The data-driven physical-based equations discovery using evolutionary approach

Mikhail Maslyaev
Alexander Hvatov
ITMO University
St. Petersburg, Russia
maslyaitis@gmail.com

ABSTRACT

The modern machine learning methods allow one to obtain the data-driven models in various ways. However, the more complex the model is, the harder it is to interpret. In the paper, we describe the algorithm for the mathematical equations discovery from the given observations data. The algorithm combines genetic programming with the sparse regression.

This algorithm allows obtaining different forms of the resulting models. As an example, it could be used for governing analytical equation discovery as well as for partial differential equations (PDE) discovery.

The main idea is to collect a bag of the building blocks (it may be simple functions or their derivatives of arbitrary order) and consequently take them from the bag to create combinations, which will represent terms of the final equation. The selected terms pass to the evolutionary algorithm, which is used to evolve the selection. The evolutionary steps are combined with the sparse regression to pick only the significant terms. As a result, we obtain a short and interpretable expression that describes the physical process that lies beyond the data.

In the paper, two examples of the algorithm application are described: the PDE discovery for the metocean processes and the function discovery for the acoustics.

CCS CONCEPTS

• **Computing methodologies** → **Genetic programming**; *Combinatorial algorithms*; • **Applied computing** → Environmental sciences.

KEYWORDS

generic programming, equation discovery, PDE discovery, data-driven models, sparse regression

ACM Reference Format:
Mikhail Maslyaev and Alexander Hvatov. 2020. The data-driven physical-based equations discovery using evolutionary approach. In *Genetic and Evolutionary Computation Conference Companion (GECCO '20 Companion), July 8–12, 2020, Cancún, Mexico.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3377929.3389943

1 INTRODUCTION

The modern machine learning methods utilize data-driven models for various purposes. It could be sophisticated surrogate-assisted models [6] as well as the complex model identification using approaches based on the evolutionary optimisation [2].

Nevertheless, the question of the interpretability of the models arises in the applications. Generally, we follow the extensive definition of the model interpretation provided [3]. Unfortunately, the complexity of the model and interpretability of it, in most cases, require trade-off to obtain good quality and the understanding of how the given model works [5].

Physics-based models could be the good examples of the interpretable models [8]. However, physical laws are mostly obtained manually by an expert in the field. We could try to derive them automatically in the closed form of the function [11], ordinary differential equation (ODE) , as well as the partial differential equations (PDE) [1, 10]. However, actual realizations require much preliminary work, such as a library of possible terms collection for symbolic regression [9].

In the paper, we try to extend the method of the PDE discovery [4] that, in our opinion, allows us to combine the transparency of the physical-based models and flexibility of genetic programming. Moreover, the utility of the sparse regression makes the resulting model form as concise as possible. The method is also similar to the symbolic regression. However, genetic programming allows us to build a flexible library of terms for regression.

2 THE ALGORITHM DESCRIPTION

The algorithm consists of three main elements: the building blocks, which we will call tokens below, selection, the evolutionary step, and the sparse regression step. We describe them consequently in this section. The general workflow of the algorithm is presented in Alg. 1.

2.1 The tokens selection

The tokens could be chosen arbitrary and do not have any restrictions on their nature. However, we stop on the applications of the homogeneous (in terms of the origin) set of the tokens. It means that we take only basic functions or only single derivative terms for evolution.

As an example, it can be all derivatives of the field up to the order k. An example of the first derivative token is shown in Eq. 1.

$$c(x, 1) = \frac{\partial u}{\partial x} \tag{1}$$

GECCO '20 Companion, July 8–12, 2020, Cancún, Mexico
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7127-8/20/07.
https://doi.org/10.1145/3377929.3389943

As seen token encodes the atomic expression. The form of the expression, as we said above, could be chosen arbitrarily.

From the set of tokens T, we compose the words of the length k, which is the first hyperparameter of the algorithm. We assume that every token in the word has the weighting coefficient and could be replaced with another one without model corruption.

2.2 The evolutionary part

The main goal of the evolutionary part of the algorithm is the detection of token combinations set (which can be denoted as $C = \{c_m = t_1 \cdot t_2 \cdot \ \dots \ \cdot t_k | \ t_i \in T, m = \overline{1, \ M}\}$, where M is the maximum number of terms in the equation, and belongs to the class of all possible token combination sets **C**), that is able to form the nontrivial linear combination with the minimum absolute value. This approach represents the task to detect structure of function or equation, which can be viewed as the minimization of functional $|\sum_{i=0}^{N} a_i c_i| \to 0 \ : \ \exists j : a_j \neq 0$, where c_i take roles of the the equation terms, and a_i - weights of the terms.

