Teammate: Ke Lin, Jingzhou Ni
Professor: Nik Bear Brown
November 4, 2018

Final Project Proposal

Kaggle Competition: Leaf Classification

Description

Leaves, due to their volume, prevalence, and unique characteristics, are an effective means of differentiating plant species. They also provide a fun introduction to applying techniques that involve image-based features.

The objective of our project is to use binary leaf images and extracted features, including shape, margin & texture, to accurately identify 99 species of plants.

Background

There are estimated to be nearly half a million species of plant in the world. Classification of species has been historically problematic and often results in duplicate identifications. Based on leaves which is the unique characteristics of every plants, we can solve these problems and develop many applications that will be helpful for industry and agriculture such as: species population tracking and preservation, plant-based medicinal research, crop and food supply management.

Data Source

Leaves' image file and shape, margin and texture from kaggle

https://www.kaggle.com/c/leaf-classification#

Plants' specimen records and images from Royal Botanic Gardens's database

http://apps.kew.org/herbcat/navigator.do

One hundred plant species leaves data from UCI machine learning repository

https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set

Algorithm

Naïve Bayeshttp://scikit-learn.org/stable/modules/naive_bayes.html

Random Foresthttp://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Logistic Regressionhttp://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

The K-Means clustering algorithmhttps://mubaris.com/posts/kmeans-clustering/

Principal component analysis (PCA) http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

The choice of the algorithms requires further evaluation.

Reference

VanderPlas, J. (2018). In Depth: Principal Component Analysis | Python Data Science Handbook. Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Visualising high-dimensional datasets using PCA and t-SNE in Python. (2018). Retrieved from https://medium.com/@luckylwk/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

Galarnyk, M. (2018). PCA using Python (scikit-learn) – Towards Data Science. Retrieved from https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

Principal Component Analysis. (2018). Retrieved from https://plot.ly/ipython-notebooks/principal-component-analysis/

Group Member

Ke Lin   NUID: 001817984

Jingzhou Ni   NUID: 001235201