Project Milestone

John Randis

11/7/2016

Abstract

This paper explores the early stages of an application project intended to classify different species of leaves.

Introduction

The implications and benefits of automatic plant classification are wide and far reaching. Human error often leads to misclassified or duplicate classifications. Automated classification can help discoveries in biology and medicine. This paper will include the background and related works, the methodology of the approach taken for this project, and a careful analysis of results.

Background

In very recent history, biological information has never been more readily available. Databases exists across the web full of biological information on plant life. Online biodiversity databases such as <http://www.gbif.org/> and <http://www.pfaf.org/> have made this information more accessible than ever. Biologists have been classifying plant and animal species for years. This is why introducing machine learning to this problem could be so revolutionary. The problem of plant classification has not been tackled much in the data science community. On Kaggle, many submissions involve the KNeighbors classifier or the DecisionTreeClassifier of sklearn. These built-in classifiers are nice, but accuracy is often not very good and leaves room for improvement.

Methodology

The approach being used for my classifier will be through logistic regression. Logistic regression is a simpler approach to this problem and is less prone to overfitting. The python libraries pandas and sklearn are being implemented to complete this task.