SYMBOLIC REGRESSION AS AN INTERPRETABLE MACHINE LEARNING APPROACH FOR PARTICLE PHYSICS

D.M. Klekots[1,2], O.A. Bezshyyko[1], L.O. Golinka-Bezshyyko[1]
[1]Taras Shevchenko National University of Kyiv, Ukraine;
[2]Université Paris-Saclay, CNRS/IN2P3, IJCLab, Orsay, France
E-mail: denys.klekots@knu.ua

The high-energy physics experiments are gathering a large amount of data, and reliable algorithms are needed to select the signal from the background in the measured data. This study provides an overview of the application of symbolic regression as another machine learning algorithm that can be used for high-energy physics. The text describes training a symbolic regression model on the "Higgs Boson Detection 2025" open dataset from the ATLAS collaboration. Also, this work describes the advantages and weaknesses of the symbolic regression algorithm compared with boosted decision trees.

PACS: 29.85.-c; 07.05.Kf

INTRODUCTION

Experimental high-energy physics historically was on the front edge of big data analysis. At first, the analysis was mostly implemented with a cut-based approach, which implied sequential selection on observational physical variables (such as kinematic variables, like momentum or directional angle, and fit reconstruction quality variables). That approach is completely transparent, but does not account for more complex relations between variables.

With the commissioning of new collider experiments and increasing luminosity, the approach shifted towards the application of multivariable analysis. The first successful application of multivariable analysis was neural networks [1] used in the experiments on the Tevatron [2] and LEP [3]. A significant change happened with the introduction of Boosted decision trees [4] (BDT), which helped to discover the Higgs boson [5, 6], and became a standard due to incredible efficiency in signal/background separation.

An analysis of modern methods shows that machine learning enables the acquisition of state-of-the-art sensitivity in high-energy physics [7]. Not limited to high-energy physics, machine learning is important to other fields of physics too, as evidenced by the 2024 Nobel Prize in physics, which recognized pioneering research that enabled machine learning [8, 9].

Despite the high efficiency, the modern machine learning approaches (like boosted decision trees or neural networks) allow limited transparency in their decision-making, which is putting a limitation on their validation. For example, oftentimes it is unclear if a trained machine learning model senses the physical effect, or just exploits artefacts in the simulation data it was trained on. Those models are also not guaranteed to give reliable results if applied to a pace space different from the one on which they were trained. The models used in the research are validated, but it is preferable to have a transparent model, which would be interpretable by the human eye and written in formulas, similar to what physics uses. This work presents an approach that allows to obtain an analytical formula as a machine learning model.

1. THE SYMBOLIC REGRESSION

Symbolic regression proposes an alternative to widely used machine learning techniques. Instead of optimizing parameters of a fixed model architecture, it finds the mathematical formula of the function that describes the data the best. The symbolic regression relies on a multipopulational genetic algorithm for optimization. It simultaneously optimizes for the quality of data description (via minimization of the loss function) and model size (while penalizing model complexity), ensuring that the final model will be simple and effective.

The symbolic regression algorithm [10] internally works in an evolve-simplify-optimize loop. At the evolve stage, the formulas get randomly changed, their metrics are evaluated, and better-performing models are prioritized. The formulas then get mathematically simplified, by applying algebraic rules to the symbolic formulas, which helps the algorithm to keep formula complexity in check. The last stage is optimization, at which constants in the formula get optimized by fast gradient descent methods.

One of the advantages of the symbolic regression model is a lower chance of overfitting and biasing due to noise in training data, as the final model is less complex. Additionally, this makes symbolic regression work fine with small training datasets as well. Also, the prediction time of the trained model of the symbolic regression can be faster compared to neural networks or decision trees, but on the other hand, it takes much longer to train the model.

Though the symbolic regression framework was initially designed for data approximation, i.e., to find the function f(x) that would relate experimental data y=f(x), it can also be applied for signal/background discrimination in high-energy physics data. The trained boosted decision trees model gives a real number as an output, with a higher number predicting a signal and a lower number predicting a background. Instead of boosted decision trees, it could be a function f(x) from symbolic regression, giving the output for signal/background separation.

