

Using Ontologies to Enhance Human Understandability of Global Post-hoc Explanations of Black-box Models (Extended Abstract)

Roberto Confalonieri¹, Tillman Weyde², Tarek R.Besold³ and Fermín Moscoso del Prado Martín⁴

¹Faculty of Computer Science, Free University of Bozen-Bolzano, Domenikanerplatz 3, Bolzano-Bozen, Italy

²Dept. of Computer Science, City, University of London, GB-EC1V 0HB London

³Philosophy & Ethics, Faculty of IE/IS, Eindhoven University of Technology, 5600 MB Eindhoven

⁴Lingvist Technologies OÜ, Tallinn, Estonia

1. Extended Abstract

This extended abstract overviews the work presented in [1] where an extension of TREPAN is proposed. TREPAN is a seminal global explanation approach that extracts surrogate decision trees from black-box models. TREPAN was extended to take into account explicit knowledge, modeled by means of ontologies, to extract human-understandable explanations.

TREPAN is a tree induction algorithm that recursively extracts decision trees from black-box classifiers [2]. The algorithm is model-agnostic, and it can be applied to explain any black-box classifier (e.g., Multi-Layer Perceptron, Random Forest). TREPAN combines the learning of the decision tree with a trained machine learning classifier (the oracle).

The proposed extension of the TREPAN algorithm, called TREPAN RELOADED, uses a modified information gain that, in the creation of split nodes, gives priority to features associated with more general concepts defined in a domain ontology. This was achieved by means of an information content measure defined using the refinement operators [3].

The perceived understandability of the extracted explanations by humans was tested by means of a user study with four different tasks. Results were evaluated in terms of response times and correctness, subjective ease of understanding and confidence, and similarity of free text responses. The results showed that decision trees generated with TREPAN Reloaded, taking into account domain knowledge, were significantly more understandable than those generated by standard TREPAN. The enhanced understandability of post-hoc explanations was achieved with little compromise on the accuracy with which the surrogate decision trees replicate the behaviour of the original neural network models.

3rd International Workshop on Data meets Applied Ontologies in Explainable Artificial Intelligence (DAO-XAI 2021)


✉ roberto.confalonieri@unibz.it (R. Confalonieri); t.e.veyde@city.ac.uk (T. Weyde); tarek.besold@gmail.com (T. R.Besold); fermosc@gmail.com (F.M. d.P. Martín)

🌐 <http://www.inf.unibz.it/~rconfalonieri/> (R. Confalonieri)

🆔 0000-0003-0936-2123 (R. Confalonieri)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

References

- [1] R. Confalonieri, T. Weyde, T. R. Besold, F. Moscoso del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, *Artificial Intelligence* 296 (2021).
- [2] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: *NIPS 1995*, MIT Press, 1995, pp. 24–30.
- [3] N. Troquard, R. Confalonieri, P. Galliani, R. Peñaloza, D. Porello, O. Kutz, Repairing Ontologies via Axiom Weakening, in: S. A. McIlraith, K. Q. Weinberger (Eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 1981–1988.