Formal Reasoning about Input-Output Relations of Tree Ensembles

John Törnblom
john.tornblom@saabgroup.com

Description

AI advances are now being applied in critical systems where software defects may cause harm to humans and the environment. Machine learning models with large sets of parameters are difficult to interpret, which calls for computer-aided reasoning, both for the purpose of verification (using deductive reasoning), and for explaining predictions (using abductive reasoning).

Motivation

Safety

Several industries are now offering products equipped with machine learning based components, advertised with super-human functionalities. To be applicable in safety-critical domains, however, we need new methods that can provide convincing arguments that such components behave correctly.

Explainability

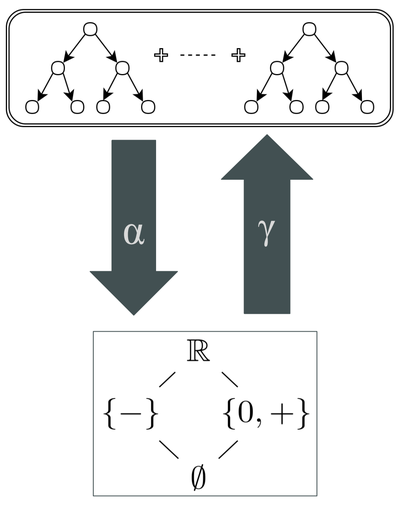
Critical decision support systems must sometimes provide their operators with explanations of its decisions, e.g., for medical diagnosis and in flight management systems. Formal methods can be used to synthesize such explanations that are both provably correct, and without redundant information.

Research Goal & Questions

Our goal is to develop rigorous and scalable formal methods tailored specifically for analyzing input-output relations of tree ensembles, e.g., robustness against input perturbations.

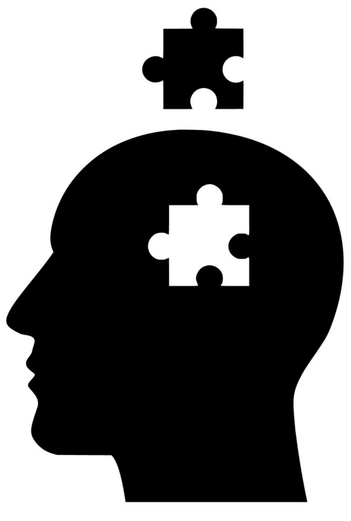
- what intrinsic characteristics of tree ensembles affect the scalability of algorithms that reason about their input-output relations?
- Is it possible to formally analyze input-output relations of non-trivial tree ensembles with reasonable amounts of computing resources?

Contributions



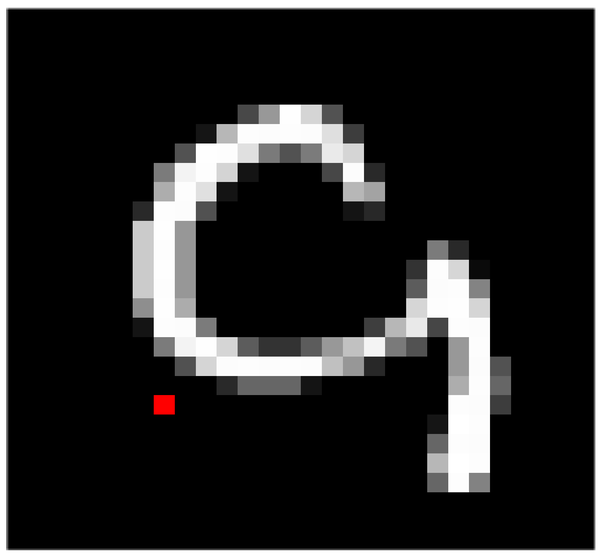
- A mathematical formalization of tree ensembles in the abstract interpretation framework.
- An abstraction-refinement method that counteracts combinatorial explosion when exploring path combinations in tree ensembles.
- VoTE, a fast and memory-efficient implementation of the methods, tailored specifically for random forests and gradient boosting machines.

Research Insights



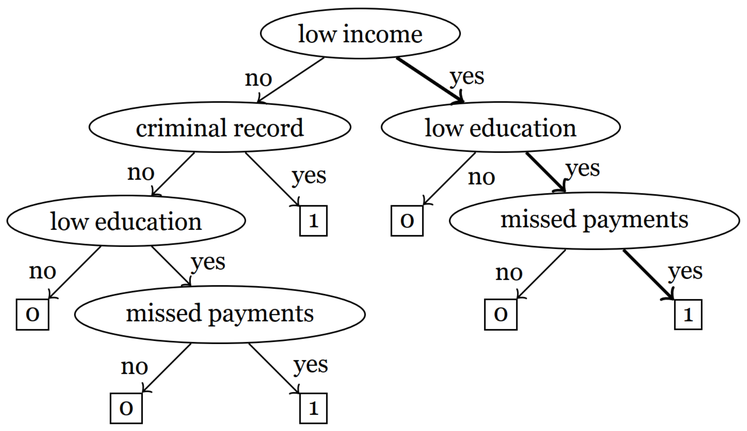
- Complexity is primarily determined by the number of trees and their depths.
- Input space is partitioned into hyper rectangles, which can be captured efficiently and precisely in the interval domain.
- Low-dimensional input spaces contain infeasible path combinations. Detecting these early improves performance significantly.

Verify Robustness



- Robustness against certain type of additive noise can be verified formally for complex tree ensembles.
- When noise can be captured in the interval domain, formal methods outperform point-wise testing by several orders of magnitude.

Explaining Predictions



A fictive bank-loan scenario in which the applicant is denied the loan, where the explanation “low education” and “missed payments” is both correct and without redundant information.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM