Machine Learning Classifiers in Honeynets - A Critical Review

Ben Baker[1], Kelsey Quinn[1], Jason M. Pittman[2]

[1] *Student, Dept. of Computer Science, High Point University, High Point NC USA*
[2] *Associate Professor, Dept. of Computer Science, High Point University, High Point NC USA*

---***---

Abstract - *In this paper, we present a critical review of machine learning classifiers applied to honeynet logs. A honeypot is a deceptive technology used to study adversary behavior during computer attacks. Further, a honeynet is a collection of honeypots used to increase the scope of adversary observability. In both incarnations, a fundamental problem has been the efficient and effective analysis of collected data. Contemporary research has focused on applying machine learning to these datasets. Thus, the timing is advantageous for comparative, replication, and reproduction of such literature. This work provides a serious examination of the replicability of seminal research in this topic. To that end, the review revealed replicability issues in the structure and results of two primary sources. We then offer recommendations and ideas for future work based on the findings and state of the field.*

Key Words: **Honeypots, Honeynet, Machine Learning, Replication, Critical Review, Classification Algorithms**

1.INTRODUCTION

1.1 Honeypots

Honeypots are systems whose purpose is to be attacked, probed, or compromised in order to collect data on blackhat activity [1]. By allowing the system to be used in this manner, researchers and practitioners can identify trends and behaviors [1][2]. This is of course a broad description of how a honeypot works and the intent of deploying a honeypot as a deceptive technology.

More specifically, honeypots can be categorized according to type, deployment form, or more commonly by interaction level [2]. Furthermore, a honeypot can be deployed in three different deployment modes: deception, intimidation, and reconnaissance [2][3][4]. The deception mode tries to lead the hacker to believe that the responses they are receiving are from a real system. If deception mode is in place, the system ensures that the hacker is aware of the measures taken to secure the system. While in reconnaissance mode, the honeypot determines the tools and techniques used by the intruder [2].

Research suggests that honeypots are primarily used to study malicious behavior [4]. To a lesser degree, honeypots can be used as a deceptive technology in enterprise environments [3]. Regardless, two oft-repeated problems with honeypots is that the systems receive little traffic and are easy to detect [1][2]. As a result, the *honeynet* was proposed.

1.2 Honeynets

A honeynet is a collection of honeypots [1][3] with the purpose of improving deception and increasing captured traffic yields [5]. Broadly speaking, there are two types of honeynets: generational and virtual [4][6]. Like honeypots, the types are defined by variance in how the technology functions in relation to its purpose as a deceptive technology.

Honeynets understandably produce large amounts of data based on the multiplicative number of interactions with attackers. While addressing the shortcomings of honeypots, honeynets introduce new challenges insofar as the large volumes of data can be difficult to sort through using traditional means [7]. Thus, the field has turned to machine learning for honeypots and honeynets.

1.3 Machine Learning in Honeynets

Machine learning represents a broad category of algorithms and the implementation of such technologies follows suite. In the context of honeypots and honeynets, the field tends to envision two applications of machine learning: one application creates dynamic systems and the other enhances post-mortem data analysis [8]. According to existing research [9][10], dynamic honeypots utilize machine learning to adapt to changing operational environments. This includes removal or additions of services in adjacent systems, deviations in network traffic, and so forth. Here, reinforcement learning and regression algorithms are popular. The benefit is derived from a honeypot or honeynet which ought to be more deceptive and thus less detectable.

Alternatively, machine learning is useful for rendering data analysis of honeynet data more efficient and potentially more accurate [11]. This is largely the province of classification algorithms since the intended output is a match or no match type decision between honeynet data known to contain malicious data and another data source (e.g. a production server). The benefit here is derived from being able to identify (classify) malicious behavior amongst non-malicious behavior [12]. However, there has not been much research to validate which specific machine learning classification algorithms generate the least false positive and false negatives.

2. METHODS

Accordingly, the original purpose of this study was to replicate the comparative analysis of Naive Bayes, Support Vector, and Random Forest classification algorithms as such were used to analyze honeynet log data [12]. However, in the process of establishing a replication environment, we discovered fundamental information necessary to empower replication was not present. Davison [13], alluded to this general problem insofar as replicability is difficult to achieve because there is not full transparency in research. Accordingly, we pivoted from a replication study to a critical review to fill in the gaps by drawing from related literature.

In order to achieve this aim, we employed a standard critical review method [14]. This was an appropriate selection based on the goal of assessing the literature through an analytical lens. Further, the chief divergence from other types of reviews is that critical reviews intend to produce insight into where future concepts may exist [14]. The intent of this work then was to illuminate areas in the target study [12] where information necessary for replication was absent or incomplete and provide ideas for how to move forward.

2.1 Literature Search

There were two phases to our literature search. Both phases leveraged common academic research databases through standard indexes such as Google Scholar, EBSCO, as well as professional societies such as ACM and IEEE.

