Lightweight Symbolic Regression with Interaction-Transformation Representation

Guilherme Seidyo Imai Aldeia   Prof. Fabricio Olivetti de França

Federal University of ABC
Center for Mathematics, Computation and Cognition (CMCC)
Heuristics, Analysis and Learning Laboratory (HAL)

12 de Julho de 2018

Topics

1. Introduction
2. Interaction-Transformation
3. Lab Assistant
4. Some experiments
5. Conclusion

Introduction

Regression Analysis

Regression Analysis studies the relationship between a **dependent variable** (y) and one or more **independent variables** ($\mathbf{x}$)

Linear Regression

Models the relationship as a linear combination:

$$\hat{f}(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x}.$$

Linear Regression

- Easy to understand the impact of every variable
- How can we fit a wave function?

Multi-layer Perceptron

The Multi-layer Perceptron, with one hidden layer, adjusts the function:

$$\hat{f}(\mathbf{x}, \mathbf{w}) = \mathbf{w_2} \cdot g(\mathbf{w_1} \cdot \mathbf{x}).$$

where g is an *activation function*.

Multi-layer Perceptron

- It is a **universal approximator**
- Though conceptually closed form, the topology can be evolved, thus exploring the function form a bit
- What is the meaning of tanh(tanh(tanh(tanh(tanh(. . . ?

Symbolic Regression

Symbolic Regression searches for a function form and adjust the free parameters at the same time.

A secondary objective is that this function assumes the simplest form possible.

Genetic Programming

- Evolutionary algorithms: Genetic Programming, Gene Expression, etc.
- Explore the whole mathematical expressions search space
- Expression trees, linear data, grammar, etc.

Genetic Programming

Problems:

- Huge search space
- Many local and global optima (equivalent expressions)

Symbolic Regression

Example:

$$f(x) = \frac{x^3}{6} + \frac{x^5}{120} + \frac{x^7}{5040}$$
$$f(x) = \frac{16x(\pi - x)}{5\pi^2 - 4x(\pi - x)}$$
$$\mathbf{f}(\mathbf{x}) = \sin(\mathbf{x}).$$

Genetic Programming

Solutions:

- Introduce a complexity measure in the objective
- Restricted search space

Interaction-Transformation

Interaction-Transformation

Restrict the function form as a **linear combination** of the application of different **transformation functions** to **interactions** of the original variables.

Interaction-Transformation

$$\hat{f}(x) = \sum_i w_i \cdot t_i(p_i(x))$$

$$p(x) = \prod_{i=1}^{d} x_i^{k_i}$$

$$t_i = \{id, \sin, \cos, \tan, \sqrt{}, \log, \ldots\}$$

Interaction-Transformation

Valid expressions:

- $w_1 \cdot x_1 + w_2 \cdot x_2$
- $3.5 \sin\left(x_1^2 \cdot x_2\right) + 5 \log\left(x_2^3 / x_1\right)$

Interaction-Transformation

Invalid expressions:

- $\tanh\left(\tanh\left(\tanh\left(w \cdot x\right)\right)\right)$
- $\sin\left(x_1^2 + x_2\right)/x_3$

SymTree

Simple algorithm to find an IT expression, e.g., given $x = \{x_1, x_2\}$, starts from a Linear Regression:

$$it = w_1 \cdot id(x_1^1 \cdot x_2^0) + w_2 \cdot id(x_1^0 \cdot x_2^1)$$

SymTree

Create new terms to evaluate by interacting pairs of terms:

$$t_1 = id(x_1^1 \cdot x_2^1)$$
$$t_2 = id(x_1^1 \cdot x_2^{-1})$$
$$t_3 = id(x_1^{-1} \cdot x_2^1)$$

Create new terms by changing the current transformation functions:

$$t_4 = \sqrt{x_1^1 \cdot x_2^0}$$
$$t_5 = \sin(x_1^1 \cdot x_2^0)$$
$$t_6 = \sqrt{x_1^0 \cdot x_2^1}$$
$$t_7 = \sin\left(x_1^0 \cdot x_2^1\right)$$

SymTree

Create one or more IT expressions by adding these terms to the current expression:

$$it = w_1 \cdot id(x_1^1 \cdot x_2^0) + w_2 \cdot id(x_1^0 \cdot x_2^1) + w_3 \cdot \sqrt{x_1^0 \cdot x_2^1}$$

SymTree

In (de França, 2018)[1] SymTree was shown to be lightweight and capable of outperform different Symbolic Regression, Linear and Nonlinear Regression approaches.

