AAAI Fall Symposium on Integrated Approaches to Computational Scientific Discovery

Bayesian probabilistic grammar-based equation discovery

Jure Brence, Ljupčo Todorovski, Sašo Džeroski

We present methods for the automated discovery of closed-form equations from data. The methods use probabilistic grammars to define and constrain the space of mathematical expressions. They use an iterative Bayesian algorithm to search this space.

Equation discovery and symbolic regression address the task of finding a symbolic mathematical model that best describes observed data. Models can be as simple as an algebraic equation or as complex as a system of differential equations. Equation discovery methods seek to automate the identification of equation structure as well as parameters. The advantage of discovering closed-form equations over black-box models, popular in machine learning, lies in the inherent interpretability of equations and the connections that can be made between equations and domain knowledge.

Following the generate-and-test paradigm, we use a probabilistic generator (grammar) to generate candidate expressions for the right-hand side of an equation and use numerical optimization to fit numerical parameters in the equation to the data. Probabilistic context-free grammars (PCFGs) are useful for generating mathematical expressions, constraining the space of expressions, and encoding domain knowledge. Crucially, PCFGs define probability distributions over mathematical expressions. In addition to hard constraints imposed by the production rules of context-free grammars (CFGs), production rule probabilities in PCFGs allow us to impose soft constraints on the search space of mathematical expressions. Through these probabilities, PCFGs can also inherently and intuitively parametrize the parsimony principle.

Although PCFGs are a powerful tool for encoding domain knowledge, their context blindness can limit their applicability to more complex types of knowledge. We improve the expressivity of PCFGs by equipping symbols and production rules with attributes. The resulting probabilistic attribute grammars (PAGs) can encode complicated constraints, such as dimensional consistency and interactions between coupled differential equations that describe dynamical systems.

The simplest way of searching the space of expressions defined by a probabilistic grammar (PCFG or PAG) is a Monte-Carlo algorithm that samples random mathematical expressions from the distribution imposed by the grammar. To pave the way toward more sophisticated grammar-based equation discovery algorithms, we present a novel Bayesian algorithm for sampling mathematical expressions from a probabilistic grammar. The algorithm iteratively updates grammar probabilities based on the error-of-fit and prior probability of generated expressions.

This feedback loop guides the search towards more promising areas of the space of mathematical expressions and conditions the grammar to the data. The Bayesian algorithm improves the equation discovery performance and enables the estimation of the posterior distribution. We can interpret the resulting grammar as a representation of domain knowledge, updated by considering new evidence.

References

- Jure Brence, Ljupčo Todorovski, Sašo Džeroski. “Probabilistic grammars for equation discovery”. *Knowledge-Based Systems* 224 (2021).
- Jure Brence, Sašo Džeroski, Ljupčo Todorovski. “Dimensionally-consistent equation discovery through probabilistic attribute grammars”. *Information Sciences* 632 (2023).
- Nina Omejc, Boštjan Gec, Jure Brence, Ljupčo Todorovski, Sašo Džeroski. "Probabilistic grammars for modeling dynamical systems from coarse, noisy, and partial data." *Machine Learning, Special Issue on Discovery Science 2022* (2024).
- Jure Brence. “Probabilistic grammar-based equation discovery”. Doctoral dissertation, 2024.