Knowledge Discovery in Malware Datasets using Formal Concept Analysis

Ángel Mora Bonilla, Domingo López-Rodríguez, Manuel Enciso and Pablo Cordero
Universidad de Málaga, Bulevar Louis Pasteur 35, 29071 Málaga (Spain)
e-mail: {dominlopez,pcordero,enciso,amora}@uma.es

Abstract: Intelligent malware detection [4] is a problem that is generating growing interest in the industry due to the increase in the diversity of threats and attacks suffered by small users to large organisations or governments, in many cases compromising sensitive information and without ruling out possible economic consequences.

Among the different problems that arise in this area, the homogenisation of the nomenclature of malware threats [5] stands out, as different antivirus engines or applications often use different names for the same threat or the same family of threats, which is related to the problem of malware family classification [7].

Another big open problem in this field is the definition of methodologies that allow optimising the detection process itself of new threats, since the different engines have different detection capabilities and no single software can detect all the threats at one point, thus there is a need of determining which combination or possible combinations of engines cover the majority of detection and which features present in malicious software allow us to detect it at an early stage [2,1].

In this paper, we propose the use of formal concept analysis (FCA) [3] to exploit the existing knowledge in previous threat and malware databases by different detection engines. In this formal framework, based on lattice theory and logic, we can build a lattice where threat sets are organised hierarchically according to specialisation-generalisation criteria, which provides us with a direct approach to setting up a unified taxonomy of malware.

On the other hand, the use of FCA itself enables the discovery of logical rules and the application of automated reasoning methods [6] whose objective is to simplify the detection process without losing information or threat detection capacity and even increasing this capacity.

In this sense, our proposal differs from previous [4] ones in that it does not use statistical criteria, but rather an exhaustive analysis and mathematical modelling of the knowledge contained in malware databases, so that the models obtained are based on logical and algebraic tools and offer a greater degree of interpretability and explainability than previous proposals.

Keywords: Logic programming · immediate consequence operator · generalized quantifiers

Acknowledgement: Partially supported by the Spanish Ministry of Science, Innovation, and Universities (MCIU), State Agency of Research (AEI), Junta de Andalucía (JA), Universidad de Málaga (UMA) and European Regional Development Fund (FEDER) through the projects PGC2018-095869-B-I00 and TIN2017-89023-P (MCIU/AEI/FEDER), and UMA2018-FEDERJA-001 (JA/UMA/FEDER).

References

1. Escudero Garcia, D., DeCastro-Garcia, N.: Optimal feature configuration for dynamic malware detection. Computers & Security **105**, 102250 (2021)
2. Firdaus, A., Anuar, N.B., Karim, A., Razak, M.F.A.: Discovering optimal features using static analysis and a genetic search based method for android malware detection. Frontiers of Information Technology & Electronic Engineering **19**(6), 712–736 (2018)
3. Ganter, B., Obiedkov, S.: Conceptual Exploration. Conceptual Exploration pp. 1–315 (2016)
4. Kouliaridis, V., Kambourakis, G.: A comprehensive survey on machine learning techniques for android malware detection. Information **12**(5) (2021)
5. Maggi, F., Bellini, A., Salvaneschi, G., Zanero, S.: Finding non-trivial malware naming inconsistencies. In: International Conference on Information Systems Security. pp. 144–159. Springer (2011)
6. Mora, A., Cordero, P., Enciso, M., Fortes, I., Aguilera, G.: Closure via functional dependence simplification. International Journal of Computer Mathematics **89**(4), 510–526 (2012)
7. Walker, A., Shukla, R.M., Das, T., Sengupta, S.: Ohana means family: Malware family classification using extreme learning machines. In: 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC). pp. 534–542 (2022)