[1] de Franca, Fabricio Olivetti. "A Greedy Search Tree Heuristic for Symbolic Regression." Information Sciences (2018).

Lab Assistant

Lab Assistant

Objective: proof of the concept of SymTree as practical tool for regression analysis.

Client-side Web tool for Symbolic Regression developed with HTML + JavaScript.

SymTree in your Browser!

https://galdeia.github.io/

Lab Assistant

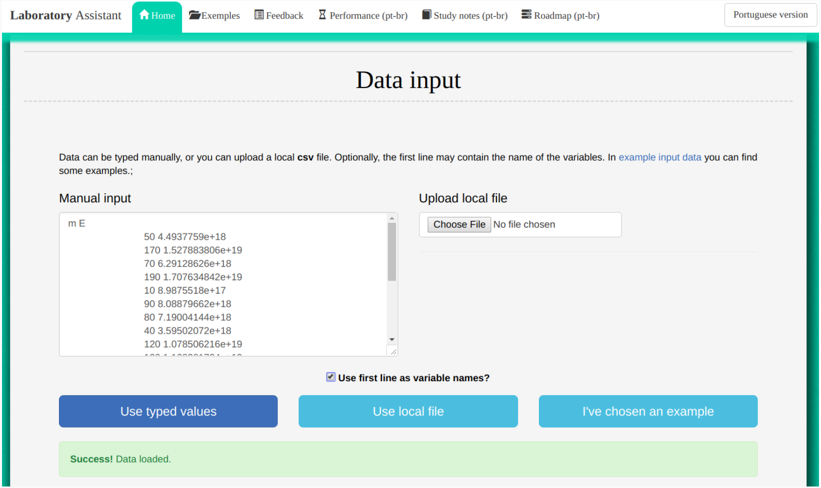
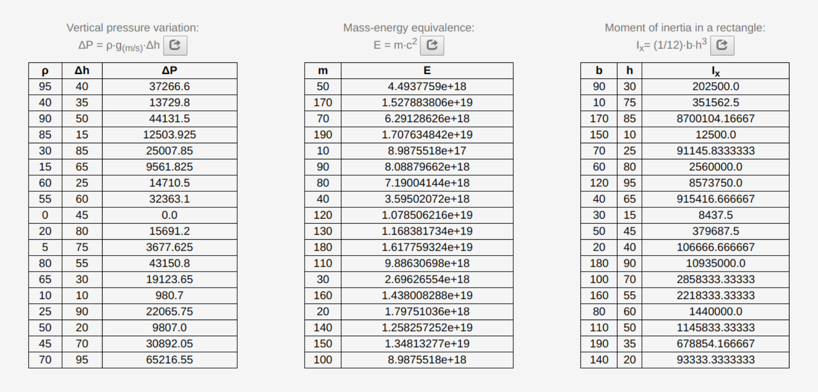


Figura 1: Main Interface



Vertical pressure variation:
$\Delta P = \rho \cdot g_{(m/s)} \cdot \Delta h$

ρ	Δh	ΔP
95	40	37266.6
40	35	13729.8
90	50	44131.5
85	15	12503.925
30	85	25007.85
15	65	9561.825
60	25	14710.5
55	60	32363.1
0	45	0.0
20	80	15691.2
5	75	3677.625
80	55	43150.8
65	30	19123.65
10	10	980.7
25	90	22065.75
50	20	9807.0
45	70	30892.05
70	95	65216.55

Mass-energy equivalence:
$E = m \cdot c^2$

m	E
50	4.4937759e+18
170	1.527883806e+19
70	6.29128626e+18
190	1.707634842e+19
10	8.9875518e+17
90	8.08879662e+18
80	7.19004144e+18
40	3.59502072e+18
120	1.078506216e+19
130	1.168381734e+19
180	1.617759324e+19
110	9.88630698e+18
30	2.69626554e+18
160	1.438008288e+19
20	1.79751036e+18
140	1.258257252e+19
150	1.34813277e+19
100	8.9875518e+18

