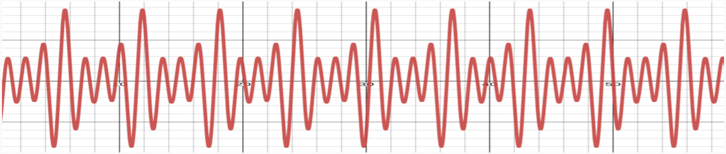
It’s time to get cereal – a whole grain of truth about GP for SR



Prof. Fabrício Olivetti de França

Federal University of ABC

11 July, 2024

Genetic Programming for Symbolic Regression

A small introduction

> Symbolic regression (SR) is an approach to machine learning (ML) in which both the parameters and structure of an analytical model are optimized.

– *Contemporary Symbolic Regression Methods and their Relative Performance, William La Cava et al.*

> Symbolic regression (SR) is a type of regression analysis that searches the space of mathematical expressions to find the model that best fits a given dataset, both in terms of accuracy and simplicity.

– *Wikipedia*

A small introduction

It is an important tool for regression analysis as it can deliver a high accuracy while keeping the possibility of statistical inference.

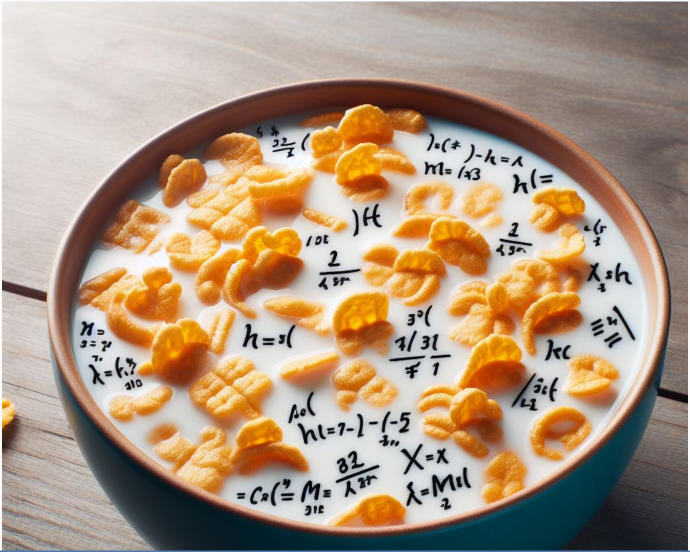
We can observe a growing interest in SR from other fields, but we are still far from the widespread adoption that we see with different regression models.

A small introduction

- How did we get where we are right now?
- How can we show our value to other communities?
- How can we prevent a bad first impression?
- How can we join forces to increase our popularity?

A small introduction

It’s time to get cereal!



Benchmarking

In the last episodes of GP

In 2012, the paper Genetic Programming Needs Better Benchmarks[1] started a discussion about a lack of standards in GP benchmarks:

> We argue that such benchmarks do little to move the field forward, and in fact may hold it back as they reward techniques that are effective at rapidly solving trivial problems rather than performing as well as possible on hard ones

[1] McDermott, James, et al. "Genetic programming needs better benchmarks." Proceedings of the 14th annual conference on Genetic and evolutionary computation. 2012.

In the last episodes of GP

They aimed at the following criteria:

- Tunably difficult
- Varied
- Relevant
- Fast
- Acommodating to Implementors
- Easy to Interpret and Compare
- Representation Independent
- Current

In the last episodes of GP

10 years later, in[2] the authors analysed what has changed so far and what still required some work. Among other things, the authors expressed some concerns about:

- Goodhart's law
- Difficulty in comparing reformulated problems (e.g., Copilot x GP)
- Mass benchmarking does not reflect real-world scenario

[2] McDermott, James, et al. "Genetic programming benchmarks: looking back and looking forward." ACM SIGEVOlution 15.3 (2022): 1-19.

Synthetic problems as unit tests

They argue that synthetic problems can be used as unit tests to verify correctness of new approaches:

- Verify whether it can solve easy problems
- Ensure that the solution is contained in the search space
- Test the influence of noise or number of samples or dimensionality in the capabilities of reaching a good solution

Having a controlled environment is helpful when studying all the aspects of the algorithm, but real-world datasets are still required to measure the expected performance compared to other algorithms.

Real-world problems

Real-world problems should be selected avoiding the endpoints of being too easy or too hard where all methods equally succeed or fail. Finding the sweetspot is not easy… These endpoints could be used as a baseline or to measure progress / milestones.

Another issue is whether to pre-process the data or leave that to GP? Both are valid benchmarks, as argued by the authors.

