Project Report

Elise Mol, Wendy Nieuwkamer, and Isobel Smith

21 December 2016

1 Problem

After three months of learning about several machine learning techniques and classifying algorithms it was time to put our newly gained skills into practice. We decided to try our hand at one of the competitions hosted by Kaggle. The challenge we chose is Leaf Classification; the objective of the competition is to use binary leaf images and their extracted features to predict what species each leaf belongs to. The dataset which was provided includes around 1584 images of leaf specimens; these samples include 99 species. From every image three sets of features are extracted: a shape contiguous descriptor, an interior texture histogram and a fine-scale margin histogram. Every feature is represented as a 64-attribute vector.

As the competition started at 30 August 2016, there were already quite some submissions made and kernels active. One of them stood out, namely "10 Classifier Showdown in Scikit-Learn". In this kernel Jeff Delaney compared ten classifiers which are part of the python library Scikit-Learn. The results of his experiment are shown in figure 1. In his experiment Delaney used the default setup for the classifiers , combined with semi-random parameters. He notes that the performance of the classifiers could be improved by tuning the hyper-parameters. Thus, our research question is whether we can improve the accuracy found by Delaney by tuning the hyper-parameters of the classifiers.

As we did not have time to run experiments for all ten classifiers, we decided to look into five of them. We chose the K-neigbors classifier, decision tree classifier and Gaussian Naive Bayes as they had relatively decent accuracy at respectively 88%, 65%, and 57%. Also, we had some background in these algorithms as they had been taught in our Machine Learning course. Additionally, we decided to choose either Ada boost or quadratic discriminant analysis as they could improve the most with found accuracies of 4.5% and 2.5%.

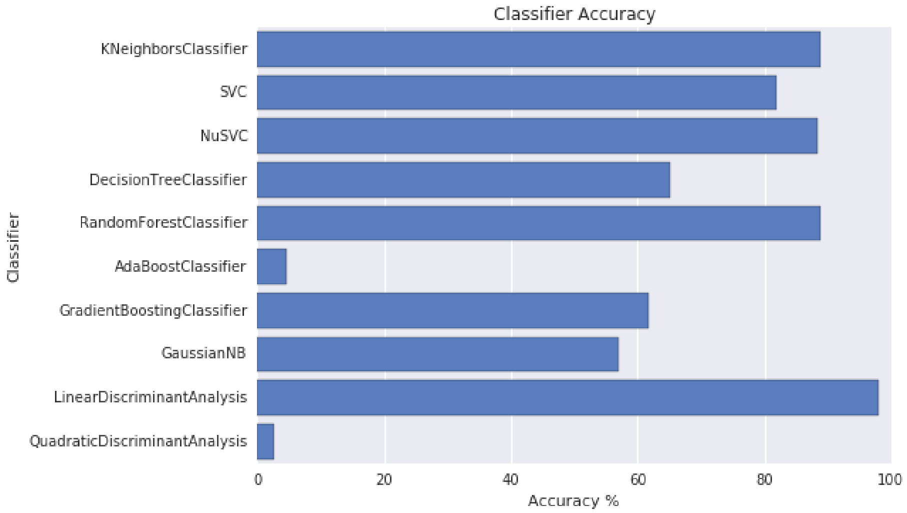

Figure 1: The 10 Classifier Showdown

2 The algorithms

Before we could start optimizing the hyper-parameters we had to research the algorithms we chose. In order to discover the most relevant information for our experiment we aimed to answer the following questions:

- How does the classifier work?
- Which hyper-parameters exist for each classifier??
- Which hyper-parameters have the most effect on the classifier?

The answers to these questions are specific to the Scikit-Learn package.

2.1 K-neighbors

The k-neighbors algorithm classifies an object according to the k nearest objects from the training set. For a higher k noise might be filtered out, but class boundaries might become more vague and less accurate. Thus, it would be interesting to find out which value of k gives the highest accuracy for our dataset. There are two ways of computing the class from the neigbors. The first one is by giving all the neighbors equal weight, no matter how far or close they are. The second is to give each neighbor a weight relative to its distance. The weighted version of the

2.2 Decision Tree Classifier

2.3 Gaussian Naive Bayes

2.4 Ada Boost Classifier

2.5 Quadratic Discriminant Analysis

3 Choices made and justifications

Which classifiers you chose ; why? Describe the relevant features of the classifiers.

4 Solution

How do we optimize them? k-fold cv, gridsearch

5 Evaluation of Approach and Solution

All the curves, the results

6 Sources

https://www.kaggle.com/jeffd23/leaf-classification/10-classifier-showdown-in-scikit-learn