Explaining Random Forests Using Bipolar Argumentation And Markov Networks

Nico Potyka , Xiang Yin , Francesca Toni
Department of Computing, Imperial College London, London, UK
{n.potyka, x.yin20, f.toni}@imperial.ac.uk

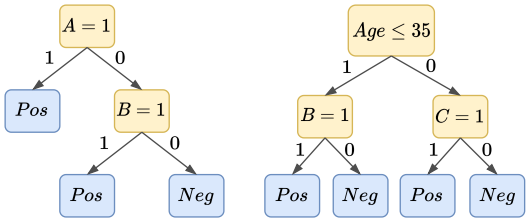

Figure 1: A simple random forest with two decision trees.

Random forests (RFs) (Breiman 2001) are machine learning models with various applications in areas like E-commerce, Finance and Medicine. They consist of multiple decision trees that use different subsets of the available features. Figure 1 shows a simple RF with two decision trees. The boolean features A, B, C are symptoms that can be observed, Age is a numerical feature for the age of a patient and the classification corresponds to a diagnosis that can be positive or negative. Given an input, every tree makes an individual decision and the output of the RF is obtained by a majority vote. RFs have low risk of overfitting; support both classification and regression tasks and come equipped with some feature importance measures (Breiman 2001). However, feature importance measures can be too simplistic as they can represent neither joint effects of features (e.g., multi-drug interactions) nor non-monotonicity (e.g., increasing the weight may be healthy for an underweight person, but not for an overweight person).

In recent years, a variety of other explanation methods has been proposed. Model-agnostic feature importance measures like LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017) and MAPLE (Plumb, Molitor, and Talwalkar 2018) have similar limitations like the feature importance measures defined for RFs. Counterfactual explanations explain how an input can be modified to change the decision (Wachter, Mittelstadt, and Russell 2017), but mainly explain the model locally. Another interesting family of explanation methods are abductive explanations, also called prime implicant explanations (Shih, Choi, and Darwiche 2018; Izza and Marques-Silva 2021; Wäldchen et al. 2021). Roughly speaking, abductive explanations are sufficient reasons for a classification. For example, $(B = 1, Age = 20)$ is sufficient for a positive diagnosis with respect to the RF in Figure 1. Recently, SAT encodings have been applied to compute abductive explanations in tree ensembles (Izza and Marques-Silva 2021; Ignatiev et al. 2022) and many other logic-based explanation approaches have been investigated for this purpose (Marques-Silva and Ignatiev 2022; Cyras et al. 2021; Vassiliades, Bassiliades, and Patkos 2021).

Existing reasoning approaches for explaining RFs are based on classical reasoning formalisms. While every individual tree in a RF can be seen as a collection of classical Horn rules, the trees are often jointly inconsistent. Therefore, it seems natural to investigate non-classical reasoning formalisms to explain RFs. In (Potyka, Yin, and Toni 2023), we studied non-monotonic and probabilistic reasoning approaches to explain RFs, namely bipolar argumentation frameworks (BAFs) (Amgoud et al. 2008; Oren and Norman 2008; Boella et al. 2010; Cayrol and Lagasquie-Schiex 2013) and Markov networks (MNs) (Koller and Friedman 2009). While explanations based on classical logic often require a task-specific translation (e.g., for abductive explanations), we showed that RFs can be directly translated into

- BAFs under bi-stable semantics (Potyka 2021) with quadratic time and space complexity in such a way that there is a 1-1-correspondence between inputs of the forest and bi-stable labellings of the explanation BAF and
- MNs with linear time and space complexity in such a way that there is a 1-1-correspondence between inputs of the forest and the support of the MN (random variable assignments with non-zero probability).

Our translations allow reducing explanation tasks like finding sufficient and necessary reasons for classes and analytical tasks like computing the number of non-ambiguous inputs (inputs for which an unambiguous majority decision can be made) to well studied reasoning tasks in these frameworks. Explanations can be generated by finding sufficient and necessary reasons in argumentation frameworks (Borg and Bex 2021) for BAFs and by applying exact and approximate probabilistic reasoning algorithms for MNs (Koller and Friedman 2009). MNs additionally support more general δ-sufficient ($100 \cdot \delta$ % of all inputs that are compatible with the reason yield a particular class decision) and δ-necessary reasons ($100 \cdot \delta$ % of inputs classified in a particular way are

compatible with the reason).