SRBench

SRBench[3][4] is a milestone in GP benchmarking for SR as it established a standard for comparison of old and new algorithms. Apart from the large selection of datasets, it brings:

- A standard API in Python based on scikit-learn, but the participants are free to use any language as the backend
- A standard benchmark procedure with k-folds using a fixed set of seeds, ensure reproducibility, fairness in comparison, and can be updated incrementally (except for runtime comparison)
- In a short time it is already well adopted in many comparisons

[3] La Cava, William, et al. "Contemporary Symbolic Regression Methods and their Relative Performance." Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2021.

[4] https://cavalab.org/srbench/

This benchmark can be split in two parts[5], almost half of it is composed of variations of Friedman datasets. A few instances of this class of problem shows some difference between algorithms.

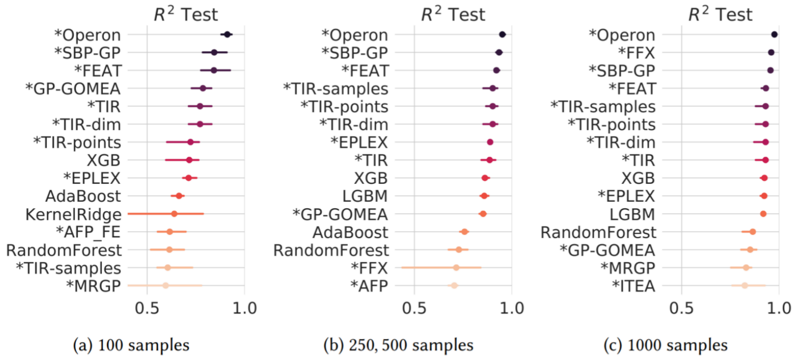
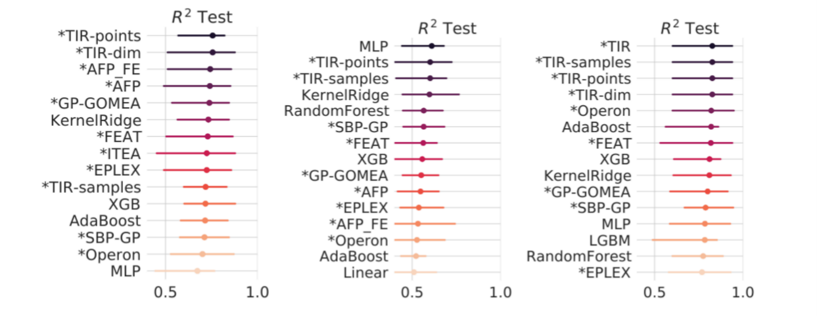


Fig. 9. Errorbar plots of the Friedman benchmarks grouped by number of samples.

[5] De França, Fabrício Olivetti. “Transformation-interaction-rational representation for

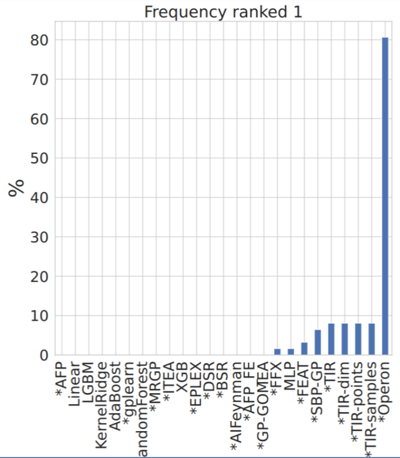
In "non-Friedman" datasets there's barely any difference. So the final ranks are determined by a subset of the Friedman sets.

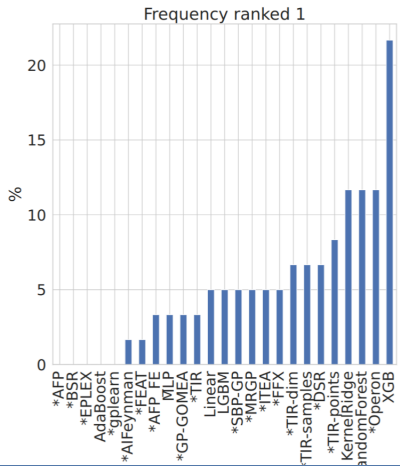


(a) [0, 5] (b) [6, 10] (c) [11, 20]

