Critical Review:

Explainable AI for Classification using Probabilistic Logic Inference

Despite the advancements in AI, the uncertainty and lack of transparency of the models act as an obstacle when implementing these models in the health and medicine sector. Therefore, these “black boxes” are to be replaced with “glass-box” models that are based on explainable AI. Explainable AI leads toward building a trustworthy system as the decision-making of AI models becomes more explainable. In this paper, the author proposes explainable AI for classification, which employs probabilistic logic inference on a knowledge base to explain predictions. Intrinsically interpretable methods and model-agnostic methods are two existing approaches that are discussed. Intrinsic models use the same model for prediction and explanation and thus are easy to understand however are entitled to low performance whereas the agnostic approach is based on different yet accurate models for decision and explanation, yet explanation model is considered untrustworthy in miscellaneous scenarios. Therefore, the authors present a classification approach that produces accurate predictions and explanations and supports domain knowledge incorporation. A knowledge base comprised of a set of disjunction clauses is the essence of this approach. The paper highlights two main approaches to formulate this knowledge base i.e. the tree method and the direct method. However, the obtained knowledge base can be inconsistent resulting in elevated computational complexity. Therefore, a set of constraints are introduced that are solved through linear programming. Furthermore, to reduce complexity, the knowledge base is constructed in relevance to the query fed. Lastly, to further enhance accuracy, decisive features are computed from each query by dividing it into a sub-query and identifying the decisive clauses. The authors also tested the proposed methodology on various datasets and compared it with other existing models and concluded that the approach returns satisfactory performance. For explain-ability, the method was compared with SHAP.

The paper concisely highlights the need for this classification explainable AI model and draws a clear comparison with the already existing approaches. The algorithms to design a knowledge base are the imperative aspects of this paper. The paper is not only well-structured but also has a smooth flow as the authors initially provide relevant background information followed by the constraints and details for each algorithm. Each algorithm is an improvised version of the previous one. Throughout the paper, all necessary assumptions were clearly stated and supported through either mathematical deductions or facts. Furthermore, the stated algorithms were explained using examples and the shortcomings were effectively discussed followed by a proposed solution. To further support the theory, the approach was tested on both synthetic and non-synthetic datasets and several models were tested and the results were effectively compared before drawing any conclusion. The acquired results were compiled adequately and were well presented through visualizations. Besides that, the authors cited several references from authentic and well-grounded researchers and have included both recent and old papers hence making the research more reliable.

The proposed approach provides one with insight into the prediction made. The performance (prediction and explanation) achieved through this approach is satisfactory as seen when it was compared with a state-of-the-art system (SHAP). Moreover, the approach makes its own knowledge base from the training data therefore inconsistencies can be catered effectively. Likewise, domain knowledge can be incorporated (relevant query knowledge base). Furthermore, the approach is non-parametric thus it doesn’t require any tuning. The time complexity is reduced due to the addition of a polynomial-time inference algorithm and the reduction of input size depending on the query. On a whole, the approach is commendable in terms of the accuracy achieved and thus is a massive contribution to the field of explainable AI. This paper has set the base for future studies in the field.

However, explainable AI is quite a complex approach and has high computational cost whereas a large amount of AI problems desire less extensive explanations and higher prediction accuracy. Moreover, the performance deteriorates as the number of features or clauses increases. Likewise, the complexity also elevates and the computational cost will be immense many domains don’t require such costly explain-ability therefore they will let go of this approach and focus more on enhancing the existing machine learning or deep learning models. By increasing the features, the knowledge base would also increase exponentially and might become inconsistent.

Overall, the paper was well-expressed however all the test examples were binary, and even the implemented datasets had two classes hence the complexity it would encompass when the number of classes was to be increased is not featured. Likewise, this paper lacks to highlight the issues and limitations of the proposed approach. Furthermore, the authors were required to explain the mathematical notations in detail keeping. The authors can improve the paper by further explaining the notations used and stating concisely about SHAP before drawing the comparison.

In conclusion, the authors have surely presented a well-structured paper clearly stating the importance of explainable AI, the proposed approach, and the output. However, the above-mentioned concerns regarding the under-emphasis of issues surrounding the approach need to be addressed. Moreover, to pursue future research practical implementation in detail is required and shortcomings need to be dealt effectively.