A comparison of metalearning strategies

John Liddell

jl0000285@gmail.com

Abstract

Determining what algorithm to use when analyzing a dataset is a problem as old as machine learning itself. In some cases, the individuals wishing to perform an analysis have access to an expert, possibly themselves, that can simply tell them which algorithm is best in the given situation. In other situations, the individuals wishing to perform analysis may not have the budget neccessary to acquire access to such an expert, in which case the usage of a metalearner becomes appropriate. With a metalearner one feeds the metalearner a dataset and it returns to the user what it thinks is the most appropriate machine with which to perform analysis. To get to the point wherein a decision can be made on new datasets the metalearner itself must first be trained, and this training itself requires some sort of learning strategy. This fact suggests thath the decision of what metelearning strategy to use for some given body of datasets should be succeptable to Wolpert's "No free lunch" theorem, that is to say that some metalearning strategies will work better on some given set of databases than others. The confirmation or denial of this theorem in this context is the goal of this thesis. The act of utilizing a metalearning strategy generally involves the following activities: one must take a set of datasets, obtain a measure of how well those datasets are analyzed by a given body of algorithms then use the analysis information contained within this "metabase" to run a decision algorithm that will then generate a resulting machine for consideration. It is the modification of the measuring and deciding portions of this process that can be considered to represent differing metalearning strategies. The system that I will build will first iterate thru the set of lcomparison strategies, building metabases for each of them and recording how long it took to build each of these metabases. It will then take a set of test datasets and have machines generated with a set of deciding algorithms containing at least K-means clustering with other algorithms to be included if time allows. The combination of deciding algorithm, dataset, and actual best algorithm will be recorded as a given metabase makes its decision. Training time vs accuracy analysis will then be performed to conclude the validity of this projects core statement.