Symbolic Regression for model selection: the LH transition as a case study

E. Peluso\textsuperscript{a}, A. Murari\textsuperscript{b}, M. Gelfusa\textsuperscript{a}, M. Lungaroni\textsuperscript{a}, P. Gaudio\textsuperscript{a}

EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK

\textsuperscript{a} University of Rome “Tor Vergata”, Via del Politecnico 1, 00133 Rome, Italy
\textsuperscript{b} Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy.
Symbolic Regression (SR) via Genetic Programming (GP)

• Basically the Genetic Programming (GP) is a technique that can be used to find out the best “computer program” to achieve a specific goal;

• It takes inspiration by the theory of natural selection of Charles Darwin;

• Symbolic Regression (SR) uses GP to find out a pool of mathematical individuals (so “functional forms” made up by functions, constants and variables), not computer programs, for describing a specific problem. The actual application is a Genetic Algorithm (GA).

• In other words, SR produces models which answer the question:

  “what are the best models describing data without any “a priori” assumption?”
Logical placement of the technique

---

Mined Data

...Model formulation....

Model
Logical placement of the technique

1. Traditional approach
2. Symbolic Regression

...Model formulation....

Mined Data

Model
Focusing on the two different approaches: Traditional one

- Suppose we have a direct or not direct physical quantity $\tilde{y}$ related to a physical process.

- Our “friends” theoreticians can provide a theoretical model of this physical quantity as a function of other relevant physical quantities $\tilde{x}$ “weighted” using appropriate parameters $\tilde{\alpha}$. Then we have to obtain the best $\tilde{\alpha}$ fitting linearly or nonlinearly the function $f(\tilde{x}; \tilde{\alpha})$.

- Example:

\[
y_{\text{theoretical}} = \sin(\sqrt{x_1}) + \cos(x_1 \cdot x_2)
\]

\[
y_{\text{to be fitted}} = \alpha_1 \cdot \sin(\alpha_2 \cdot x_1^{\alpha_3}) + \alpha_4 \cos(\alpha_5 \cdot x_1^{\alpha_6} \cdot x_2^{\alpha_7})
\]
Focusing on the two different approaches: SR’s one

- Once we have selected the relevant physical quantities $\tilde{x}$, that we suppose to be relevant to depict the quantity $(\tilde{y}_{data})$, the best mathematical model, without any “a priori” assumption on its structure, can be provided by SR $(\tilde{y}_{mod \ by \ SR})$

$$y_{mod \ by \ SR} = \alpha_1 \cdot \sin(\alpha_2 \cdot x_1^{\alpha_3}) + \alpha_4 \cos(\alpha_5 \cdot x_1^{\alpha_6} \cdot x_2^{\alpha_7})$$
Focusing on the two different approaches: comparison with literature
Symbolic Regression (SR) via Genetic Programming (GP): Overlook of the methodology adopted

Run Phase:
Analysis via GA and sifting of a pool of models
Symbolic Regression (SR) via Genetic Programming (GP):
Overview of the methodology adopted

Run Phase:
Analysis via GA and sifting of a pool of models

Post-run Phase:
Selected models are non linearly fitted and qualified using a suitable set of statistical estimators
Run phase: main steps of SR via GP and similarities with biological concepts

1. An initial population of mathematical expressions is built, each described by one or more trees.

2. A quality criterion is applied to test the fitness of each individual respect to data, so respect to the “environment”, this qualifier is called Fitness Function:

The FF used for obtaining the results achieved is

\[ FF = 2k + n \cdot \ln(MSE) \]

Example of a mathematical expression

\[ y = \frac{\delta - a}{2} \cdot \frac{R}{\varepsilon} \]
Run phase: main steps of SR via GP and similarities with biological concepts

3. The population evolves mutating or crossovering or directly reproducing some of its individuals to a new population; best models reproduce more likely.

