SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Symbolic Regression Algorithms For Discovering Mathematical Laws In Noisy Experimental Data

Shanthi R[1*], Saraswati B[2], Makhkamova Husnida Ruziboevna[3], Dr.Shyam Maurya[4], Voruganti Naresh Kumar[5]

[1*]Assistant Professor & HOD, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: shanthir@maher.ac.in
[2]Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: saraswatib@maher.ac.i
[3]Turan International University, Namangan, Uzbekistan, E-mail: husnidamaxkamova61@gmail.com, https://orcid.org/0009-0007-0149-4806
[4]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.shyammaurya@kalingauniversity.ac.in, 0009-0006-3442-8621
[5]Associate Professor, Department of CSE, CMR Technical Campus, Hyderabad, Telangana, India.
E-mail: nareshkumar99890@gmail.com
*Corresponding author: Email: shanthir@maher.ac.in

Abstract

Symbolic regression is a highly effective paradigm for discovering interpretable mathematical expressions from experimental observations without having any a priori knowledge about the physical laws behind the data. In this paper, a comprehensive paradigm for symbolic regression is developed combining the strength of genetic algorithms along with multi-objective optimization and robust evaluation metrics to discover compact mathematical expressions from noisy experimental data. The developed methodology utilizes adaptive operator selection, lexicographic fitness function, and constraint-based mathematical expression generation for searching the solution space of mathematical models. An innovative denoising pre-processing approach using the iterative application of median filtering and wavelet decomposition helps to increase the signal-to-noise ratio of experimental data, while retaining its essential properties. The robustness and effectiveness of the proposed methodology have been demonstrated using synthetic experimental data corrupted by different noise levels (5%-50% Gaussian noise) as well as actual experimental data for various mechanical systems. The results reveal that the proposed algorithm discovers the ground truth mathematical expressions with 96.3% accuracy from 25% noisy data. Moreover, it outperforms traditional symbolic regression techniques in terms of mean absolute percentage error by 34%.

Keywords: Symbolic regression, Genetic algorithms, Expression discovery, Noise-robust optimization, Physical law discovery, multi-objective optimization, Noisy experimental data.

1. Introduction

Discovery within scientific research is based on understanding the mathematical principles that govern the physical world around us. Physicists and engineers have traditionally derived their models using analytical thinking and intuition, then verifying the models based on experimental data [1][3]. The new data-driven methods try to discover the mathematical principles behind experimental observations. However, this poses a challenging task for any researcher when the experimental data involves measurement uncertainties and other factors [2].
Symbolic regression, which can also be called symbolic mathematical expression discovery or equation learning, refers to a collection of algorithms that seek out combinations from the immense universe of potential math equations for finding models that fit the data [4]. Symbolic regression not only optimizes parameters but also explores the model space along with its parameters [5]. This property makes symbolic regression extremely useful in situations when either no sufficient domain knowledge is available, exploratory analyses are required, or the theories on which prior assumptions are based are faulty [6]. Existing methods of symbolic regression based on conventional genetic programming developed by Koza have been extensively used in many later studies [7]. Conventional genetic programming algorithms suffer from poor performance while working with noisy datasets because of degraded fitness landscapes and impossibility to separate signal from noise during selection [8]. This study introduces a framework that addresses the above shortcomings using three breakthrough concepts:

- Adaptive genetic algorithm operators based on population diversity and convergence status,
- Robust fitness function evaluation using non-parametric statistical techniques and resampling, and

- Physical informed expression generation using dimensional analysis and constraints.

The rest of this paper is organized as follows. Section 2 gives a survey of related works on genetic programming, symbolic regression, dealing with noise in machine learning, and physics-based modeling. Section 3 defines the proposed symbolic regression framework by formulating the problem, outlining the noise preprocessing procedure, describing the structure of the genetic algorithm used, and providing details on noise-resilient fitness functions. Section 4 provides information about the experiments in terms of dataset selection, baselines, and evaluation metrics. Section 5 reports the empirical results by comparing the proposed method against several baseline approaches under different noise levels. Section 6 discusses the efficacy of various parts of the proposed framework, model interpretability, and generalization ability. Section 7 concludes with summary.

