

Theses for the Disputation

Interpretable Substitutions for Predictive Artificial Neural Networks

Patrick Suwinski

supervised by Prof. Dr.-Ing. Rafael Felix Schaefer

Faculty of Electrical and Computer Engineering

TUD Dresden University of Technology

to achieve the academic degree

Doktoringenieur (Dr.-Ing.)

Predictive Artificial Neural Networks (ANNs) offer enormous potential for complex control tasks in automation. However, their inherent black box character prevents their direct use in time- and safety-critical environments. Substituting ANNs with interpretable surrogates that offer the same performance as ANNs while reducing the requirements for model-inference hardware proposes a solution for applications in automation.

This dissertation assesses the following claims:

- I. Predictive ANNs do not meet the requirements of traditional, rule-based automation. Their stochastic character and lack of formal verifiability categorically rule out their direct use in traditional, rule-based automation systems due to a lack of interpretability.
- II. The use of evolutionary algorithms for symbolic regression (SR) transforms the interpretability problem into a data-driven form of system identification. As a machine learning method, SR enables the generation of surrogates as intrinsically interpretable symbolic models from trained predictive ANNs, e.g., as mathematical equations.
- III. Established compression methods, such as pruning or quantization, only reduce computational requirements without addressing the fundamental transparency problem. In contrast, complete substitution with symbolic models enables formal mathematical verification for safety-critical applications while significantly reducing volatile and non-volatile memory usage.
- IV. Substituting ANNs through mathematical equations significantly reduces both memory footprint and computational requirements. In typical industrial hardware such as micro-controllers and programmable logic controllers, severe memory constraints are a far greater bottleneck than raw computing performance. Drastically lowering volatile and non-volatile memory requirements through symbolic surrogates ultimately enables the deployment of predictive models at the lowest level of the automation pyramid.
- V. The interpretable surrogates generated do not guarantee that the overall system can operate in real time on its own. However, they do provide the necessary determinism and inferability on (micro-)processors as a basic prerequisite for real-time hardware and software architectures to be able to fulfill real-time guarantees.

- VI. The substitution of trained ANNs with mathematical equations fundamentally changes the operational error modes in unpredictable system states. Analytical surrogates extrapolate strictly deterministically outside their training domain, whereas predictive ANNs generate probability distributions that are difficult to quantify.
- VII. External monitoring mechanisms for ANNs mask the underlying uncertainty problem rather than solving it. The intrinsic transparency of surrogates generated using SR eliminates the need for error-prone runtime monitoring, thanks to a priori analyzable, deterministic system structure.
- VIII. The high energy consumption of ANN inferences contradicts the strict efficiency targets of modern production facilities. The significant reduction in mathematical operations enabled by generated surrogates maximizes energy efficiency on target industrial hardware, such as microcontrollers and programmable logic controllers.