4. The new population is qualified according to the same fitness criterion used for the previous population;

5. The process goes on until a convergence criterion is satisfied

6. Individuals are studied using the Pareto Frontier: among them there is the final best model we are looking for.
Run phase: Pareto Frontier

- A pool of individuals, each representing a possible solution for the problem, is provided computing the Pareto Frontier (PF).

- The PF has to minimize two objectives: complexity (k standing for the number of nodes) and fitness to data (FF).

- Models selected are finally the best fitting for each class of complexity, so they are BUEMS (Best Unconstrained Empirical Models Structures)

An example of a pareto frontier (PF) in yellow respect to the whole pool of elements in blue. Individuals have been selected according to the two objective criteria.
Models are then non linearly fitted using an iterative algorithm \[1\] aiming at finding the best parameters, exponents and constants, represented by the vector \( \vec{c} \), minimizing the sum of squares of the residuals \( \vec{E} \):

\[
\min_{LB \leq \vec{c} \leq UB} \left[ \sum (\vec{E}(\vec{x}; \vec{c}, \vec{c}_0))^2 \right] = \min_{LB \leq \vec{c} \leq UB} \left[ \sum (\vec{y}_{\text{data}}(\vec{x}) - \vec{y}_{\text{mod}}(\vec{x}; \vec{c}, \vec{c}_0))^2 \right]
\]

\( \vec{c}_0 \) stands for the vector of initial conditions (by GA), the Upper Bounds (UB) and the Lower Bounds (LB) define the boundaries corresponding to the 99\% of confidence level of \( \vec{c}_0 \).

The recursive procedure requires the stability of constants and exponents to reach the convergence, so a tolerance of \( 10^{-6} \) has been set for the difference between the values assumed by the fitted parameters between two successive iterations.

Post run phase: Non linear fits and statistical estimators

- Once fitted, models’ quality is assessed using a pool of statistical estimators, like the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Kullback Leibler Divergence (KLD).


- The AIC and BIC parameterization used are:

\[
AIC = n \log(MSE) + 2k \\
BIC = n \log(\sigma^2) + k \log(n)
\]

\[k \equiv \text{number of parameters, used for penalizing the complexity of models}
\]

\[n \equiv \text{number of observations} \]

The Kullback-Leibler divergence (KLD) allows getting in touch with the quality of reproducing the distribution of data using the model.

For distributions $P$ and $Q$ of a continuous random variable, KL-divergence is defined to be the integral:

$$KLD = D_{KL}(P \parallel Q) = \int p(x) \ln(p(x)/q(x))$$

The Kullback-Leibler divergence is always non-negative and zero if and only if $p = q$

$p$ is interpreted as the reference pdf (data), $q$ the model.

The Kullback-Leibler divergence (KLD) requires the probability density functions (pdf), both for data and models, provided by the Kernel Density Estimation (KDE).

---

Post run phase: Non linear fits and statistical estimators

- KDE[5] is a non parametric technique used for estimating a not known a priori probability density function (pdf) \( f(x) \) for a random variable “x” (univariate in our case) using a kernel density estimator \( \hat{f}(x, \lambda) \), where \( \lambda \) is the bin width (i.e smoothing factor). The easier way to describe a pdf is via histogram so using:

\[
\hat{f}(x, \lambda) = \frac{1}{nh} \sum w \left( \frac{x-x_i}{h} \right) = \frac{1}{n} \sum \frac{1}{h} \begin{cases} 
1/2 & \text{if } |x-x_i| < h \\
0 & \text{otherwise}
\end{cases}
\]

- Better descriptions can be achieved using a continuous function (kernel), so \( \hat{f}(x, \lambda) \) takes the proper name of kernel density estimator:

\[
\hat{f}(x, \lambda) = \frac{1}{nh} \sum K \left( \frac{x-x_i}{h} \right) / \int K(u)du = 1
\]

Short introduction
Genetic Programming for Symbolic Regression
SR for Te @ LH transition using ITPA DB

Conclusions

Summarizing the methodology: related flow chart

General description
Run phase
Post run phase

RUN PHASE
- Generation of the population of models
- Computing the FF for each model

