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

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

1. 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
* Cottonwood/Willow
* Lodge pole Pine
* Ponderosa Pine
* Aspen
* Douglas-fir
* 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 lodge pole pine, spruce/fir and aspen. Cache la Poudre would tend to have Ponderosa pine, Douglas-fir, and cottonwood/willow.

1. Exploratory Data Analyss

The given dataset has 581012 data and 55 features. From these features 45 are categorical and 10 are numeric. From 55 features, 10 are numeric and 45 are categorical. From 45 categorical, 40 are of soil types and 4 of Wilderness Area and Cover type. On analysing, all the features have 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. 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 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 Hillshades features can take on a max value of 255. On performing the basic explorations, we can infer the following : The dataset does not contain any missing values. The distribution of the cover types is unequal.The forest cover type, Lodgepole Pine has the highest number of observations followed by Spruce. Only 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). The features which are strongly correlated with a threshold above 0.5 can be found below: a.Elevation, Horizontal\_Dist\_To\_Fire\_Points, Wilderness\_Area1 & 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, Hillshades have weak correlation with target cover type; have strong correlations among themselve. d. Wilderness\_Area3 has higher correlation with Aspen cover type. e. Soil types are distinct and unrelated to each other (mostly grey/ light color) f. Certain soil types have higher correlations with certain cover types.

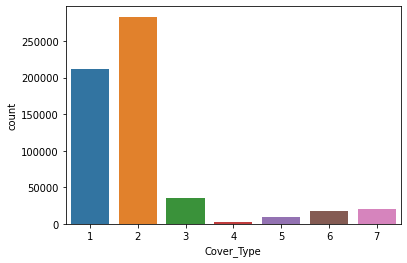


Fig. 1. Bar chart differentiating the spread of various cover types

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 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.



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

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whereas Fig. 3 shows an example of an image with adequate resolution. Check that the resolution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

* the colors used in each figure contrast well,
* the image used in each figure is clear,
* all text labels in each figure are legible.

1. Figure Captions

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.



Fig. 2 Example of an unacceptable low-resolution image



Fig. 3 Example of an image with acceptable resolution

1. Table Captions

Tables must be numbered using uppercase Roman numerals. Table captions must be centred and in 8 pt Regular font with Small Caps. Every word in a table caption must be capitalized except for short minor words as listed in Section III-B. Captions with table numbers must be placed before their associated tables, as shown in Table 1.

1. Page Numbers, Headers and Footers

Page numbers, headers and footers must not be used.

1. Links and Bookmarks

All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

1. References

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. Do not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Reference [3] shows …”. Multiple references are each numbered with separate brackets (e.g. [2], [3], [4]–[6]).

Examples of reference items of different categories shown in the References section include:

* example of a book in [1]
* example of a book in a series in [2]
* example of a journal article in [3]
* example of a conference paper in [4]
* example of a patent in [5]
* example of a website in [6]
* example of a web page in [7]
* example of a databook as a manual in [8]
* example of a datasheet in [9]
* example of a master’s thesis in [10]
* example of a technical report in [11]
* example of a standard in [12]

1. Conclusions

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: “original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)”.

Acknowledgment

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

References

1. S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
4. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.
6. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
7. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
8. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
9. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
10. A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
11. J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
12. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.