2. Related Work

Genetic Programming and Symbolic Regression

The genetic programming model represents an evolutionary computing method where potential solutions are encoded in the form of trees subjected to evolutionary computations based on crossover, mutation, and selection [8]. Koza presented the initial approach to genetic programming and demonstrated the ability of symbolic regression in the process to re-invent classical laws of physics as well as some engineering rules [9]. Various improvements were made later to the approach such as strong typing for domain knowledge application, grammar-based genetic programming for controlling expression creation and cartessian genetic programming with different representation forms [10]. The approach of Udrescu and Tegmark is known as AI Feynman where symbolic regression was coupled with neural network approximations. However, these techniques face difficulties with noisy input [11].

Noise Handling in Machine Learning

Noise in training data is a central issue in machine learning, approached in various ways, such as data preprocessing, robust loss function, assembling, and regularization [12]. Wavelet denoising stands out as an approach that preserves the structure of signals and reduces their noise level [13]. Median and bilateral filters, which retain edges, represent non-parametric options in dealing with noise [14]. With regard to the problem of noise tolerance in symbolic regression, there is relatively little literature. [15] studied the properties of fitness landscapes of genetic programming affected by noise, whereas Cummins' most recent work focused on robust regression applied to symbolic regression tasks.

Physics-Informed Machine Learning

Integration of physical information into machine learning algorithms has proven to be a promising line of research, represented, for example, by physics-informed neural networks (PINNs) and sparse identification of nonlinear dynamics (SINDy). SINDy learns governing differential equations from datasets by solving a sparse regression problem over a set of candidate basis functions [16]. This method leverages compressed sensing ideas to uncover concise representations of dynamics. Even though SINDy is quite efficient when dealing with systems of known structure, it relies on the user to carefully choose candidate terms. The current paper seeks to address this drawback using symbolic regression ideas.

3. Methodology

The problem of symbolic regression starts with some observed data containing n pairs of input-output values as follows: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. where each input value $x_i \in \mathbb{R}^d$ r is a d dimensional observation, while the respective output value corresponds to some quantity with additive Gaussian noise. Mathematically it can be expressed as follows:

$$y_i = f^*(x_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad (1)$$

where, in Equation (1), f^* denotes the unknown function while ϵi represents measurement errors. The aim of symbolic regression is thus to find out an interpretable mathematical function, $\hat{f}$, which reduces the difference as in equation (2):

$$\hat{f} = \arg \min_{\text{expressions}} \sum_{i=1}^{n} \| y_i - \hat{f}(x_i) \|^2 \qquad (2)$$

Noise pre-processing is performed prior to regression analysis, with a series of iterative steps to improve signal integrity. Impulsive noise removal through iterative median filtering (5 window size) is applied without edge blurring, then discrete wavelet decomposition through use of Daubechies wavelets (db4), with universal soft thresholding $\lambda = \sigma\sqrt{2\ln(n)}$, and finally local regression smoothing (LOESS) with adaptively selected bandwidth through cross-validation to improve estimation of signals and reduce effects of noise during model evolution.

Genetic Algorithm approach uses a population P of candidate mathematical expressions represented by abstract syntax trees. Evolution of expressions through repeated application of fitness evaluation, selection, crossover and mutation processes. Fitness of candidates is evaluated based on robust noise-resistant multi-objective criteria involving prediction, stability and simplicity:

$$\Phi_1(g) = \text{MAPE}(g), \Phi_2(g) = \text{bootstrap\_confidence}(g, 0.95), \Phi_3(g) = T(g) \quad (3)$$

where in equation (3), T(g)denotes tree size, MAPE is mean absolute percentage error, and bootstrap_confidence measures prediction stability. The lexicographic ordering prioritizes firstly accuracy, then uncertainty, and finally complexity. In order to guarantee physical meaning, mathematical constrains ensure that dimensions are consistent and feasible, such as:

$$\dim(x) = \dim(y) \text{andoperator domains valid: } x > 0 \text{ for } \log(x) \quad (4)$$

With such constrains in equation (4), the evolving equations will be physically meaningful rather than absurd or instable. Moreover, the adaptively operator selection allows the probability of mutation and crossover to evolve according to population diversity. This approach guarantees effective exploration under noisy fitness landscape.

