Dos and Don'ts of Machine Learning in Computer Security
Paper Review
Shivam Bhat
bhat41@purdue.edu
09/19/2022

Summary

With the advent of cheap and readily available computing resources, along with the recent breakthroughs in the domain, machine learning has today found its application in many areas including computer security. However, Machine learning's application to the security domain still continues to be a black-box, containing various pitfalls and shortcomings. This is the argument the authors try to make in the given paper. The paper, through its analysis of 30 top papers from last decade, tries to identify pitfalls, study their security implications and then suggest recommendations to help improve them. The author claims the research to be first of its kind, in terms of doing a general analysis of the misuse of Machine Learning in Computer Security. The review methodology adopted also appears to be robust with 2 independent reviews, followed by a 3rd reviewer to resolve any disputes. As part of the process, feedback also was taken from the authors of the concerned research papers which must have helped correct any misinterpretations and reinforce the validity of legitimate pitfalls. However , I think the paper could have been more broad. The current sample size only considers 30 papers from the last 10 years which could often lead to authors taking hasty conclusions. At certain points the author seems to be contradicting himself specifically regarding complexity of machine learning models and limitations of the data used. The paper, nonetheless, is very much relevant which is highlighted by the survey feedback from the authors and the impact analysis performed on real world examples. The detailed process, my comments and suggestions are listed in the next section.

Comments

In this paper, the authors conduct first of its kind research to 'generally explore' the pitfalls present in top 30 papers trying to leverage machine learning for the Computer Security domain. All the papers have been selected from top-tier security conferences, published over the last decade. While I understand that there are logistical difficulties of conducting such a thorough review for a higher number of pertinent research papers, 30 seems to be a very small sample size to draw generic conclusions for the entire computer security publishing community. Although the percentage might seem high, in terms of absolute numbers a single

digit count of papers with a particular kind of pitfall might not be adequate to jump to a conclusion.

The paper does a great job of breaking down the machine learning workflow into 4 sections and then identifying pitfalls in each of them. In total, 10 pitfalls have been identified. For each of the pitfalls authors do a good job of explaining the pitfall, analysing its security implications and then suggesting easy to overcome it. 'Sampling Bias' in data collected is found to be the most widespread pitfall present in 90% of the reviewed papers. Most of the sampling bias is said to have occurred due to usage of data(often synthetic or combined from different sources) which does not represent the input space or follows a different distribution. While the author recommends avoiding doing so, he also at the same time recognizes the limitations of acquiring good security data and goes on to suggest that we have to often synthesise data but one should try to maintain compatibility between data sources when doing so. Author acknowledges that this pitfall can't be completely removed and also does not provide any metric for a benchmarking. Given the sensitivity of mobile and consumer network data, getting such data is difficult and often purged or anonymized. This I believe forces researchers to use the limited open sourced data available.

Similarly authors suggest using simple models instead of complex ones specially in situations where no benchmark model is determined. To demonstrate this Mirai dataset by Mirsky et al. is analysez by two models namely "KITSUNE" and "Boxplot Method". Although the results seem to suggest a better performance by simpler Boxplot, this does not negate the fact that more sophisticated models like the neural networks have been found to perform exceptionally better than other models. Further given logistical and financial limitations it might be more intuitive for a researcher to directly select a more sophisticated model for analysis. Model selected is also determined by the data and operation being performed. For example for pixel data, Neural Networks are found to better handle that than simple models like Decision Trees.

Further the pitfall of "Lab-only evaluation" might be difficult to avoid given that its difficult to predict or catch a novel attack live. Even when found , a black hack operation or exploit would be under scrutiny of the government and agencies given that sensitive data of common people could be involved. Often whenever there is an exploit discovered it's reported to the concerned agency so that a patch could be sent before it could be exploited via an adversary algorithm.Since one cant possibly test at the scale a real attack could execute , the experiments are often limited to the small constraints of a lab environment.

Recommendations and takeaways

Given the real world limitations discussed above , the authors should try to quantify the pitfalls. Having a minimum feasible metric which incorporates the above limitations would help to evaluate these approaches better.

The research has a smaller sample size and could be expanded to a more exhaustive list of papers.

In the “Source Code Author Attribution” analysis the approach uses GCJ which is not only limited in size but also in the nature of the code. Instead of that one could leverage the open source code available on platforms like github like it's been done for GitHub Co-Pilot AI tool. The data would not only be more diverse but also very encompassing of real world software projects.

Lastly, having a more detailed review of the researcher’s intuition behind the dataset and model selection could help validate the assumptions . It would help provide the context in which those choices were made for the purpose of experiments. This data could be collected via that feedback survey.