Using Image Classification to label plant species, Project 2 for Big Data & Analysis

Henry Stuklis (a1706223)

October 14, 2019

Abstract

1 Introduction

This project for Big Data & Analysis required us to complete one of four Kaggle competitions. Kaggle is a website that lists interesting data sets along with problems to solve using that data. Mainly these problems involve building a statistical model to perform prediction. On the Kaggle website you are allowed to submit your statistical models predictions and see how much error it has in predicting the true known values. Kaggle then displays a leaderboard of all submitted models and the main aim of competing in these Kaggle competitions is to try and rank as high as possible. The less error present in a model the better and more robust the model is in prediction. So the workflow pipeline for a Kaggle competition is similar to as follows.

- Clean and pre-process the base data (given as training and testing data)
- Complete some feature engineering (transforming, dropping or creating variables)
- Pick a reasonable statistical model to predict the required values
- Train the statistical model on the training data
- Analyse the error in predicting the testing data (either by submitting on Kaggle or calculating the error yourself)

My detailed and commented Python code that goes through this entire problem pipeline can be found attached separately to this report in plant-species-analysis-code.ipynb.

The Kaggle competition that I chose out of the four is a multilevel image classification problem. The given data contains 4750 pictures (in `.png` format) of twelve different plant species. The aim of the predictive model to be constructed is to correctly classify each plant species just based on the provided image. To do so we will need to split the images provided under `train.zip` on Kaggle into a separate training and testing data. As we have not been given a percentage split in the problem statement and we have a very large amount of images, I have elected to make the training/testing split 90% to 10%. So our predictive model will need to be able to learn the differences between these twelve plant species based on the provided visuals in the training data.

The type of predictive model I will be using for this problem is known as a convolutional neural network. This is a very commonly used neural network in image problems as it is able to progressively simplify each image down to a certain image and then compare with each canonical image the trained model has. It then returns the species label of the canonical image it matches the most.

2 Methodology

3 Experiment

4 Discussion & Conclusion