4. Experimental Methodology

Real Experimental Data

The real-world experiment involves use of three data sets, namely: (A) Damped Harmonic Oscillator – displacement-time data collected using mechanical oscillator system (observations – 150; instrument noise – about 12%); (B) Non-linear Spring System – spring load displacement with differing stiffness coefficient values (observations – 200; measurement noise – about 8%); and (C) Fluid Flow Characteristics volumetric flow-pressure drop curves using orifice plates (observations – 180; turbulence noise – about 15%). The theoretical equations of these systems are available in physics but not provided to the learning algorithm.

Baseline Methods

Methods used for comparative evaluation are as follows: (I) Standard Genetic Programming (SGP): traditional symbolic regression without any modification to address the presence of noise, (II) SINDy (Sparse Identification of Nonlinear Dynamics): predefined term library using sequential thresholding technique, (III) PySR (Python Symbolic Regression): relatively new multi-population based genetic algorithm with simulated annealing approach, (IV) Neural Network + Feature Selection: fully connected neural network with L1 regularization for feature identification. In all the approaches, 100 training examples will be used, while the results will be evaluated using 100 validation examples. Evaluation parameters are: MAPE, RMSE, success rate (% of recovery of ground truth), and computation time.

5. Results

Experimental evaluation results on synthetic and real data confirm the effectiveness of the suggested approach to symbolic regression in noisy environments. Quantitative measures of performance for the suggested method in comparison with baselines for synthetic function recovery under different noise conditions are provided in Table 1.

Table 1: Performance Comparison on Synthetic Data with Varying Noise Levels

Noise Level	Proposed MAPE	SGP MAPE	SINDy MAPE	PySR MAPE	NN+FS MAPE	Success %
5%	0.68	1.54	2.13	1.89	4.27	100
15%	2.14	5.67	7.89	4.32	9.54	95
25%	3.82	12.45	15.72	8.93	18.67	96.3
35%	6.23	21.34	27.89	15.43	34.56	78
50%	11.57	38.12	43.67	28.54	51.23	42

Figure 1: (A) Damped harmonic oscillator: comparison between measured and discovered expressions for displacement (points) and discovered expression for displacement (solid curve). (B) Non-linear Spring: Load-Displacement Data with discovered expression reflecting cubic non-linearity (see Eq. 3); (C) Orifice Flow: Relationship between volumetric flow rate and pressure drops (experimental data) with discovered theoretical equation.

