SYMBOLIC REGRESSIONS IN NON-PHYSICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We apply symbolic regressions, the Machine learning approach of recovering models from data, in non-physical systems. Specifically, we present a data set consisting of equations that cover a broad range of topics in financial economics. These equations are built off a common set of mathematical symbols but importantly have new variations in functional forms. We test the joint performance of deep learning and genetic programming symbolic regression systems in recovering these non-physical equations.

1 INTRODUCTION

Symbolic regression is a branch of machine learning that aims to recover the mathematical expression underlies a given data set. While this is a meaningful and challenging problem that is essential in building explainable AI, this is also at the heart of scientific communities more generally: Scientists aim to find accurate yet relatively simple equations that describe the system of interest (Anderson, 1972).

Recent progress in symbolic regression has been promising since Schmidt & Lipson (2009). Although SR has been theoretically proven to be an NP-hard problem in general (Virgolin & Pissis, 2022), literature has proposed several viable approaches, including Genetic programming and Deep Learning. Current SofA methods achieve high recovery rates for physical equations, for example, AI Feynman (Udrescu & Tegmark, 2019) recovered all 100 equations from Feynman lectures in physics. GSR (Tohme et al., 2022) further achieved perfect recovery rates in a range of datasets. On the other hand, the performance of symbolic regression methods on non-physical equations is still largely unknown. This is at least partly due to the lack of a benchmark dataset for non-physical systems. In this paper, we present a ground-truth database that complements existing benchmarks.

Our main contribution is extending the scope of SR from physical equations into other domains, in particular, human-made systems including financial markets. A closely related contribution is to provide a novel out-of-sample test for existing methods. As seen in Table 1, the exact functional forms of our new equations are not particularly similar to physical equations. Therefore our new dataset includes some unique variations that algorithms have not seen so far.

2 RELATED WORK

Our paper is closest to two strands of literature. First, our paper is related to the papers providing SR benchmarks (Udrescu & Tegmark (2019), Cava et al. (2021), Romano et al. (2020)). Our paper differs and complements this line of prior work by focusing on the set of equations that comes from financial systems and not the physical world.

Second, we connect with the literature of applications of SR in social sciences. Duffy & Engle-Warnick (2002) applied symbolic regression to the data from the repeated ultimatum game in an effort to unveil the players' strategies based on past actions. Their algorithm is able to uncover certain simple strategies while encountering difficulties with recovering more random strategies. Yang et al. (2015) model oil productions and Pan et al. (2019) adopted this method in discovering the most critical factors that influence the carbon emissions in OECD countries, which yields country-dependent results that can aid policy-making towards lower carbon emissions. Truscott & Korns

(2014) use symbolic regressions in predicting unemployment rates. (Balla et al., 2022) recovered classical economic models.

3 METHODS

Inspired by AI feynman (Udrescu & Tegmark, 2019), we select 14 equations that are suitable for SR from the appendix of the seminar textbook by Ross, Westerfield, and Jordan. We select non-linear equations that are economically meaningful and foundational.

We randomly sample inputs from [0,1) interval and fit them into the mathematical expression to obtain the target. We generate 1000 samples for each equation listed in 1.[1]

4 EXPERIMENTS

We evaluate the performance of two representative SR methods. For genetic programming (GP) -based models, we select the commercial software Eureqa Schmidt & Lipson (2009) and allow up to 300 seconds for Eureqa to solve the equation. For deep learning (DL)-based models, we use the transformer-based end-to-end model (Kamienny et al., 2022). We tested the model with 3, 7, or 20 attention heads. If the exact or mathematical equivalent expressions were recovered, we consider an equation solved.

Table 1 reports the results from the experiments. We report that both DL and GP are capable of recovering a considerable set of nonlinear equations in our newly assembled dataset. Both methods combined solved 9 out of 14 equations. Compared to the close-to-perfect recovery rates for physical equations, the evidence suggests that there is potential room for improvement for SR methods in non-physical settings. We refrain from making conclusions about the performance of DL and GP methods but instead, consider a problem solved if either method solved the equation.

