Post-hoc Model-Agnostic Explanation through Evolutionary Computing-based Methods

Internship 2021

Topic:	Genetic Programming, Machine Learning, Explainable AI
Funding:	European Project H2020 FET 952060 TRUST-AI
Team:	TAU (TAckling the Underspecified), INRIA and Université Paris-Saclay
Advisors:	Marc Schoenauer (marc.schoenauer@inria.fr)
Duration:	5 to 9 months, starting Feb. or March 2021
Location:	LRI, Paris-Sud University – Building 660 – Shannon
Level:	Master or last year Engineering School

1 Context & Motivation

Explanation and interpretability are crucial features required from machine learning (ML) users to increase their confidence in the models [1]. Indeed, the General Data Protection Regulation (GDPR) introduced them the option to demand explanations of automated-decisions.

Interpretability in this case means the ability to explain the properties of a model employing understandable terms to a human in the function of only some particular data and model outputs [2]. An explanation, on the other hand, comprises a set of self-contained features from the interpretable domain which leads a given input to produce a specific decision, without requiring any further explanation [3, 4, 5]. Model interpretability can help the users to detect and correct bias in the training data, to point out potential data perturbation that might lead to change a model output, to ensure that only the correct features influence the output, and to understand the underlying phenomenon such as in physics and social sciences [6, 7].

On the one hand, state-of-the-art ML methods can achieve high accuracy performance. Therefore, it usually hard to understand the mechanisms by which they work due to their huge parameter space. These methods are considered black-box models [5]. On the other hand, some machine learning models can offer high interpretability, but with a drop in their predictive performance. Clearly, there is a trade-off choice between accuracy, explanation, and interpretability of ML models [8, 9].

In contrast, genetic programming (GP), an evolutionary computation (EC) technique, follows a bio-inspired strategy by which a population of solutions representing models is evolved during a determined number of generations. A fitness function is then used to evaluate the performance of the models and to probabilistically select them to go through mutation and crossover operations. Evolutionary computing-based techniques can work without requiring beforehand from users the structure of the solution. Moreover, they are domain-independent and can solve problems by starting from high-level statements. Furthermore, evolutionary computing-based methods can be combined with other heuristic techniques (e.g., local search) to improve the quality of the solutions [10] by taking into account the semantic of each generated solution [11, 12]. Finally, genetic programming evolved models are normally interpretable by humans [13].

2 Goal

The aim of this internship is to work on one of the following topics.

1. **Post-hoc explanation through evolutionary computing-based techniques**: explore how evolutionary computing-based methods can be used to provide both model agnostic post-hoc explanation and local explanation for the prediction of black-box models.
2. **Mixed-integer linear programming (MILP) and genetic programming**: investigate the use of MILP as a way to guide evolutionary computing methods when searching the optimal solutions, as well as to improve their explainability.
3. **Genetic programming and reinforcement learning**: investigate how domain knowledge can be used to reward an agent according to the function or operators it selects to evolve a set of GP models.
4. **Federated learning and genetic programming**: investigate how the concept of federated learning can be explored by evolutionary computing techniques.

3 Profile

The internship requires excellent machine learning skills, the ability to work on cross-disciplinary problems, and good programming experience, preferably in Python.

References

[1] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: a survey of results," in *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, 2017, pp. 1–6.

[2] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, 2019, pp. 5–22.

[3] F. C. Keil, "Explanation and understanding," *Annual Review of Psychology*, vol. 57, pp. 227–254, 2006.

[4] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, 2018.

[6] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *41st International convention on information and communication technology, electronics and microelectronics*, 2018, pp. 0210–0215.

[9] H. K. Dam, T. Tran, and A. Ghose, "Explainable software analytics," in *40th International Conference on Software Engineering: New Ideas and Emerging Results*, 2018, pp. 53–56.

[10] W. E. Hart, N. Krasnogor, and J. E. Smith, "Memetic evolutionary algorithms," in *Recent advances in memetic algorithms*, 2005, pp. 3–27.

[11] R. Ffrancon and M. Schoenauer, "Memetic semantic genetic programming," in *Annual Conference on Genetic and Evolutionary Computation*, 2015, pp. 1023–1030.

[12] ——, "Greedy semantic local search for small solutions," in *Companion Publication of the Annual Conference on Genetic and Evolutionary Computation*, 2015, pp. 1293–1300.

[13] P. G. Espejo, S. Ventura, and F. Herrera, "A survey on the application of genetic programming to classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 2, pp. 121–144, 2009.