Machine Learning Engineer Nanodegree

Capstone Proposal

Charles Brands
June 3, 2018

Introduction

This project wil use image analysis to classify twelf different species of plants. The dataset comes from the Plant Seedlings Classification competition.

Domain Background

Farmers have been spraying their fields with herbicides for decades to reduce weeds. So far one herbicide mixture was used to spray an entire field. This is not a very effective approach. Massive usage of herbicides are expensive, bad for the environment and may lead to resistance in the weeds. Also a herbicide mixture that is effective against one species of weed can have little or no effect on another species. Not to mention that the herbicide may be damaging tothe crop it is supposed to protect.

Aarhus University in Denmark is working on a project to create weed maps of a field from a tractor to pinpoint where a certain species of weed resides on a field. They have released a dataset of images of 960 unique plants belonging to 12 species at several growth stages.

Kaggle is hosting this dataset as a Kaggle competition in order to give it wider exposure and to give the community an opportunity to experiment with different image recognition techniques.

My motivation

Growing up in a rural area of the Netherlands I realize that food production is extremely labor intensive and uses significant amounts of herbicides to kill weeds. If a deep learning algorithm could distinguise the weeds from the disered plants then herbicides could be more efficiently and therefore more sparringly used.

Problem Statement

For this project we have to detect which species of plant is in the picture. Each image contains one seedling which has to be classified into one of twelve categories. The kaggle competition closed three months ago but I can evaluate any solution by measuring the mean multi-class F1 score run against the given test data set just as in the original Kaggle competition.

Datasets and Inputs

The train and test datasets can be found on kaggle. There are 4751 images for training and 794 images for testing. Each image has a filename that is its unique id. The dataset comprises 12 plant species. The list of species is as follows:

Danish	English	Latin	EPPO code
Majs	Maize	Zea mays L.	ZEAMX
Vinterhvede	Common wheat	Triticum aestivum L.	TRZAX
Sukkerroe	Sugar beet	Beta vulgaris var. altissima	BEAVA
Lugtløs kamille	Scentless Mayweed	Matricaria perforata Mérat	MATIN
Fuglegræs	Common Chickweed	Stellaria media	STEME
Hyrdetaske	Shepherd's Purse	Capsella bursa-pastoris	CAPBP
Burresnerre	Cleavers	Galium aparine L.	GALAP
Agersennep	Charlock	Sinapis arvensis L.	SINAR
Hvidmelet gåsefod	Fat Hen	Chenopodium album L.	CHEAL

Liden storkenæb	Small-flowered Cranesbill	Geranium pusillum	GERSS
Agerrævehale	Black-grass	Alopecurus myosuroides	ALOMY
Vindaks	Loose Silky-bent	Apera spica-venti	APESV
EPPO codes are computer codes developed for plants, pests (including pathogens) which are important in agriculture and plant protection.			

Solution Statement

The solution to this problem is a model trained to predict the plant species for a given image.

- Input: An image of a plant.
- Output: The label of the plant on the image.

The 794 test images Will be run through the model and the results will be combined into a single CSV file. This cvs file can than be submitted to Kaggle for evaluation of the mean multi-class F1 score. Alternatively I can calculate the F1 score myself as Kaggle has published how they evaluate this project.

Benchmark Model

The model will be evaluated through k-fold cross-validation. The data will be sepatated into k folds, and then run k times, keeping one of the folds for validation eache run.

Next the results will be benchmarked against other submissions in the Kaggle leaderboard.

Evaluation Metrics

I will use the mean multi-class F1 score as the metric to evaluate my solution.
From scikit-learn.org

> The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1

> score is:

$ \frac{n!}{k!(n-k)!} $

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

> In the multi-class and multi-label case, this is the weighted average of the F1 score of each class.

Precision is the ability of the classifier not to label as positive a sample that is negative. In formula:

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

precision = True positives / (True positives + False positives)

Recall is the ability of the classifier to find all the positive samples. In formula:

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

recall = True positives / (True positives + False Negatives)

Project Design

(approx. 1 page)

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone

project.

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?

References

Plant Seedlings Classification
A Public Image Database for Benchmark of Plant Seedling Classification Algorithms
Original dataset