Table 1: Table of Equations

Eq. Index	Equation Expression	Solved by DL	Solved by GP
7.1	$P = C\frac{1-\frac{1}{(1+r)^t}}{r} + \frac{F}{(1+r)^t}$	No	No
8.1	$P_0 = \frac{D+P_1}{1+R}$	Yes	Yes
8.4	$P_t = \frac{D_{t+1}}{R-g}$	No	**Yes**
8.5	$P_0 = \frac{D_1}{R-g_1}[1-(\frac{1+g_1}{1+R})^t] + \frac{P_t}{(1+R)^t}$	No	No
14.6	$WACC = (\frac{E}{V})R_E + (\frac{D}{V})R_D(1-T_C)$	No	**Yes**
16.1	$R_E = R_A + (R_A - R_D)\frac{D}{E}$	Yes	Yes
19A.4	$C^* = \sqrt{\frac{2TF}{R}}$	No	**Yes**
19A.5	$C^* = L + (\frac{3}{4}F\frac{\sigma}{R})^{\frac{1}{3}}$	No	No
20.4	$PV = \frac{(P-v)(Q'-Q)}{R}$	Yes	Yes
20.14	$Q^* = \sqrt{\frac{2TF}{CC}}$	No	**Yes**
21.3	$E(S_t) = S_0[1+(h_{FC}-h_{US})]^t$	No	No
21.4	$F_1 = S_0\frac{1+R_{FC}}{1+R_{US}}$	**Yes**	No
21.6	$F_1 = S_0[1+(R_{FC}-R_{US}]$	Yes	Yes
25.8	$V = E \times e^{-Rt} - P$	No	No

5 CONCLUSION

We present a novel SR groud-truth dataset that incorporates equations from financial systems instead of traditionally focused physical systems. SR models can recover some of the equations but struggle in some of the tasks. Future generations of SR methods may aim to achieve stronger performance in both physical and non-physical systems to find the true unifying theory that bridges these fields.

[1]This dataset will be publicly available at XXX.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

P. W. Anderson. More Is Different. *Science*, 177(4047):393–396, 1972. ISSN 0036-8075. doi: 10.1126/science.177.4047.393.

Julia Balla, Sihao Huang, Owen Dugan, Rumen Dangovski, and Marin Soljacic. AI-Assisted Discovery of Quantitative and Formal Models in Social Science. *arXiv*, 2022. doi: 10.48550/arxiv.2210.00563.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H Moore. Contemporary Symbolic Regression Methods and their Relative Performance. *arXiv*, 2021. doi: 10.48550/arxiv.2107.14351.

John Duffy and Jim Engle-Warnick. Evolutionary Computation in Economics and Finance. *Studies in Fuzziness and Soft Computing*, pp. 61–82, 2002. ISSN 1434-9922. doi: 10.1007/978-3-7908-1784-3_4.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *arXiv*, 2022. doi: 10.48550/arxiv.2204.10532.

Xiongfeng Pan, Md Kamal Uddin, Bowei Ai, Xianyou Pan, and Umme Saima. Influential factors of carbon emissions intensity in oecd countries: evidence from symbolic regression. *Journal of Cleaner Production*, 220:1194–1201, 2019.

Joseph D Romano, Trang T Le, William La Cava, John T Gregg, Daniel J Goldberg, Natasha L Ray, Praneel Chakraborty, Daniel Himmelstein, Weixuan Fu, and Jason H Moore. PMLB v1.0: An open source dataset collection for benchmarking machine learning methods. *arXiv*, 2020. doi: 10.48550/arxiv.2012.00058.

Michael Schmidt and Hod Lipson. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923):81–85, 2009. ISSN 0036-8075. doi: 10.1126/science.1165893.

Tony Tohme, Dehong Liu, and Kamal Youcef-Toumi. GSR: A Generalized Symbolic Regression Approach. *arXiv*, 2022. doi: 10.48550/arxiv.2205.15569.

Philip Truscott and Michael F Korns. Explaining unemployment rates with symbolic regression. *Genetic Programming Theory and Practice XI*, pp. 119–135, 2014.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: a Physics-Inspired Method for Symbolic Regression. *arXiv*, 2019. doi: 10.48550/arxiv.1905.11481.

Marco Virgolin and Solon P Pissis. Symbolic Regression is NP-hard. *arXiv*, 2022. doi: 10.48550/arxiv.2207.01018.

Guangfei Yang, Xianneng Li, Jianliang Wang, Lian Lian, and Tieju Ma. Modeling oil production based on symbolic regression. *Energy Policy*, 82:48–61, 2015. ISSN 0301-4215. doi: 10.1016/j.enpol.2015.02.016.