Topic ID: 350-IT-EN

Evolutionary Computation and Explainable AI

Supervisor: Prof. Ing. Šenkeřík Roman, Ph.D.

Consultant: Ing. Viktorin Adam, Ph.D., ---

Department: Department of Informatics and Artificial Intelligence

Programme: Information Technologies

Abstract:

This Ph.D. thesis explores the evolving domain of evolutionary computation (EC) with a strong focus on enhancing its explainability and trustworthiness or on explainable AI (XAI) itself. There are two possible directions. Firstly the explainable AI principles for evolutionary computation, and secondly evolutionary computation for explainable AI.

Evolutionary computation, which draws inspiration from natural biological evolution, employs algorithms that iteratively evolve solutions to complex problems. A critical challenge in this field is the opacity of these algorithms, making it difficult to understand and trust their decision-making processes. This research seeks to address this issue by developing methods that make evolutionary algorithms both interpretable and transparent, enhancing their reliability and acceptance. A significant portion of the thesis will be dedicated to integrating the concept of trustworthy optimization into evolutionary computation. Trustworthy optimization involves creating algorithms that are not only effective but also reliable and accountable.

The field of (XAI) seeks to address the growing need for transparency and interpretability in AI systems. This research may contribute to this endeavor by exploring the application of EC methods within XAI. The research will focus on a novel framework that applies EC techniques to develop AI models while simultaneously ensuring their explainability, i.e., propose an approach that simplifies the solutions generated by EC algorithms without substantially reducing their effectiveness. The PhD thesis may cover several directions like genetic programming for rule generation. For XAI, the programs/models can be evolved to generate a set of rules or decision trees that are easily interpretable by humans. Further, feature selection and reduction, where EC algorithms can be used for feature selection and dimensionality reduction in data sets, which is crucial for building explainable models. Possible directions also include optimization of neural network architectures, post-hoc analysis and visualization, simplifying complex solutions, and more.

Literature:

[1] BACARDIT, Jaume, et al. The intersection of evolutionary computation and explainable AI. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2022. p. 1757-1762.

[2] XU, Feiyu, et al. Explainable AI: A brief survey on history, research areas, approaches and challenges. In: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. Springer

International Publishing, 2019. p. 563-574.

[3] HOFFMAN, Robert R., et al. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608, 2018.

[4] HOLZINGER, Andreas. From machine learning to explainable AI. In: 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE, 2018. p. 55-66.

[5] SAMEK, Wojciech; MÜLLER, Klaus-Robert. Towards explainable artificial intelligence. Explainable AI: interpreting, explaining and visualizing deep learning, 2019, 5-22.

[6] BACARDIT, Jaume, et al. Evolutionary Computation and Explainable AI Towards ECXAI 2023: A Year in Review. ACM SIGEVOlution, 2023, 16.2: 1-6.