Partial Classification Forest

Jonas Faßbender

jonas@fc-web.de

Abstract

I. Introduction

Some datasets do not allow a classifier to generate a descision surface good enough to be able to predict unseen observations well enough. Well, in this case, referrs to a context dependent threshold for any quality measurement of a classifier, for example the accuracy or an information loss metric.

But for some of those problems, it may still be valuable to predict only on partitions of the feature space, in which the dataset is 'clean' enough, meaning a classifier can be found within the subset of the dataset laying inside one of those partitions which equals or exceeds the threshold.

This paper proposes a Monte Carlo based ensemble method called Partial Classification Forest (PCF), builing an ensemble of trees having a structure similar to k-d trees to partition the feature space of the dataset in order to find 'clean' partitions. In the following a tree generated by the PCF is spelled Tree with a capital T, rather than tree, which is used to denote the data structure.

In Section II I will lay out the structure of a Tree generated by the PCF before, in Section III, describing how PCF uses the Trees and listing its parameters. After that I will continue displaying test results using PCF. In Section V I will discuss further optimizations and possible additional features before finishing with a conclusion.

II. The Tree structure

A Tree generated by the PCF is a binary search tree structure similar to k-d trees. Its purpose is to randomly generate disjoint partitions of a feature space.

It provides two operations, FIT, described with pseudo-code in Algorithm 1, building the Tree based on a provided dataset and the PREDICT operation (cmp. Algorithm 2), returning a label for an observation.

It has two types of nodes, non-leaf nodes, here denoted as Nodes and leaf nodes denoted as Leafs. The Node structure contains three properties, a split value, a left and a right successor, referencing either another Node or a Leaf.

A Leaf, on the other hand, is the structure representing a partition of the feature space, having a label and two vectors with arbitrary length containing the datapoints of a dataset laying inside the partition.