Data-Driven Discovery of Design Specifications (Student Abstract)

Angela Chen, Nicholas Gisolfi, Artur Dubrawski

Auton Lab, School of Computer Science, Carnegie Mellon University
{angelac2,ngisolfi,awd}@andrew.cmu.edu

Abstract

Ensuring a machine learning model's trustworthiness is crucial to prevent potential harm. One way to foster trust is through the formal verification of the model's adherence to essential design requirements. However, this approach relies on well-defined, application-domain-centric criteria with which to test the model, and such specifications may be cumbersome to collect in practice. We propose a data-driven approach for creating specifications to evaluate a trained model. This framework allows us to prove that the model will exhibit safe behavior while minimizing the false-positive prediction rate. This strategy enhances predictive accuracy and safety, providing insight into the model's strengths and weaknesses, and promotes trust through a systematic approach.

Introduction

Our ability to train good models outpaces our ability to understand what exactly the model learned during training. Artificial Intelligence (AI) continues to aid in human decision-making in increasingly critical application contexts, such as healthcare, where erroneous decisions can inflict serious harm (Pinsky, Dubrawski, and Clermont 2022). How can we trust a model when we do not know for certain whether it will cause harm that would otherwise be easily avoided by a trustworthy human decision-maker?

Formal verification of model adherence to critical design specifications is a growing area of research that offers one possible solution to the problem (Sato et al. 2020). However, at times it is just as hard to determine what properties to prove about a model as it is to develop a formalism capable of verifying those properties. Current practice involves collaboration with subject matter experts (SMEs) who codify their domain expertise into a set of requirements that must govern the operational behavior of any model. This process can be confusing and time-consuming, in part because it requires thinking of inherently probabilistic systems in logical, contractual terms. We propose a data-driven framework for generating candidate specifications to use as proxy requirements for a trained model. We demonstrate the utility of the framework by proving the extent to which a model's false positive rate can be reduced while simultaneously adhering to the specification. This not only lets us strike a balance between predictive performance and safety, but it also helps us understand the strengths and weaknesses of the model.

Methodology

We describe our framework for generating candidate specifications and then expand upon our experimental pipeline.

Formalism

The Tree Ensemble Accreditor (TEA) (Gisolfi 2021) (Figure 1) is a SAT-based formalism for verifying properties of random forests. TEA encodes a random forest model and its design specifications as Boolean formulas and then converts them into Conjunctive Normal Form (CNF) to solve for Boolean satisfiability. The solver outputs one of two possible outcomes: satisfiable (SAT) or unsatisfiable (UNSAT) depending on whether the solver discovers a violation of the selected specification.

Data-Driven Design Specifications

Our algorithm generates candidate design specifications to use as inputs to TEA. They comprise input-output mappings that prescribe the prediction the ensemble must make for a contiguous range of inputs.

The Data-Driven Discovery of Design Specifications (D4SPEC) algorithm trains another random forest model on the same data. It can limit the depth of the trees in this auxiliary forest to reduce the risk of overfitting and the complexity of design specifications. For each tree in the auxiliary ensemble, we check leaf nodes against a minimum support and minimum purity requirements. Paths to the leaves that pass this check become the list of piece-wise constant threshold rules that form the bounds on inputs for our input-output mapping. The majority class label in the leaf becomes the prescribed output of our input-output mapping.

Experimental Design

We sourced our data from the publicly available Breast Cancer Wisconsin Diagnostic dataset. This data set contains 30 continuous features and two categorical outcomes. Our random forest classifier, from scikit-learn library, was trained with 80%/20% train/test data split, a maximum depth of ten, and a minimum five samples per leaf, for 100 trees. In the

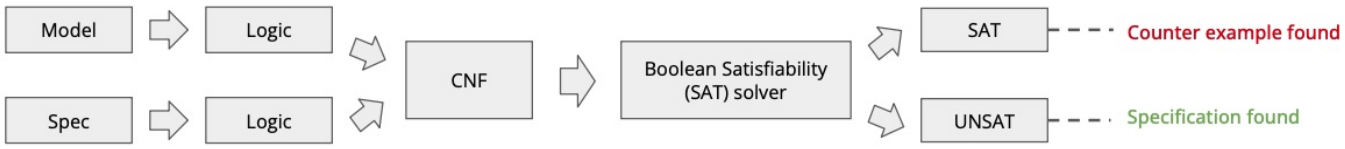

Figure 1: Tree Ensemble Accreditor

D4SPEC step, the auxiliary classifier was trained with identical parameters, except for a maximum depth of five.

