Critical Review: Explainable AI for Classification using Probabilistic Logic Inference

Explainable AI deals with assigning responsibility to features that are causal to a prediction by a model. Such models, although needed, are prone to many issues including a trade-off between accuracy and interpretability, determining the quality of explanation, and the computational complexity of the problem itself. This paper by *Fan et al.* proposes an explainable AI model that employs probabilistic logic inference to provide a reasonable framework to classify and explain predictions (hence an intrinsically interpretable model). A knowledge base (a set of clauses, each with a measure of probability) is at the crux of this method. The authors propose two different methods to create this knowledge base i.e., the tree method and the direct method. An Algorithm to query this knowledge base (in order to *classify* predictions) involves an originally computationally hard problem relaxed to a system of constraints and objectives that could be solved through linear programming in polynomial time. Furthermore, to reduce time complexity, the knowledge base fed to the linear solver is reduced in size depending on the query being solved. This is done by extracting clauses relevant to the query and discarding others before setting up constraints in the linear program. Finally, for each query, one can iterate sub-queries over the program to select those that yield probability closest to the ground truth (0 for negative and 1 for positive). This way, one can identify which path/clause produces the most *decisive* result, concluding that features in that clause are the *decisive* features (that *explain* the prediction of the said query). The authors test their proposed method for both prediction performance and explain-ability. The direct method has satisfactory performance on a variety of datasets. For explain-ability, this method is compared with SHAP revealing promising results.

This paper concisely reiterates the need for explainable models in AI, before proposing a method that caters to the shortcomings of previous relevant work while also adding efficiency, with presumably little cost to accuracy. Each algorithm introduced is preceded by either a parent algorithm or relevant background information that allows for a smoother flow. Intuition is clearly stated where required, and deductions taken from that intuition are properly and impartially reasoned. Furthermore, the paper also employs continuous-themed examples that procedurally demonstrate the proposed method on a scale suitable for written work without being trivial. In addition to examples, the theory is adequately supported by comparative tests with well-known algorithms on well-known datasets. Where needed (e.g., owing to the lack of the concept of *ground truth* in explain-ability), datasets were augmented with synthetic data that provided a more analytical approach to address issues out of the scope of this paper. Results and visualizations were effectively compiled, and prevalent issues combined into concrete future paths for research.

The proposed method has the advantage of providing explanations comparable to state-of-the-art systems (such as SHAP), without major loss to performance (as verified by extensive tests on commonly known datasets). The method is also non-parametric, thereby not needing any form tuning (a computationally expensive process). Moreover, the proposed querying algorithm caters for inconsistencies in the knowledge base with respect to specific queries. Although, such inconsistencies are the product of the proposed method of knowledge base construction, a way to tolerate them allows for a more robust model. Furthermore, since the knowledge base is just a set of clauses with probabilities, one can essentially incorporate relevant domain (professional) knowledge in the form of intuitive clauses to the knowledge base. Since the proposed knowledge base is both independent (from the rest of the system, much like a database) and generic (clauses are general and easy to understand), it allows an effective and efficient way to add insights from a variety of sources ─ an integral concept in various industries.

While this paper reasonably advocates for explainable AI, it does so in a more universal manner. A sizeable amount of problems in AI do not require extensive explanations, especially when these explanations come at the cost of performance. In light of that, the concern of explain-ability is as

important as the concern for accuracy, only for some problems (such as the medical diagnosis example provided in the paper). Given that, it may be worthwhile to note that since the requirement of explain-ability is specific, explainable models should benefit from intuitions specific to those problems, instead of a general algorithm as proposed in this paper.

Furthermore, the proposed method is not scalable, in the sense that it cannot cope with developments in machine learning (especially deep learning, transfer learning, and even online learning). This again poses the question of the trade-off between explain-ability and performance. For this specific method, the trade-off seems to be acceptable. However, it may be noted that with an increasing number of features and data points, the computational complexity of this proposed method overwhelms its chances of explaining the model. Furthermore, almost all features employed in this paper were binary; and while the mathematics is satisfied for categorical variables having any number of values, the computational complexity that would entail is not addressed. With an increasing number of features and an increasing number of categories they can hold; the knowledge base exponentially increases in size. This is concerning, as this greatly increases the complexity of the linear program. In essence, large models and/or large datasets can make the method practically infeasible without extensive hardware.

Moreover, the tests to verify explain-ability are creative but nevertheless insufficient. Firstly, the comparison is with a state-of-the-art method, which is known to have its own issues, and even without them, is not perfect. To counter that, synthetic data was used, and the results were decent. The data was also successfully tested for performance to prove that it was not a trivial classification problem. However, it is important to note that even though it may not be a trivial *classification* problem, it may be an *easy* if not a trivial *explanation* problem. One may successfully argue that classification and explanation in an inherently interpretable explainable AI model are somehow related, but this would mean that the trade-off to accuracy may be detrimental to the overall goal to understand *black-box* models. Furthermore, performance and explain-ability are judged separately, and results are inconclusive for that reason. This justifies the inclusion of a new performance metric (similar to the F1 score) that concurrently judges the performance and interpretability of the model. However, this is out of the scope of this paper (and also with its own set of issues), and therefore, just speculation for future research.

To conclude, it may be noted that the above-mentioned concerns were inevitable considering the novelty of the research area and the extent of work carried out. Overall, the authors have presented an efficiently structured paper highlighting the importance of explainable AI, proposing an improved model, and verifying it. Future research direction specified is natural, with proper consideration given to practicality of the proposed method and improving interpretability metrics (e.g., user studies and knowledge incorporation).