This is achieved by applying the custom loss function for symbolic regression. In this study, the following loss function was applied:

$$L[f] = \frac{1}{N}\sum_{i=1}^{N} ln\left(1 + e^{-(1-y_i f(x_i))}\right).$$

Here N is the number of entries in the training dataset; y_i is the training labels (1 for signal events and -1 for background); x_i collectively represents a set of features of the i-th event that is used for multivariable analysis. $f(x)$ is the function that the algorithm is optimizing. The loss function $L[f]$ depends on fitted function $f(x)$, and has lower values for a more precisely predicted signal/background separation.

2. TRAINING THE MODEL ON ATLAS HIGGS BOSON DATA

The described symbolic regression approach was applied for signal/background discrimination of the "Higgs Boson Detection 2025" open dataset [11]. The dataset was generated by the Monte Carlo method and are close to the real events of Higgs boson decay inside the ATLAS detector at the Large Hadron Collider. The overall dataset contains of 50000 labeled events, with 28 features that of the trajectories of particles. For training time optimization, in this study, only 25% of the events in the dataset were used during training of the model.

Histograms of the symbolic regression model outputs on training and validation datasets are presented in Fig. 1. The similarity of histograms on both datasets confirms that the model is not overfitting.

It is also important to note that even though the 28 features were provided for the symbolic regression algorithm to combine into a formula, due to the simplicity of the symbolic regression models, only 8 are used in the final trained model.

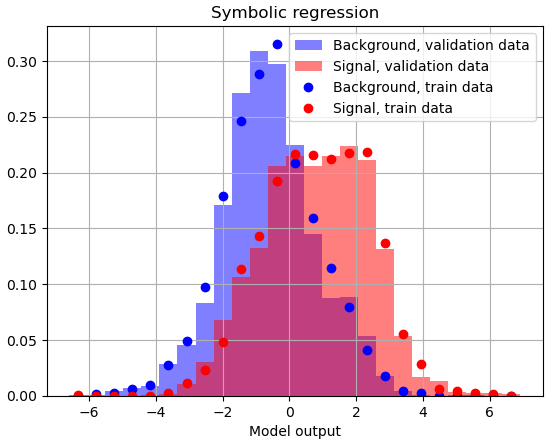

Fig. 1. The output of the symbolic regression model for validation and training datasets

The receiver operator characteristic (ROC) curve was calculated as a metric of model quality. The area under ROC curve for training data was found to be 74.18% for validation dataset, and 74.896% for training data. For comparison, the boosted decision trees model was trained on the same dataset, and estimated on validation dataset, it scored 77.482% of the area under the ROC curve, at the same time, the evaluation on the training dataset indicates significant overfitting.

Histograms of the boosted decision trees model outputs on training and validation datasets are presented in Fig. 2.

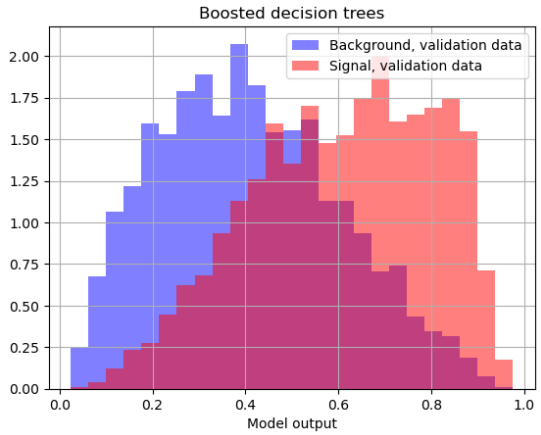

Fig. 2. The output of the boosted decision trees model for validation dataset

The ROC curves of symbolic regression and boosted decision trees are presented in Fig. 3.

