Quantifying expression dissimilarity

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

(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

Equation discovery, a fundamental task of scientific discovery, seeks to identify concise mathematical expressions that accurately model given data. However, prevailing equation discovery methods encounter a common challenge: they explore a discrete space of expressions, where subtle modifications of an expression can produce a new expression with dramatically different behavior, which hinders the search for optimal solutions to equation discovery tasks. This challenge can be addressed by generative models that map expressions to a latent space, where small moves from one expression lead to expressions that behave similarly (Mežnar et al., 2023). New behavior-based expression dissimilarity metrics are needed to train such models, as using established, syntax-based metrics is at the core of the above-mentioned challenge. To this end, we propose BBED, a behavior-based expression distance (Mežnar et al., 2024).

Quantifying dissimilarity between expressions raises several issues that need to be resolved. Firstly, behavior-based dissimilarity depends on the considered domain of the expression variables. Two expressions might behave similarly when observed over one domain, while behaving dissimilarly over another. Because of this, we must consider the domain of the expression variables when discussing the dissimilarity of behavior between expressions. Secondly, candidate expressions in equations discovery include constants with unknown values to be fitted against data. The same expression might behave significantly differently when we change the values of constants. Because of this, we need to find a way to incorporate constants with unknown values into a behavior-based metric.

A possible approach to resolve these issues is to view mathematical expressions as a probability distribution of the expression value (output) over the domain variables and constants. This perspective allows us to use metrics from probability theory, specifically the Wasserstein distance. This option is particularly well-suited for our problem because it considers the relative importance of different output values rather than simply comparing their means or variances. Additionally, it is shown that the Wasserstein distance closely correlates with human perception of similarity. However, we must take care of some boundary cases when using the Wasserstein distance. These occur because an expression might not contain constants or be defined over the whole domain. The final metric is stochastic since we sample from the domain and the space of constants, which are both continuous.

We evaluate two aspects of our metric. First, we test the consistency of our stochastic metric with different parameter settings. We run the calculation of BBED between 100 expressions 100 times and calculate the Spearman-rho correlation coefficient between the results of different runs for each expression. The results show that BBED is robust: we can achieve a correlation of 0.99 or more with a modest number of samples and runs. Secondly, we test the smoothness of the error manifold. Here, we expect that the expressions at a low distance will also have a similar error on a given data set. We compare BBED to two syntax-based metrics (edit distance and tree-edit distance) and an optimal metric, where optimal constant values are fitted to the data set before calculating the distances. We find that in most cases, our metric outperforms both syntax-based metrics. Additionally, when the correct domain is chosen, the performance of BBED is comparable to the performance of the optimal metric.

Mežnar, S., Džeroski, S., and Todorovski, L. (2023). Efficient generator of mathematical expressions for symbolic regression. Mach Learn 112, 4563–4596. DOI: 10.1007/s10994-023-06400-2

Mežnar, S., Džeroski, S., and Todorovski, L. (2024). Quantifying behavioural distance between mathematical expressions. arXiv. DOI: 10.48550/arXiv.2408.11515