: **a.** Elevation, ‘Horizontal Distance To Fire Points’, ‘Wilderness\_Area1’ and ‘Wilderness\_Area4’ have stronger correlations with the target cover type. **b.** Aspect, Slope and ‘Distances To Hydrology’ have weak correlations with the target cover type. **c.** While, Hill shades have weak correlation with target cover type; have strong correlations among themselves. **d.** Wilderness\_Area3 has higher correlation with Aspen cover type. **e.** Soil types are distinct and unrelated to each other (mostly grey/ light colour) **f.** Certain soil types have higher correlations with certain cover types.

1. DATA VISUALIZATION

In the next steps, we visualize some important features of the dataset, to get an better understanding, as inferred, it consists of a mix of both categorical and continuous variables, one of the important features are the soil types, there are 40 types, in the four wilderness areas. In order to identify which of the top 20 soil types that affect the accuracy of the model. Implementing, ExtraTreesClassifier, an ensemble learning method based on decision trees. That randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. There are 40 soil types in the given data, by using feature importance we have plotted the top 20 soil types which has high importance for predicting the cover type.

(WRITE SOMETHING HERE)

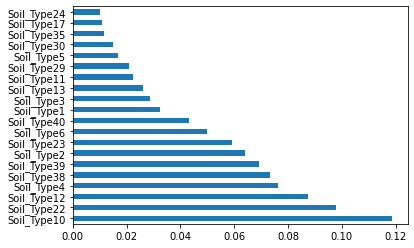


Fig. 2 Example of an unacceptable low-resolution image

On analysing, soil types are related to cover types and wilderness area. In figure(3), visualizing that the cover type Spruce/Fir, Lodgepole Pine and Krummholz are present in most of the top 10 soil types.

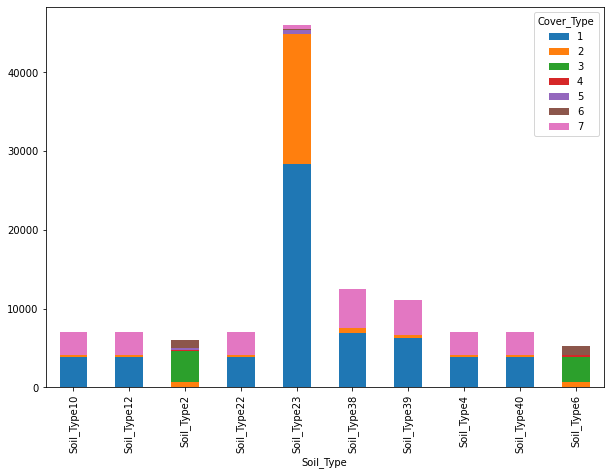


Fig.3 Example of an unacceptable low-resolution image

Fig. 4 There are a total of four wilderness area. On visualizing we identify, wilderness area 1 and 3 has higher percentage of cover type than the wilderness area 2 and 4. And cover type 1 and 2 contribute highest number to wilderness area 1 and 3.

(WRITE SOMETHING HERE)

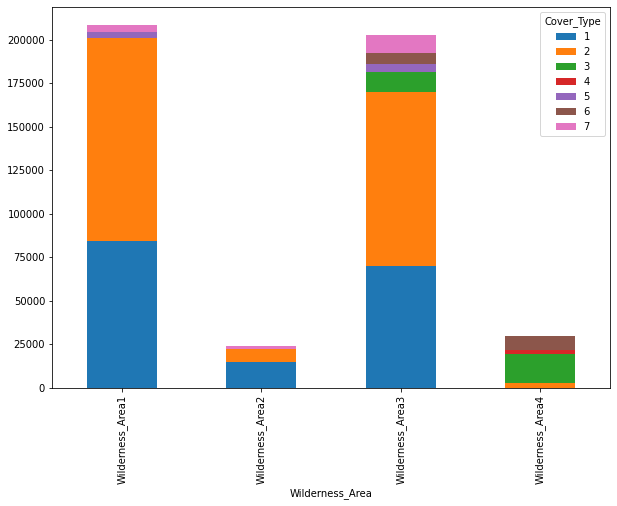


Fig.4 Example of an unacceptable low-resolution image

1. FEATURE SELECTION

Identifying the skewness of the features, Soil\_Type15 has the highest positive skewness meaning the mass of the distribution is concentrated to the left and has long tail to the right followed by Soil\_Type7, 36, 38. This is also called right skewed distribution, where mode of the feature is to the left most followed by median and mean. This means that mostly all of the observations have will have 0 value for this feature. Elevation and Hillshade's having negatively skewed distribution, it's the opposite of the positively skewed distribution, where mode is to the right most followed by median and mean.

In general looking at skew scores of Soils it seems like we can reduce our dimensions by removing some Soil Types only if they don’t have any different information to give our models and improving its performance.

We remove the following soil features to improve the performance,

('Soil\_Type7','Soil\_Type8','Soil\_Type14','Soil\_Type15','Soil\_Type21','Soil\_Type25','Soil\_Type28','Soil\_Type36','Soil\_Type37)

(WRITE SOMETHING HERE)

1. MODELLING
2. Conclusions

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.
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.
4. Chapman, Pete. "Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler),"." CRISP-DM 1.0. Step-by-step data mining guide (1999).