Review of Master Thesis:
Explainability of malware classsifiers

Thesis author: Bc. Ondřej Vereš
Reviewer: RNDr. Petr Somol, Ph.D.
21 June 2024

Explainability is one of the crucial yet often neglected problems in Machine Learning. The need for explainability comes from the fact that complex learned models can be perceived as black boxes. The lack of understanding of where do verdicts come from hinders acceptance of Machine Learning in many areas of human endavour where humans so far keep responsibility of decision making. The problem is perceived also by model creators – it is often hard to get sufficient certainty that a model indeed performs correctly, even if statistical measures suggest no problem is present. I myself have seen overfitted models seemingly performing well, where the problem got only revealed by explaining a sample of verdicts.

Explainability as a field of study can be still considered underdeveloped, although decent methods are available at least for models build on top of vector spaces (LIME, Shapley values, etc). The situation is, however, even less satisfactory in the special class of learning problems from structural multi-type data. Yet, learning from structural data is at the core of many fields involving machine-generated data, with cybersecurity being one of the prominent ones.

The author took the goal of developing a new explanation method, that should ideally bring better (or at least comparable) results to prior art, while bringing potentially other advantages, e.g., in terms of speed, or compactness of explanations.

Chapter 1 gives good overview of explainability as research area. Chapter 2 summarizes reasonably well the concept of HMIL model which is particularly suitable for learning from tree structured data. Chapter 3 gives reasonably good summary of the existing HMIL explanation framework provided by HMIL authors in ExplainMill.jl. Here the text is harder to follow and would benefit from inclusion of a simplified summary of the method as a whole, perhaps illustrated visually. At any rate, this chapter is hard to read if the reader does not have prior understanding of the context. Chapter 4 gives an overview over some of the important existing explanation methods in general, not restricted to those applicable to structural data only. Chapter 5 introduces the author's contribution, the method TreeLIME and its variants. The chapter is reasonably well written and contains a lot of technical detail, studied in depth. Chapter 6 covers the performed experiments and evaluations in much detail and with scientific rigor. Chapter 7 correctly discussess the observed limits of TreeLIME performance. Chapter 8 concludes the thesis and outlines the next steps. It correctly focuses on the need of a new masking system.

The contributions of the thesis did not result in a method notably better than prior art. Nevertheless, the thesis succeeds in uncovering in more depth some of the principal difficulties of explaining structural data and pointed at their sources. The thesis overall presents sound scientific non-trivial results and I consider it a success for that reason. That said, it is not without issues.

- In section 2.5 HMIL is put into perspective of other options that practitioners in cybersecurity have for learning from tree structural data. I believe Graph Neural Networks should be covered here as well, especilly given the later reference to GNN specific explanation in section 4.5
- chapter 4 feels imbalanced. I agree with LIME getting the most detailed treatment in section 4.1 as it is intended as the basis for author's own work later in the thesis. However, sections 4.2 and 4.3 are so short that the respective methods remain mostly unclear to the reader. The thing I miss the most is a critical comparison of the various listed methods.
- The biggest drawback in the thesis is the lack of evaluation using stability, robustness and other criteria, although their existence is mentioned. The thesis focuses entirely on the size of explanation as the

criterion of success. I understand that the amount of work kept growing as the author progressed and that eventually it was found impossible to cover more than size criterion within given deadlines. In that case I recommend to take it as a lesson for the future. The key problem I see here is the fact, that size in itself does not give strong guarantee that the result is good as such. It is just assumed, that shorter explanation is more convenient for humans to verify. (That assumption might have been mentioned)

- Some of the fundamental questions that do affect the results would deserve more detailed discussion. Case in point: confidence gap threshold seems to be a constant to be set by the user. It is clear however that wrong setting can cause the method to fail. What is the optimal setting, or can the optimum be at least approximated in some way?
- The text would be easier to read if all symbols would be introduced or commented on the context where they appear with important role. E.g. in formula (1) I have to guess what y_i means, and that β are model parameters. Overall the text makes sense but does not make it easy for the reader to follow quick.
- The term "relative tolerance" is introduced in page 32, way after it has been used in text, e.g. in pages 19, 20, 21.
- Section 5.6 title "TreeLIME visualization" is a bit confusing, but one quickly realizes that you mean visualization of the process of computing the explanation (not the explanation itself)
- The fundamental question of how correct the explanation really is, is not addressed. I admit it is hard and most methods get evaluated though various proxy criteria only in literature. However, for practical use, the notion of correctness needs to be at least discussed. I do not have a good definition myself, but I do see at least two related questions: is it possible to evaluate how well the set of explanations really corresponds to the original model accuracy on the respective samples? Can there be additional criteria that a good explanation should fulfill, like a semantic meaning? This is often the case in cybersecurity, where human analysts tend to reject explanations that lack interpretability within constraints given, e.g., by correctness of executable code (in case of malware detection)

For my interest I would also ask questions:

- In TreeLIME, it is assumed that perturbations can be generated from distribution, with constant, uniform and normal distributions tested. Is there any reasoning why a particular distribution should be better? Would the same reasoning apply to LIME as well?
- in section 5.5 you interpret the surrogate model: if beta coefficients >0, the respective predictor is put to explanation. For beta equal 0 the predictor is ignored. For negative beta you claim that "In that case, we do not want to put the predictor with index into the explanation, as it is important according to the model"; this claim is confusing and I am not sure if it is a typo. Anyway, my question is, are there other meaningful interpretations of the relations between surrogate model and explanation possible here?
- I fully agree that the disconnect between individual layers processing in layered TreeLIME is likely the reason of its results being lesser than hoped for (section 7.2.2). Would not there be some straightforward solution like releasing the requirement to stay above the threshold in each layer, and keep the threshold requirement only for the last layer? Under which condition could that work?
- Some of the problems that you describe for flat and layered TreeLIME, namely dependencies among predictors and problems with optimization path, are actually not the problems of the proposed method but are problems of explaining tree-structured data in general. It would be good, at leart in the future work, to address the problem of masking in general; this would lead to improvements in both TreeLIME and the prior art HMIL Explainer.

Conclusion

The author shows great promise to continue improving as a talented scientist. In next iterations of author's publishing efforts I recommend to continue the good work in terms of formally correct handling of problems and verification of results, but I do recommend to pay more attention to the higher-level meaning and implications of particular solutions to a problem, especially if the problem solved can have practical implications. Do question whether the selected criteria of success are sufficient to make actual impact. Do question which details of a solution are more important than others, and make sure not to spend disproportional time with the less important while omitting the more important ones. That said, I am sure we can expect inspired research from the author in the future and I encourage the author to continue with a scientific career.

I recommend to accept this thesis as a Master thesis and the author to obtain the respective engineering title. Having hasitated a bit between suggesting mark A or B (due to the unfinished evaluation of criteria beyond explanation size), **I finally lean towards A** to praise the exceptional formal qualities of the core body of the text.

RNDr. Petr Somol, Ph.D.
petr.somol@gmail.com (preferred), tel 603 719 429

Research Fellow
Institute of Information Theory and Automation
Czech Academy of Sciences
Pod vodárenskou věží 4
18208 Praha 8
somol@utia.cas.cz

AI Research Director
Gen Digital (former Avast Software)
Pikrtova 1737/1A
140 00 Praha 4-Nusle

petr.somol@gendigital.com