We presented a probabilistic approximation scheme for computing δ-sufficient and δ-necessary reasons (sufficient and necessary reasons are obtained for the special case $\delta = 1$) and demonstrated its effectiveness on some standard benchmark datasets. In our experiments, the fraction of non-ambiguous inputs was usually around 99%. For the Iris dataset (classification of Iris flowers), $(\text{petallength} \in (5.0, 5.14])$ was an example of an almost sufficient reason ($\delta \approx 1$) for the class Virginica. $(\text{sepallength} \in (5.45, 5.5], \text{petallength} \in (2.64, 2.75])$ was an almost sufficient reason for the class Versicolor. For the Mushroom dataset (classifications of mushrooms as poisonous or edible), $\text{Odor\_Foul} = 0$ was 0.98-necessary for Edible. Our Python implementation is available in the *Uncertainpy*[1] library in the folder *examples/explanations/randomForests*.

We believe that the connection between RFs, BAFs and MNs is particularly interesting for KR researchers working on sufficient and necessary reasoning (explanation) and counting (non-ambiguous inputs) in argumentation frameworks or probabilistic inference (explanations) and computing the partition function (non-ambiguous inputs) in probabilistic graphical models because the connection may allow applying their results immediately to explaining and analyzing random forests more efficiently.

Acknowledgments

This research was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

References

Amgoud, L.; Cayrol, C.; Lagasquie-Schiex, M.-C.; and Livet, P. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems* 23(10):1062–1093.

Boella, G.; Gabbay, D. M.; van der Torre, L.; and Villata, S. 2010. Support in abstract argumentation. In *International Conference on Computational Models of Argument (COMMA)*, 40–51. Frontiers in Artificial Intelligence and Applications, IOS Press.

Borg, A., and Bex, F. 2021. Necessary and sufficient explanations for argumentation-based conclusions. In Vejnarová, J., and Wilson, N., eds., *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, volume 12897 of *LNCS*, 45–58. Springer.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2013. Bipolarity in argumentation graphs: Towards a better understanding. *International Journal of Approximate Reasoning* 54(7):876–899. Publisher: Elsevier.

[1] https://github.com/nicopotyka/Uncertainpy

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In Zhou, Z., ed., *International Joint Conference on Artificial Intelligence (IJCAI)*, 4392–4399. ijcai.org.

Ignatiev, A.; Izza, Y.; Stuckey, P. J.; and Marques-Silva, J. 2022. Using maxsat for efficient explanations of tree ensembles. In *AAAI Conference on Artificial Intelligence (AAAI)*, 3776–3785. AAAI Press.

Izza, Y., and Marques-Silva, J. 2021. On explaining random forests with SAT. In Zhou, Z., ed., *International Joint Conference on Artificial Intelligence, (IJCAI)*, 2584–2591.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 4768–4777.

Marques-Silva, J., and Ignatiev, A. 2022. Delivering trustworthy AI through formal XAI. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Oren, N., and Norman, T. J. 2008. Semantics for evidence-based argumentation. In *International Conference on Computational Models of Argument (COMMA)*, 276–284. IOS Press.

Plumb, G.; Molitor, D.; and Talwalkar, A. 2018. Model agnostic supervised local explanations. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2520–2529.

Potyka, N.; Yin, X.; and Toni, F. 2023. Explaining random forests using bipolar argumentation and markov networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, in press. AAAI Press.

Potyka, N. 2021. Generalizing Complete Semantics to Bipolar Argumentation Frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2021)*, Lecture Notes in Computer Science, 130–143. Springer.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In Lang, J., ed., *International Joint Conference on Artificial Intelligence, IJCAI*, 5103–5111. ijcai.org.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* 31:841.

Wäldchen, S.; MacDonald, J.; Hauch, S.; and Kutyniok, G. 2021. The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res.* 70:351–387.