POST RUN PHASE
- Linear or Non linear Fits
- Selection of the best Performing model.
- Comparison with literature

Convergence criterion satisfied
- Pareto Frontier

Convergence criterion NOT satisfied
- Application of Genetic Operators:
  - Reproduction
  - Crossover
  - Mutation
- New population
Short introduction
Genetic Programming for Symbolic Regression
SR for $Te @ LH$ transition using ITPA DB
Conclusions

SR for MCNF

SR for Nuclear Fusion

$Te @ core$ for LH transition using the ITPA DB
IGDBTH4V6B [6]


Description of mined variables
Main work
Extrapolation @ ITER
Possible development
### Description of mined variables

<table>
<thead>
<tr>
<th>Quantity</th>
<th>$[\text{Min, Max}]$</th>
<th>$[\mu, \sigma]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ne</td>
<td>$[0.84, 49.18] \times 10^{19} \text{m}^{-3}$</td>
<td>$[12.74, 9.58] \times 10^{19} \text{m}^{-3}$</td>
</tr>
<tr>
<td>Rgeo</td>
<td>$[0.66, 2.95] \text{m}$</td>
<td>$[1.27, 0.74] \text{m}$</td>
</tr>
<tr>
<td>q95</td>
<td>$[2.11, 10.50]$</td>
<td>$[4.15, 1.15]$</td>
</tr>
<tr>
<td>B</td>
<td>$[0.39, 8.01] \text{T}$</td>
<td>$[3.92, 1.88] \text{T}$</td>
</tr>
<tr>
<td>k</td>
<td>$[1.41, 2.02]$</td>
<td>$[1.68, 0.08]$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$[0.07, 0.78]$</td>
<td>$[0.38, 0.20]$</td>
</tr>
<tr>
<td>a</td>
<td>$[0.21, 1.12] \text{m}$</td>
<td>$[0.44, 0.27] \text{m}$</td>
</tr>
<tr>
<td>V</td>
<td>$[0.80, 114.90] \text{m}^3$</td>
<td>$[18.12, 32.81] \text{m}^3$</td>
</tr>
<tr>
<td>Te</td>
<td>$[0.36, 11.0] \text{keV}$</td>
<td>$[2.33, 1.26] \text{keV}$</td>
</tr>
</tbody>
</table>

- 316 shots [AUG (107), CMOD (158), JET (39), NSTX (11), TCV (1)] having a temporal window of less than 50 ms from the transition.
Leading term obtained

• The analysis performed has stressed the primary importance of the term:

\[ T_e^c = 1.59^{1.65}a^{0.94}B^{0.99} \]

• While the best performing model on data is:

\[ T_e^{GIL43} = 7.05^{7.41}a^{1.35}B^{0.58}V^{0.76} \]

<table>
<thead>
<tr>
<th></th>
<th>MSE [keV^2]</th>
<th>AIC</th>
<th>BIC</th>
<th>KLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIL43</td>
<td>0.52</td>
<td>-187.60</td>
<td>-160.28</td>
<td>0.20</td>
</tr>
<tr>
<td>( T_e^c )</td>
<td>0.62</td>
<td>-141.34</td>
<td>-125.62</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Comparison with literature

• The coupled term in the models reported is actually the same coupling term that the theoretical model predicted by Guzdar and Hassam[7] shows.

• This result is an independent proof for the theory behind the model of GH [7].

\[ T_e^{GH} \propto \frac{B^{6/7} a^{8/7}}{(Rq)^{4/7}} \text{ tuned on data} \rightarrow T_e^{GH} = 5.71^{5.92}_{5.50} \frac{B^{6/7} a^{8/7}}{(Rq)^{4/7}} \]

<table>
<thead>
<tr>
<th></th>
<th>MSE [keV²]</th>
<th>AIC</th>
<th>BIC</th>
<th>KLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIL43</td>
<td>0.52</td>
<td>-187.60</td>
<td>-160.28</td>
<td>0.20</td>
</tr>
<tr>
<td>T_e^c</td>
<td>0.62</td>
<td>-141.34</td>
<td>-125.62</td>
<td>0.34</td>
</tr>
<tr>
<td>T_e^{GH}</td>
<td>0.70</td>
<td>-99.32</td>
<td>-76.25</td>
<td>0.31</td>
</tr>
</tbody>
</table>

• The additional term depending on the volume and on the triangularity is not a spurious one. In fact all tests performed excluding V and δ produced models fitting worse on data.

