Critical Review – 1

by

Fizza Tauqeer

A research study titled “Explainable AI for Classification using Probabilistic Logic Inference” headed by Fan et al. aims to further bridge the gap between calculated insights gained from artificially engineered systems and the extended set of rationale behind these insights that enables them to augment the problem solving such systems intend to tackle. The study builds upon the workings and shortcomings of methodologies of previous notable researches, such as SHapley additive exPlanations (SHAP) which is considered the industry standard for descriptive learning provision, to introduce a new implementation that provides an equally sophisticated set of explanations but with better performance via probabilistic predicate knowledge bases aided by linear programming in order to avoid NP-Hard time complexities. The initial sub-algorithms generate clauses using the paths of a Decision Tree, from which feature-value pairs are extracted while retaining information on the positive class balance. These clauses populate an initial knowledge base. The intermediate sub-algorithms concentrate on generating an extensive separate knowledge base based on direct exhaustive key (combination of feature-value syntactic tags)-label pairs instead, which yield almost the same explanatory sets. The last leg of sub-algorithms concentrates on exclusively consuming a linear programming-based weighting objective function. This enables querying from the earlier computed knowledge base, downsizing it to a relevant domain in polynomial time consideration and then calculating explanations based on the ‘less than’ or ‘greater than’ probability equivalent measure value of 0.5. This research was tested on standardized datasets – such as the UCI-famous Titanic and UK Parliament Bill amongst others – with comparison to benchmark implementations, such as CART and SVM. The results of this study indicated that it was able to recreate and even extend the performance metrics prevalent to these standard algorithms and specified datasets, thus instantiating a new and effective way to aid the growing need of understandable artificial intelligence.

When reviewing it critically, the first point of contention seen in this paper is that the results the implementation provides is deemed as similar to that of SHAP – the authors lag in justifying then why their work was necessary enough to be carried out when their reproduction has no novel value over existing methods (most results only approach the findings of previous methods). Though relying on multiple researches of similar notions is preferred when exploring potential fields, each research is expected to individually offer some establishment of novelty – be it in terms of better resources or time complexities, as a basic example – unless it is merely a review paper. The paper in question does not offer any such advances or self-categorization in this aspect. The claim of producing similar results to SHAP is also not as practically justified, as SHAP is applicable to any predictive algorithm while the author implementation solely relies on

decision tree and probabilistic methodology for reverse-engineering the AI explanation. However, it is seen later on that there are indeed some improvements over existing methodology, such as the Non-Linear Probabilistic Logic Solver (NILS) approach.

Additionally, though the paper excels in proving the problem as an exponentially increasing one with stellar baseline explanation, it certainly can incorporate more publicly understood terms (for example, there is no explanation in the paper of why the same representation can be presented alternately as 'disjunction clauses' in the knowledge base construction – a reader unacquainted with predicate logic may well be unable to understand what the term entitles). There also appears to be an over-reliance on explaining underlying pseudocodes and their associated examples via set theory, which does undermine the overall understanding that the research could have better provided. The study also claims that their approaches generalize to multi-class problems as well; however no such experimentation is done to verify this. The study utilized the 'Nursery' dataset which happened to host itself as a multi-class problem, providing the researchers with the perfect opportunity to test their generalization. Instead, the authors binned the classes into two while discarding the rest, thus keeping the experimentation restricted to binary classification as a whole. Moreover, a large part of the paper's understanding is built upon affirmation-negation theory, yet they do not assess the question of whether explainable AI is as required as it is implying – had it been a more exponentially growing field, many such implementations would also have been in the research pipeline. Exploratory Data Analysis (EDA), as a simplistic example, directs the end consumer to visually assess how certain attributes can better influence the predictor variable with the added benefit of easier explanation and information capture, a premise on which this paper relies on. Exploring why these nuances were not deliberated on would have added an additional touch of comparative realism to this study.

At best, the implementation – though unique – is able to yield nearly the same amount of performance or benchmark result metrics as previous systems or standardized state of the art modulations, with an improvement over only a minor set of systems. The only upper-hand that the research – in terms of results – seems to employ is when utilizing synthetic datasets for experimentation, which could have been unconsciously modelled to solve the problem, thus cannot be considered as reliable or beneficial as utilizing randomly produced real-life datasets or standardized datasets, in which its performance is approachable at most.

Moreover, the datasets utilized do not house "hundreds of variables," which was the premise explained in the study for exploiting the basis of utilizing linear programming in order to curb the implementation to a polynomial time run. The experimental results thus are not as dependable – as one would think – in justifying their methodology. This could very well be the reason why the findings seem to approach SHAP results, as they did not exactly face issues from larger scales of data to truly deploy the benefits or gains of the proposed implementation. The trade-offs between linear programming and more traditional probabilistic methods are, hence, never fully explored – it quickly becomes clear that as the clause set may tend to increase, the

objective function will begin to exponentially decay thus pulling back the implementation on the complexity-saving premise it had built for algorithm justification.

There is also no explicit tradeoff mentioned in the algorithmic implementations, even when there is clarity over which implementation was more time complex (such as building the knowledge base directly). Hence, no motivating factor is specified to reason with as to why the implementation keeps building knowledge bases via different approaches in the same study alone. Furthermore, the authors claim their approach is fundamentally different from Decision Trees, which it happens to model quite significantly in terms of both implementation and results. If a Breadth First Search (BFS) traversal is instead added to the initial knowledge bases on each node split, probabilities from other clauses are also considered via averaging.

Neglecting these severe critiques, it is evident that this is a well-researched and equally well-curated article. An exponential objective that is achieved by this research is that their querying system can somewhat tackle missing values, which is another gradually growing field of interest in the goal to minimize the negative effects of having to rely on flawed data, while furthermore acknowledging the fact that such inconsistency is not completely handled by its implementation. The grey area where the research seems to lag is addressed rather than letting the reader discover it; hence the authors are cognizant of the gaps in their work which is a good research habit to showcase. Another strong feature of this study is the due reasoning of each sub-algorithm by way of examples and justifications. Algorithm 5, concerned with generating only relevant knowledge bases, happens to be worthwhile in generating a less exhaustive yet equally effective knowledge base. Data curation to tackle ground truth for explanations was also carried out, which is a time-consuming task in itself. Moreover, the related work is not merely a traditional retelling of the literature representing similar types of solutions but rather an excellent comparative analysis to their techniques as a whole.

This paper happens to resolve the classification ambiguity and numerical complexity that the NILS implementation incurs, by restricting time complexity using inequality equations and handling inconsistencies by way of alternating when facing missing values in the algorithmic tree sub-queries. A futuristic aspect of this research's outline is that it does not rely on older legacy implementations but does sketch out credible works in case the reader intends to explore the relative regions of logic and probability more so, which can be appreciated. Lags in previous literature are also fully examined in order to provide the intrinsic sense of motivation that may have propelled this work. All implemented viewpoints of explainable AI – from the secular to the expansive methods – are revisited such that the growing need of such methods and variations of each such implementation (since the idea's inception) can thoroughly reinstate the requirement of such systems to be continuously set up and improvised frequently.

This paper, thus, can serve as an excellent starting point for baseline execution using traditional machine learning approaches – such as probabilistic procedures – with the certified expectation of receiving promising results.