Symbolic Regression via Genetic Programming to derive Empirical Models and Scaling Laws as Monomial or Polynomial Expansions.

E. Peluso[1], A. Murari[2], I. Lupelli[1,3], M. Gelfusa[1] and P. Gaudio[1]

[1]*Associazione EURATOM-ENEA - University of Rome "Tor Vergata", Roma, Italy*

[2]*Consorzio RFX-Associazione EURATOM ENEA per la Fusione, I-35127 Padova, Italy*

[3]*EURATOM/CCFE Fusion Association, Culham Science Centre, Abingdon, Oxfordshire, OX14 3DB, UK*

Abstract

Many processes in plasma physics are inherently complex and highly nonlinear. Typically their behaviour is difficult to interpret with theoretical models based on first principles. To perform high-quality inferences, these processes have to be modelled starting directly from the experimental data. In this contribution we study and analyse the capabilities of Symbolic Regression via Genetic Programming as a tool for advanced data mining in Nuclear Fusion to derive Empirical Models. Whereas traditional linear and non-linear regression techniques simply try to find the best parameters of predefined model by fitting the available data, Symbolic Regression via Genetic Programming searches for the Best Unconstrained Empirical Model Structure. This implies deriving the significant variables, the functional form of the model and its parameters. A set of synthetic problems are used to assess some important capabilities of SR tools: over-fitting avoidance, extrapolation properties, identification of model constants, scalability to higher-dimensional problems and capacity to handle noisy data. As an example of application to Nuclear Fusion research, the method has been applied to the ITPA database of the energy confinement time of Tokamak plasmas in H mode.

1. Algorithm for Symbolic Regression

Symbolic Regression (SR) via genetic programming (GP) takes inspiration from the biological criteria of "natural selection" and "evolution", since the aim of an algorithm implemented for SR is to provide the best "individual" among many for solving a specific problem. For applications in physics, this individual is typically the function best fitting the data. Whereas traditional linear and non-linear regression techniques simply try to find the best parameters of predefined model by fitting the available data, Symbolic Regression searches for the Best Unconstrained Empirical Model Structure (BUEMS)[1,2,3]. The algorithm implemented includes modified parts of a free open source program called GPTIPS[4].

Starting from an initial collection of models (population), each consisting of numbers or variables or function (nodes) and represented by a linear combination of tree structures well defined in the graph theory ($y_{\mathrm{model}} = \sum c_i\, t_i$, where c_i are the constants of each tree t_i), their fitness to the data is computed using a criterion. We chose to use the following form of the AIC estimator, $AIC = 2 \cdot k + n \cdot \ln(RMSE/n)$; where $RMSE$ is the Root Mean Square Error, k is the total number of nodes of the model and n the number entries provided. Once ranked, the population is used to build a new one and the process continues until a convergence criterion is satisfied [1]. One criterion among those implemented requires the best individual not to change before reaching a fixed percentage of a controlled parameter called "maturity". If the fitness of the best model does not vary in double precision, the "maturity" increases of a fixed value (FPVI). In a post run phase, the best models, the second and third elements of each iteration are used to provide a possible solution. using the Pareto Frontier (PF) where the number of nodes and the fitness of each model are considered for their ranking. The PF allows identifying, for each subgroup of models with the same number of nodes, the individual best fitting to data. Moreover, models on the PF are also classified using a Bayesian criterion (BIC) parameterized as $BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n)$, where ϵ are the residuals and the other symbols defines the same quantities of the AIC. Two methodologies of analysis have been developed. The first one, that has been called "population convergence", consists of launching more runs, fixing for each the same maximum number of trees and the FPVI value, while the population parameter differs..On the other hand, the methodology consisting of launching more runs, each with a different maximum number of trees that can be used, but all with the same population and FPVI value, has been called "multitree convergence". In both cases the results of different runs are compared and the BUEMS belonging to the saturation part of the PF and more frequently found is selected.

Dealing with the paper, in Section 2 the classical Koza and Nguyen-6 examples and the three Maclaurin expansions have been solved using the "population convergence" methodology; while in Section 3 an application to the scaling laws for the scaling of the confinement time in Tokamak plasmas is provided using the "multitree convergence". Finally in Section 4 conclusions are drawn.

2. Benchmark applications

The Koza-1[5] function $\left(y = \sum_{i=1}^{4} x^i\right)$, as been chosen as a test for the polynomial class of models and the Nguyens-6 [5] one $\left(y = \sin(x) + \sin(x + x^2)\right)$,for the dependence of the

trigonometric function on polynomials arguments. The functions allowed [5] are F:{+,-,*,/,sin, cos, ln, exp}, no constants have been used and the variable range is $.U[-1,1;20]$ where U stands for uniform random sample between $[-1,1]$ having twenty points.The stopping RMSE value has been fixed to 10^{-5}. Results are reported in Table I.

Another example of application of our algorithm consists of finding a Maclaurin expansion of three functions (sin(x), cos(x), exp(x)). In this case the algorithm found the expansions up to the order which leads to an RMSE value under the tolerance criterion of 10^{-2}. For completeness, the same conditions of the previous tests have been chosen, except for the fact that the variable runs over $U[-\pi,\pi;20]$ and only the foundamental mathematical operators have been used. The excellent behaviour of the expansion obtained have been graphically plotted in Figure1

Table II. Population convergence.For the Koza-1 function (K) and the Nguyens-6 (N) one.