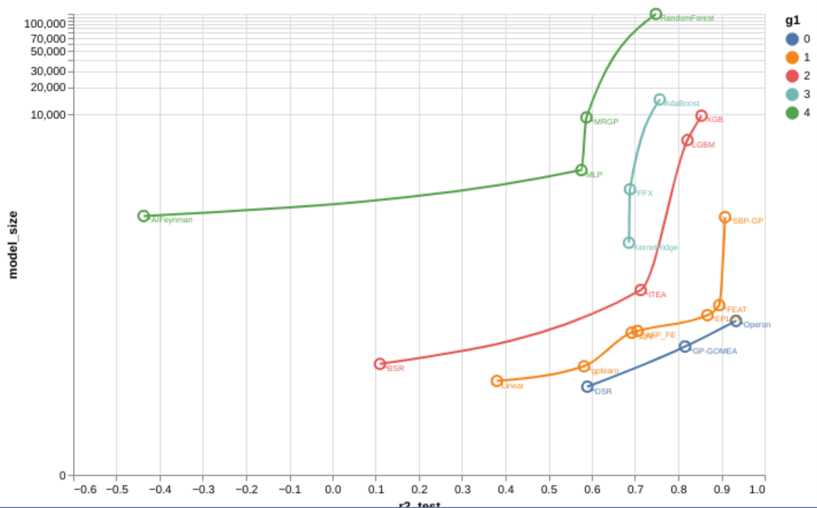
SRBench

How we present results is also important, possibly there's no **best** visualization. Maybe instead of a grand champion, we can look at the minimal set that gets the best results.





Or the set of best compromises between accuracy and model size, or accuracy and runtime.



Competitions

Following SRBench, we hosted two competitions with a smaller set of synthetic benchmarks and real-world datasets. The main goal was to investigate the capabilities of SR algorithms to:

- Retrieve the exact generating function
- Select relevant features
- Avoid deceptive (noisy) shortcuts
- Behave correctly when extrapolating
- Be robust against noise

Lessons to be learned

Having analysed just a small set of algorithms on a small number of datasets:

- There are many fine details that becomes obscure during analysis chiefly since we *demand* a single solution from the algorithms, how to choose a model can be very important
- We cannot agree on an interpretability measure. Interpretability depends on context and it also depends on supporting tools. For example, a large model may be more interpretable if there are less adjustable parameters.
- Artificially increasing difficulty is not that simple

Lessons to be learned

- Checking whether we reach the exact solution is an unsolved problem, also, should we really care for that?
- Extrapolating without additional information is shot in the dark; from many possible models that fit the data, how should we know what lies beyond?
- Noise and feature selection are still a challenge for SR.

Different and refreshing ideas

A byproduct of SRBench and corresponding competitions were the many different ideas proposed by the participants:

- Different representations inspired by nonlinear regression literature (interactions, rational of polynomials, continued fraction, piecewise regression).
- Use of nonlinear optimization to fit the parameters.
- Island model to promote diversity
- Popularization of multi-objective GP
- Bayesian model selection
- Use of local search for the combinatorial part of the problem
- Ensembles
- Deep Learning / Transformers

SRBench 3.0

We are currently organizing a new edition of SRBench with a better selection of datasets, different tracks, and better analysis of the results.

We need your help! If you want to get involved or want to share some thoughts:

folivetti@ufabc.edu.br

Open an issue / discussion at https://github.com/cavalab/srbench/

Accuracy is not all you need

I beat you by 0.001!

As we seen in the previous slides, many current SR algorithms show a comparable performance.

We can safely say that, for tabular data, we are competitive with opaque models.

If everything is the same, what reasons would we have to choose SR instead of XGBoost, MLP, or anything else?

Beyond accuracy

SR offers a lot more than a simple predictive model, we should start exploring and *packaging* additional features with our main algorithm:

- Interpretability
- Confidence and prediction intervals
- Statistical tools in general
- Use of prior knowledge
- Model simplification

Interpretability

The elephant in the room! There is no consensus, there will never be!

Interpretability and explainability is context dependent. We won't have a step-by-step procedure that will work to explain every SR model[6].

Instead, we should work together with colleagues from other fields to help them[7] generate symbolic models that are accurate, useful, and interpretable…for them!

This requires a long and manual work.

---

[6] Nadizar, Giorgia, et al. "An analysis of the ingredients for learning interpretable symbolic regression models with human-in-the-loop and genetic programming." ACM Transactions on Evolutionary Learning 4.1 (2024): 1-30.

[7] Russeil, Etienne, et al. "Multi-View Symbolic Regression." arXiv preprint arXiv:2402.04298 (2024).

Confidence intervals and statistical tools

As a regression model, SR is compatible with the many stats tool available for nonlinear regression. One example is the calculation of confidence intervals and profile likelihood[8].

In particular, profile likelihood can reveal not only the associated uncertainties of the parameters and predictions, but it can help to verify whether the parameters are identifiable[9] (i.e, does the available data and the choice of model can uniquely identify the parameter?). This is related to interpretability!

[8] de Franca, Fabricio Olivetti, and Gabriel Kronberger. "Prediction Intervals and Confidence Regions for Symbolic Regression Models based on Likelihood Profiles." arXiv preprint arXiv:2209.06454 (2022).

[9] Raue, Andreas, et al. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood." Bioinformatics 25.15 (2009): 1923-1929.

Prior knowledge

Unlike most regression models that departs from a fixed structure and adjusts only the parameters, SR has the additional degree-of-freedom of choosing the best function that describes the data.

