Automatic Evolutionary Learning of Composite Models With Knowledge Enrichment

Anna V. Kalyuzhnaya
ITMO University
Saint Petersburg, Russia
anna.kalyuzhnaya@itmo.ru

Nikolay O. Nikitin
ITMO University
Saint Petersburg, Russia
nnikitin@itmo.ru

Pavel Vychuzhanin
ITMO University
Saint Petersburg, Russia
pavel.vychuzhanin@gmail.com

Alexander Hvatov
ITMO University
Saint Petersburg, Russia

Alexander Boukhanovsky
ITMO University
Saint Petersburg, Russia

ABSTRACT

This paper provides the main concepts of the knowledge-enriched AutoML approach and shortly describes the current results of the proof of concept implementation within the FEDOT framework. By knowledge enrichment, we mean the insertion of domain-specific models and expert-like meta-heuristics. Also, we involve multi-scale learning as a part of complex models identification. The proposed concepts make it possible to create effective and interpretable composite models.

CCS CONCEPTS

• **Theory of computation → Evolutionary algorithms**; • **Computing methodologies → Model development and analysis**;

KEYWORDS

AutoML, evolutionary learning, domain knowledge, machine learning

ACM Reference Format:
Anna V. Kalyuzhnaya, Nikolay O. Nikitin, Pavel Vychuzhanin, Alexander Hvatov, and Alexander Boukhanovsky. 2020. Automatic Evolutionary Learning of Composite Models With Knowledge Enrichment. 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/10.1145/3377929.3398167

1 INTRODUCTION

The automated machine learning (AutoML) approaches have become highly demanded in recent years. Nowadays, there are several popular frameworks for AutoML [6]: H2O, TPOT, AutoSklearn, MLBOX, AutoKeras. The modern AutoML is mostly focused on relatively simple tasks of hyperparameters optimization, input data preprocessing, selecting a single model or a set of models [1] (this approach is also referred to as the Combined Algorithm Selection and Hyperparameters optimisation - CASH), since the overall learning and meta-learning process is extremely expensive. On the other hand, pure ML-based solutions cannot solve the wide range of real-world problems (with a satisfactory trade-off between quality and resources consumption) and are mostly focused on specific classes of problems (image recognition, text mining, etc.). Today, the ML-based solutions often lose to solutions based on domain-specific models. For example, a lot of physics-based problems can be solved more efficiently with equation-based models: both data-driven and domain-specific numerical models. Also, the simulation quality can be increased using the hybrid approach [4].

So, integration of the domain knowledge to the ML-based solutions is a promising way to achieve a better quality of predictions, as well as the identification of data-driven models with custom structure. For example, a quite promising evolutionary approach named DarwinML that allows to combine and tune various models as a part of the custom-shaped pipeline is proposed in [5]. Also, the expert knowledge involvement allows improving the quality of auto-generated models [3] and evolutionary optimisation.

For these reasons, we decided to propose the concept of a framework for an evolutionary-based generative identification of custom multi-model pipelines, which support various types of models (ML, data-driven equation-based, domain-specific models).

2 THE KNOWLEDGE-ENRICHED AUTOML

Despite the fact that the idea of automated learning is extremely attractive there are several stumbling stones that should be overcome for the wide integration of ML and AutoML solutions into real-world systems:

1. AutoML algorithms learning is too expensive for most of the cases.

2. ML and AutoML solutions may give lower quality than domain-specific solutions in fields with the rich legacy of theory-based models (e.g. climate models, geological models, economic models, etc.). However, due to the acceleration and amplification of technological progress in these domains, science and industry want to improve the decision-making time, quality of forecasts, and so on.

3. Complex ML models may give quite good quality results but they unavoidably lose in clearly reproducing causal relationships and general regularities of simulated objects. This fact leads to the limited ability of such models to produce new knowledge and the low level of trust from decision-makers or customers due to their non-interpretability.

As a response to existing problems, we are developing a hybrid **knowledge-enriched** approach that aims to join the advantages

GECCO '20 Companion, July 8–12, 2020, Cancún, Mexico

ACM ISBN 978-1-4503-7127-8/20/07.
https://doi.org/10.1145/3377929.3398167

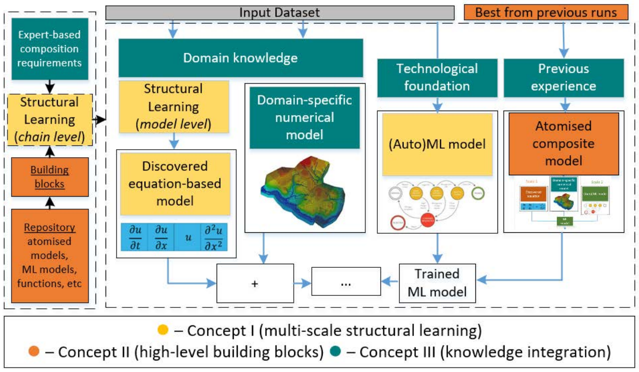
Input Dataset
Best from previous runs
Expert-based composition requirements
Structural Learning (*chain level*)
Building blocks
Repository atomised models, ML models, functions, etc
Domain knowledge
Structural Learning (*model level*)
Discovered equation-based model
$\frac{\partial u}{\partial t}$ $\frac{\partial u}{\partial x}$ u $\frac{\partial^2 u}{\partial x^2}$
Domain-specific numerical model
Technological foundation
(Auto)ML model
Previous experience
Atomised composite model
+
...
Trained ML model
● – Concept I (multi-scale structural learning)
● – Concept II (high-level building blocks) ● – Concept III (knowledge integration)

Figure 1: The proposed concepts of knowledge-enriched AutoML and their contribution to the structural blocks of composite model and learning environment.

of automated learning systems and domain-specific solutions to improve the quality of the whole solution.

