Bachelor Project - Research description
Version 1

Timo Wahl (s3812030)

February 2021

Problem Statement

The research I will be doing will first and foremost be based on the research done by Bench-Capon (1993). The project essentially consists of two stages: in the first stage I will be replicating the work done by Bench-Capon (1993) to extend my knowledge on the subject and to reaffirm the conclusion of the paper. In the second stage I will be building on his work; extending it with different techniques. Therefore there are multiple research questions for this project, a single research question for the paper by Bench-Capon (1993) and multiple others for directions I could take with my extension of his research (There will be only a single one actually taken). The research questions for my extension are currently not set in stone as it will be quite a long while before I start working on phase two of the project.

The research by Bench-Capon (1993) is essentially about how well a neural network performs when given the task of recovering the rules that define the data set. This research sets out to prove that a neural network can be very good at classifying data points correctly based on a multi-valued input variable variable, but figuring out what that actually means, e.g. which of the variables cause the outcome is an entirely different story.

Theoretical background of the project

As mentioned earlier, my research will be based on that of Bench-Capon (1993). However there are also some other papers that give information on the topic. The paper by Atkinson et al. (2020) gives a short overview on the topic, also citing other papers like Možina et al. (2005) and Wardeh et al. (2009) that tried to replicate the results but with different approaches, inductive logic programming and data mining association respectively. Aside from those there are also other interesting articles: articles on the introduction of LIME and Shap or articles on how reasoning from humans differ from that of machines.

Research questions

Research questions for the replication:

1. Can neural networks perform well in a domain, and if so, did they learn the correct rationale?
2. Can the rationale be improved by adjusting the training data, based on the evaluation of the rationale?

Research questions for the extension:

1. Comparing two neural networks when extracting a rationale from a single data set
2. Evaluating the rationale of a single neural network, when trained on two related data sets
3. Comparing two neural networks when extracting a rationale from a single data set using XAI techniques
4. Possibly more

References

Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in ai and law: Past, present and future. *Artificial Intelligence*, page 103387. [1]

Bench-Capon, T. (1993). Neural networks and open texture. In *Proceedings of the 4th international conference on Artificial intelligence and law*, pages 292–297. [1]

Možina, M., Žabkar, J., Bench-Capon, T., and Bratko, I. (2005). Argument based machine learning applied to law. *Artificial Intelligence and Law*, 13(1):53–73. [1]

Wardeh, M., Bench-Capon, T., and Coenen, F. (2009). Padua: a protocol for argumentation dialogue using association rules. *Artificial Intelligence and Law*, 17(3):183–215. [1]