SR for MCNF: validation of theoretical models

Short introduction
Genetic Programming for Symbolic Regression
SR for Te @ LH transition using ITPA DB
Conclusions

Description of mined variables
Main work
Extrapolation @ ITER
Possible development

SR for Te @ LH transition

1st TM IAEA/2015
SR for MCNF: validation of theoretical models

Residuals are however an order of magnitude smaller than the best performing theoretical model.
SR for MCNF: validation of theoretical models

Model GIL43

- GIL43
- GIL43 ± 1σ
a = 2 m
B = 5.3 T
V = 830 m³

• GIL43 model in the range of validity provided for the plasma triangularity δ

• Caution should be applied for considering the extrapolation at ITER due to the number of entries in the mined DB, however a core temperature threshold around 5keV is predicted.

Short introduction
Genetic Programming for Symbolic Regression
SR for Te @ LH transition using ITPA DB
Conclusions

Description of mined variables
Main work
Extrapolation @ ITER
Possible development
Euclidean distance

The value of the data points is considered as “infinitely” precise.

How to handle error bars?
How to handle outliers?

SR and different “metrics”: can geodesic distance [8] based on gaussians help us?

SR and different “metrics”: can geodesic distance [8] based on gaussians help us?

The Geodesic distance (GD).

Data points are considered as gaussians.
So they are characterized by an expected value (\(\mu_i \sim y_i\)) and a standard deviation (e.g. \(\sigma_i \sim \%y_i\) both for data and models).

Tests performed have shown a better capability of GD respect to MSE to find out outliers in data, confirming what stated in literature [9], but more work is needed.

The distance between two Gaussians whose means are $\mu_1$ and $\mu_2$ and standard deviations $\sigma_1$ and $\sigma_2$ can be computed according to Rao [8] as:

$$GD(g_1 \parallel g_2)_i = \sqrt{2} \ln \left( \frac{1 + \delta_i}{1 + \delta_i} \right) = 2\sqrt{2} \tanh^{-1}(\delta_i) \quad \text{where} \quad \delta_i = \sqrt{\left(\frac{\mu_1^i - \mu_2^i}{\mu_1^i - \mu_2^i} + 2\left(\frac{\sigma_1^i - \sigma_2^i}{\sigma_1^i + \sigma_2^i}\right)^2\right)}$$

SR and different “metrics”: implementation adopted for the geodesic distance

• Using the GD, we have modified the SR’ FF in the following way:

\[
FF = 2k + n \cdot \ln(MSE) + \sum_{i} GD_i
\]

Visualization of the distributions \((p_1, p_2)\) along the geodesic parameterized by \(t\) [10].

Conclusions

• The method allows finding mathematical models.

• It has furthermore demonstrated a great flexibility in assessing the quality of theoretical models, like in the case shown in this presentation for the electron temperature scaling for the LH transition [11], where the leading term is a clear evidence supporting the Guzdar and Hassam’s theoretical interpretation.

• Many models have however been tested (e.g. Chankin, Kernel’s collisional and collisionalless, Scott, Rogister and Shaing&Crume), and the variables included in those models used for the analysis. No data driven models performed better than the GIL43 one, stressing the importance of triangularity and plasma volume.

• We are managing to refine our “preliminary” tests on JET’s data with ILW.

• Other implementations and applications have been considered and are actually in progress, like the use of different non euclidean metrics (e.g the GD).

Short introduction
Genetic Programming for Symbolic Regression
SR for Te @ LH transition using ITPA DB
Conclusions

Thank you for your attention