Moment of inertia in a rectangle:
$I_x = (1/12) \cdot b \cdot h^3$

b	h	I_x
90	30	202500.0
10	75	351562.5
170	85	8700104.16667
150	10	12500.0
70	25	91145.8333333
60	80	2560000.0
120	95	8573750.0
40	65	915416.666667
30	15	8437.5
50	45	379687.5
20	40	106666.666667
180	90	10935000.0
100	70	2858333.33333
160	55	2218333.33333
80	60	1440000.0
110	50	1145833.33333
190	35	678854.166667
140	20	93333.3333333

Figura 2: Main Interface

Lab Assistant

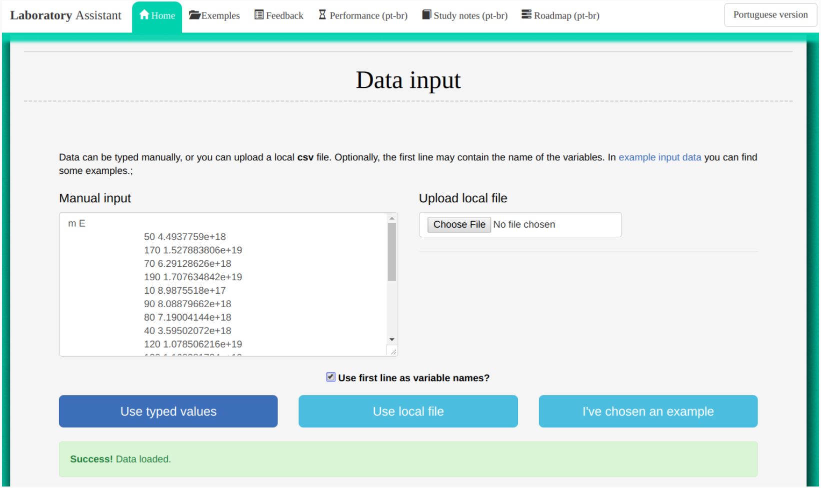


Figura 3: Main Interface

Lab Assistant

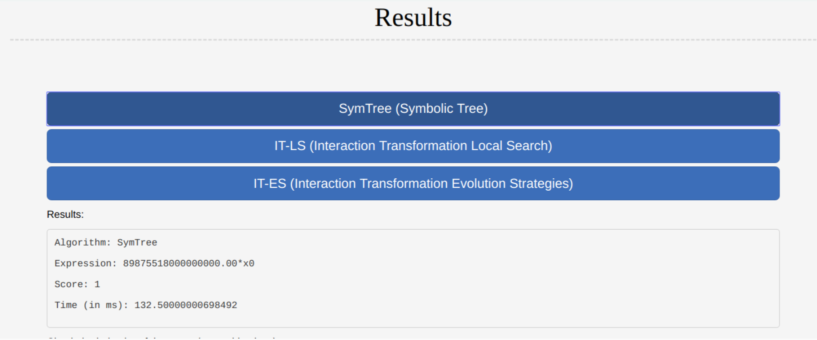


Figura 4: Main Interface

Lab Assistant

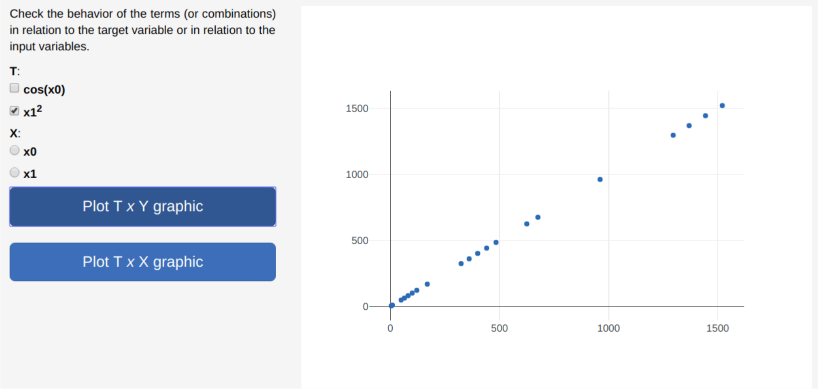


Figura 5: Main Interface

Lab Assistant

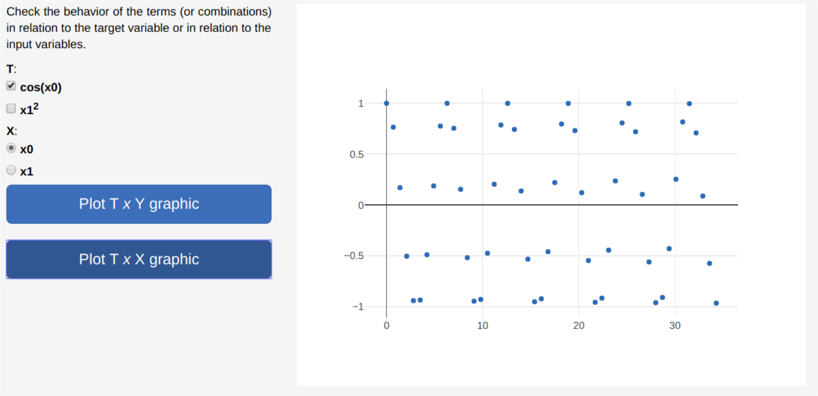


Figura 6: Main Interface

Some experiments

Methodology

- 20 different Physics and Engineering equations
  - 14 can be represented as an IT-expression
- 30 executions of each algorithm
  - Comparison between SymTree and Eureqa
  - IT-LS and IT-ES results in the paper
