Special issue on explainable AI in evolutionary computation.

BACARDIT, J., BROWNLEE, A., CAGNONI, S., IACCA, G., MCCALL, J. and WALKER, D. (eds.)

2024

Introduction to the Special Issue on Explainable AI in Evolutionary Computation

Explainable Artificial Intelligence (**XAI**) has recently emerged as one of the most active areas of research in AI. While **Evolutionary Computation** (**EC**) is also a very active research area, the intersection between XAI and EC is still rather unexplored. This topic was the subject of our Workshops on **Evolutionary Computing and Explainable Artificial Intelligence** (**ECXAI**) organized at GECCO 2022 and GECCO 2023. This special issue collects four articles further exploring the intersection between XAI and EC, including both, the use of EC for XAI as well as the use of explainability techniques to better understand EC methods.

In "Multi-objective Feature Attribution Explanation For Explainable Machine Learning", Ziming Wang, Changwu Huang, Yun Li, and Xin Yao formulate the **feature attribution-based explanation** (**FAE**) as a multi-objective learning problem that simultaneously considers multiple explanation quality metrics, such as faithfulness, sensitivity, and complexity. Their approach, compared with six state-of-the-art FAE methods on eight datasets, is able to provide a diverse set of explanations with different tradeoffs in terms of higher faithfulness, lower sensitivity, and lower complexity.

In "A Multi-Objective Evolutionary Approach to Discover Explainability TradeOffs when Using Linear Regression to Effectively Model the Dynamic Thermal Behaviour of Electrical Machines", Tiwonge Msulira Banda, Alexandru-Ciprian Zăvoianu, Andrei Petrovski, Daniel Wöckinger, and Gerd Bramerdorfer propose a multi-objective strategy for creating Linear Regression models of the heat transfer in rotating electrical machines. Their approach provides decision makers with a clear overview of the optimal tradeoffs between data collection costs, the expected modeling errors, and the overall explainability of the generated thermal models.

In "Exploring the Explainable Aspects and Performance of a Learnable Evolutionary Multiobjective Optimization Method", Giovanni Misitano explores the combination of learnable evolutionary models, namely a class of optimization algorithms that combine evolutionary algorithms with machine learning models (where the latter are utilized to learn a hypothesis describing what characterizes a desired solution, and how to generate it) with interactive indicator-based evolutionary multiobjective optimization, to create a learnable evolutionary multiobjective optimization method. The proposed method also leverages interpretable machine learning, to provide decision makers with potential insights about the problem being solved in the form of rule-based explanations.

In "An analysis of the ingredients for learning interpretable symbolic regression models with human-in-the-loop and genetic programming", Giorgia Nadizar, Luigi Rovito, Andrea De Lorenzo, Eric Medvet, and Marco Virgolin study a recently-introduced human-in-the-loop system that allows the user to steer the generation process of **Genetic programming** (**GP**) to their preferences,

which are online-learned by an artificial neural network. Focusing on symbolic regression problems, they propose an incremental experimental evaluation aimed at assessing the effectiveness of a human-in-the-loop approach to discover interpretable machine learning models with GP.

The intent of this special issue is to stimulate researchers from the EC field to consider the explainability dimension in their research, hence fostering further studies that explicitly address this aspect. There are many untapped areas, from providing explanations about the characteristics of the optimization problems (e.g., in terms of fitness landscape analysis), to searching for tradeoffs between explainability, fairness, and accuracy in machine learning. Our view is that, in the long term, explainability can provide EC methods with the necessary foundation to further broaden their applications in real-world domains, while, in turn, EC can add a rich dimension to XAI research.

Jaume Bacardit
Newcastle University, Newcastle upon Tyne, UK

Alexander Brownlee
University of Stirlin, Stirling, UK

Stefano Cagnoni
University of Parma, Parma, Italy

Giovanni Iacca
University of Trento, Povo, Italy

John McCall
Robert Gordon University, Aberdeen, UK

David Walker
University of Exeter, Exeter, UK

Guest Editors