Using Machine Learning To Design A Robust Computerized Adaptive Test for Post Traumatic Stress Disorder Diagnosis

Ishanu Chattopadhyay (ishanu@uchicago.edu)
Robert Gibbons (ishanu@uchicago.edu)

Data set

No. of samples	304
No. of items (features)	211

The dataset consists of 211 items to which 304 subjects responded. The responses are integer valued in the range $[0, 5]$. Each respondant is either has a positive or a negative PTSD diagnosis.

Problem Description

Given the dataset as described above, we aim to infer a model that predicts the diagnosis given the responses to a model-directed choice of subset from the item bank. Our model has the following a priori constraint:

- At most 6 items can be used for a single subject

under which we aim to maximize performance measured by standard metrics such as the area under the reciever-operating characteristics curve (AUC).

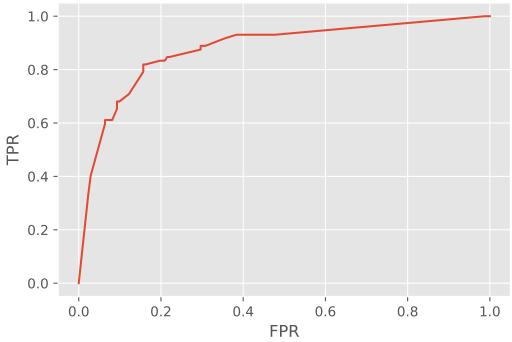
Additionally, we require that our approach has the ability to generate a plurality of distinct tests of comparable performance, i.e., two subjects taking the test are not necessarily given the same items to respond to.

Solution: Extremely Randomized Trees

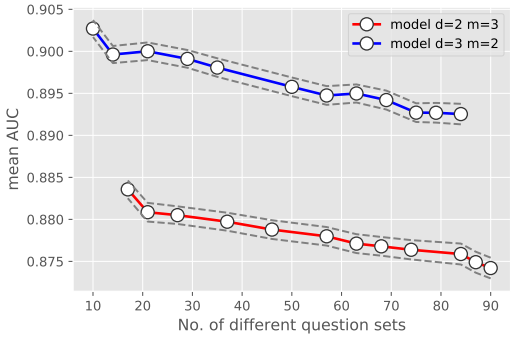
A search of the space of possible classification algorithms indicated that the *extra-trees algorithm* performs best, i.e. maximizes AUC under the constraint described above, while allowing for the generation of hundreds of distinct test sets. The Extra-Trees method (standing for extremely randomized trees) was proposed in,[1] with the objective of further randomizing tree building in the context of numerical input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree.

With respect to random forests, the method drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random. This idea is rather productive in the context of many problems characterized by a large number of numerical features varying more or less continuously: it leads often to increased accuracy thanks to its smoothing and at the same time significantly reduces computational burdens linked to the determination of optimal cut-points in standard trees and in random forests. From a statistical point of view, dropping the bootstrapping idea leads to an advantage in terms of bias, whereas the cut-point randomization has often an excellent variance reduction effect. This method has yielded state-of-the-art results in several high-dimensional complex problems. From a functional point of view, the Extra-Tree method produces piece-wise multilinear approximations, rather than the piece-wise constant ones of random forests.

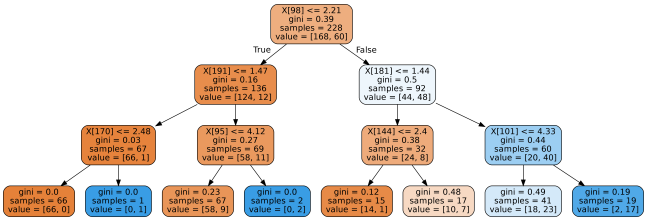
A. ROC Curve



B. AUC vs No. of Distinct Tests



C. Generated Test Example (Estimator 1)



D. Generated Test Example (Estimator 2)

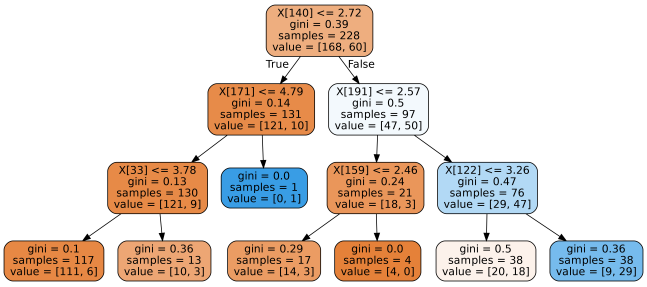


Fig. 1. Plate A shows a representative ROC curve obtained during training with random test-train split (30/70 split). Plate B shows the variation of the median performance against the number of such models obtained, which reflects the number of distinct test sets that can be generated. Plate C and D illustrate a single test, with two estimators, each being a decision tree of depth 3, implying that the maximum number of items presented: 6.

Results & Discussion

Software

All software sources are available at the Github Repository link: https://github.com/zeroknowledgediscovery/zcad

References

[1] P. Geurts, D. Ernst, and L. Wehenkel, *Extremely randomized trees*, Mach. Learn., 63 (2006), pp. 3–42.