- Score $= \frac{1}{1+MAE}$

Eureqa

- Without any preprocessing (same as Lab Assistant)
- With an execution time budget of 3 minutes (more than Lab Assistant)

Results

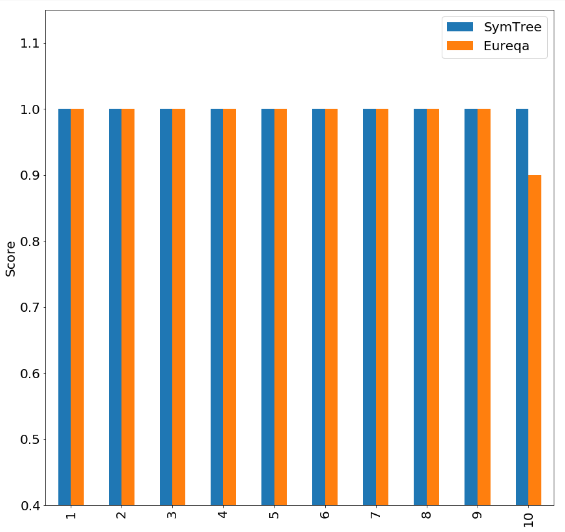


Figura 7: Score for the first 10 functions

Results

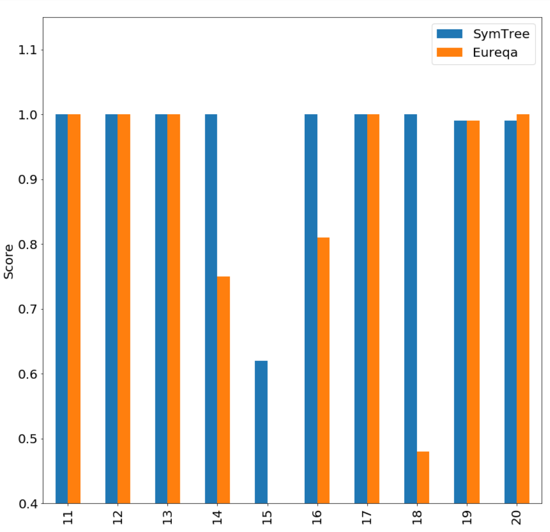


Figura 8: Score for the next 10 functions

Conclusion

Conclusions

- Lab Assistant is a proof of concept of how SymTree algorithm can be used in low-cost devices to find good approximation models.
- The Models are usually simpler than those generated by black box approaches and more accurate than linear models.

Future Research

Next in line:

- Create a prototype for board computers (Raspberry Pi)
- Expand IT expressions to include even more expressions
- Test the performance on real world regression problems

Acknowledgement

The authors would like to thank UFABC for their support.

Try it!

You can try it yourself! It works even on mid-range Smartphones!

https://galdeia.github.io/