Leaf Classifier

CAPSTONE PROJECT PROPOSAL
KAGGLE COMPETITION

Abdulrahman Abuzaid | Machine Learning Udacity Nanodegree | Jan 29, 2019

Domain Background

Plants are an important aspect of life on the planet, without them, life on earth will go extinct, as they provide the necessary Oxygen for our existence [1]. Since there are around 400,000 types of plants in the planet [2], formulating an automatic classifier and identifier is an important endeavor. Such a classifier will improve humanity's efforts to preserve all plants, conduct scientific floricultural studies, and possibly reduce global warming. I chose this project since my life partner is a floriculturist and I want to impress her and introduce her to the power of Machine Learning, Deep Learning, and Artificial Intelligence in general.

Problem Statement

The leaf classification project is based on a Kaggle competition [4] where participants are challenged to produce a machine learning model capable of learning to classify 99 different plant species based on leaf images and features extracted from these images [3]. The challenge is to have the minimum classification error possible. The ultimate solution should be able to classify all leaves correctly. The classification model can be further improved in future extensions by including more plant species.

Datasets and Input

The data for this project has been provided by Kaggle [4] and it is freely available for download. The data consists of two types: image data and tabular data. The image data consists of 16 images for each one of the 99 species under study. The images are in black and white to emphasize the shapes of these leaves rather than their colors. The tabular data consists of three feature vectors of size 64 each which were previously extracted from the images as detailed in this article [3]. The extracted features are for the shape, margin, and texture.

To summarize, the available data is as follows:

Image Data (test and train datasets)

- 16 x 99 = 1584 images
- Varying pixel dimensions
- Binary pixel data (black and white)

Tabular Training Data

- 991 rows x 194 columns
- One row for each image with the following columns

- 64 columns for each extracted feature, margin, shape, and texture, in that order. (64 x 3 = 192)
- Unique id column for each image
- Species column

Tabular Testing Data

- 595 rows x 193 columns
- One row for each image with the following columns
  - 64 columns for each extracted feature, margin, shape, and texture, in that order. (64 x 3 = 192)
  - Unique id column for each image

Solution Statement

This problem can be modeled as a supervised classification problem and can be solved through applying this class of machine learning algorithms. The problem is considered as supervised since we know the label that we want to predict: the plant species. The problem is a classification problem since we are interested in predicting a categorical label, and not a value.

First, using the tabular data, a number of classification algorithms can be tested and compared, such as Multi-Layer Perceptron (Neural Network), Decision Tree, Random Forest, XGBoost, and K-Nearest Neighbors. Moreover, since there is a large number of features in the provided data, dimensionality reduction can be applied by means of Principal Component Analysis, which might reduce the classification model complexity. Finally, the image data can be used to enhance the model through two methods: extracting more features then adding them to the tabular data features, and applying Convolutional Neural Networks directly on the 2D images to capture the spatial data. The final solution can either be the best performer of the previously stated models, or better yet, the models can be combined into one, more powerful model.

Benchmark Model

One possible benchmark model is to reshape the image data into a 1d vector and feed it to an MLP Neural Network without any extracted features. The final model will be compared with this benchmark model using the evaluation metric as explained in the next section.

Evaluation Metric

The Kaggle competition specifies the evaluation metric as the multi-class logarithmic loss as given below.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\, log(p_{ij})$$

where N is the number of images in the test set, M is the number of species label, log is the natural logarithm, y_{ij} is *one* if observation *i* is in class *j* and *zero* otherwise, and finally, p_{ij} is the predicted probability that observation *i* belongs to class *j*.

Predicted probabilities are to be replaced by the following values to avoid the extremes of the *log* function.

$$p' = \max(\min(p, 1 - 10^{-15}), 10^{-15})$$

Project Design

I propose to tackle this project through the following steps.

Data Preprocessing

The data is provided through the Kaggle platform. The tabular data has already been cleaned and preprocessed. The image data, however, has varying dimensions and some preprocessing and resizing might be required. The data will be used as is inside Kaggle kernels or alternatively, it will be downloaded and analyzed on a local machine.

Data Exploration

Exploratory Data Analysis will be conducted, in particular for the provided feature columns. Histograms for the features will be plotted in addition to the cross-correlations between the different features.

Classifier Using the Provided Features

A classifier will first be implemented only on the tabular data in order to see how much the model can learn to classify based on the already extracted features. Different classifier algorithms will be assessed such as MLP, Decision Tree, Random Forest, XGBoost, and KNN.

Further Feature Extraction

More features can be extracted directly through the images, such as width and length of the leaf. The classifier performance will be recorded as a result of adding such features.

Dimensionality Reduction Consideration

Dimensionality reduction techniques such as PCA will be applied to all the features: the original and the extracted ones. The motivation for this is to assess the performance, in terms of accuracy and speed of the classifier in case of using less number of dimensions as features.

Exploring Convolutional Neural Networks

CNN will be applied directly on the images to capture the spatial data in the images. The performance of CNN will be compared with the classifiers applied on the tabular data.

Optimum Model

The final and optimum model can either be the best performer of the previously stated models, or better yet, the models can be combined into one, more powerful model.

References

[1]

https://www.nature.com/scitable/blog/our-science/no_trees_no_humans

[2]

https://news.mongabay.com/2016/05/many-plants-world-scientists-may-now-answer/

[3]

https://www.researchgate.net/publication/266632357_Plant_Leaf_Classification_using_Probabilistic_Integration_of_Shape_Texture_and_Margin_Features

[4]

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