We used a SPEC-ROC (Gisolfi 2021) plot to obtain the maximum level of predictive performance out of the model while simultaneously adhering to the design specification of interest. SPEC-ROC is based on the Receiver Operating Characteristic (ROC) curve, used to display the trade-off between minimizing the rate of false positives and maximizing the rate of true positives.

Analysis

We plotted a SPEC-ROC (Figure 2) for three specifications (Table 1) to understand the extent to which our model's false positive rate (FPR) can be reduced while simultaneously adhering to the specifications. Based on the plot, the lowest FPR threshold while adhering to Spec_1 is 0.336 (rounded up to three decimal places). This means our model fails to adhere to Spec_1 under any FPR lower than 0.336. Likewise, the lowest FPR thresholds for adherence to Spec_2 and Spec_3 are 0.001 and 0.026, respectively. This implies that if we want to enforce a select specification that maps inputs to alarms as a design requirement, we must set the threshold for positive predictions at or below these levels. For voting tree ensembles, an increase in FPR indicates that strictly more leaf nodes are producing positive predictions. Thus, to express this behavior, the threshold for producing an alarm needs to be reduced.

The complexity of specifications could explain the variation in FPR thresholds. For example, Spec_2, which contains the most complex rules (i.e., a combination of five feature values), has the lowest FPR threshold, whereas Spec_1, containing the simplest rules (i.e., a combination of three feature values), has the highest FPR threshold. This is because adhering to complex rules is more challenging. As a result, Spec_2 has the largest safety region (any FPR greater than the threshold). These three modalities illustrate different possible outcomes from our experiments. The model will satisfy Spec_2 regardless of the prediction threshold set by a user. On the other hand, the model can only satisfy Spec_1 if it severely limits its discriminative capabilities—this exemplifies what happens when a candidate specification does not align with the learned decision logic of the model. Spec_3 is interesting because its satisfaction or violation depends on the setting of the predictive threshold, providing users an option to prioritize predictive performance or safety, meaning adherence to explicit design specifications.

Attribute	Spec_1	Spec_2	Spec_3
Perimeter Error	<2.57		
Worst Area		<811.05	<552.95
Worst Perimeter			<105.95
Worst Concave Points	<0.12		
Worst Texture		<32.0	
Area Error			<79.83
Worst Radius	<17.34		
Concave Points Error		>0.01	>0.01
Worst Concave Points		<0.16	
Symmetry Error		>0.02	

Table 1: The three specifications comprise different combinations of features and value ranges.

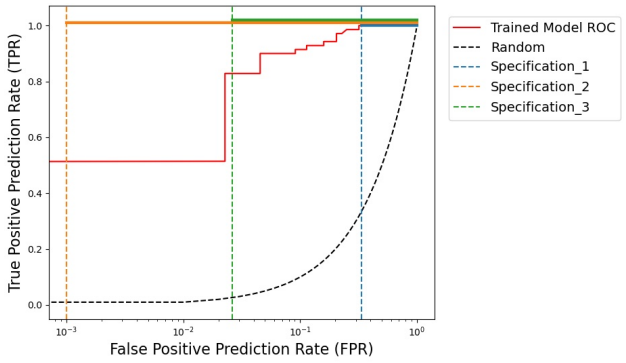

Figure 2: A log-scaled ROC for our classifier along with different predictive thresholds for the design specifications outlined in Table 1.

Acknowledgements

This work was partially supported supported by a Space Technology Research Institutes grant from NASA's Space Technology Research Grants Program and by the U.S. Army Research Office and the U.S. Army Futures Command under contract W911NF-20-D-0002.

References

Gisolfi, N. 2021. *Model-centric verification of artificial intelligence*. Ph.D. thesis, Carnegie Mellon University.

Pinsky, M. R.; Dubrawski, A.; and Clermont, G. 2022. Intelligent Clinical Decision Support. *Sensors*, 22(4): 1408.

Sato, N.; Kuruma, H.; Nakagawa, Y.; and Ogawa, H. 2020. Formal verification of a decision-tree ensemble model and detection of its violation ranges. *IEICE TRANSACTIONS on Information and Systems*, 103(2): 363–378.