In many large ML solutions, it is desirable to avoid structural learning, because of the expense of combinatorial identification of a model structure with dozen of variants. However, the explicit structure is often required in solutions that need high reliability and interpretability.

Therefore we suggest the idea of **learning the structure of the model**, which can be implemented on different levels (multi-scale learning). In addition, the paradigm of meta-heuristic approaches makes it possible to develop algorithms that could partially borrow patterns from the behaviour of model development experts. In can be achieved by the involvement of the behavioral meta-heuristics or the automated analysis of the known models' structure patterns.

To achieve the high quality and flexibility of structural learning, we decided to use the hybrid algorithm [2] that combines the flexibility of evolutionary algorithms and the interpretability of sparse regression to obtain the shortest mathematical expressions describing the data. This approach can be applied to various types of mathematical objects (as an example, to the partial differential equations that can be identified for the metocean data).

One of the ways to reduce resource intensity (and to increase interpretability) is to learn the structure of the composite model (as a graph-based composition of separate models) [5], where each node includes a previously obtained low-level model. In this case the variability of the structure is usually much lower than when assembling all of the models from scratch, this makes structure learning feasible.

To eliminate a gap between domain specific solutions and data-driven solutions, we develop algorithms for (1) for building **accurate composite models** which include **domain specific models**; (2) **discovering models** in terms of inherent domain notation.

The other concept in the frame of the hybrid knowledge-enriched approach is a form of representation of each node (the part of the composite model) as a fully independent model with the possibility of transformation of the whole composite model into the form of a model-node the can be used in subsequent runs of the framework (**fractality property**). These concepts are summarised in Fig. 1.

3 EVOLUTIONARY LEARNING OF THE COMPOSITE MODEL STRUCTURE

We can highlight two main types of evolutionary algorithms that are used in the described approach: algorithms for learning the structure of the composite model as a graph (composing algorithm); and algorithms for discovering of data-driven node-models. Nowadays the existing AutoML solutions are based on quite basic optimization techniques.We decided to base our composer on a genetic algorithm with custom evolutionary operators and a DAG-based genotype. It allows generating a chain with several ML models obtained from the repository according to the specified requirements.

4 THE FEDOT FRAMEWORK

To combine the proposed concepts and methods with the existing state-of-the-art approaches and share the obtained experience with the community, we decided to develop the FEDOT framework.

The framework kernel can be configured for different classes of tasks: predictive modeling, system dynamics, simulation modeling. The framework includes the library with implementations of intelligent algorithms to identify data-driven models with different requirements; and composite models (chains of models) for solving specific subject tasks (social and financial, metocean, physical, etc.).

It is possible to obtain models with given parameters of quality, complexity, interpretability; to get an any-time result; to pause and resume model identification; to integrate many popular Python open source solutions for AutoML/meta-learning, optimization, quality assessment, etc.; re-use the models created by other users.

5 CONCLUSION

This paper is devoted to the idea and main concepts of the knowledge enriched AutoML approach. The implementation of all proposed features will be done within the FEDOT framework. At the current moment, we implemented two types of structural learning algorithms with real-world examples. FEDOT framework, as well as separate algorithms, is open-source and available in the repository[1].

ACKNOWLEDGMENTS

This research is financially supported by The Russian Scientific Foundation, Agreement #19-71-00150.

REFERENCES

[1] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2019. AutoML: A Survey of the State-of-the-Art. *arXiv preprint arXiv:1908.00709* (2019).

[2] Alexander Hvatov and Mikhail Maslyaev. 2020. The data-driven physical-based equations discovery using evolutionary approach. (2020). arXiv:cs.NE/2004.01680

[3] Doris Jung-Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *Data Engineering* 58 (2019).

[4] Fearghal O'Donncha, Yushan Zhang, Bei Chen, and Scott C James. 2018. An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts. *Journal of Marine Systems* 186 (2018), 29–36.

[5] Fei Qi, Zhaohui Xia, Gaoyang Tang, Hang Yang, Yu Song, Guangrui Qian, Xiong An, Chunhuan Lin, and Guangming Shi. 2018. DarwinML: A Graph-based Evolutionary Algorithm for Automated Machine Learning. *arXiv preprint arXiv:1901.08013* (2018).

[6] Marc-André Zöller and Marco F. Huber. 2019. Benchmark and Survey of Automated Machine Learning Frameworks. (2019). arXiv:cs.LG/1904.12054

[1] https://github.com/ITMO-NSS-team/FEDOT.Algs/wiki/Knowledge-enriched-AutoML