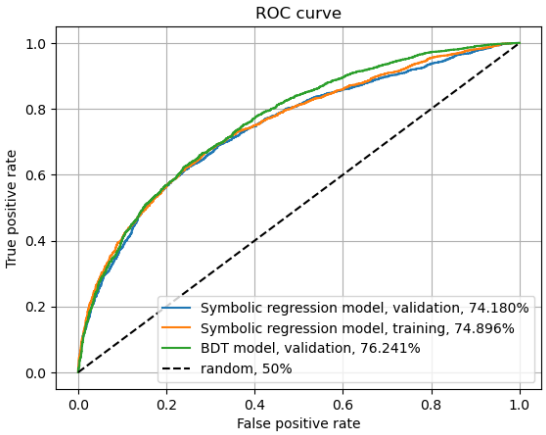

Fig. 3. ROC curves characterizing the trained model evaluated on training and validation datasets

CONCLUSIONS

In this work, the symbolic regression was tested for signal background discrimination in comparison to existing methods of boosted decision trees. The model was tested on the Higgs boson detection open dataset of the ATLAS collaboration with generated proton collision data.

The results of the study show that symbolic regression can be used for signal/background discrimination, it is simpler and uses fewer features compared to boosted decision trees, and is also less sensitive to overfitting.

At the same time, the signal/background discrimination efficiency of symbolic models, estimated with the area under the ROC curve, is comparable with boosted decision trees, but it is still lower. Also, the computation time needed to train the model is much larger for symbolic regression, due to the application of the evolutionary algorithm, instead of fast gradient algorithms.

Symbolic regression models are comparatively simple and consequently fast to use for prediction. The symbolic regression potentially could have an application in trigger systems of collider experiments, where the speed of prediction is crucially important, and more prioritized than accuracy.

REFERENCES

1. L. Teodorescu. Artificial neural networks in high-energy physics // *Inverted CERN School of Computing*. 2008, 2005 and 2006 edition, p.13-22.

2. Pushpalatha C. Bhat, Advanced Analysis Methods in High Energy Physics // *AIP Conf*. 2001, (583) 1, p. 22-30; https://doi.org/10.1063/1.1405257

3. P. Abreu, W. Adam, T. Adye, et al. Classification of the hadronic decays of the Z^0 into b and c quark pairs using a neural network // *Physics Letters B*. 1992, v. 295, p.383-395;
https://doi.org/10.1016/0370-2693(92)91580-3

4. Byron P. Roe, Hai-Jun Yang, Ji Zhu, et al. Boosted decision trees as an alternative to artificial neural networks for particle identification // *NIM-A*. 2005, (543) 2-3, p. 577-584;
https://doi.org/10.1016/j.nima.2004.12.018

5. ATLAS Collaboration et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC // *Physics Letters B*. 2012, (716) 1, p. 1-29;
https://doi.org/10.1016/j.physletb.2012.08.020

6. CMS Collaboration, et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC // *Physics Letters B*. 2012, (716) 1, p. 30-61;
https://doi.org/10.1016/j.physletb.2012.08.021

7. A. Radovic, M. Williams, D. Rousseau, et al. Machine learning at the energy and intensity frontiers of particle physics // *Nature*. 2018, 560, p. 41-48;
https://doi.org/10.1038/s41586-018-0361-2

8. J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities // *Proceedings of the National Academy of Sciences.* 1982, (79) 8, p. 2554-2558.

9. D.H. Ackley, G.E. Hinton, T.J. Sejnowski. A learning algorithm for Boltzmann machines // *Cognitive Science.* 1985, (9) 1, p. 147-169.

10. M. Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, *arXiv:2305.01582*, 2023;
https://doi.org/10.48550/arXiv.2305.01582

11. Haopeng Zhang. Higgs Boson Detection 2025, *Kaggle*, 2025; https://kaggle.com/competitions/higgs-boson-detection-2025

ВИКОРИСТАННЯ СИМВОЛЬНОЇ РЕГРЕСІЇ ЯК ІНТЕРПРЕТОВНОЇ МОДЕЛІ МАШИННОГО НАВЧАННЯ У ФІЗИЦІ ЧАСТИНОК

Д.М. Клекоць, О.А. Безшийко, Л.О. Голінка-Безшийко

Експерименти у фізиці високих енергій збирають величезні обсяги даних, і для відокремлення сигналу від фонового шуму в отриманих даних потрібні надійні алгоритми. Проведено огляд застосування символьної регресії як алгоритму машинного навчання, який можна використовувати у фізиці високих енергій. Описується навчання моделі символьної регресії на відкритому наборі даних “Higgs Boson Detection 2025” від колаборації ATLAS. Також описано переваги та недоліки алгоритму символьної регресії порівняно з посиленими деревами рішень.