Efficient Generator of Algebraic Expressions for Symbolic Regression

Sebastian Mežnar[1,2], Sašo Džeroski[1,2], and Ljupčo Todorovski[1,3]

(1) Jožef Stefan Institute, Department of Knowledge Technologies, Jamova 39, Ljubljana, Slovenia
(2) Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
(3) University of Ljubljana, Faculty of Mathematics and Physics, Jadranska 21, Ljubljana, Slovenia

Domain-specific modeling knowledge is commonly used in symbolic regression and equation discovery systems to constrain the space of candidate equations. Typical knowledge-based systems, such as grammar-based equation discovery or process-based modeling, require the user to manually encode the modeling knowledge into grammars or libraries of generic model components, such as entities and processes. The system transforms the formalized knowledge into a generator of candidate equations.

Here, we present an approach where the generator of candidate equations is trained from a corpus of expressions (Mežnar et al., 2023). With an appropriate selection of the expressions in the corpus, the generator can be tailored to the domain of interest. It uses a hierarchical variational autoencoder (HVAE) of algebraic expression trees. State-of-the-art sequential variational autoencoders (Gómez-Bombarelli et al., 2018; Kusner et al., 2017) suffer several inefficiencies when generating expressions. First, they often generate syntactically incorrect expressions. Moreover, they require large amounts of training data (expressions) and high-dimensional latent vector spaces to represent the expressions. In contrast, we empirically show that HVAE can be efficiently trained to generate syntactically correct expressions from small corpora of expressions while embedding them into low-dimensional spaces. We also show that syntactically similar expressions have similar representational vectors in the latent space.

The embedding of expressions in a low-dimensional vector space enables the use of various methods for numerical optimization to select an optimal equation structure. In a series of computational experiments, we show that the system based on combining the HVAE with genetic algorithms outperforms alternative symbolic regression systems (Mundhenk et al., 2021) on standard symbolic regression benchmarks, such as the Feynman dataset and Nguyen equations.

The ability to train the generator of equations from corpora of expressions opens several venues for further development of equation discovery systems. First, we can train generators from equations found in the relevant literature (textbooks and scientific articles) in the domain of interest. Moreover, by training the generators on sets of accurate equations discovered in one data set, we can transfer the knowledge from one modeling experiment to other experiments and modeling problems in the same domain. Finally, the generator of expressions can be iteratively trained on the results of a series of equation discovery experiments. In each iteration, the focus of the generator is narrowed down to the space of accurate candidate equations.

Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science 4(2), 268–276*. DOI: 10.1021/acscentsci.7b005

Mežnar, S., Džeroski, S., and Todorovski, L. (2023). Efficient Generator of Algebraic Expressions for Symbolic Regression. *arXiv*. DOI: 10.48550/arXiv.2302.09893

Mundhenk, T.N., Landajuela, M., Glatt, R., et al. (2021). Symbolic Regression via Neural-Guided Genetic Programming Population Seeding. *arXiv*. DOI: 10.48550/ARXIV2111.00053

Kusner, M.J., Paige, B., and Hernández-Lobato, J.M. (2017). Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning (pp. 1945–1954)*. DOI: 10.48550/ARXIV.1703.01925