The Use of Decision Trees to Defend Against Specific Cyberattacks

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jordan Crawley

Spring, 2023

Technical Project Team Members

Jordan Crawley

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Sebastian Elbaum, Department of Computer Science

The Use of Decision Trees to Defend Against Specific Cyberattacks

CS4991 Capstone Report, 2022

Jordan Crawley
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
jlc9cnu@virginia.edu

ABSTRACT

In recent years, society has seen an increase in large-scale data breaches, which demonstrates the importance of the cybersecurity field and the necessity of thorough practices within it. One such important practice is the use of algorithms to detect cyberattacks and malware. Particularly, algorithms which use decision trees are a valuable tool in that they are useful in certain attack detection methods. This is because decision tree algorithms can be used to identify commonalities between malicious attacks in order to identify similar likely malicious activity. Additionally, decision tree algorithms can be used to decide how to respond to specific user actions. In the future, exploring the further use of decision tree algorithms will doubtlessly be valuable in helping cybersecurity professionals adapt to new types of attacks.

1. INTRODUCTION

A decision tree is the name given to refer to a flowchart or flowchart-like structure wherein each node represents a specific condition or variable to test on an object or scenario. In a decision tree, branches split based on the outcome of conditions in previous nodes; and each leaf node represents a final outcome, classification, or decision taken after evaluating all variables and conditions. Decision trees are commonly used in a number of areas with regards to data management, and are especially useful in data analytics and machine learning due to their simplicity and how they help visualize flow of data or the way that a given algorithm should run in specific circumstances.

A decision tree algorithm, also called decision tree learning, specifically refers to an algorithm based on a decision tree, or which uses a decision tree-like structure to process some scenario and arrive at a decision. However, while generally less accurate, it is also possible to construct decision trees unsupervised using machine learning.

2. RELATED WORKS

A report by Peddabachigari et al., (2012) does an excellent job of summarizing applications of decision trees in the area of intrusion detection. This report clearly defines an intrusion as “any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource” (p. 3). The report also does a great job of realistically evaluating the state of intrusion detection efforts, mentioning that a completely secure network is virtually impossible, but also that decision tree algorithms are one of the strongest and most impartial ways for security algorithms to classify user behavior and detect malicious activity in accordance with past or expected activity.

A 2018 web article by Chib provides information about the recent shift in cybersecurity from using decision trees to using deep learning, an automated process which focuses on allowing findings and results to be generated from data without explicit programming. Chib states that decision tree algorithms need to be manually set up and therefore have inbuilt human limitations.

3. PROPOSED DESIGN

I have highlighted existing ways that decision trees are used and can be used in cybersecurity. However, the current use of decision trees can be expanded as well. Chib (2018) proposes that deep learning is strictly more efficient than decision tree algorithms because decision tree algorithms by nature must be created by a human, and thus have the limitation of human foresight and potential error. While this is traditionally true, it is also possible for cybersecurity programs to create their own decision trees using machine learning to classify malicious behavior and act on their own.

3.1 Autonomous Construction

Decision trees offer an advantage in that they provide an easily followable, human-understandable approach to evaluating scenarios and possibilities and reaching a decision, classification, etc. As such, there is value in maintaining and expanding the use of decision trees in cybersecurity, as they are not only useful to cybersecurity from a machine perspective, but can also allow site administrators, cybersecurity workers, etc., to more easily understand vulnerabilities in their systems, and thus address them more easily.

As an example, suppose that a malicious user gains access to a site by using an injection attack or a similar intrusion method that involves incorporating user input into a website in order to achieve some functionality. Suppose also that this attack ultimately succeeds, the website is shutdown, and data is compromised. By logging the steps the user took during this attack, or even comparing the code the website executed leading to the attack with the site's standard unaltered code, the site could add these steps as individual nodes to a new or existing "malicious activity" decision tree entirely built or added to by the site or program itself. Effectively, each leaf node at the end of the tree would represent a malicious attack, and thus could be used to decide how to respond to future malicious activity of the same kind, by stopping user access or reverting the state of the code.

