**Explainable AI for Classification using Probabilistic Logic Inference**

**Paper Review**

**Summary:**

The authors work on producing models that explains ‘why’ such prediction is made as made by the model. This is brief description of explainable AI. Two previous approaches include interpretable models that explains its prediction itself and the other one is model-agnostic in which a wrapped process outside the model is used to explain results. The authors worked on the first approach. So, the algorithm has to classify as well as tell the features that caused it to make such prediction.

The author’s approach is to create a probabilistic knowledge base either using decision trees or directly from data. However, the authors point out that knowledge base may be inconsistent when a query is made. They solved it as optimization problem so the inconsistencies can be tolerated. They also point out this can be done using linear programming. The approach also allows knowledge incorporation by adding clauses that you know matters in predicting the right output.

For testing, the authors used 4 other classifiers to compare results named CART (a decision tree algorithm), multi-layer perceptron (MLP) neural networks (with two hidden layers with 12 and 10 nodes, respectively), random forest (with 100 trees) and support vector machine. The models applied to 6 non-synthetic and 4 synthetic datasets and f1 scores were calculated. It seemed direct KB approach scored similar to random forest which scored very well in all of the datasets. They also compared the results to state of the art explainable AI named SHAP. The authors pointed out that their approach’s accuracy of pointing out right features on synthetic datasets was higher than SHAP which is quite commendable.

In the end there is a discussion of future research that can be done on this promising approach. The topics include development of underlying classification technique, semantics for inconsistent KB, explanation of logic inference and other suitable representations for knowledge incorporation.

**Critical Analysis:**

The paper can be seen as progress in the field of explainable AI. The strength of this paper surely is comparison of the proposed approach with the existing ones. The writers used 4 other mainstream classifiers from both machine learning and deep learning and compared the results on 6 different datasets that includes binary classification problem dataset such as titanic, multi class problem such as nursery and image dataset called vehicle. Not only this they created 4 more synthetic datasets. The idea of these dataset is very well thought and can be used by others while developing future explainable AI models. The datasets include a string of n length and query has to have 5 similar characters to be considered positive so characters act as features so dataset help us finding in whether the model explaining the right features. The comparison with SHAP approach which is seen as state of the art in the field of explainable AI. The approach is shown to be superior as it achieves higher accuracy on the synthetic datasets. As an intrinsic interpretable approach this algorithm has great accuracy as well as answering the question of understanding the results. Also, this approach being non-parametric, there is no need of extensive training is required. So, this paper seems to contribute a lot in the field of explainable AI in the future.

As all research papers has some issues this research paper although great has some issues. The biggest one being not comparing the computational cost of the algorithms. The authors have not even discussed in the paper while this may be a big problem to implement this algorithm practically. If this takes too long to process compared to simple models like random forest or SVM than this is practically of no use even though linear programming is mentioned in the paper and running time can another thing that can be added to this paper as other models only take once to train while in this approach relevant KB is created for each query. Another domain where more insight was required is the ability of the algorithm to work on large datasets with loads of features then knowledge bases may become too large for computational resources to hold and this method may not stay effective. Also, the topic of knowledge incorporation is not sufficiently touched as it was not explained in what representation and in what meaning we can use the domain knowledge in our model.

**Conclusion:**

Overall, this is a great research paper with the right flow of explanation of the algorithm and great comparisons. The mathematical proofs highlighted in the paper are great too which means a great deal of work is done in research and the quality of research is also good as previous papers that included inference logic are also referred with around 40+ links given in the end. The pseudo code can be written in a manner that is more self-explanatory but with enough mathematical knowledge it is understandable. I think with the results it produced in comparison to SHAP, it has shown that this can spearhead the field of explainable AI in the future.