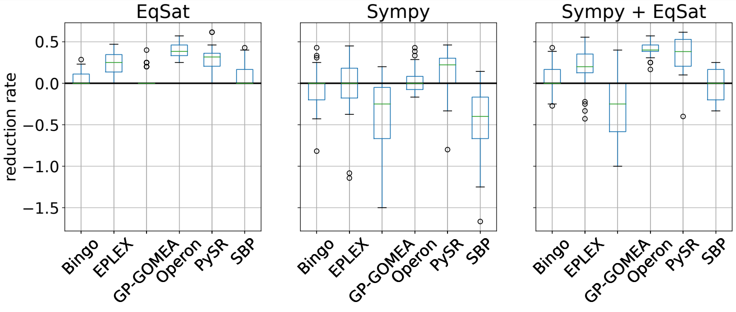
As such, we can easily integrate prior knowledge[10][11] into the search process by limiting the search space or penalizing non-conformant solutions. We can request for models that:

- Have a bounded co-domain (for a bounded domain)
- Is monotonic
- Is Symmetric
- many more…

[10] Kronberger, Gabriel, et al. "Shape-constrained symbolic regression—improving extrapolation with prior knowledge." Evolutionary computation 30.1 (2022): 75-98.

[11] Kubalík, Jiří, Erik Derner, and Robert Babuška. "Symbolic regression driven by training data and prior knowledge." Proceedings of the 2020 Genetic and Evolutionary Computation Conference. 2020.

Some SR algorithms have a bias towards overparameterized models with redundant parameters that can lead to unidentifiability and convergence issues in numerical optimization. Simplification can help to alleviate this problem and reduce model complexity.



Model simplification

An interesting approach is Equality Saturation[12][13], it allows to rewrite the expression into a more convenient form as the context ask for.

We can also infer general and local equivalence rules from our data [14].

[12] Willsey, Max, et al. "Egg: Fast and extensible equality saturation." Proceedings of the ACM on Programming Languages 5.POPL (2021): 1-29.

[13] de Franca, Fabricio Olivetti, and Gabriel Kronberger. "Reducing Overparameterization of Symbolic Regression Models with Equality Saturation." Proceedings of the Genetic and Evolutionary Computation Conference. 2023.

[14] Aldeia, Guilherme Seidyo Imai, Fabricio Olivetti de Franca, and William G. La Cava. "Inexact Simplification of Symbolic Regression Expressions with Locality-sensitive Hashing." arXiv preprint arXiv:2404.05898 (2024).

Closing Remarks! Let us preach!

We need to grow our community

It's time to make people aware of SR:

- Collaborate in real-world application
- Grow our ecosystem of supporting libraries
- Offer good end user experience
- Incorporate tools beyond prediction
- Write more books
- Teach and advertise
- Stimulate your students to create awesome repositories and tutorials

Supporting tools - SRTree-Opt

I'm currently developing a support tool for post-processing SR expressions:

- Parse different formats
- Refit parameters with support to different distributions
- Simplify the expressions with equality saturation
- Calculate the confidence interval of parameters and predictions
- Display stats about the model and different performance measures

https://github.com/folivetti/srtree-opt

Grad course experience

From February till May, I teached a 12 weeks graduate course on Symbolic Regression. It covered the essential of nonlinear regression analysis, statistics, symbolic regression and genetic programming, and related topics.

The audience was mixed: from CS students to humanities, so a good support of easy-to-use tools was important.

Students impression

In summary, the students were very interested about the possibilities of creating a nonlinear regression model. By the end of the course they got a draft paper for interesting applications:

- Non-intrusive blood pressure measurement
- Understanding what influences students grade on an online course
- Symbolic regression with human-in-the-loop for data science
- Symbolic regression with federated learning
- Understanding what attracts tourists

Students impression

They did have some difficulties, tho:

- No MS-Windows support!! (MS-Windows users installed HeuristicLabs)
- Not always easy to install even on Linux
- Too many hyper-parameters to finetune, none of them worked well with defaults
- Lack of documentation
- Lack of customization

Slides

The slides are freely available and every collaboration is welcome!

The SR algorithms part is lacking! Write some lecture notes about your own algorithm!

https://github.com/folivetti/symreg-course

COIL

Universidade Federal do ABC participates of the Collaborative Online International Learning (COIL). We can apply to create an online course on SR!

https://collab-edu.com/hub/coil

Book

There's a new book on Symbolic Regression hitting the shelves soon. This book offers a more practical approach aimed at data scientists.

Link to book

Conclusion

- SR field has made major advances in the last years.
- Accuracy is on par with other regression methods.
- We need to invest on:
  - a better enduser experience
  - a complete ecosystem
  - advertising through tutorials, repositories, courses
  - include post-processing reports and customizations in our own tools

Thank you!

You can download these slides in PDF format from:

https://folivetti.github.io/