Report

Introduction and background

In this paper I present the implements and results of classification challenge on the Plant Phenotyping Dataset. The objective is to build a predictive model which can distinguish Arabidopsis plant images from tobacco plant images, with a dataset including 165 images for Arabidopsis plant and 62 images for tobacco plant.

As a brief overview, the phenotype of a plant contains the number of leaves, architecture, visual age or maturity level, height, leaf shape and so on. These features can be used to distinguish different plant.

Color Histogram

Colour Histogram represent the global distribution of pixel colours in an image.

Scale-invariant Feature Transform

SIFT feature describes the texture features in a localised region around a keypoint, SIFT descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes.

Histogram of Oriented Gradients

The histogram of oriented gradients (HOG) describes the distributions of gradient orientations in localized areas and does not require initial segmentation.

Random Forest(RF)

A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x,\theta k),k=1,...\}$ where the $\{\theta k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

RF works efficiently on large datasets and may avoid overfitting compare to decision tree, but less interpretable than decision tree.

Support Vector Machine(SVM)

A support vector machine constructs a hyperplane or set of hyperplanes in a high-or infinite-dimensional space. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class.

SVM is effective when the number of features is larger than the training data size and classes are separable, but may take long time to process when dataset become larger.

K Nearest Neighbors(KNN)

Given a query instance x_q to be classified, take vote among its k nearest neighbors based on some distance measure.

In comparison to other classifiers, KNN is more simple and easier to implement, but may perform badly when the number of variables grows.

Method

In order to find the most suitable features to classify the two group plants, I try to acquire the features of plant from 4 aspects(its original pixels features, color histogram features, texture features and shape features) and use 3 models(RF, SVM, KNN) to classify these features.

The select four features:

The features extract from images' original pixels(Original features)

The original feature is the content of image which can express the information of the image and show differences with the other kind of images.

The color histogram(RGB)

The histogram of an image is a plot of the gray level values or the intensity values of a color channel versus the number of pixels at that value. The shape of the histogram provides us with information about the nature of the image, or subimage if we are considering an object within the image.[1] Therefore, color histogram can distinguish different images

The texture feature

Textures play an important role in the field of Image Classification. It is one among the significant features used for identifying regions of interest in an image.[2]

The shape feature

The shape of image in one group may have a huge difference with that in the other group, thus shape features can be used to distinguish two group images.

Experiment

Image acquisition

I use glob.glob to get image files and apply cv2.imread to read each image.

Feature extraction

Original features: resize pixel array to (50,50), then flatten it.
The color histogram: separately extract R, G, B color features then flatten them in one array.
The SIFT feature: convert image to gray image and extract it sift descriptor, then use k-means model to extract its feature.
The HOG feature: use the defined function hog to extract hog feature from each image.

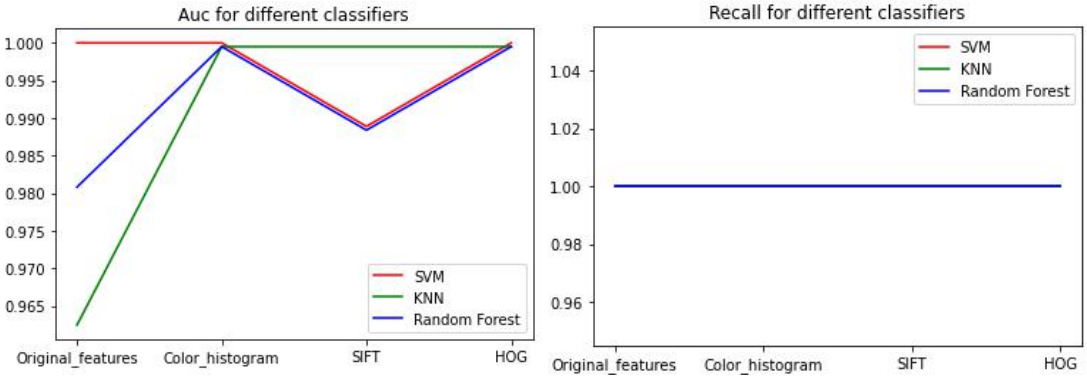
Learning algorithm

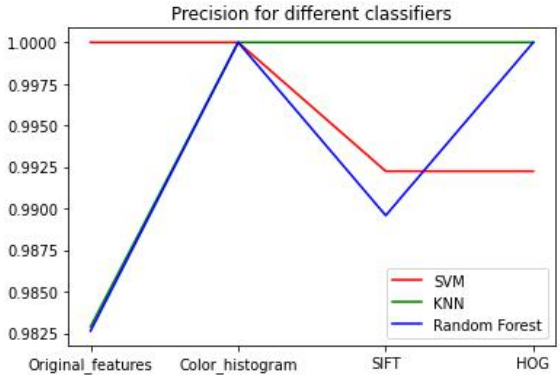
After acquire features, I shuffle the data first, then I separately use 80%, 75%, 70% of data to train and use 20%, 25%, 30% of data to test. Specially for KNN, I choose k =3,5,7,9 to find the best k.

System evaluation

Roc_auc_score: An ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive.Rate AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).
Recall: Fraction of the true object that is correctly segmented.
Precision: Fraction of the segmented object that is correctly segmented.

Result





From above three graph about Auc, recall, precision for three classifiers, I find all three classifiers have relatively good performance, that can be explained by the obvious difference among two groups of plant images. For the three classifiers, SVM have the best performance than the rest two in all three aspect considering four different features.

Reference:

[1] Szabolcs Sergyan Budapest Tech John von Neumann "Color Histogram Features Based Image Classification in Content-Based Image Retrieval Systems" 2008 6th International Symposium on Applied Machine Intelligence and Informatics p221 2008

[2] Aviral Kumar Gupta1, Abhishek Dabas2,"Image Classification Using Textures" Ranendu Ghosh3 Dhirubhai Ambani Institute of Information and Communication Technology April 2013