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
1. Introduction

2. Interaction-Transformation

3. Lab Assistant

4. Some experiments

5. Conclusion
Introduction
Regression Analysis studies the relationship between a dependent variable ($y$) and one or more independent variables ($x$)
Linear Regression

Models the relationship as a linear combination:

\[
\hat{f}(x, w) = w \cdot x.
\]
- Easy to understand the impact of every variable
- How can we fit a wave function?
The Multi-layer Perceptron, with one hidden layer, adjusts the function:

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

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 \( \tanh(\tanh(\tanh(\tanh(\tanh(\ldots) \) ? \)
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.
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)
Example:

\[
f(x) = \frac{x^3}{6} + \frac{x^5}{120} + \frac{x^7}{5040}
\]

\[
f(x) = \frac{16x(\pi - x)}{5\pi^2 - 4x(\pi - x)}
\]

\[
f(x) = \sin(x).
\]
Genetic Programming

Solutions:

- Introduce a complexity measure in the objective
- Restricted search space
Interaction-Transformation
Restrict the function form as a **linear combination** of the application of different **transformation functions** to **interactions** of the original variables.
\[ \hat{f}(x) = \sum_i w_i \cdot t_i(p_i(x)) \]

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

\[ t_i = \{id, \sin, \cos, \tan, \sqrt{}, \log, \ldots\} \]
Valid expressions:

- $w_1 \cdot x_1 + w_2 \cdot x_2$
- $3.5 \sin (x_1^2 \cdot x_2) + 5 \log (x_2^3 / x_1)$
Invalid expressions:

- $\tanh (\tanh (\tanh (w \cdot x)))$
- $\sin (x_1^2 + x_2)/x_3$
Simple algorithm to find an IT expression, e.g., given $x = \{x_1, x_2\}$, starts from a Linear Regression:

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

\[
t_4 = \sqrt{x_1^1 \cdot x_2^0}
\]
\[
t_5 = \sin(x_1^1 \cdot x_2^0)
\]
\[
t_6 = \sqrt{x_1^0 \cdot x_2^1}
\]
\[
t_7 = \sin(x_1^0 \cdot x_2^1)
\]
Create one or more IT expressions by adding these terms to the current expression:

\[ it = w_1 \cdot id(x_1^1 \cdot x_2^0) + w_2 \cdot id(x_0^0 \cdot x_2^1) + w_3 \cdot \sqrt{x_1^0 \cdot x_2^1} \]
In (de França, 2018)⁠¹ SymTree was shown to be lightweight and capable of outperform different Symbolic Regression, Linear and Nonlinear Regression approaches.

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/
Data input

Data can be typed manually, or you can upload a local csv file. Optionally, the first line may contain the name of the variables. In example input data you can find some examples.

Manual input

```
50 4.4937759e+18
170 1.527838306e+19
70 6.29128626e+18
190 1.707634842e+19
10 8.9857518e+17
90 8.09879662e+18
80 7.19004144e+18
40 3.59502072e+18
120 1.078506216e+19
```

Upload local file

Choose File

No file chosen

Use first line as variable names?

Use typed values

Use local file

I've chosen an example

Success! Data loaded.

**Figura 1:** Main Interface
**Figura 2: Main Interface**

Vertical pressure variation:
\[ \Delta P = \rho \cdot g \cdot (\text{mass}) \cdot \Delta h \]

