Interpretable Machine Learning

Kazim Topuz
The Tulsa University, Collins College of Business
Kazim-Topuz@utulsa.edu

Akhilesh Bajaj
The Tulsa University, Collins College of Business
Akhilesh-Bajaj@utulsa.edu

Ismail Abdulrashid
The Tulsa University, Collins College of Business
Ismail-Abdulrashid@utulsa.edu

Abstract

The machine learning scientific community has concentrated on interpretable machine learning (IML) in terms of algorithmic interpretability, explainability, and transparency. This minitrack includes applications of IML in the fields of transportation and cybersecurity. Six high-quality articles were submitted in the initial year of the minitrack's existence, but we could only approve three of them after a rigorous review process.

Keywords: Interpretable machine learning, interpretable decision support system, explainable artificial intelligence.

Machine learning (ML) has sparked considerable attention in recent years due to its ability to correctly assess a wide range of complex situations. However, there is a growing recognition that, in addition to producing predictions, ML models may provide knowledge about the domain relations included in data, which is commonly referred to as interpretations. The scientific community in artificial intelligence (AI) has focused on explainable AI (XAI) in terms of algorithmic interpretability, explainability, transparency, and, finally, accountability of algorithmic decisions (Kim et al., 2020). While white-box techniques, such as rule learners and inductive logic programming, provide explicit modeling that is inherently interpretable, black-box techniques, such as (deep) neural networks, provide veiled models (Rai, 2020). With the increasing use of machine learning, there have been major social concerns about using black-box models for high-stakes choices in fields such as healthcare, medicine, finance, and criminal justice. The capacity to represent information in human-comprehensible language -interpretability- has piqued the interest of academics and industry alike. These interpretations have been used in healthcare (Topuz et al., 2018), education (Coussement et al., 2020), finance (Fu et al., 2021), and transportation (Topuz et al., 2021).

This minitrack will be the go-to place for timely and in-depth presentations on the most recent advances in interpretable ML (IML). We intend to tackle these difficulties from the standpoints of modeling and learning, to develop interpretable methodologies and models that explain themselves and their output. As a result, this minitrack presents papers on advancements in IML from the modeling and learning perspectives.

As a direct consequence of this, our call covers a wide range of different topics. We searched for unique papers of the highest possible quality that provide research on the following (by no means complete) topics:

- Probabilistic graphical model applications
- Rule learning for interpretable machine learning
- Interpretation of black-box models
- Interpretability in reinforcement learning
- Interpretable supervised and unsupervised models
- Interpretation of neural networks and ensemble-based methods
- Interpretations of random forests and other ensemble models
- Causality of machine learning models
- Novel applications requiring interpretability
- Methodologies for measuring the interpretability of machine learning models
- Interpretability-accuracy trade-off and its benchmarks

In the initial year of the minitrack's existence, there were a total of six high-quality papers that were submitted for consideration, but we could only accept three of them. These examples demonstrate the breadth of possible applications and research problems that might be centered on interpretable machine learning. Additionally, they offer a diversity of viewpoints on the significance of interpretability, explainability, and transparency. Among the subjects that will be discussed are the detection of cyberattacks

URI: https://hdl.handle.net/10125/102782
978-0-9981331-6-4

and understanding aviation accidents, as well as continual representation learning.

The three papers that are included in the minitrack are as follows:

1. Business Inferences and Risk Modeling with Machine Learning: The case of Aviation Incidents (Burak Cankaya, Kazim Topuz, Aaron Glassman)
   This paper presents a machine-learning framework that can be used to estimate airplane damage. Furthermore, it describes patterns of flight parameters discovered using a modeling program and sheds insight into the underlying causes of certain aircraft mishaps. In summary, this paper suggests they can predict aircraft damage with an accuracy of 85% and an in-class accuracy of 84%.

2. Bayesian Networks for Interpretable Cyberattack Detection (Barnett Yang, Matthew Hoffman, Nathanael Brown)
   The authors used Bayesian Networks to identify cyberattacks using Bayes-Server API-based automated workflows. Their proposed methodology delivers an interpretable and easier-to-understand pipeline. The Sandia National Laboratories host-based log data collection is utilized to create the technique and compare it to other well-known machine learning algorithms. The outcomes are compared to the widely used random forest classification approach.

3. Hebbian Continual Representation Learning (Paweł Morawiecki, Andrii Krutsylo, Maciej Wołczyk, and Marek Smieja)
   Continual Learning seeks to place machine learning in a more realistic context, where tasks are learned sequentially and the i.i.d. assumption is not maintained. Although this environment is natural for biological systems, it presents significant challenges for machine learning models such as artificial neural networks. To close the performance gap, this study looks at whether biologically inspired Hebbian learning is effective for dealing with ongoing issues. The authors additionally modify the technique for the supervised case and get promising results in class-incremental learning.

Due to the fact that the subject has garnered an even greater amount of attention throughout the past few years, we want to arrange the minitrack once more the following year, and we sincerely hope that we will be able to organize the session in person once more in breathtaking Hawaii.

References

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2020.113325

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, *32*(1), 72–92. https://doi.org/10.1287/isre.2020.0990

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*(July 2019), 113302. https://doi.org/10.1016/j.dss.2020.113302

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems*, *106*, 97-109.

Topuz, K., & Delen, D. (2021). A probabilistic Bayesian inference model to investigate injury severity in automobile crashes. *Decision Support Systems*, *150*, 113557.