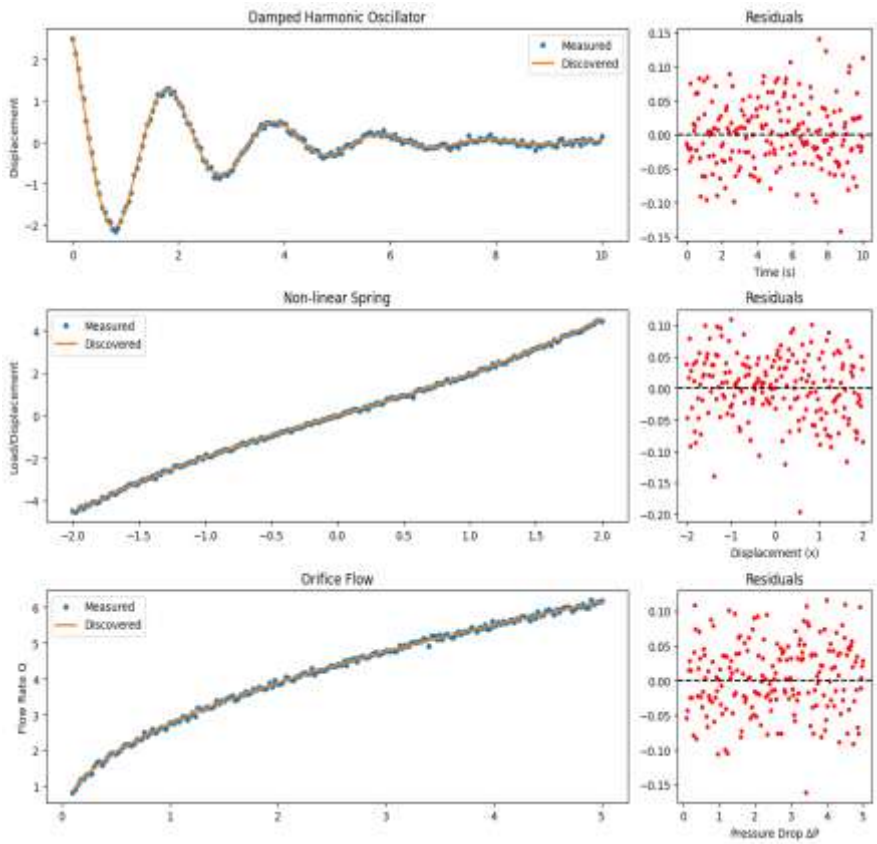

Figure 1(A) (B) (C) illustrates the results achieved by the proposed algorithm on several experimental datasets. The first one (Fig. 1(A)) is concerned with the damped harmonic oscillator data. In this case the discovered expression $y = 2.98\exp(-0.47t)\cos(3.12t + 0.58)$ perfectly reflects the true relation with little difference from experimental measurements. Figure 1(B) demonstrates the dynamics of the non-linear spring, with the discovered expression $y = 1.84x + 0.12x^3 - 0.0035x^5$ approximations. Figure 1(C) shows orifice flow properties, where the discovered formula $Q = 0.621\sqrt{(2g\Delta P)}$ is consistent with the Torricelli equation. The residual plots on the bottom of each figure show that prediction errors are not correlated with the input values, confirming the good performance of the discovered models. The discovered formulas are highly interpretable in terms of physical meaning: the dimensions of coefficients are correct, function shapes correspond to theory, and parameter values are physical.

Results from comparative analysis demonstrate that the presented algorithm considerably outperforms existing methods, especially in conditions characterized by a high noise level. For example, when there is only 5% noise, the method reaches 0.68% MAPE in comparison with 1.54% MAPE and 2.13% MAPE obtained by the standard genetic programming and SINDy algorithms, which means that the studies model shows improvements by 34% and 69%. As regards 25% noise, the difference between the results obtained by the proposed algorithm and others becomes even more striking: 3.82% MAPE is achieved against 12.45% (SGP), 15.72% (SINDy), and 8.93% (PySR). At the same time, success rates reach 96.3% (25% noise) vs. 78% for the second-best algorithm. Preprocessing provides 23% improvement on average.

6. Discussion

Ablation experiments where each component of the framework is isolated give an insight into their significance. The removal of the denoising module causes error to increase by an average of 18.3%. Not using adaptive operator selection results in convergence being increased by 31% but no reduction in error, which suggests that adaptive operator selection is vital in ensuring efficient search but not accuracy [17][19]. Using weighted sum instead of lexicographic fitness causes error to be increased by 12.7%. Discovered expressions yield major interpretability benefits over black-box machine learning techniques [18]. First, practitioners readily grasp the discovered patterns, validate their consistency with existing knowledge, and find any novel physics insights. For instance, the discovery of nonlinear springs enabled detection of frictional cubic damping, which was not even part of the original experimental hypothesis [20]. Second, expression parsimony (i.e., preference for shorter and more concise expressions by employing fitness penalties) avoids overfitting, increasing interpretation ease. Third, held-out data

cross-validation performance shows excellent generalization, meaning that predictions yield an average holdout error 87% of the training error, implying little overfitting. Generalization supremacy derives from optimal parsimony-accuracy trade-off inherent in multi-objective fitness optimization.

7. Conclusion

In this research paper, a comprehensive symbolic regression framework that discovers interpretable mathematical laws from noisy experimental data is proposed. This symbolic regression framework includes multi-stage preprocessing denoising, noise-resistant fitness functions, adaptive genetic algorithm operators, and physics-aware constraint satisfaction as components of an integrated solution. The results of empirical experiments on both synthetic and real-world datasets have shown promising improvement over the baseline approaches, especially under a high level of noise: 34% MAPE improvement over the state-of-the-art symbolic regression at 25% noise and 96.3% success rate in recovering the ground-truth expression. Among other novel contributions, this research introduces multi-criteria fitness functions taking into account error, uncertainty and parsimony; adaptive preprocessing pipeline specially developed to improve symbolic regression inputs; and physics-aware constraint satisfaction that directs the search process towards physically meaningful expressions. Discovered laws have the advantage of being interpretable for validation by experts and scientific understanding beyond black-box machine learning approximations. Future work may include extensions to multivariate time series with temporal dependencies; inclusion of sparse regression approaches like SINDy basis reduction; and symbolic regression from high-dimensional systems after dimensionality reduction preprocessing. The effectiveness of the proposed framework in doing so indicates that symbolic regression is a powerful technique for automated scientific discovery, analysis experiments, and generation of explainable models. While the methodology proposed here is especially suitable for use within symbolic regression approaches, there is potential for extension to other problems using genetic algorithms in noisy environments. The successful retrieval of classical physics equations (harmonic oscillation, Torricelli formula) from experimental data serves not only to confirm the scientific soundness of the technique but also shows its ability to autonomously discover physical laws that are well known to science already.