The first search phase occurred in support of the original research goal of replicating a specific study [12]. This first phase search consisted of keywords or phrases such as honeynet, machine learning, dynamic honeypot, and so forth with the goal of establishing a baseline of existing research. Here, we selected 13 articles from an original set of 54 consisting of various elements related to applying machine learning to honeynets in some manner. Selection criteria prioritized whether the literature included machine learning, involved honeynets, and used the former to analyze log output from the latter.

A second search was necessary once we pivoted from a replication study to a critical review methodology. This second search phase consisted of expanded searches intended to uncover missing elements in the core studies such as cowrie AND machine learning or honeypot log analysis AND WEKA. The search was necessarily narrow to prevent contaminating replication with unrelated information. For example, the type of honeypot was critical since not all honeypots operate similarly. As well, the type of machine learning classifiers employed matter as does the operational scientific methodology. Ultimately, we found a single study [Dumont] containing related material.

2.2 Replication Criteria

Overarchingly, the goal of replication is to redo a study using the same setting, method, and instrumentation [15]. Achieving this goal is vital to computational research because doing so uncovers latent assumptions [15] and adds weight to a scientific baseline [13]. Within that, replication is intended to produce consistent results across efforts, not necessarily identical results [16]. Fortunately, there are definitive criteria associated with designing and executing a replication study beyond the broad considerations of setting, method, and instrumentation. Such criteria serve as a form of rubric that, when adhered to, ensure a level of integrity and consistency in the resulting research.

The rubric is most visible when examining associated literature in collection. Foremost, Drummond [2009] noted that replication is possible when (a) transparent description of the design, setup, and execution; (b) complete sharing of instruments (code where applicable); (c) availability of raw data. Similarly, Brandt [17] defined replication criteria as (a) test the assumed underlying theoretical process; (b) assess the average effect size of a reported effect; (c) and test the robustness of that effect. Likewise, Anderson [16] summarized replication criteria as consisting of (a) repeating data analysis and (b) measuring statistical equivalence between the original and replication studies. These steps presuppose the presence of the Drummond [18] criteria while providing a more directional or procedural context.

In view of this, we first reached to the original authors [12] to confirm details of the research. Absent any replies, we then turned to critically reviewing the study to establish precisely what would be necessary to perform a future replication.

3. RESULTS

The four sections below present our findings after conducting a critical review of the source study [12]. We organized the results into a linear sequence following the standard research presentation framework found in most literature. While these results are oriented towards what lacked transparency in the source study, this is a convention to expedite future replication. The results do not imply the source study is without scientific merit.

On the contrary, Marydas [12] includes an actionable research problem and outlines relevant background material. The work also includes a reasonable discussion of the underlying research method and experimental environment. Further, the study makes clear what technologies are used and what the purpose of the research is as such pertains to the overall study design. However, data collection, format of data, instrumentation, and data analysis were insufficiently presented.

3.1 Data Collection

The first element limiting replication is consideration for what log files were collected from the honeypots. For example, cowrie outputs three different log files. These files do not contain identical information and, in some cases, may seem to conflict if not identified correctly prior to data collection. Thus, any replication effort is left to assume which log file must be used to train a machine learning classifier to identify malicious behavior.

3.2 Data Format

On a related but separate point, the format of the collected data is critical to replication because of the tight coupling to the data analysis phase. Here, the source study provided contradictory information. On one hand, the data format is indicated as Attribute-Relation File Format (ARFF). On the other hand, later the source study discussed cowrie data in Comma Separated Value (CSV) format. However, one limitation for replication is that cowrie logs are in plain text and JavaScript Object Notation (JSON) formats only. Another limitation is that the source study does not demonstrate how the cowrie log files were converted between file formats.

3.3 Instrumentation

In the context of replication, the most important instrumentation would be the simulated adversarial behavior employed to generate the honeypot raw data. While the source study specified the use of Metasploit, there is a limitation insofar as the specific adversarial protocol is not discussed. That is, there is no indication of what specific exploits were executed, what payloads delivered, or what instrumentation results correspond to certain log file data. Without this information, replication is limited in ability to assess to what degree results are similar as well as test and compare effects.

3.4 Data Analysis

A final element in our critical review evaluated the data analysis procedure in the source study. Here, we uncovered several limitations impacting potential replication. For instance, although the authors were clear about using WEKA as the data analysis platform, there is no discussion of specifically what data was loaded into WEKA. Furthermore, WEKA is a robust machine learning platform and has a myriad of functions. Yet, the source study does not outline a procedure covering what function or functions were used. Perhaps most critically, the source study does not detail what *features* in the data were used to build the classifier models. The authors mention using the *Message* field but nothing more specific. Collectively, these limitations rather completely inhibit replication because there is no means to compare results.

4. CONCLUSIONS

Honeypots and, by extension, honeynets facilitate learning about adversary behavior [1]. That is, by allowing a system to be compromised, researchers and practitioners can study trends and techniques used by attackers [1][2]. However, these deceptive technologies have constraints and challenges. For instance, honeynets generate an enormous amount of data which can be difficult to analyze manually [12].

