Tony Worm (Binghamton, Computer Science)

Prioritized Grammar Enumeration for Fitting Equations to Data

Abstract for the Combinatorics Seminar 2013 September 10

Symbolic Regression (SR) is the task of discovering the best-fitting equations for a given data set. Unlike in linear and non-linear regression, the functional form is unknown. In this sense, SR can be viewed as a form of data mining.

Kenneth Chiu and I have introduced a deterministic method for SR, called Prioritized Grammar Enumeration (PGE) (our paper is linked here). We formulate the SR problem as a search within the grammar of Mathematics. Equations are represented as their parse trees and grammatical rules for equations are used as generating rules. Starting from a small initial set of simple equations, the generating rules are recursively applied to current members of the set, producing new members.

This process expands the set at an exponential rate. We combat this with several techniques. First, the fact that addition and multiplication are commutative and associative makes for two combinatorial reductions in the size of the space to be explored and enables us to create a canonical tree for each equation. Second, since the same generating rules, when applied in different orders, can produce the same equation or tree, we use dynamic programming to memorize and detect these situations. This transforms the search tree into a search graph, with equation trees at the vertices and edges that represent the localized application of a generating rule. The problem then becomes analogous to finding the shortest path on the graph defined by the initial set of equations (or trees), a set of generating rules, and metrics for determining edge weights and node fitness.

From:
http://www2.math.binghamton.edu/ - **Department of Mathematics and Statistics, Binghamton University**

Permanent link:
http://www2.math.binghamton.edu/p/seminars/comb/abstract.201309wor

Last update: **2020/01/29 19:03**