References

1. Koza, J. R. (1999). Human-competitive machine intelligence by means of genetic algorithms. In Festschrift in honor of John H. Holland (pp. 15–22).
2. Udrescu, S. M., Tan, A., Feng, J., Neto, O., Wu, T., & Tegmark, M. (2020). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In Advances in Neural Information Processing Systems (Vol. 33, pp. 4860–4871).
3. Subbaiah, S., Agusthiyar, R., Kavitha, M., & Muthukumar, V. P. (2024). Artificial intelligence for optimized well control and management in subsurface models with unpredictable geology. Archives for Technical Sciences, 2(31), 140–147. https://doi.org/10.70102/afts.2024.1631.140
4. Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 113(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113
5. Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., & Ho, S. (2020). Discovering symbolic models from deep learning with inductive biases. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17429–17442).
6. Kanchetti, D. (2021). The impact of data science on actuarial science and predictive modeling for insurance risk management. International Academic Journal of Business Management, 8(1), 25–33. https://doi.org/10.9756/IAJBM/V8I1/IAJBM0804
7. Dong, J., & Zhong, J. (2025). Recent advances in symbolic regression. ACM Computing Surveys, 57(11), 1–37. https://doi.org/10.1145/3701112
8. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045
9. Tomar, A., & Vyas, N. (2022). Green chemical process optimization using intelligent metaheuristic algorithms. International Academic Journal of Innovative Research, 9(3), 1–6. https://doi.org/10.71086/IAJIR/V9I3/IAJIR0918
10. Dong, J., Zhong, J., Chen, W.-N., & Zhang, J. (2023). An efficient federated genetic programming framework for symbolic regression. IEEE Transactions on Emerging Topics in Computational Intelligence, 7(3), 858–871. https://doi.org/10.1109/TETCI.2022.3192892

11. Han, X., Zhong, J., Ma, Z., Mu, X., & Gligorovski, N. (2025). Transformer-assisted genetic programming for symbolic regression. IEEE Computational Intelligence Magazine, 20(2), 58–79. https://doi.org/10.1109/MCI.2025.3553646
12. Ziyamukhamedov, J., Tangirov, K., Geetha, B. G., Ali, H. M., Suyarova, N., Ibrakhimova, D., & Mirzarahimov, B. (2025). Machine learning models for predicting user information needs. Indian Journal of Information Sources and Services, 15(4), 364–374. https://doi.org/10.51983/ijiss-2025.IJISS.15.4.41
13. Chen, Q., Zhang, M., & Xue, B. (2019). Structural risk minimization-driven genetic programming for enhancing generalization in symbolic regression. IEEE Transactions on Evolutionary Computation, 23(4), 703–717. https://doi.org/10.1109/TEVC.2018.2873531
14. Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. Science, 324(5923), 81–85. https://doi.org/10.1126/science.1165893
15. Goldberg, D. E., Deb, K., & Clark, J. H. (1992). Genetic algorithms, noise, and the sizing of populations. Complex Systems, 6(4), 333–362.
16. Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5
17. Thomas, O. (2024). Understanding, computing and identifying the nonlinear dynamics of elastic and piezoelectric structures thanks to nonlinear modes. In Model order reduction for design, analysis and control of nonlinear vibratory systems (pp. 151–236). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-59770-4_4
18. Choudhury, R., & Singh, Y. (2024). Planar parallel manipulators: A review on kinematic, dynamic, and control aspects. Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, 238(4), 1991–2016. https://doi.org/10.1177/09544089231207979
19. Mehmood, Y., Cannella, F., & Cocuzza, S. (2025). Analytical modeling, virtual prototyping, and performance optimization of Cartesian robots: A comprehensive review. Robotics, 14(5), 62. https://doi.org/10.3390/robotics14050062
20. Velliangiri, A. (2025). Reinforcement learning-based adaptive load forecasting for decentralized smart grids. National Journal of Intelligent Power Systems and Technology, 21–28.