2.3 The regression part

While the previously discussed evolutionary part of the algorithm was developed to discover the best set of token combinations, which will represent the desired structure of the model, the regression methods are utilized to calculate the weights for these terms. Not only the best but also some redundant token combinations may be present in the best discovered set C'. Therefore, the task of set filtering is also bestowed to the regression element of the algorithm. The primary method that can perform these jobs is the sparse (regularized) regression, performed with LASSO operator.

3 CONCLUSION

In the paper, we describe the algorithm for the physical-based equations discovery. We want to outline the following properties of it:

- It does not depend on the form of the equation: it could be a polynomial, differential equation, and potentially the other models. However, additional work for the adaptation for each type of the equation is required;
- The genetic programming can be used to obtain an optimal bag of the terms from the small set of the building blocks and preliminarily defined mutation and crossover rules;
- The sparse regression step allows one to filter out the non-descriptive terms that lead to a robust model. As an additional advantage the resulting model has the short form of the expression, which makes the interpretation process easier;
- PDE discovery implementation is noise stable even for multi-dimensional data cases. The overall performance of the algorithm implementation allows reproducing tempo-spatial physic fields correctly.

The code and extended version of the article are publicity available at GitHub [7].

ACKNOWLEDGMENTS

This research is financially supported by The Russian Scientific Foundation, Agreement #19-11-00326.

```
Input: set of elementary tokens $T$
Parameters: $M$ - number of token combinations in a single
            individual; $k$ - number of elementary tokens
            in a combination; $n\_pop$ - number of
            candidate solutions in the population;
            evolutionary algorithm parameters: number
            of epochs $n_{epochs}$, mutation $r_{mutation}$ &
            crossover rates $r_{crossover}$, part of the
            population, allowed for procreation $a_{proc}$,
            number of individuals, refrained from
            mutation (elitism) $a_{elite}$; sparse regression
            parameter - sparsity constant $\lambda$
Result: set of token combinations $C^{best}$ (if required, with
        accompanying weights), representing best
        model/equation for the data
Generate population P of individuals of size $n\_pop$, with $M$ -
  random permutations of $k$ tokens to form sets $C^j$;
for epoch = 1 to $n_{epochs}$ do
    for individual in population do
        Apply sparse regression to individual to calculate
          weights;
        Calculate fitness function to individual;
    end
    Hold tournament selection and crossover;
    for individual in population except $n\_pop \times a_{elite}$ "elite"
      ones do
        Mutate individual;
    end
end
Select the individual with highest fitness function value as
  the final solution to the problem;
```

Algorithm 1: The pseudo-code of the algorithm operation

REFERENCES

[1] Jens Berg and Kaj Nyström. 2019. Data-driven discovery of PDEs in complex datasets. *J. Comput. Phys.* 384 (2019), 239–252.

[2] Sergey V Kovalchuk, Oleg G Metsker, Anastasia A Funkner, Ilia O Kisliakovskii, Nikolay O Nikitin, Anna V Kalyuzhnaya, Danila A Vaganov, and Klavdiya O Bochenina. 2018. A Conceptual Approach to Complex Model Management with Generalized Modelling Patterns and Evolutionary Identification. *Complexity* 2018 (2018).

[3] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[4] Mikhail Maslyaev, Alexander Hvatov, and Anna Kalyuzhnaya. 2019. Data-Driven Partial Derivative Equations Discovery with Evolutionary Approach. In *Computational Science – ICCS 2019*. Springer International Publishing, 635–641.

[5] Christoph Molnar. 2019. *Interpretable machine learning*. Lulu. com.

[6] Nikolay O Nikitin et. al. 2019. Deadline-driven approach for multi-fidelity surrogate-assisted environmental model calibration: SWAN wind wave model case study. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 1583–1591.

[7] NSS Team. 2020. Fedot E* algotirhms. https://github.com/ITMO-NSS-team/FEDOT.Algs.

[8] M. Raissim. 2018. Deep hidden physics models: Deep learning of nonlinear partial differential equations. (2018). https://arxiv.org/abs/1801.06637

[9] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. 2017. Data-driven discovery of partial differential equations. *Science Advances* 3, 4 (2017), e1602614.

[10] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher. 2017. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 473, 2197 (2017), 20160446. https://doi.org/473(2197):20160446

[11] Michael Schmidt and Hod Lipson. 2009. Distilling free-form natural laws from experimental data. *science* 324, 5923 (2009), 81–85.