Ontological Representation of the EMBER Dataset

Peter Švec[1], Štefan Balogh[1], Martin Homola[2], and Ján Kľuka[2]

[1] Institute of Computer Science and Mathematics, Faculty of Electrical Engineering and Information Technology Slovak University of Technology, Ilkovičova 3, 81219 Bratislava, Slovakia

[2] Comenius University in Bratislava, Mlynská dolina, 84248 Bratislava, Slovakia

Malware detection is an important problem in information security. Historically, large number of diverse methods have been applied on this problem, including some AI methods such as machine learning [3]. To facilitate research in this area there are several publicly available datasets in this domain such as EMBER [1] and SOREL [2].

In order to apply symbolic AI tools on these data, a suitable ontological representation is required. For instance, Švec et al. [4] were able to to obtain malware characterizations in form of structured concept descriptions, based on an ontology.

We present an updated version of EMBER ontology previously developed by Švec et al. [4]. We expect this updated version to improve the concept learning results. At the same time the ontology was reconstructed using current ontology engineering guidelines and we hope it will be more universal and reusable also by other symbolic or neural-symbolic methods, or any application that needs to process Windows malware related data. The ontology is compatible with both EMBER and SOREL datasets.

Acknowledgents. This work was partially supported by projects ORBIS, funded by Slovak SRDA agency under contract No APVV-19-0220, and by TAILOR, funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Anderson, H.S., Roth, P.: EMBER: An open dataset for training static PE malware machine learning models. arXiv preprint 1804.04637 (2018). https://doi.org/10.48550/ARXIV.1804.04637
2. Harang, R., Rudd, E.M.: SOREL-20M: A large scale benchmark dataset for malicious PE detection. arXiv preprint 2012.07634 (2020). https://doi.org/10.48550/ARXIV.2012.07634
3. Husák, M., Komárková, J., Bou-Harb, E., Celeda, P.: Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun. Surv. Tutorials **21**(1), 640–660 (2019)

4. Švec, P., Balogh, S., Homola, M.: Experimental evaluation of description logic concept learning algorithms for static malware detection. In: Proceedings of the 7th International Conference on Information Systems Security and Privacy, ICISSP 2021, February 11-13. pp. 792–799. SCITEPRESS (2021)