Pop	RMSE	Time	Func
250	10^{-15}	16min, 42s	K
500	10^{-15}	7min,21s	K
750	10^{-14}	21s	K
1000	10^{-10}	1min, 49s	K
250	10^{-15}	22s	N
500	10^{-15}	41s	N
750	10^{-13}	1min,21s	N
1000	10^{-12}	51s	N

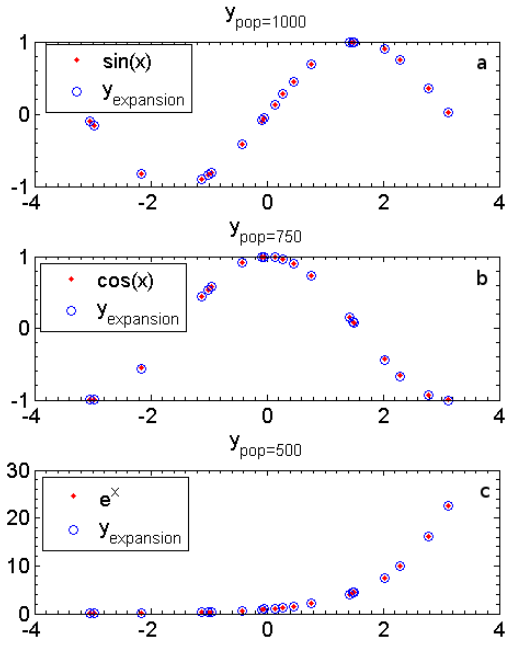

Figure 1.Analytic data and expansions.

3. Scaling of the confinement time in non power law form

The extrapolation of the energy confinement time to the next generation of devices has been investigated both theoretically and experimentally for several years. Dimensional or dimensionless scaling laws have been proposed, but the most widely accepted in the community are in power law (PWL) form. PWL can be unsatisfactory for several reasons, such as. no saturation effects (even when variables grow to infinity or go to zero), or monotonic behaviour, or overestimation of the relevance of the variables with the longest tails. To investigate this assumptions, the SR approach presented in the first section has been applied to the ITPA database DB3v13. In line with the previous literature[6], the same independent quantities have been considered to be good candidate regressors in the present work. One of the best functional forms for the confinement time given by SR, in terms of dimensional quantities, is reported in eq(1). Also the power laws typically used as reference [6] by the community are reported: IPB98(y,2) (PL1) in eq.(2) and EIV (PL2) in eq.(2).

$$NPL = 0.070^{0.071}_{0.069} \cdot I^{1.071^{1.079}_{1.062}} R^{1.706^{1.727}_{1.685}} \cdot \kappa_a^{1.250^{1.290}_{1.211}} P^{-0.715^{-0.707}_{-0.723}} \frac{n^{0.100^{0.109}_{0.091}}}{1 + e^{-0.408^{-0.390}_{-0.426} \cdot n^{1.036^{1.108}_{0.996}}}} \quad (1)$$

$$PL1 = 5.62 \cdot 10^{-2} I^{0.93} B^{0.15} n^{0.41} M^{0.19} R^{1.97} \epsilon^{0.58} \kappa_a^{0.78} P^{-069} \quad (2)$$

$$PL2 = 5.55 \cdot 10^{-2} I^{0.75} B^{0.32} n^{0.35} M^{0.06} R^{2.0}\ \epsilon^{0.76} \kappa_a^{1.14} P^{-062} \quad (3)$$

Table II. Comparison of eq.(1,2,3). After the non linear fit, AIC has been recomputed using the RSS instead of the RMSE; the complexity (k) of the model has been considered as the number of parameters (p) plus one, so k=p+1 both for AIC and BIC. The Kullback Leibler divergenge (KLD) have been computed in a $\pm 6\sigma$ range from the mean value of the distribution of the original data (τ).

	NPL	PL1	PL2
k	9	10	10
AIC	-19610.81	-19416.86	-19084.36
BIC	-19556.55	-19362.86	-19203.68
MSE [s^2]	$1.753 \cdot 10^3$	$1.866 \cdot 10^3$	$2.077 \cdot 10^3$
KLD	0.0255	0.0337	0.0802

The most important aspect of the BUEMS in eq.(4) in the non power law (NPL) functional form, is the presence of a squashing term in the density. The physical interpretation involves the analysis of the behaviour of the smaller devices for which the squashing term introduces more flexibility in fitting the region of the smaller densities, at the same time allowing the use of less favourable exponents for the power law part of the scaling. The superior properties of the NPL are reported in Table II. Eq.(4) predicts a confinement time at ITER of $2.83^{3.31}_{2.42}\,s$, while other models of similar behavior predict even a more pessimistic extrapolation.

4. Conclusions

The results obtained using the genetic algorithm implemented, show how it can be used to find hidden functions in the form of McLaurin expansions as well as BUEMS problems. A major role is played by the combined use of the information criterion AIC and of the statistical estimator BIC which allow finding a good compromise between limited complexity and good data fitting. Finally the application of the technique to the ITPA database of the energy confinement time shows the relevance of relaxing the assumption of power law scalings.

References

[1] A.Murari et al, 2012, Nuclear Fusion, Vol 52, 063016-18.

[2] A. Murari et al 2013 Nucl. Fusion 53 043001

[3] I. Lupelli et al 2013, Fusion Engineering and Design, Volume 88, Issue 6, Pages 738-741

[4] Searson D.P., Leahy D.E. & Willis M.J., 2010, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2010 (IMECS 2010), Hong Kong, 17-19 March.

[5] McDermott J. et al, 2012, Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference (GECCO '12), pp 791-798 .

[6] McDonald D.C et al, 2007, Nuclear Fusion,Vol 47, pp 147-174