Capstone Project Proposal

Mohan Prasath Chinnasamy

January 6, 2017

1 Domain background

Our Earth is an unique planet in the solar system holding the key to sustain Life. Even now many parts of the oceans, and forests remain unexplored. Several species are discovered periodically in remote forests across the globe. Under further genetic studies it is revealed that we humans do posses some characteristics of the DNA of the mammals!. Exploration of such remote regions can provide answers for many unanswered questions regarding origin of life, evolution, and life in unpolluted environments[1].

However such explorations are costly and even dangerous for humans. Aerial surveillance have improved much during these last decades. These technological improvements can assist in human explorations by capturing images of unexplored regions reducing cost and human effort. But most forests are dense with vegetation and such aerial surveillance can provide only the top view. We can then use the surveillance data to predict and restrict the search space for human explorations. From these explorations an extrapolation[3] can be done about missing species in similar areas, thus reducing the search space further.

2 Problem statement

This project aims to use surveillance data with cartographic information to train models that can predict and identify the forest cover types. The project also aims to create a classification model with more than 90% accuracy which can classify the forest cover type.

The performance of the model can be evaluated by calculating the prediction accuracy. For testing, 20% of the total data set will be used which is not used to train the model All classification models will be trained with 80% of the data. This huge amount of data used to train the models can increase the prediction accuracy over the unseen data. For example, a k-Nearest Neighbor model that can classify the features into k target variable can be applied here.

3 Datasets and inputs

The cover type data set was acquired from UCI Machine Learning Repository[2]. The data was collected in four major areas of wilderness in Roosevelt National Forest of northern Colorado. This region is affected minimally by human intervention. So this holds information about ecological changes occurring over time.

The data set has 12 major features with one multivariate target variable with information about forest cover type. Two features Ẃilderness_Area; and Śoil_Typeáre further divided into 4 and forty binary valued sub features respectively. Thus adding up all features to a count of 54. There are 581012 instances of the data with no missing values.

4 Solution statement

The data has a multivariate target variable, so many multivariate classification models can be trained to predict the forest cover type. Also a huge training set of 464809 samples can help the models to achieve a higher prediction rate possible. A comparative study of various classifiers can help in studying the data and the models in detail.

5 Benchmark model

The comparison of various classifiers helps us further study the maximum achievable prediction accuracy by a model using the same training data. The 20% of the data(116203 samples) will be used to test the classifier models. This is unseen data, and is readily available with associated target variable. Evaluating the models against this data can help us understand their performance. This performance can be quantified and visualized.

6 Evaluation metrics

The models can be evaluated against their prediction accuracy. This quantification can help us identify suitable model to use against any future data.

7 Project design

The forest cover type data is divided into training and test data. Many multivariate classifiers are initialized and trained using the training data. Then the test data without the target variable is used by the classifier models to predict the test data target variable. The accuracy of that prediction will be measured.

References

[1] Christopher W. Dick and W. John Kress. Dissecting tropical plant diversity with forest plots and a molecular toolkit. *BioScience*, 59(9):745–755, 2009.

[2] M. Lichman. UCI machine learning repository, 2013.

[3] Rosanne Tackaberry, Nicholas Brokaw, Martin Kellman, and Elizabeth Mallory. Estimating species richness in tropical forest: the missing species extrapolation technique. *Journal of Tropical Ecology*, 13(3):449–458, 005 1997.