

Convergence of distributed symbolic regression.

Ben Cardoen
ben.cardoen@student.uantwerpen.be

April 29, 2017

1 Abstract

Symbolic regression (SR) fits symbolic expression to a dataset of expected values. Amongst its advantages over other techniques are the ability for a practitioner to interpret the resulting expression, determine important features by their usage in the expression, and an insight into the behavior of the resulting model (e.g. continuity, derivation, extrema). SR combines a discrete combinatoric problem (combining base functions) with a continuous optimization problem (selecting and mutating constants). One of the main algorithms used in SR is genetic programming. The convergence characteristics of SR using GP are still an open issue. In this work we will use a distributed GP-SR implementation to evaluate the effect of topologies on the convergence of the algorithm. We introduce and evaluate a new topology with the aim of finding a new balance between structured and random diffusion. We introduce a variation of k-fold cross validation to estimate how accurate a generated solution is in predicting unknown datapoints. This validation technique is implemented in parallel in the algorithm combining both the advantages of the cross validation with the increase in covered search space for each instance. We validate our work on several test problems, and a real world use case.

2 Introduction

3 Design

4 Experiments

5 Related Work

6 Conclusion

7 Future work