3.2 Limitations

Of course, this approach would not be perfect. This would require that the site, network, or program be sophisticated enough to log individual states during an attack and potentially recognize individual behaviors that could be categorized as malicious. Though this is a lot, it is not impossible and deep learning could play a role in categorizing these behaviors effectively without outside human input. The nature of this proposition also leaves a possibility that user behavior could be incorrectly categorized, and innocent users could be incorrectly identified as malicious or find their activity restricted (Goeschel, 2016).

Additionally, this approach would require that the entity in question also recognize the end state of a successful attack or intrusion, many of which are such that they are intentionally difficult to identify by programs and sometimes even by humans. However, this could be counteracted somewhat in that a human could predefine some known types of malicious attacks relevant to the entity in question, providing examples of the potential end state after such a completed attack.

4. ANTICIPATED RESULTS

One of the main benefits of decision trees, in addition to their actual usefulness, is their ability to help humans analyze and understand what is going on with the state of their systems. With more focus on the continued and expanded use of decision trees within the cybersecurity industry, it is reasonable to expect a continued high standard of security, as well as cybersecurity systems that are easy to follow and build upon over time.

Having a central decision tree that is built or added to autonomously in particular will ensure that a given entity's security will continue to adapt to modern cybersecurity threats and appropriately classify or respond to these threats. This point is especially important as cybersecurity is a field which is constantly evolving with the threat of new types of attacks and new security methods. Though there is a reasonable argument that deep learning is altogether a better direction for cybersecurity to take going forward, it is worth noting that decision tree classification can perform similarly to deep learning (Alam et al., 2020). Both methods generally operate at very similar accuracies, with deep learning working only around 30% faster for the tradeoff of less human understandability. In some cases, as in the proposed method, decision tree algorithms are even capable of working alongside deep learning for maximum efficiency.

5. CONCLUSION

Decision tree algorithms are a useful and important cybersecurity medium that should not be overlooked and still have a lot to contribute to the industry. Through the use of decision tree algorithms, cybersecurity specialists can both identify commonalities between attacks in order to identify likely malicious activity, and pick ways to respond to such activity.

Going forward, a more sophisticated approach to decision tree learning may involve decision trees being autonomously constructed or built upon based on logging of past user activity or an entity's past states before an identified attack. Though some advocate for the use of deep learning as a replacement for decision trees, this approach sacrifices the decision trees' benefit of high understandability. Furthermore, as the two approaches can work together, decision trees should not be wholly abandoned in modern cybersecurity.

6. FUTURE WORK

Since cybersecurity is such a complex and case-sensitive field, it is difficult to immediately define the next steps needed across the industry to make realizing this proposed usage of decision trees actionable. Many entities across the internet are currently using decision trees in some way, as they have a reputation for being useful, easy to understand and reliable.

Beyond this, the proposed method of utilizing decision trees that are autonomously constructed or built upon can only truly be realized by entities with a certain level of sophistication. Depending on the size of the system, constructing or adding to these trees and responding accurately to nodes could require a high amount of computing power which may not be feasible for smaller entities. Ultimately, as with any other tool in the cybersecurity industry and the internet as a whole, the key to making it more popular is likely to get more reputable companies and sites to use and facilitate discussion around this method.

REFERENCES

Alam, F., Mehmood, R., Katib, I. (2020). Comparison of Decision Trees and Deep Learning for Object Classification in Autonomous Driving. *Smart Infrastructure and Applications. EAI/Springer Innovations in Communication and Computing.* https://doi.org/10.1007/978-3-030-13705-2_6

Chib, H. (2018, April 27). How security vendors are moving away from using decision trees to Deep Learning. *Intelligent Tech Channels.* Retrieved September 1, 2022, from https://www.intelligenttechchannels.com/2018/04/27/how-security-vendors-are-moving-away-from-using-decision-trees-to-deep-learning/

Goeschel, K. (2016). Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive bayes for off-line analysis. *SoutheastCon 2016.* https://doi.org/10.1109/secon.2016.7506774

Peddabachigari, S., Abraham, A., & Thomas, J. (2012). Intrusion Detection Systems Using Decision Trees and Support Vector Machines. *Department of Computer Science, Oklahoma State University, USA.*