Temporal Logic Explanations for Dynamic Decision Systems Using Anchors and Monte Carlo Tree Search (Abstract Reprint)

Tzu-Yi Chiu[1], **Jerome Le Ny**[1], **Jean-Pierre David**[1]

[1]Electrical Engineering Department, Ecole Polytechnique Montreal

Abstract Reprint. This is an abstract reprint of a journal article by Chiu, Ny, and David (2023).

Abstract

For many automated perception and decision tasks, state-of-the-art performance may be obtained by algorithms that are too complex for their behavior to be completely understandable or predictable by human users, e.g., because they employ large machine learning models. To integrate these algorithms into safety-critical decision and control systems, it is particularly important to develop methods that can promote trust into their decisions and help explore their failure modes. In this article, we combine the anchors methodology with Monte Carlo Tree Search to provide local model-agnostic explanations for the behaviors of a given black-box model making decisions by processing time-varying input signals. Our approach searches for descriptive explanations for these decisions in the form of properties of the input signals, expressed in Signal Temporal Logic, which are highly likely to reproduce the observed behavior. To illustrate the methodology, we apply it in simulations to the analysis of a hybrid (continuous-discrete) control system and a collision avoidance system for unmanned aircraft (ACAS Xu) implemented by a neural network.

References

Chiu, T.-Y.; Ny, J. L.; and David, J.-P. 2023. Temporal logic explanations for dynamic decision systems using anchors and Monte Carlo Tree Search. *Artificial Intelligence*, 318: 103897.