Forest Cover Type Prediction Using Cartographic Variables

Nikki Fernandez, Jeanne Nicole Magpantay, and Ian Benedict Valencia
Department of Computer Science
College of Engineering
University of the Philippines, Diliman

***Abstract*—Given elevation, hydrologic, soil, and sunlight data, this project attempts to predict the predominant type of tree in sections of wooded area. The dataset used in this research was transformed through data cleaning, standardization, and Principal Component Analysis (PCA). Model selection was done through Stratified K-Fold Cross-Validation in which combinations of parameters were tested as well for both Artificial Neural Network (ANN) model and Support Vector Machine (SVM) model. The SVM model outperformed the ANN model w/ or w/o PCA. The SVM model produced 90% accuracy on the testing set provided.**

I. Introduction

Understanding a forest's composition is important in the maintenance and management of forest areas. Knowing the cover type aids in finding the appropriate actions to prevent common forest related problems such as wildfires, infestations, and deforestation. However, processes that are often used in acquiring forest cover type data are time and resource intesive. As such, an alternative to such processes is beneficial.

Our research aims to predict the predominant family of tree in a given forest area using the area's cartographic data as alternative to the usual processes done in acquiring forest cover type data.

II. Short of Review of Related Studies

The problem at hand is a classification type problem. Therefore, machine learning techniques focused on classification are to be discussed in our research. For processing the data, we made use of Artificial Neural Networks and Support Vector Machines.

A. Artificial Neuron Network

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output. ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found.

B. Support Vector Machine

On the other hand, a Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVMs theory. Therefore, the optimal separating hyperplane maximizes the margin of the training data.

C. Principal Component Analysis

Principal Component Analysis (PCA) is a method of reducing the dimensionality of data while, usually, maintaining most of the datas variance. All dimensions of the reduced data are linearly uncorrelated, meaning the original data is projected into a space where each component is orthogonal to the others.

D. Stratified K-Fold Cross Validation

In some cases, there may be a large imbalance in the response variables. For example, in dataset concerning price of houses, there might be large number of houses having high price. Or in case of classification, there might be several times more negative samples than positive samples. For such problems, a slight variation in the K Fold cross validation technique is made, such that each fold contains approximately the same percentage of samples of each target class as the complete set, or in case of prediction problems, the mean response value is approximately equal in all the folds. This variation is also known as Stratified K Fold.

E. Standardization

Standardizing means to rescale your data to have a mean of zero and a standard deviation of one. A standardized variable is sometimes called a z-score or a standard score. Using variables without standardization can give variables with larger ranges greater importance in the analysis. Transforming the data to comparable scales can prevent this problem.

$$Z = \frac{X - \mu_x}{\sigma_x} \tag{1}$$

Equation for standardizing

III. Methodology and Results

A. Dataset

The dataset used in this research is a subset of a forest cover type dataset from the UCI Machine Learning Repository. It is composed of 50,400 samples of 30x30 meter cell from four wilderness areas in Roosevelt National Forest of northern Colorado. Each sample has 54 attributes including the cover type. Dataset is split using the 60-40 train-test split.

B. Preprocessing

The cover type is dropped from the training and testing data set and was used as labels instead.

1) Data Cleaning: Attributes that had a standard deviation equal to zero were dropped to prevent them from making the standardization later on behave poorly. The dropped attributes are *Soil_Type7* and *Soil_Type15*.

2) Standardization and Feature Selection: Standardization was applied to the dataset in order to make the weighing between the features more reliable. Moreover, principal component analysis was applied to the dataset. The transformed dataset contains only 10 principal components.

C. Model and Parameters Selection

Since the problem at hand is a classification problem, three initial candidates came to our minds: - *Naive Bayes*, *Artificial Neural Networks*, and *Support Vector Machines*. Due to the Naive Bayes known to have strong feature independence assumptions, it was the first among the three to be dropped since some attributes in the given dataset are related to each other like the compass direction of the slope face and the sunlight intensity at a given hour. Stratified K-Fold cross-validation, with K=10, is done to both ANN and SVM models in order to know which of the two would be able to generalize with an independent or unseen dataset. In addition, combinations found in a list of possible parameter values are tested. The ANN model had 216 parameter combinations while the SVM model had 60. The following tables contain the top 10 parameter combinations to both models.

kernel	gamma	degree	Mean Train Score	Mean Test Score
poly	0.125	3	0.90	0.7619
poly	0.25	3	0.93	0.7412
poly	0.0625	3	0.86	0.73
poly	0.25	2	0.84	0.7282
poly	0.5	2	0.85	0.7281
poly	0.5	3	0.95	0.727
poly	1	2	0.86	0.726
poly	0.125	2	0.83	0.723
poly	1	3	0.96	0.713
rbf	0.5	1	0.91	0.712

TABLE I
PARAMETERS-SCORE TABLE OF SVM MODEL W/O PCA

From the validation method done, it can be observed that the ANN model w/o PCA outperformed the SVM model w/ or w/o PCA. As such, it will be the model to be used in predicting the cover type of the testing set.

kernel	gamma	degree	Mean Train Score	Mean Test Score
poly	0.5	3	0.84	0.7138
poly	1	3	0.85	0.7137
poly	0.25	3	0.82	0.70
rbf	0.5	1	0.83	0.689
rbf	0.5	2	0.83	0.689
rbf	0.5	3	0.83	0.689
poly	0.125	3	0.78	0.687
rbf	1	1	0.87	0.686
rbf	1	2	0.87	0.686
rbf	1	3	0.87	0.

TABLE II
PARAMETERS-SCORE TABLE OF SVM MODEL W/ PCA

activation	hidden_layer_sizes	max_iter	Mean Train Score	Mean Test Score
tanh	(52,52)	200	0.94	0.7528
tanh	(52,52)	1000	0.96	0.7527
tanh	(52,52)	900	0.95	0.7521
tanh	(52,52)	400	0.95	0.7518
tanh	(52,52)	700	0.95	0.7517
tanh	(52,52)	500	0.94	0.7515
tanh	(52,52)	800	0.95	0.7498
tanh	(52,52)	600	0.94	0.7474
tanh	(52,26)	700	0.93	0.7471
tanh	(52,52)	300	0.95	0.7469

TABLE III
PARAMETERS-SCORE TABLE OF ANN MODEL W/O PCA

activation	hidden_layer_sizes	max_iter	Mean Train Score	Mean Test Score
tanh	(52,52)	500	0.84	0.7063
tanh	(52,52)	1000	0.84	0.7053
tanh	(52,52)	900	0.84	0.7034
tanh	(52,52)	700	0.84	0.7008
tanh	(52,52)	200	0.83	0.6998
tanh	(52,52)	600	0.84	0.6990
tanh	(52,52)	400	0.84	0.0.6975
tanh	(52,26)	1000	0.81	0.6964
tanh	(52,52)	800	0.83	0.6956
tanh	(52,26)	300	0.81	0.6956

TABLE IV
PARAMETERS-SCORE TABLE OF ANN MODEL W/ PCA

D. Performance

The model was able to achieve 90% accuracy, precision, and recall. This was way lower than the accuracy found in the validation step however, when compared to other model-parameter combination, this produce the highest accuracy.

X	1	2	3	4	5	6	7
1	3524	1205	1	0	29	2	58
2	531	13671	29	0	122	27	5
3	0	11	1760	69	17	303	0
4	0	0	29	2111	0	20	0
5	7	228	29	0	2118	14	0
6	0	7	200	31	8	1914	0
7	17	0	0	0	0	0	2143

TABLE V
CONFUSION MATRIX

E. Pitfalls

While finding for the optimal parameters for the SVM model, certain combinations of parameters result to a very slow run time. One combination was done 40 minutes while other combinations did it for less than 10 seconds and even with higher a higher mean score. As such, parameter values were bounded. There is a possibility that the parameter values not used in this study might produce better results.

IV. Conclusion

Our research was able to produce a model that is capable of predicting the cover type of a given woodland area given its cartpgraphic attributes. However, the resulting accuracy on the given test set is still average.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Kevin Crain and Graham Davis, *Classifying Forest Cover Type using Cartographic Features* Stanford University - CS 229: Machine Learning, December 2014