A Trustworthy View on XAI Method Evaluation - Appendix

Ding Li, Jun Huang, Zerui Wang, Yan Liu
Department of Electrical and Computer Engineering, Concordia University, Montreal,CA

■ **ABSTRACT** As the demand grows to develop end-user trust in AI models, practitioners start to build and configure customized XAI (Explainable Artificial Intelligence) methods. The challenge is lack of systematic evaluation of newly proposed XAI method. As a result, it limits the confidence of XAI explanation in practice. In this paper, we follow a process of XAI method development and define two metrics in terms of consistency and efficiency in guiding the evaluation of trustworthy explanation. We demonstrate the development of a new XAI method in feature interactions called Mean-Centroid Preddiff, which analyzes and explains the feature importance order by a clustering algorithm. Following the process, we perform cross-validation on Mean-Centroid Preddiff with existing XAI methods and show comparable consistency and gain in the computation efficiency. The practice helps to adopt the core activities in the trustworthy evaluation of a new XAI method with rigorous cross-validation on consistency and efficiency.

Appendix

Case study: Code Vulnerability Detection

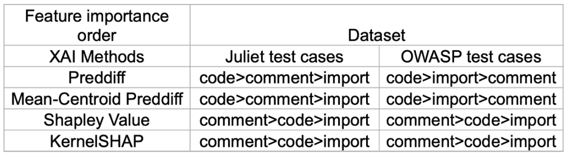
A case study examines a code vulnerability detection problem that leverage NLP models to identity the Common Weakness Enumeration (CWE) in software repositories. We identify three features in the code file that comment, package importing and the code. The goal is examining if the code itself affects most on the model decision to identify the CWEs. The consistency and efficiency of proposed XAI method also be evaluated comparing with state-of-the-art XAI methods.

Code Vulnerability Dataset The datasets are from The Open Web Application Security Project (OWASP) Benchmark [1] and Juliet test suite [2] for Java from The National Institute of Standards and Technology (NIST)'s Software Assurance Reference Dataset. The OWASP Benchmark test suite contains 2740 test cases with 52% files of vulnerable code and 48% non-vulnerable files. The 52% vulnerable files contain 11 CWE (Common Weakness Enumeration) labels. The Juliet test suite contains 217 vulnerable files and 297 non-vulnerable files that 112 different CWE labels inside. The CWE labels are the ground-truth for multi-classification. We deploy the XLNet model [3] to output the log-odd probability of CWE label.

The XLNet Model The Natural Language Processing (NLP) model XLNet [3] is levered to classification the code to the particle CWEs labels. It is a model that could capture bi-directional text information and reach to state-of-the-art. The code file is considered as the text-based data to feed in the XLNet model to generate the target CWE label.

Experiment Design The objective of code vulnerability detection case study is examining the state-of-the-art XAI methods' consistency and efficiency in Juliet and OWASP datasets. Juliet and OWASP test suite are treated as two sub-datasets to perform the experiments. Code files are identified as three features, comments, package importing and code. The features are removed before input the state-of-the-art XLNet model [3] to output the log-odd probability of the ground truth CWE label. Totally two groups of prediction log-odd possibility are output for each XAI method. Finally, two groups of feature importance order for three features are derived.

Experiment Results We summarize in Figure 1 observe that Preddiff and Mean-Centroid Preddiff derive the same feature importance order, code is the most importance. The order of comment and import varies from datasets. Shapley Value and KernelSHAP have a consistent result but value comment more than code and the import.



Feature importance order	Dataset	
XAI Methods	Juliet test cases	OWASP test cases
Preddiff	code>comment>import	code>import>comment
Mean-Centroid Preddiff	code>comment>import	code>import>comment
Shapley Value	comment>code>import	comment>code>import
KernelSHAP	comment>code>import	comment>code>import

Figure 1: Feature importance order summary of code vulnerability detection case study

Table 1: Time consumption (Seconds) of XAI methods for Juliet and OWASP dataset in case study 2

XAI Methods	Juliet	OWASP
Preddiff	6.335	18.161
Mean-Centroid Preddiff	6.430	18.624
Shapley Value	18.072	56.285
kernelSHAP	123.694	313.414

The time consumption examination in case study 2 keeps insistent that unchanging trend with study case 1 according to Table 1. KernelSHAP is significant longer consuming than others and Mean-Centroid Preddiff achieve the moderate results between Preddiff and Shapley Value.

REFERENCES

1. T. I. V. OWASP, "Url https://www. owasp. org/index. php," *Top IoT Vulnerabilities*, 2016.
2. P. E. Black and P. E. Black, *Juliet 1.3 test suite: Changes from 1.2.* US Department of Commerce, National Institute of Standards and Technology, 2018.
3. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.