Review by Floris Rossel

Main Objective

The main objective of this paper is to explore the relationship between Shapley values, used in eXplainable AI (XAI), and causal inference methods based on graphical models. The authors aim to reconcile both approaches by reinterpreting them from a maximum entropy perspective and propose variations of Shapley values that incorporate prior information about the model. This prior knowledge allows the resulting explanations to align better with the underlying causal structure of the data, thereby improving the accuracy of the explanations.

Proposal of the Paper

The authors propose two main variations of Shapley values that incorporate additional information about the underlying causal structure of the data:

1. Markov Blanket Shapley values, which consider orderings where the set of spouses follows the sets of parents and children when computing the weighted Shapley values. This approach incorporates the "explaining away" effect.
2. Ancestor Shapley values, which compute interventional contributions based on a set of candidate graphs representing prior knowledge about the causal structure of the variables.

Evidence Given

The authors provide substantial theoretical evidence to support their proposed variations of Shapley values. They draw parallels between the computation of Shapley values and the Pearl-Verma Theorem from the area of graphical models, which demonstrates the dependence of the resulting graph on the chosen ordering of variables.

The authors introduce theorems that establish how the weights in Weighted Shapley values can be used to promote sparsity, leading to the definitions of Markov Blanket Shapley values and Ancestor Shapley values. They provide theoretical proofs for the properties of these proposed variations. However, it should be noted that the numerical experiments comparing Markov Blanket Shapley values with sparsest Shapley values are only presented in the appendix. While the appendix can offer additional support, the main body of the paper should ideally be self-contained and present the key results.

The paper demonstrates the computation of Ancestor Shapley values in different scenarios, including cases with unmeasured variables.

Related Work

The paper builds upon previous work on Shapley values in XAI (Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014; Datta et al., 2016) and causal inference algorithms like IC, SGS, and PC. It also relates to recent works on incorporating causal knowledge into Shapley values (Janzing et al., 2020; Heskes et al., 2020; Frye et al., 2020a; Frye et al., 2020b).

The authors discuss alternative payoff functions for Shapley values, such as the expected value taken over the interventional distribution using *do*-calculus (Pearl, 2009) proposed by Janzing et al. (2020) and Heskes et al. (2020), which aims to align Shapley values with a causal perspective. They also mention the work of Frye et al. (2020a), who criticize both interventional and observational Shapley values and propose an alternative approach called "Asymmetric Shapley values."

The authors also highlight the Pearl-Verma Theorem (Verma & Pearl, 1990) to derive a graphical structure representing the conditional independence relations.

Reproducibility

1. The paper provides clear descriptions of the proposed algorithms and their mathematical settings.
2. Assumptions are clearly stated and explained.
3. Proofs for theoretical claims are provided in the appendix.
4. While the paper's main focus is theoretical, the lack of concrete examples using specific datasets makes it difficult to assess the practical implications of the proposed Shapley values for explainability. Providing intuitive examples would have strengthened the paper's contributions by demonstrating how these new Shapley values improve explainability in real-world scenarios.
5. Numerical experiments are described in detail.
6. Computing infrastructure used is not specified.
7. The paper does not provide a detailed analysis of the computational time and space complexity of the proposed algorithms. However, it does mention some aspects related to computational complexity.

Feedback

Strengths:

1. The paper introduces novel variations of Shapley values that incorporate prior information and causal knowledge, supported by well-constructed proofs and numerical experiments. It creates a bridge between causal discovery algorithms and explainable artificial intelligence (XAI) using Shapley values.
2. The authors provide a comprehensive discussion of the properties preserved by their proposed Shapley value variations.

Critiques:

1. The paper could benefit from a clearer outline of its structure at the start of the paper, namely that the authors propose two distinct methods for calculating Shapley values: Markov Blanket Shapley values and Ancestor Shapley values. This would provide readers with a clearer understanding of the paper's structure. Currently, it's only stated clearly in the conclusion, and the headings do not provide enough clarity.
2. While the target audience of reviewers is expected to be familiar with the technical concepts surrounding causal inference and the concept of 'maximum entropy', providing clearer explanations and intuitive examples alongside the technical content would enhance the paper's accessibility and clarity. In particular, the authors could better define entropy in the context of their work, justify its maximization, and explain the relevance of the maximum entropy perspective in the context of the proposed Shapley value variations. This would help readers better understand the significance of the work and potentially foster cross-disciplinary collaborations.
3. While the paper provides numerical results from a minimalistic linear SEM, it would be beneficial to include more extensive empirical evaluations on real-world datasets to demonstrate the practical implications of the proposed methods.
4. Although the authors have stated that this work focuses on theoretical aspects, a brief discussion of the computational complexity of the proposed algorithms would provide valuable insights for future research on practical applications. However, it is understandable that addressing all practical considerations may be beyond the scope of this article.

5. (Extending on the previous point) Although the authors did not consider it necessary, exploring the potential applications and societal implications of the proposed methods would be highly valuable, particularly in the context of interpretable machine learning and decision-making systems, given the growing influence of AI on society.
6. Minor: the term ‘model’ is used for two different concepts; (1) the underlying causal structure of the variables, and (2) the model that is investigated with the purpose of interpretability. This can be confusing to a reader not-too-familiar in the field.

Questions

1. Have you considered extending your approach to non-linear models or other types of machine learning algorithms beyond linear SEMs?
2. How sensitive are the proposed methods to inaccuracies in the prior causal knowledge provided as input?
3. Are there other aspects of explainable AI, beyond feature attribution, where prior knowledge of the models could be applied? Do you see potential for extending your work related to causal graphs to other areas of interpretability?
4. Have you considered ways to approach situations where the causal structure is not faithfully represented by a set of Directed Acyclic Graphs (DAGs), e.g., cyclic graphs?
5. How do the proposed methods handle scenarios where the causal structure may change over time?
6. Could you elaborate on the limitations of incorporating causal knowledge into feature attribution methods?
7. Are there any plans for releasing code for the algorithms?

In conclusion, this paper makes a valuable contribution to explainable AI by proposing novel variations of Shapley values that incorporate prior causal knowledge. The authors successfully address the crucial challenge of integrating causal information into feature attribution methods, providing novel solutions backed by rigorous theoretical work. The main strengths include a solid theoretical foundation, well-constructed proofs, and illustrative numerical experiments.

However, the paper could be further strengthened by more extensive empirical evaluations on real-world datasets and a deeper discussion of the algorithms' computational complexity and limitations. Despite these areas for improvement, the paper's contributions are significant, and its technical quality is high.

Considering the paper's strengths and potential impact, I recommend accepting this paper with minor revisions.
Score: 8