For that reason, researchers [12][13] have turned to machine learning as a means of analyzing the volumes of data collected by a honeynet. As the field works to establish protocols for implementing the variety of machine learning algorithms in this context, the significance of replication work should not be underestimated. Computational research benefits from replication research because it helps establish valid baselines and helps correct errors [13][15]. Thus, we originally set out to conduct a replication study of a seminal piece of research [12].

Ultimately, we were not able to conduct a replication though because necessary elements lacking detail or missing. Rather than give up, we shifted our focus to performing a critical review with the goal of illuminating specific areas for improvement such that replication can take place in the future. To that end, we uncovered four areas- data collection, format of data, instrumentation, and data analysis- requiring additional information before replication can occur.

4.1 Recommendations

Based on our findings, we recommend two potential means to address the limitation stemming from data collection. Most simply, the authors [12] could clarify what log file was used to train the machine learning classifiers. However, future work could also quasi-experimentally validate each type of log file as a proper training data source. As well, related research [2] may provide some idea as to how cowrie log files may be preprocessed into a usable file format. Alternatively, future work may independently develop a preprocessing instrument for use with the same honeypots.

Furthermore, more work is needed on applying machine learning to honeynet log data beyond replication of existing research. Marydas [12] and Dumont [19] studied classifiers but there are other categories of machine learning (e.g. regression) that may add to researcher and practitioner repertoires. Likewise, existing work with classifiers may be comparatively applied to other honeypots and honeynets.

REFERENCES

[1] Curran, K., Morrissey, C., Fagan, C., Murphy, C., O'Donnell, B., Fitzpatrick, G., & Condit, S. (2005). Monitoring hacker activity with a Honeynet. International Journal of Network Management, 15(2), 123-134.

[2] Campbell, R. M., Padayachee, K., & Masombuka, T. (2015). A survey of honeypot research: Trends and opportunities. In 2015 10th international conference for internet technology and secured transactions (ICITST) (pp. 208-212).

[3] Spitzner, L. (2003). The honeynet project: Trapping the hackers. IEEE Security & Privacy, 1(2), 15-23.

[4] Park, B., Dang, S. P., Noh, S., Yi, J., & Park, M. (2019). Dynamic Virtual Network Honeypot. 2019 International Conference on Information and Communication Technology Convergence (ICTC). doi:10.1109/ictc46691.2019.8939791

[5] Sochor, T., & Zuzcak, M. (2014, June). Study of internet threats and attack methods using honeypots and honeynets. In International Conference on Computer Networks (pp. 118-127). Springer, Cham.

[6] Silva, D. V., & Rafael, G. D. R. (2017). A review of the current state of Honeynet architectures and tools. International Journal of Security and Networks, 12(4), 255-272.

[7] Nawrocki, M., Wählisch, M., Schmidt, T. C., Keil, C., & Schönfelder, J. (2016). A survey on honeypot software and data analysis. arXiv preprint arXiv:1608.06249.

[8] Huang, L., & Zhu, Q. (2020). Strategic Learning for Active, Adaptive, and Autonomous Cyber Defense. In Adaptive Autonomous Secure Cyber Systems (pp. 205-230). Springer, Cham.

[9] Zakaria, W. Z. A., & Kiah, M. L. M. (2013). A review of dynamic and intelligent honeypots. ScienceAsia 39S, 1-5.

[10] Zakaria, W. Z. A., & Kiah, M. L. M. (2012). A review on artificial intelligence techniques for developing intelligent honeypot. In 2012 8th International Conference on Computing Technology and Information Management (NCM and ICNIT) (Vol. 2, pp. 696-701).

[11] Chowdhary, V., Tongaonkar, A., & Chiueh, T. C. (2004, December). Towards Automatic Learning of Valid Services for Honeypots. In ICDCIT (pp. 469-470).

[12] Marydas, M., and Varsha Priyah, J. N. (2019). A cloud based honeynet system for attack detection using machine learning techniques. International Research Journal of Engineering and Technology (IRJET), 6(7), p. 330-335.

[13] Davison, A. (2012). Automated capture of experiment context for easier reproducibility in computational research. Computing in Science & Engineering, 14(4), 48-56.

[14] Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91-108.

[15] Feitelson, D. G. (2015). From repeatability to reproducibility and corroboration. ACM SIGOPS Operating Systems Review, 49(1), 3-11.

[16] Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods, 21(1), 1.

[17] Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. Journal of Experimental Social Psychology, 50, 217-224.

[18] Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. Retrieved from: http://cogprints.org/7691/7/ICMLws09.pdf

[19] Dumont, P., Meier, R., Gugelmann, D., & Lenders, V. (2019, May). Detection of Malicious Remote Shell Sessions. In 2019 11th International Conference on Cyber Conflict (CyCon) (Vol. 900, pp. 1-20). IEEE.