<table>
<thead>
<tr>
<th>p</th>
<th>∆h</th>
<th>∆P</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>40</td>
<td>3726.6</td>
</tr>
<tr>
<td>40</td>
<td>35</td>
<td>1372.9</td>
</tr>
<tr>
<td>90</td>
<td>50</td>
<td>4413.1</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
<td>12503.9</td>
</tr>
<tr>
<td>30</td>
<td>85</td>
<td>25007.8</td>
</tr>
<tr>
<td>15</td>
<td>65</td>
<td>9561.8</td>
</tr>
<tr>
<td>60</td>
<td>25</td>
<td>14710.5</td>
</tr>
<tr>
<td>55</td>
<td>60</td>
<td>32363.1</td>
</tr>
<tr>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>15691.2</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>3677.6</td>
</tr>
<tr>
<td>80</td>
<td>55</td>
<td>43150.8</td>
</tr>
<tr>
<td>65</td>
<td>30</td>
<td>19123.6</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>980.7</td>
</tr>
<tr>
<td>25</td>
<td>90</td>
<td>22065.75</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>980.7</td>
</tr>
<tr>
<td>45</td>
<td>70</td>
<td>30892.05</td>
</tr>
<tr>
<td>70</td>
<td>95</td>
<td>65216.55</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>m</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4.4937759e+18</td>
</tr>
<tr>
<td>170</td>
<td>1.527883806e+19</td>
</tr>
<tr>
<td>70</td>
<td>6.29128626e+18</td>
</tr>
<tr>
<td>190</td>
<td>1.707634842e+19</td>
</tr>
<tr>
<td>10</td>
<td>8.9875518e+17</td>
</tr>
<tr>
<td>90</td>
<td>8.08879662e+18</td>
</tr>
<tr>
<td>80</td>
<td>7.19004144e+18</td>
</tr>
<tr>
<td>40</td>
<td>3.59502072e+18</td>
</tr>
<tr>
<td>120</td>
<td>1.078506216e+19</td>
</tr>
<tr>
<td>130</td>
<td>1.168381734e+19</td>
</tr>
<tr>
<td>180</td>
<td>1.617759324e+19</td>
</tr>
<tr>
<td>110</td>
<td>9.88630698e+18</td>
</tr>
<tr>
<td>30</td>
<td>2.69626554e+18</td>
</tr>
<tr>
<td>160</td>
<td>1.43808288e+19</td>
</tr>
<tr>
<td>20</td>
<td>1.79751036e+18</td>
</tr>
<tr>
<td>140</td>
<td>1.25825725e+19</td>
</tr>
<tr>
<td>150</td>
<td>1.34813277e+19</td>
</tr>
<tr>
<td>100</td>
<td>8.9875518e+18</td>
</tr>
</tbody>
</table>

Moment of inertia in a rectangle:
\[ I_x = \frac{1}{12} \cdot b \cdot h^3 \]

<table>
<thead>
<tr>
<th>b</th>
<th>h</th>
<th>I_x</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>30</td>
<td>202500.0</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>351562.5</td>
</tr>
<tr>
<td>170</td>
<td>85</td>
<td>8700104.16667</td>
</tr>
<tr>
<td>150</td>
<td>10</td>
<td>12500.0</td>
</tr>
<tr>
<td>70</td>
<td>25</td>
<td>91145.8333333</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
<td>2560000.0</td>
</tr>
<tr>
<td>120</td>
<td>95</td>
<td>8573750.0</td>
</tr>
<tr>
<td>40</td>
<td>65</td>
<td>915416.66667</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>8437.5</td>
</tr>
<tr>
<td>30</td>
<td>45</td>
<td>37988.7</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>106666.66667</td>
</tr>
<tr>
<td>180</td>
<td>90</td>
<td>1093500.0</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>2858333.3333</td>
</tr>
<tr>
<td>160</td>
<td>55</td>
<td>2218333.3333</td>
</tr>
<tr>
<td>80</td>
<td>60</td>
<td>1440000.0</td>
</tr>
<tr>
<td>110</td>
<td>50</td>
<td>1145833.3333</td>
</tr>
<tr>
<td>190</td>
<td>35</td>
<td>678854.16667</td>
</tr>
<tr>
<td>140</td>
<td>20</td>
<td>93333.3333333</td>
</tr>
</tbody>
</table>
Data input

Data can be typed manually, or you can upload a local csv file. Optionally, the first line may contain the name of the variables. In example input data you can find some examples.

Manual input

<table>
<thead>
<tr>
<th>m</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4.4937759e+18</td>
</tr>
<tr>
<td>170</td>
<td>1.52783806e+19</td>
</tr>
<tr>
<td>70</td>
<td>6.29128626e+18</td>
</tr>
<tr>
<td>190</td>
<td>1.707634842e+19</td>
</tr>
<tr>
<td>10</td>
<td>8.9875518e+17</td>
</tr>
<tr>
<td>90</td>
<td>8.08879662e+18</td>
</tr>
<tr>
<td>80</td>
<td>7.19004144e+18</td>
</tr>
<tr>
<td>40</td>
<td>3.59502072e+18</td>
</tr>
<tr>
<td>120</td>
<td>1.078506216e+19</td>
</tr>
</tbody>
</table>

Upload local file

Choose File  No file chosen

Use first line as variable names?

Use typed values  Use local file  I've chosen an example

Success! Data loaded.

Figura 3: Main Interface
Figura 4: Main Interface
Check the behavior of the terms (or combinations) in relation to the target variable or in relation to the input variables.

T:
- $\cos(x_0)$
- $x_1^2$

X:
- $x_0$
- $x_1$

**Figura 5: Main Interface**
Check the behavior of the terms (or combinations) in relation to the target variable or in relation to the input variables.

T:
- $\cos(x_0)$
- $x_1^2$

X:
- $x_0$
- $x_1$

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 = \( \frac{1}{1 + MAE} \)
- Without any preprocessing (same as Lab Assistant)
- With an execution time budget of 3 minutes (more than Lab Assistant)
Figura 7: Score for the first 10 functions
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
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/